Thesis for the degree of Doctor of Philosophy

Methods to identify novel disease genes and uplift diagnosis rates in rare diseases

Eleanor G. Seaby

MMedSc (Hons) BMBS (Hons)

December 2023

Student ID: 22493484

ORCID: 0000-0002-6814-8648

University of Southampton

This PhD thesis is dedicated to Ralph Martin Seaby.

Abstract

Methods to identify novel disease genes and uplift diagnosis rates in rare diseases

by

Eleanor G. Seaby

Since the advent of next generation sequencing technologies, the ability to diagnose rare diseases has improved considerably. Yet despite advances, most rare diseases remain undiagnosed. In part, this is due to a demand for more efficient methods to interpret genomic sequencing data, in addition to the need to establish the phenotypic consequence of variants in genes not yet associated with disease.

This thesis describes the development and testing of novel methods to improve diagnostic efficiency in patients with rare diseases, in addition to the discovery of novel disease-gene relationships. Herein describes the DeNovoLOEUF method, which identifies putative pathogenic *de novo*, loss-of-function variants in both known disease and putative disease genes. The gene-agnostic HiPPo protocol is further described, which prioritises variants identified in sequencing data. Finally, application of the GenePy dimensionality reduction algorithm to identify missed biallelic diagnoses is discussed.

DeNovoLOEUF was applied in established disease genes to ~14,000 trios recruited to the 100,000 Genomes Project (100KGP). In total, 98% of all variants identified were proven diagnostic, including 39 new diagnoses missed by 100KGP. DeNovoLOEUF was then applied to novel genes to the same 100KGP cohort. A total of 18 putative disease genes were identified, whereby 12/18 (67%) of these genes have since been functionally validated. For the remaining 6 genes, case series are underway and two of these with supportive functional evidence are presented in this thesis: *DDX17* (comprising 11 patients with *de novo* monoallelic variants and neurodevelopmental phenotypes, named Seaby-Ennis Syndrome); and *HDLBP* (comprising 7 patients with *de novo* monoallelic variants and neurodevelopmental phenotypes).

Finally, application of the HiPPo protocol was demonstrated to be an effective, efficient, alternative method to interpret genomic data, capable of outperforming strategies used by the NHS Genomic Medicine Service

(GMS). The GMS utilises gene panels to analyse sequence data, whereas HiPPo is a panel-agnostic method that prioritises variants using *in silico* metrics. HiPPo had a superior diagnostic rate per number of variant assessed when compared with the GMS (20% vs 3% respectively). HiPPo further identified all pathogenic variants reported by the GMS and identified an additional missed pathogenic variant.

Data presented in this thesis demonstrate how novel methods applied to genomic sequencing data can efficiently enhance diagnosis rates for patients with rare diseases and identify new disease-gene relationships. In turn, these can improve patient outcomes by better elucidating mechanistic understanding of disease, identify novel therapeutic targets, and tailor treatments to specific diseases and individuals. To fully realise the potential of novel methods, additional research is needed. Future plans will involve the use of artificial intelligence to refine methods and models for improved clinical outcomes.

Table of Contents

List of Abbreviations

Abbreviation	Description
100KGP	100,000 Genomes Project
AB	Allele balance
ACMG-AMP	American College of Genetics and Genomics-Association for Molecular Pathology
AD	Autosomal dominant
AF	Allele frequency
AI	Artificial intelligence
AnVIL	Genomic Data Science Analysis, Visualisation, and Informatics Lab
API	Application programming interface
AR	Autosomal recessive
ARVCM	Arrhythmogenic right ventricular cardiomyopathy
ASD	Autism spectrum disorder
ATAC-seq	Assay for transposase-accessible chromatin using sequencing assay
BBS	Baratela-Scott syndrome
BMI	Body mass index
bp	Base pair
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
ClinGen	Clinical Genome Resource
CM	Cardiomyopathy
CNV	Copy number variant
CPU	Central processing unit
dbGAP	Database of genotypes and phenotypes
DCM	Dilated cardiomyopathy
DNA	Deoxyribonucleic acid
DNASE-seq	Dnase I hypersensitive sites sequencing
ENCODE	The Encyclopaedia of DNA Elements
eQTL	Expression quantitative trait loci
ER	Endoplasmic reticulum
ESE	Exonic splicing enhancer
ExAC	Exome Aggregation Consortium
EXRC	European Xenopus Resource Centre
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements sequencing
FDA	Food & Drug Administration
FISH	Fluorescence in situ hybridisation
GB	Gigabyte

GC	Guanine-Cytosine
GCS	Google Cloud Services
GDPR	General Data Protection Regulation
GeCIP	Genomics England Clinical Interpretation Partnership
GEL	Genomics England
GenCC	Gene Curation Coalition
GLH	Genomic laboratory hub
GMC	Genome Medicine Centre
GMS	Genomic Medicine Service
gnomAD	Genome aggregation database
GoF	Gain of function
GQ	Genotype quality
GRCh37	Genome Reference Consortium Human Build 37
GRCh38	Genome Reference Consortium Human Build 38
GREGoR	Genomics Research to Elucidate the Genetics of Rare Diseases
GTEx	Genotype tissue expression portal
HCM	Hypertrophic cardiomyopathy
HET	Heterozygote
HiPPo	High Pathogenic Potential
HLA	Human leucocyte antigen
HOM	Homozygote
HPC	High performance cluster
HPO	Human phenotype ontology
IGV	Integrative Genomics Viewer
IMPC	International Mouse Phenotyping Consortium
IRAS	Integrated Research Application System
kbp	Kilobase pair
KOMP	Knockout Mouse Project
LOEUF	Loss of function Observed/Expected Upper-bound Fraction
LoF	Loss of function
LOFTEE	Loss of function transcript effect estimator
LRSeq	Long read sequencing
LV	Left ventricle
LVNCM	left ventricular noncompaction cardiomyopathy
MAF	Minor allele frequency
MANE	Matched annotation from the NCBI and EMBL-EBI
MEDIP-seq	Methylated DNA immunoprecipitation sequencing
MGD	Mouse Genome Database
MGI	Mouse Genome Informatics

miRNA	MicroRNA
MLPA	Multiplex ligation-dependent probe amplification
MME	Matchmaker Exchange
MNV	Multi-nucleotide variant
MRSD	Minimum required sequencing depth
NGS	Next generation sequencing
NHS	National Health Service
NMD	Nonsense medicated decay
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase chain reaction
pext	Proportion expression across transcript
pLI	Probability of loss of function intolerance
pLoF	Predicted loss of function
PPV	Positive predictive value
QC	Quality control
RAM	Random access memory
RE	Research Environment
REC	Research Ethics Committee
RGD	Rare genetic disorders
RNA	Ribonucleic acid
sgRNA	Single guide RNA
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SRSeq	Short read sequencing
STR	Short tandem repeats
SV	Structural variant
TB	Terabyte
TF	Transcription factor
TRE	Trusted research environment
tx	Transcript
UCSC	University of California, Santa Cruz
UDN	Undiagnosed diseases network
USD	US Dollars
UTR	Untranslated region
VCF	Variant call file
vCPU	Virtual central processing unit
VEP	Ensembl Variant Effect Predictor
VM	Virtual machine
VUS	Variant of uncertain significance

WES	Whole exome sequencing
WGS	Whole genome sequencing

List of Figures

List of Tables

List of Supplementary Figures

List of Supplementary Tables

List of Supplementary Datasets

List of Appendix Papers

Declaration of Authorship

Print name: **Eleanor G. Seaby**
Title of thesis: **Methods to identify novel disease genes and uplift diagnosis rates in rare diseases**

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

SEABY, E. G., Leggatt, G., Cheng, G., Thomas, N. S., Ashton, J. J., Stafford, I., … & Ennis, S. (2023). A gene pathogenicity tool 'GenePy' identifies missed biallelic diagnoses in the 100,000 Genomes Project - *in press at Genetics in Medicine*

Singer-Berk, M., Gudmundsson, S., Baxter, S., **SEABY, E. G**., England, E., Wood, J. C., ... & O'Donnell-Luria, A. (2023). Advanced variant classification framework reduces the false positive rate of predicted loss-of-function variants in population sequencing data. *The American Journal of Human Genetics*.

SEABY, E. G., Turner, S., Bunyan, D. J., Seyed-Rezai, F., Essex, J., Gilbert, R. D. and Ennis, S. (2022), A novel variant in *GATM* causes idiopathic renal Fanconi syndrome and predicts progression to end-stage kidney disease. *Clinical Genetics*.

SEABY, E.G., Thomas, N.S., Webb, A., Brittain, H., Taylor Tavares, A.L., Genomics England Consortium, Baralle, D., Rehm, H.L, O'Donnell-Luria, A., Ennis, S. (2023). Targeting de novo loss of function variants in constrained disease genes improves diagnostic rates in the 100,000 Genomes Project. *Human Genetics.*

SEABY, E. G., Baralle, D., Rehm, H. L., O'Donnell-Luria, A., & Ennis, S. (2022). Response to Ramos et al. *Genetics in Medicine*.

SEABY, E.G., Smedley, D., Taylor Tavares, A.L., Brittain, H., van Jaarsveld, R.H., Baralle, D., Rehm, H.L., O'Donnell-Luria, A., Ennis, S. A gene-to-patient approaches uplifts novel disease gene discovery and identifies 18 putative novel disease genes. (2022). *Genetics in Medicine*. **[Editor's choice]**

Janssen, B., van den Boogaard, M., Lichtenbelt, K., **SEABY, E. G**., et al. (2022). Loss-of-function variants in TAF4 are associated with a neurodevelopmental disorder. *Human Mutation*.

Gudmundsson, S., Singer-Berk, M., Watts, N., Phu, W., Goodrich, J. K., Solomonson, M., … **SEABY, E.G**., … & Odonnell-Luria, A. (2021). Variant interpretation using population databases: lessons from gnomAD. *Human mutation*.

Smedley, D., Smith, K.R., … **SEABY, E.G**., … Caulfield, M. (2021). "The 100,000 Genomes Pilot on rare disease diagnosis in healthcare preliminary report." *New England Journal of Medicine*.

Gudmundsson, S., Karczewski, K.J., Francioli, L.C., … **SEABY, E. G.**, … & MacArthur, D. (2021). Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*.

SEABY, E. G., Rehm, H. L., & O'Donnell-Luria, A. (2021). Strategies to uplift novel Mendelian gene discovery for improved clinical outcomes. *Frontiers in Genetics*, 12, 935.

Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., **SEABY, E. G.**, Watts, N. A., Singer-Berk, M., … & MacArthur, D. (2020). Transcript expression-aware annotation improves rare variant interpretation. *Nature*, 581, 452-458. https://doi.org/10.1038/s41586-020-2329-2.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., … **SEABY, E. G.**, … & MacArthur, D. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581, 434-443. https://doi.org/10.1038/s41586-020-2308-7.

SEABY, E. G., & Ennis, S. (2020). Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Briefings in Functional Genomics*. **[Editor's choice]**

Signed: Eleanor Grace Seaby

Date: 12/12/23

Acknowledgements

Firstly, I would like to extend my sincere thanks to all the patients and families who have generously contributed to the research presented in my thesis.

I would like to thank all my colleagues in Southampton, Imperial College London, and the Broad Institute of MIT and Harvard. Particularly, I would like to thank Dr Simon Thomas, Dr David Hunt, Professor Matt Guille, Dr Annie Godwin, and Nikki Graham for their time, expertise, and contributions to work that has been crucial to progression of my PhD. I am grateful to the many PhD students and postdocs who have been a joy to work with over the last 3 years, notably Dr James Ashton, Dr Gary Leggatt, Dr Lynn Win, Dr Carolina Jaramillo-Oquendo, Dr Imogen Stafford and Dr Guo Cheng. I would like to extend a special thank you to Professor Mike Levin at Imperial College for always believing in me, supporting me, and inspiring me for over 20 years.

My PhD would not have been possible without my incredible supervisory team in the UK and the US, who have given me the freedom to build a PhD around my own research questions. Thank you to Drs Heidi Rehm and Anne O'Donnell-Luria at the Broad Institute who have been fantastic mentors, inspirational advocates for open and collaborative science, and have given me opportunities to work with world-leading experts. Thank you to Professor Diana Baralle for her kindness, advice, humour, and guidance especially with my clinical academic career. My sincerest thanks go to Professor Sarah Ennis for her unwavering support, endless encouragement, kindness, patience, and wisdom over the last 13 years.

I would like to thank my friends and family for their unconditional love and support. Thank you especially to Wendy and Andy for being a second family and to Gemma for being my best friend and a constant rock of support. Thank you to my sister Alice for letting me sleep on her sofa after night shifts; to Florence, my youngest sister, for tenaciously advocating for me; and to my big brother, Tom, for always being a calm and rational presence. Most of all, thank you to my parents for everything; I have no words to express my gratitude for all you have done for me. Particularly, thank you to you, Mum, for all the sacrifices you made for me growing up and during my illness. All I have achieved is a credit to you and Dad.

Funders

I am generously supported by the Gerald Kerkut Charitable Trust, a Southampton Presidential Scholarship, Imperial College BRC, the Foulkes Foundation, and a DRAGON grant.

Published papers (and preprints) during PhD

During my PhD I have contributed to a variety of projects, many of which extend beyond the results presented in this thesis. I undertook my PhD during the Covid-19 pandemic, and consequently, a considerable amount of my time was spent on SARS-CoV-2 related projects. Herein is a list of all publications and preprints I have contributed to during my doctoral degree.

1. **SEABY, E.G**., Thomas, S., Hunt, D., Baralle, D., Rehm, H.L., O'Donnell-Luria, A.L., & Ennis, S. (2023). A panel-agnostic strategy" HiPPo" improves diagnostic efficiency in the UK Genome Medicine Service. *In press at Healthcare*.

2. Chen, S., Francioli, L., Goodrich, J., … **SEABY, E.G**., … & Karczewski, K. (2023). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *Nature*.

3. Guo, M. H., Francioli, L. C., Stenton, S. L., … **SEABY, E. G**., … & Samocha, K. E. (2023). Inferring compound heterozygosity from large-scale exome sequencing data. *Nature Genetics*.

4. Willsey, H. R., **SEABY, E. G**., Godwin, A., Ennis, S., Guille, M., Grainger, R. M. (2023). Modeling Human Genetic Disorders in Xenopus: A Practical Guide - *in press at Disease Models and Mechanisms*

5. **SEABY, E. G**., Leggatt, G., Cheng, G., Thomas, N. S., Ashton, J. J., Stafford, I., … & Ennis, S. (2023). A gene pathogenicity tool 'GenePy' identifies missed biallelic diagnoses in the 100,000 Genomes Project - *in press at Genetics in Medicine*

6. **SEABY, E. G**., Godwin, A., Clerc, V., Meyer-Dilhet, G., Grand, X., … Ennis, S. (2023). Monoallelic *de novo* variants in *DDX17* cause a novel neurodevelopmental disorder. *medRxiv*, 2023-09.

7. Singer-Berk, M., Gudmundsson, S., Baxter, S., **SEABY, E. G**., Wood, J. C., Son, R. G., … & O'Donnell-Luria, A. (2023). Advanced variant classification framework reduces the false positive rate of predicted loss of function (pLoF) variants in population sequencing data. *American Journal of Human Genetics.*

8. Wojcik, M. H., Lemire, G., Zaki, M. S., Wissmann, M., Win, W., White, S., ... **SEABY, E. G**., … & ODonnell-Luria, A. (2023). Unique Capabilities of Genome Sequencing for Rare Disease Diagnosis. *medRxiv*, 2023-08.

9. Leggatt, G. P., **SEABY, E. G**., Veighey, K., Gast, C., Gilbert, R. D., Ennis, S. A Role for Genetic Modifiers in Tubulointerstitial Kidney Diseases. *Genes*. 2023; 14(8):1582. https://doi.org/10.3390/genes14081582

10. Sirvent, S., Vallejo, A. F., Corden, E., Teo, Y., Davies, J., Clayton, K., **SEABY, E.G**., … & Polak, M. E. (2023). Impaired expression of metallothioneins contributes to allergen-induced inflammation in patients with atopic dermatitis. *Nature Communications.* https://doi.org/10.1038/s41467-023-38588-1

11. Ashton, J. J., Gurung, A., Davis, C., **SEABY, E. G**., Coelho, T., Batra, A., … & Beattie, R. M. (2023). The paediatric Crohn's disease morbidity index (PCD-MI); development of a tool to assess long-term disease burden using a data driven approach. *Journal of Pediatric Gastroenterology and Nutrition*.

12. Pagnamenta, A. T., Jing, Y., Willis, T. A., Hashim, M., **SEABY, E. G**., Walker, S., … & Taylor, J. C. (2023). A palindrome-like structure on 16p13. 3 is associated with the formation of complex structural variations and SRRM3 haploinsufficiency. *Human Mutation*.

13. Koenig, Z., Yohannes, M. T., Nkambule, L. L., Goodrich, J. K., Kim, H. A., Zhao, X., … **SEABY, E.G**., … & Martin, A. R. (2023). A harmonized public resource of deeply sequenced diverse human genomes. *bioRxiv*, 2023-01.

14. Channon-Wells, S*., Vito, O*., McArdle, A. J*., **SEABY, E. G**., Patel, H., Shah, P., … & Lau, Y. L. (2023). Immunoglobulin, glucocorticoid, or combination therapy for multisystem inflammatory syndrome in children: a propensity-weighted cohort study. *The Lancet Rheumatology*.

15. **SEABY, E.G**., Thomas, N.S., Webb, A., Brittain, H., Taylor Tavares, A.L., Genomics England Consortium, Baralle, D., Rehm, H.L, O'Donnell-Luria, A., Ennis, S. (2023). Targeting de novo loss of function variants in constrained disease genes improves diagnostic rates in the 100,000 Genomes Project. *Human Genetics.*

16. Verseput, J., Rots, D., Venselaar, H., Innes, A. M., Stumpel, C., Ounap, K., **SEABY, E.G**., … & Kleefstra, T. (2021). A clustering of missense variants in the crucial chromatin modifier WDR5 defines a new neurodevelopmental disorder. *Human Genetics and Genomics Advances*, *4*(1), 100157.

17. **SEABY, E.G.,** Turner, S., Bunyan, D.J., Seyed-Rezai, F., Essex, J., Gilbert, R.D. and Ennis, S. (2022), A novel variant in *GATM* causes idiopathic renal Fanconi syndrome and predicts progression to end-stage kidney disease. Clinical Genetics. https://doi.org/10.1111/cge.14235

18. **SEABY, E. G**., Baralle, D., Rehm, H. L., O'Donnell-Luria, A., & Ennis, S. (2022). Response to Ramos et al. *Genetics in Medicine.*

19. Küry, S., Zhang, J., Besnard, T., Caro-Llopis, A., Zeng, X., Robert, S. M., … **SEABY, E.G**., … & Isidor, B. (2022). Rare pathogenic variants in WNK3 cause X-linked intellectual

disability. *Genetics in Medicine*.

20. Ashton, J., Cheng, G., Stafford, I. S., Kellermann, M., **SEABY, E.G.**, Fraser Cummings, J. R., Coelho, T., Batra, A., Afzal, N. A., Beattie, R. M. & Ennis, S. (2022). Prediction of Crohn's disease stricturing phenotype using a NOD2-derived genomic biomarker. *Inflammatory Bowel Diseases.*

21. Ashton, J. J., **SEABY, E. G**., Beattie, R. M., & Ennis, S. (2022). NOD2 in Crohn's disease-unfinished business. *Journal of Crohn's and Colitis*.

22. Janssen, B., van den Boogaard, M., Lichtenbelt, K., **SEABY, E. G**., et al. Loss-of-function variants in TAF4 are associated with a neurodevelopmental disorder. *Human Mutation*.

23. Patel, H., McArdle, A., **SEABY, E**., Levin, M. and Whittaker, E. (2022). The immunopathogenesis of SARS-CoV-2 infection in children: diagnostics, treatment and prevention. *Clin Transl Immunol*, 11: e1405.

24. **SEABY, E.G***., Melgar, M*., McArdle, A. J., Young, C. C., Campbell, A. P., Murray, N. L., … & BATS Consortium and the Overcoming COVID-19 Investigators. (2022). Treatment of Multisystem Inflammatory Syndrome in Children: Understanding Differences in Results of Comparative Effectiveness Studies. *ACR Open Rheumatology*.

25. **SEABY, E.G**., Smedley, D., Taylor Tavares, A.L., Brittain, H., van Jaarsveld, R.H., Baralle, D., Rehm, H.L., O'Donnell-Luria, A., Ennis, S. A gene-to-patient approaches uplifts novel disease gene discovery and identifies 18 putative novel disease genes. (2022). *Genetics in Medicine*. **[Editor's choice]**

26. Laricchia, K. M., Lake, N. J., Watts, N. A., Shand, M., … **SEABY, E.G**., … & Genome Aggregation Database Consortium. (2022). Mitochondrial DNA variation across 56,434 individuals in gnomAD. *Genome Research*, gr-276013.

27. Gudmundsson, S., Singer-Berk, M., Watts, N., Phu, W., Goodrich, J. K., Solomonson, M., … **SEABY, E.G**., … & Odonnell-Luria, A. (2021). Variant interpretation using population databases: lessons from gnomAD. *Human mutation*.

28. Smedley, D., Smith, K.R., … **SEABY, E.G**., … Caulfield, M. "The 100,000 Genomes Pilot on rare disease diagnosis in healthcare preliminary report." *New England Journal of Medicine* (2021).

29. Gudmundsson, S., Karczewski, K.J., Francioli, L.C., … **SEABY, E. G**., … & MacArthur, D. Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* (2021). https://doi.org/10.1038/s41586-021-03758-y

30. McArdle, A. J^{+}., Vito, O^{+}., **SEABY, E. G***., Patel, H*., Shah, P*., Wilson, C., … & Levin, M. (2021). Treatment of Multisystem Inflammatory Syndrome in Children. *New England Journal of Medicine.*

31. **SEABY, E. G**., Rehm, H. L., & O'Donnell-Luria, A. (2021). Strategies to uplift novel Mendelian gene discovery for improved clinical outcomes. *Frontiers in Genetics*, 12, 935.

32. Hoggart, C., Shimizu, C., Galassini, R., Wright, V. J., Shailes, H., Bellos, E., … **SEABY, E.G**., … & Levin, M. (2021). Identification of novel locus associated with coronary artery aneurysms and validation of loci for susceptibility to Kawasaki disease. *European Journal of Human Genetics*, 1-11.

33. Weng, P. L., Majmundar, A. J., Khan, K., Lim, T. Y., Shril, S., Jin, G., … **SEABY, E. G**., … & Sanna-Cherchi, S. (2021). De novo TRIM8 variants impair its protein localization to nuclear

bodies and cause developmental delay, epilepsy, and focal segmental glomerulosclerosis. *The American Journal of Human Genetics*, *108*(2), 357-367.

34. Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., **SEABY, E. G**., Watts, N. A., Singer-Berk, M., … & MacArthur, D. (2020). Transcript expression-aware annotation improves rare variant interpretation. *Nature*, 581, 452-458. https://doi.org/10.1038/s41586-020-2329-2.

35. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., … **SEABY, E. G**., … & MacArthur, D. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581, 434-443. https://doi.org/10.1038/s41586-020-2308-7.

36. **SEABY, E. G**., & Ennis, S. (2020). Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Briefings in Functional Genomics*. **[Editor's choice]**

37. Wai, H. A., Lord, J., Lyon, M. Gunning., … **SEABY, E. G**., … & Thomas, N. S. (2020). Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genetics in Medicine*, 1-10.

38. Beck, D. B., Petracovici, A., He, C., Moore, H. W., Louie, R. J., Ansar, M., **SEABY, E**… & Prijoles, E. J. (2019). Delineation of the First Human Mendelian Disorder of the DNA Demethylation Machinery: TET3 Deficiency. *The American Journal of Human Genetics.*

Papers in review

1. **SEABY, E. G**., Godwin, A., Clerc, V., Meyer-Dilhet, G., Grand, X., … Ennis, S. (2023). Monoallelic *de novo* variants in *DDX17* cause a novel neurodevelopmental disorder – in review at *Brain*.

2. Wojcik, M. H., Lemire, G., Zaki, M. S., Wissmann, M., Win, W., White, S., ... **SEABY, E. G**., … & ODonnell-Luria, A. (2023). Unique Capabilities of Genome Sequencing for Rare Disease Diagnosis - *in review at New England Journal of Medicine*

3. **SEABY, E. G**.*, Josephs, K. S.*, Andreou, A., Sinclair, H., Genomics England Research Consortium, Roberts, A. M., Thomas, E. R. A., Ennis, S., Ware, J. S. (2023). Genetic Diagnoses for Cardiomyopathies in a National Rare Disease Initiative – *in review at Genome Medicine*

Successful grants during PhD

- **€248,524 –** Dragon Award (*co-applicant*), European Commission – Research grant for RNA-seq on Covid-19 patients
- **£167,429 –** NIHR150393 (*co-applicant*), Exploring and grouping COVID-19 pharmaceutical interventions to determine mechanisms of action and endotypes of response
- **£4,200 –** Foulkes Foundation (*lead-applicant*), Foulkes Fellowship
- **£743,000 –** EPSRC UKRI (*co-applicant*), AI methods applied to Genomic Data for improved health (AGENDA)

Chapter 1 | Introduction

1.0 Contribution statement

This chapter is all my own work with guidance from my supervisory team and has been published as two separate literature reviews: *SEABY, E. G., Rehm, H. L., & O'Donnell-Luria, A. (2021). Strategies to uplift novel Mendelian gene discovery for improved clinical outcomes. Frontiers in Genetics, 12, 935* (**Appendix Paper 1**) and *SEABY, E. G., & Ennis, S. (2020). Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. Briefings in Functional Genomics* (**Appendix Paper 2**).[1,2]

1.1 Rare genetic diseases

Rare genetic disorders (RGD) affect 1-in-17 individuals in their lifetime.[3] They encompass some of the most severe disorders affecting patients worldwide, including childhood cancers, neurodevelopmental disorders, and muscle diseases to name a few.[4] Many are severe and life-limiting, with significant morbidity and mortality. Indeed, 30% of children with rare diseases die before their fifth birthday.[5] Many affected patients are wheelchair-bound, require respiratory support, feeding support, specialised community services, and significant hospitalisations.[6,7] This not only impacts the patients involved, but their caregivers and families.

Approximately 80% of rare diseases have a genetic basis yet many of the underlying genes have not yet been identified, nor has the wide spectrum of pathogenic variation been delineated for each gene.[6-8] As such, on average across all specialties, the causal variant(s) are only identified in ~30-40% of rare disease patients, leaving the majority of patients and their families without a reliable prognosis, rendering medical care largely supportive and often palliative.[9-12]

1.2 Next generation sequencing technologies

In the modern genomics era, there has been an exponential rise in exome and genome sequencing for the 400 million patients worldwide with rare genetic diseases (RGD). These next-generation sequencing (NGS) technologies are thriving in the diagnosis of RGD where traditional methods e.g. cytogenetics, Sanger sequencing, and linkage analysis have been limited by low resolution testing, single test throughput, and the requirement to test multiple affected families, respectively. Whole exome sequencing (WES) and whole genome sequencing (WGS) can objectively test the genome for potential disease-causing variants without an *a priori* candidate gene, which has facilitated a rapid expansion of research into the discovery of new disease genes.

WES and WGS have now transitioned from an exclusive research environment into routine clinical practice as recognised diagnostic tests.[13] Indeed, whole genome sequencing is now available on the National Health Service, delivered through their Genomic Medicine Service. However, despite their promise, current diagnostic rates are suboptimal. This is in part because NGS diagnostic testing is limited to known disease genes. Therefore, causal variants in novel genes or those outside of targeted (virtual) gene panels are largely ignored, at least until iterative reanalyses considering novel gene discoveries become commonplace. However, the genomics field is still an active research arena and national research projects such as the UK's 100,000 Genomes Project[14] and the USA's All of Us Research Program[15] have utilised large-scale sequencing efforts to diagnose patients and discover new disease gene relationships underpinning new RGD. As a result, the boundary between clinical diagnostics and novel gene discovery is increasingly blurred, with many analysts adopting a dual diagnostic and research role. Patients consented only for clinical diagnostics will have their genome tested against known disease genes, whereas patients consented additionally for research may receive a diagnosis following functional studies on candidate variants undertaken in a research setting. This approach is steadily increasing the number of new gene disorders identified, however there is a still much to be discovered about the human genome. In fact, >75% of the phenotypical consequences for variation in all ~20,000 human protein coding genes is unaccounted for, thus rendering the exome far from obsolete.[16]

1.2.1 Data volume

An average genome sequenced has ~3-4 million variants and sifting through these manually is impossible. WES produces significantly fewer variants at 20,000-22,000[17], however this number is still too large for individual variant-level scrutiny and misses non-coding variation. Bioinformatic methods can restrict data to a more manageable number, however this still leaves an overwhelmingly large number (tens to thousands) of variants to consider in more detail. This range is highly dependent on user-defined filtering parameters, whether analysis is conducted on an exome or genome, and whether trio sequencing is available.

1.3 Analysis of NGS data for diagnostics

It is necessary to apply filtering frameworks when analysing NGS data to whittle down vast quantities of variants to a more manageable list of candidates. Although general filtering strategies are recommended, there is no universal framework governing this process. Invariably, analysis is tailored to individual cases by considering an aggregate of information.

1.3.1 *De novo* analysis

Many RGD occur as one-off (*de novo*) events i.e. present in the proband but absent from unaffected parents. The rate of *de novo* mutation is between 44-82 events per genome with 1-2 affecting coding regions, and this number increases with paternal age.[18] Restricting analysis to *de novo* variants can reduce a genome significantly.[17] This is achieved using trio data (parents/offspring), which is not always available due to cost restrictions or inability to sequence related individuals due to parental death, loss of contact, unknown paternity and parental refusal. Consequently, some exomes and genomes must be analysed as singletons (proband-only) at the expense of analytical noise. Duo studies (one parent/proband) are preferable to proband-only, however many novel variants in the proband will be inherited by the other (unsequenced) parent. Overwhelmingly, singleton analysis is too noisy to identify candidates without arbitrary and restrictive filtering. Further, in rare cases, parental germline mosaicism can result in multiple affected offspring with assumed *de novo* inheritance.[19]

1.3.2 Recessive analysis

Multiple affected siblings to unaffected parents increases the probability for recessive inheritance. Where consanguinity is declared or implied through relatedness checks, regions of autozygosity (where regions of DNA are identical by descent) can be targeted as hotspots for rare homozygous variation. Biallelic inheritance of compound heterozygotes *in trans* are best identified with trio analysis, unless the variants are close enough together in the proband that visualising overlapping read data can determine phase. However, assigning pathogenicity becomes more challenging when one variant is a known pathogenic variant and a second is of low quality or has low *in silico* predictors. That said, there are cases of known pathogenic variants causing disease when co-inherited with a second more common variant (or hypomorph).[20] The trade-off here is ensuring that allele frequencies are not filtered too restrictively especially when hypomorphs can have allele frequencies as high as 7%.[21]

1.3.3 Allele frequency

Since population allele frequency (AF) data have been made publicly available, detecting rare variation has become far easier.[14,22] One of the largest repositories of WES/WGS data, depleted for rare childhood disease, is the genome aggregation database, or gnomAD (http://gnomad.broadinstitute.org), hosting 125,748 exomes and 71,702 genomes. To prevent inflation of allele frequencies in gnomAD and exclude somatic variants, the database only includes variants with an allele balance between 20% and 80% for heterozygous variants.[23]

The largest repository of whole genome sequencing data is TopMED, whereby they have sequencing data from over 180,000 participants.[24] Historically, global repositories have been biased towards European ancestry, and whilst this bias still exists, many more ancestries are now included with iterative updates.[25] With AF differing between ethnicities, it is important that rare variation is compared with the AF of the patient's ethnic group; i.e. what may be rare in Caucasians could be common in the Latino population.

Undoubtedly global repositories are immensely powerful; however, most individuals harbour several private variants absent from these repositories in their sequencing data. Even at 197,450 samples, gnomAD is underpowered to accurately catalogue rare variation on a population level.[26] That said, AF still remains a fundamental filtering tool in rare disease, but frequency cut-offs are often arbitrary and overly lenient. Whiffin *et al.*[27] developed a statistical tool (http://cardiodb.org/allelefrequencyapp/) for the frequency-based filtering of candidate variants accounting for disease prevalence, inheritance mode, allelic heterogeneity, penetrance and sampling variance. The tool was shown to reduce candidate variants by two-thirds whilst maintaining sensitivity.

1.3.4 Variant type

Variant type is an important consideration in variant analysis (**Table 1.1**). Coding variants are usually considered first, although splicing variants and indels may impact both coding and non-coding regions. Loss-of-function variants (frameshift, stop-gain, and essential splice-site) are regarded as the most pathogenic coding variants and are rarer compared to all other variant types.[28] However, they are enriched for technical artefacts and many of these error types are not accounted for by current curation guidelines.[26] Missense variants are scattered across disease genes in healthy populations, challenging their interpretation. Regional missense constraint metrics are helping to prioritise candidates.[29,30] Splicing variants (exonic and intronic) can augment regulatory domains within mRNA, particularly those controlling splicing and may also influence translation.[31,32] However, these are difficult to assess beyond the canonical splice site and many *in silico* tools perform poorly.[33,34] That said, recent splicing libraries are widening the biological understanding of splicing variants and these can be used to train splicing models.[35,36] Synonymous variants are often discarded but are not benign by default and show a degree of selective pressure. These variants can alter the consensus sequence around splice sites and may also augment translation.[31] Short indels spanning coding regions may result in loss-of-function, sequence truncation/elongation, or splicing events. Yet these are often poorly resolved and aligned.[37]

Table 1.1 | Variant classes including occurrence, consequence and considerations

VARIANT CLASS	DESCRIPTION	OCCURRENCE	CONSEQUENCE	CONSIDERATIONS
SYNONYMOUS	Single nucleotide variant	~11,800 per genome	Do not alter the amino acid sequence	- Not always benign - Can affect transcription and splicing
MISSENSE	Single nucleotide variant	~10,600 per genome	Single amino acid substitution	- Scattered across disease genes in healthy populations - Metrics available that map missense constraint at a regional gene level and highlight critical regions that may be implicated in disease
SPLICING	Single nucleotide variant (essential or non-essential splice site) or indel	~2,300 per genome	Alter the sequence around splice junctions and may result in aberrant mRNA transcription	- Variants may reduce or strengthen the natural splice site, disrupt regulatory sequences (e.g. silencers and enhancers), introduce new splice sites, or activate cryptic splice sites - Difficult to assess beyond the canonical splice site due to limitations in annotation software, particularly for indels and deep intronic variants
LOSS OF FUNCTION	Frameshift, nonsense, or essential splice site	~100 per genome	May result in protein truncation or protein loss due to nonsense medicated decay	- Predicted most pathogenic compared to all other variant types - Enriched for annotation, technical and transcript artefacts - Single exon genes escape nonsense mediated decay
SHORT INDELS	Small insertions or deletions <50 base-pairs	~420,000 per genome	- In-frame indels may elongate or truncate the protein product - Out of frame indels will result in loss-of-function - Indels across splice sites may disrupt splicing	- May have regulatory roles on DNA structure and function - Often poorly resolved and aligned particularly from short read exome data - WGS improves indel calling - Long read sequencing outperforms short read sequencing

Rates of variant occurrence were collated from 1000 genomes data, ExAC data, and a cohort of 44 Caucasian genomes.[28,38] Note that rates of variation are likely to be an underestimate as cohort sizes are still underpowered to detect all rare variation.[26] Additionally, indel rates are likely grossly underestimated due to alignment and mapping issues[39].

1.3.5 *In silico* tools

In silico tools score variants according to functional consequence (**Table 1.2**). A wealth of competing software is available, yet many informatics pipelines are incompatible with all these tools, nor do they keep up with updated versions. Discordant evidence lowers the confidence of relying on any one score. Composite scores, which utilise statistical methods, machine-learning, and deep neural networks certainly help to resolve this issue.[40,41] Yet, no single method emerges *consistently* superior and therefore a consensus approach is still recommended.[42] Although *in silico* tools can help to prioritise variants, they do not predict disease-causality.

TABLE 1.2 | SELECTION OF COMMONLY APPLIED IN SILICO PREDICTION TOOLS AVAILABLE FOR VARIANT PRIORITISATION

CATEGORY	ALGORITHM	SOURCE	PRINCIPLE
NON-SYNONYMOUS SNV PREDICTION	SIFT[43]	http://sift.jcvi.org	Evolutionary conservation
	PolyPhen-2[44]	http://genetics.bwh.harvard.edu/pph2	Evolutionary conservation and protein structure/function
	MutationTaster[45]	http://www.mutationtaster.org	Evolutionary conservation and protein structure/function
	Grantham[46]	https://gist.github.com/danielecook/	Biological consequence of amino-acid change
	REVEL[47]	https://sites.google.com/site/revelgenomics/	Ensemble learning method
	ClinPred[48]	https://bio.tools/ClinPred	Machine learning
	FunSAV[49]	http://sunflower.kuicr.kyoto-u.ac.jp/sjn/FunSAV	Random forest model
SYNONYMOUS SNV PREDICTION	FATHMM-MKL[50]	http://fathmm.biocompute.org.uk/fathmmMKL.htm	Sequence conservation within hidden Markov models
	GWAVA[51]	https://www.sanger.ac.uk/sanger/StatGen_Gwava	Integration of various genomic and epigenomic annotations
	TraP[52]	http://trap-score.org	Integrates sequence motif changes and GERP++
INDEL PREDICTION	IndelMINER[53]	http://omicstools.com/indelminer-tool	Heuristic model based on split read data
	ABRA[54]	https://github.com/mozack/abra	*De novo* assembly
	VarScan[55]	http://varscan.sourceforge.net/	Integration of read coverage, base quality, and number of strands
	DINDEL[56]	https://omictools.com/dindel-tool	Bayesian method
	GATK Haplotype Caller[57]	http://software.broadinstitute.org/gatk/	Bayesian method
	DDIG-in[58]	http://sparks-lab.org/ddig	Machine learning method
	DeepVariant[59]	https://github.com/google/deepvariant/	Deep neural network
SPLICING PREDICTION	GeneSplicer[60]	https://ccb.jhu.edu/software/genesplicer/	Markov model
	MaxEntScan[61]	http://genes.mit.edu/burgelab/maxent/	Maximum entropy model
	Human Splicing Finder[62]	http://www.umd.be/HSF/	Position dependent logic
	MutPred Splice[63]	http://mutdb.org/mutpredsplice/about.htm	Machine-learning prediction of exonic variants
	SpliceAi[64]	https://pypi.org/project/spliceai/	Deep neural network
CONSERVATION PREDICTION	PhyloP[65]	http://compgen.bscb.cornell.edu/phast/	Statistical phylogenetic tests
	GERP++[66]	http://mendel.stanford.edu/SidowLab/downloads/gerp/	Conservation across species
	PhastCons[67]	http://compgen.cshl.edu/phast/	Phylogenetic hidden Markov model
	Panther-PSEP[68]	http://pantherdb.org/tools/csnpScoreForm.jsp	Position-specific evolutionary conservation
NON-CODING PREDICTION	DeepSEA[69]	http://deepsea.princeton.edu/	Deep learning algorithm
	GenoCanyon[70]	http://genocanyon.med.yale.edu	Unsupervised learning
	SInBaD[71]	http://tingchenlab.cmb.usc.edu/sinbad/	Sequence information-based decision model
PROTEIN PREDICTION	AGGRESCAN3D[72]	http://biocomp.chem.uw.edu.pl/A3D/	Protein aggregation method
	DUET[73]	http://structure.bioc.cam.ac.uk/duet	Integrated computational approach
	HMMvar-func[74]	http://bioinformatics.cs.vt.edu/zhanglab/HMMvar/download.php	Hidden Markov models
	LS-SNP/PDB[75]	http://ls-snp.icm.jhu.edu/ls-snp-pdb	Genome-wide mapping of SNVs onto protein structures

	NeEMO[76]	http://protein.bio.unipd.it/neemo/	Residue interaction networks
	PMut[77]	http://mmb.irbbarcelona.org/PMut	Neural network
COMPOSITE SCORE	CADD[78]	http://cadd.gs.washington.edu/	Multiple genomic annotations
	M-CAP[40]	http://bejerano.stanford.edu/mcap/	Machine-learning composite score
	Eigen[79]	http://www.columbia.edu/~ii2135/eigen.html	Unsupervised spectral model
	Dann[41]	http://cbcl.ics.uci.edu/public_data/DANN	Deep neural network
	PrimateAI[80]	https://github.com/Illumina/PrimateAI	Deep neural network
	Alamut	https://interactive-biosoftware/alamut-visual	Composite score using splicing, missense and ESE prediction tools

Alamut is a commercial product and does not have a corresponding research paper. Updated versions are available from the source links. [ESE - exonic splicing enhancer; SNV – single nucleotide variant].

1.3.5.1 *Dimensionality reduction*

Genomic data are intrinsically sparse; variation can be rare, very rare, or entirely unique. With increasing application of contemporary sequencing technology, the dimensionality of observed variation is continually expanding. Most *in silico* metrics predict consequences of a single variant. However, for many individuals it is the combination of variants that can lead to adverse outcomes. Dimensionality reduction represents a critical requirement for many machine learning applications, whereby data can be collapsed into more intuitive and manageable information. GenePy (**see methods 2.7.1**) is a gene-based dimensionality reduction pathogenicity score that rescales variant level data to gene-level data. GenePy can be applied to both disease cohorts and controls and has been successful in selecting pathogenic genes for inflammatory bowel disease.[81] Whilst gene based methods show promise, no current method can link a specific phenotype to a gene based on biological function - all gene-based metrics reflect the burden of mutation within genes of equally-predicted causality.

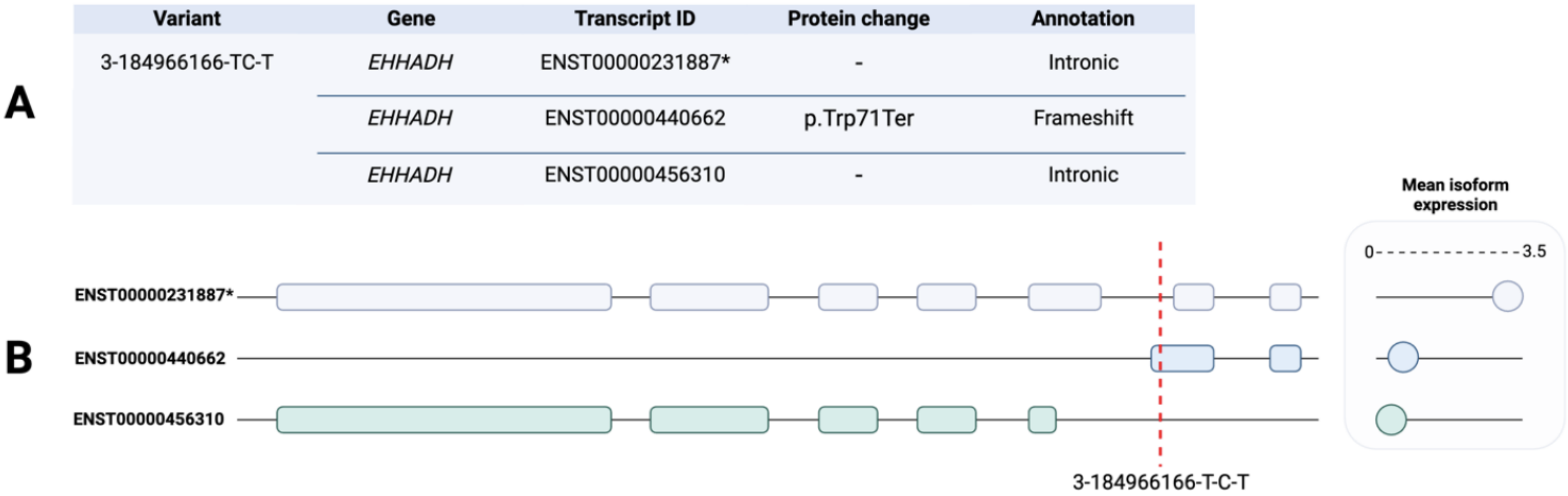
1.3.6 Tissue-specific annotation

Tissue-specific annotation has been a neglected area of variant interpretation. The genotype tissue expression (GTEx) portal has facilitated the study of tissue-specific gene expression and regulation important for variant analysis, e.g. a predicted pathogenic variant only expressed in testes in the context of a strong neurodevelopmental phenotype would likely be deprioritised. Equally, a variant strongly expressed in the kidney for a nephrological phenotype would warrant further scrutiny, even if overall tissue expression was low. That said, prudence is advised here; GTEx data were collected from cadavers, which will not necessarily represent *in utero* isoform gene expression when disease pathogenesis may have occurred.[82] Furthermore, some individuals present mosaicism and variants will be missed if the tissue of origin is not sequenced.

1.3.7 Transcript-specific annotation

Transcript-specific annotation considers how a variant affects all annotated transcripts of that given gene. Certain disorders are transcript-specific, i.e. only variants disrupting particular transcripts cause disease. Protein truncating variants in titan (*TTN)* - one of the largest genes encoding an essential protein component of striated muscle - cause dilated cardiomyopathy,[83] yet loss-of-function variants are found in healthy individuals. In healthy controls, loss-of-function variants occur in exons absent from dominantly expressed isoforms (the resultant spliced RNA as a result of a given transcript).[84] This is where nomenclature and understanding variant notation is essential. Diagnostic reports often describe variants with coding (*c.)* and protein (*p.)* notation in keeping with HGVS guidance,[85] yet often neglect to include which transcript the variant relates to. To confuse matters further, there are different annotation databases (RefSeq, GenCode, Ensembl) and transcripts vary between them. Although, disparity between annotation databases has been partially resolved by a collaborative project aiming to harmonise human gene and transcript annotation – the Matched Annotation from the NCBI and EMBL-EBI (MANE).[86] Nevertheless, reporting variants using genome (*g.)* nomenclature e.g. 1-55516888-G-A is preferable as all transcript consequences (from any preferred annotation database) can be assessed, e.g. a variant may be loss-of-function on one transcript yet intronic on all others (**Figure 1.1**). Furthermore, the same loss-of-function variant may disrupt a poorly expressed transcript in only one (clinically irrelevant) tissue. If this loss-of-function variant was reported without the transcript information e.g. p.Trp71Ter, then an analyst interpreting the data could be forgiven for assuming the variant may be pathogenic if found in a known disease gene. On the other hand, this same variant may be uniquely and moderately expressed in a tissue of interest, yet have low overall tissue expression, risking its exclusion. This is where amalgamating both tissue- and transcript-specific information is most valuable.

FIGURE 1.1 | SCHEMATIC OF THE 3-184966166-TC-T VARIANT AND ITS TRANSCRIPT-SPECIFIC ANNOTATIONS



Variant	Gene	Transcript ID	Protein change	Annotation
3-184966166-TC-T	*EHHADH*	ENST00000231887*	-	Intronic
	EHHADH	ENST00000440662	p.Trp71Ter	Frameshift
	EHHADH	ENST00000456310	-	Intronic

(A) *shows how the 3-184966166-TC-T variant causes a frameshift event on one transcript (ENST00000440662) and is intronic on the two remaining transcripts, including the canonical transcript identified by the asterisk (ENST00000231887*). If the variant were just annotated as EHHADH: p.Trp71Ter, this would neglect important transcript-specific information that would otherwise lower the pathogenicity of the variant.* ***(B)*** *Shows the individual transcripts and their corresponding exons (blue boxes). The canonical transcript is shaded in dark blue. Mean isoform expression of each transcript across GTEx tissues is shown on the right. The dotted orange line shows the position of the variant across all annotated transcripts. The transcript harbouring the p.Trp71Ter variant only disrupts the coding sequence of the ENST00000440662 isoform and has low mean expression across tissues, suggesting this transcript is not biologically important. Contrastingly, the canonical transcript (ENST0000231887*) is highly expressed approaching 3.5 transcripts per million, yet the variant is intronic for this transcript. Therefore, by using transcript level annotation, variant pathogenicity can be more accurately assessed and avoid providing a false sense of pathogenicity, particularly if the gene has a known disease association.*

To aid in the interpretation of transcript-specific annotation, Cummings *et al.*[87] developed a tool that facilitates the rapid visualisation of isoform expression values across tissues, called the proportion expressed across transcript (pext) score. Pext can differentiate between weakly and highly evolutionarily conserved exons and thus serves as a surrogate for functional importance (**see Figure 2.3 in Methods 2.6.2**). The integration of pext into the gnomAD browser now provides an opportunity to rapidly exclude variants from weakly expressed exons.

1.3.8 Incomplete penetrance

Harbouring a disease-causing variant does not always result in expression of disease, a phenomenon known as incomplete penetrance.[88] However, penetrance may increase with age, a phenomenon known as age-dependent penetrance, commonly seen in neurodegenerative diseases and hereditary cancer disorders.[89] It may well be that genes and variants thought to be incompletely penetrant are indeed penetrant at extremes of age. This contrasts with variable expressivity, whereby individuals with the *same* genotype express different degrees of the same phenotype. Whilst incomplete penetrance and variable expressivity are biologically distinct phenomena, their biological effects often overlap.[89]

Incomplete penetrance is not uncommon; several examples exist whereby pathogenic variants do not always present clinically in the individuals who carry them (**Table 1.3**).[90] To further illustrate this, gnomAD, which is depleted for individuals with rare childhood diseases, contains known pathogenic variants in haploinsufficient genes (*RB1*, *APC)*.

TABLE 1.3 | GENES MANIFESTING INCOMPLETE PENETRANCE OR VARIABLE EXPRESSIVITY

	Gene(s)	Phenotype	Paper
Incomplete Penetrance	*CX46*	Congenital cataract	Shawky *et al.*[91]
	ASXL1	Bohring-Opitz syndrome	Ropers *et al.*[92]
	CFH, CFI, CFB, C3, MCB	Atypical haemolytic syndrome	Bresin *et al.*[93]
	COL1A1, COL1A2	Osteogenesis imperfecta	Veitia *et al.*[94]
	SCN5A	Brugada syndrome	Gourraud *et al.*[95]
	DCC	Agenesis of the corpus callosum	Marsh *et al.*[96]
Variable expressivity	*NF1*	Neurofibromatosis type 1	Fahsold *et al.*[97]
	LDLR, APOB	Familial hypercholesterolaemia	Oetjens *et al.*[98]
	PKD1, PKD2	Polycystic kidney disease	Igarashi *et al.*[99]
Age-dependent penetrance	*BRCA1, BRCA2*	Breast cancer risk	van der Kolk *et al.*[100]
	TP53	Li-Fraumeni	Correa *et al.*[101]
	MLH1, MSH2, MSH6, PMS2, EPCAM	Lynch Syndrome	Biller *et al.*[102]
	LRRK2	Parkinson's disease	Alessi *et al.*[103]

Examples of genes displaying incomplete or reduced penetrance, variable expressivity or age-dependent penetrance as supported by scientific literature across a diverse range of phenotypes.

Population sequencing projects have revealed a high burden of potential protein-damaging variants in apparently healthy individuals.[26,28,88] Data extracted from the UK biobank, are beginning to evaluate rare disease-causing variants and refine penetrance estimates.[104] This challenges the analysis of NGS data when some genetic diseases may segregate through unaffected parents at allele frequencies higher than expected for a fully penetrant disease variant. On the other hand, some variants may be fully penetrant, yet modifier genes or rescue events may vary phenotype expressivity.

1.3.9 Non-coding variation

Until recently, non-coding DNA (comprising 98% of the human genome) was considered 'junk'. With the falling costs of WGS and advancing technologies (**Table 1.4**), non-coding variation is proving itself to be important for gene regulation, epigenetics and 3D genome structure.[105]

TABLE 1.4 | DESCRIPTION OF TECHNOLOGIES FOR THE DETECTION OF GENE REGULATION, EPIGENETICS AND GENOME STRUCTURE

TYPE	DESCRIPTION
CHIP-SEQ	ChIP-sequencing (ChIP-seq) is a method to analyse protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with DNA sequencing to identify DNA regions bound by proteins of interest. Histone marks indicating regulatory properties include H3K4me3, H3K27me3, and H3K27ac. Transcription factor binding sites can be analysed using ChIP-seq in a genome-wide manner.
DNASE-SEQ, ATAC-SEQ AND FAIRE-SEQ	These technologies identify 'open chromatin' regions where DNA binding proteins can access target DNA sequences.
HI-C AND CHIA-PET	These technologies are chromosome conformation capture-based methods for genome-wide analysis of 3D chromatin interactions. Hi-C can detect all possible interactions between DNA sequences in the genome, such as topological association domains. Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) incorporates ChIP-based enrichment and high-throughput sequencing to determine *de novo* long range chromatin interactions across the genome.
MEDIP-SEQ AND WHOLE-GENOME BISULFITE-SEQ	These technologies provide genome-wide analysis of methylated DNA sequences. MeDIP-seq combines ChIP and high-throughput sequencing. Whole-genome bisulphite-seq uses both treated (with sodium bisulphite) and untreated genomic DNA to identify methylated regions.

ATAC-seq – assay for transposase-accessible chromatin using sequencing assay; ChIA-PET – chromatin interaction analysis by paired-end tag sequencing; ChIP-seq – chromatin immunoprecipitation sequencing; DNASE-seq – Dnase I hypersensitive sites sequencing; FAIRE-seq – formaldehyde-assisted isolation of regulatory elements sequencing; MEDIP-seq – methylated DNA immunoprecipitation sequencing.

Efforts to understand non-coding variation have come through the ENCODE project, which has provided experimental evidence on the function of non-coding regulatory elements and gained new insights into the nature of transcription, chromatin structure and histone modification.[106] Results suggest that 80% of the genome contains elements linked to biochemical function, and that intergenic regions contain *cis*-regulatory elements such as enhancers, promoters, silencers and untranslated RNA transcripts with regulatory function (**Table 1.5**).[107] *In silico* methods that exploit ENCODE data and use machine-learning to predict the functional impact of non-coding single nucleotide variants are emerging.[50,51,79,108] However, the routine integration of these metrics is lacking, not least because the same variant may have conflicting effects in different tissues, individuals, and developmental stages and most analytical workflows are ill-equipped to interpret their effects.[51]

TABLE 1.5 | DESCRIPTION OF NON-CODING VARIANT CLASSES

TYPE	DESCRIPTION
REPRESENTATIVE FUNCTIONAL NON-CODING ELEMENTS	
UNTRANSLATED REGION	Untranslated regions (UTRs) are transcribed DNA included in the mature mRNA but not translated into protein. The 5' (upstream) UTR and 3' (downstream) UTR are important for translational control, mRNA stability and intracellular localisation. The 5' UTR contains the 5' cap and the 3' UTR contains the polyadenylation signal (AAUAAA) and miRNA-binding sites.
PROMOTER	Promoters provide binding for the promoter machinery that initiates gene transcription, typically located upstream of a gene. Epigenetic markers for active and repressive promoters include H3K4me3 and H3K27me3, respectively.
ENHANCER	Enhancers provide binding sites for proteins that help active transcription and control of expression levels of distal gene(s). H3K27ac is known as the representative mark of active enhancers.
SILENCER	Silencers provide binding sites for proteins important in the repression of transcription.
TRANSCRIPTION FACTOR BINDING SITES	TF binding sites are genomic elements bound by TF proteins that play a crucial role in regulation of gene transcription. TFs bind preferably to specific DNA sequence motifs in *cis*-regulatory regions of promoter and enhancer DNA.
INSULATORS	Insulators are DNA elements that act as a blocker of interactions between regulatory elements (e.g., enhancer-blocker insulators) and/or a barrier to propagation of repressive chromatin (barrier insulators).
TRANSCRIBED NON-CODING REGIONS	Some non-coding DNA is transcribed into functional RNA but not translated into protein. Specific examples include transfer RNA, ribosomal RNA, miRNA, long (intergenic) non-coding RNA, small nucleolar RNA, and others.

[TF – transcription factor; miRNA – microRNA; UTR – untranslated region].

1.3.10 Structural variation

Structural variants (SVs) are genomic rearrangements larger than 50 base-pairs. SVs not only affect gene dosage but can modulate the basic mechanisms of gene regulation by altering the copy number of regulatory elements and by modifying the 3D genome through topological associating domains. SV disease mechanisms are extensively reviewed elsewhere in the literature.[105]

SVs can be balanced (inversions and translocations), or unbalanced (deletions, insertions, and duplications) and their true abundance is only just being realised. Collins *et al.*[109] catalogued 433,371 distinct SVs across >14,000 genomes in gnomAD, producing 7,439 SVs on average per individual, contributing to approximately 25-29% of all rare protein truncating events. Therefore, there is likely to be an appreciable burden of rare pathogenic variation secondary to SVs undetected by current clinical testing standards. This has often been suspected, particularly in cases when one pathogenic variant is detected in a recessive gene of interest and there's suspicion for a 'second hit' somewhere else in the genome. Indeed, a Collins *et al*. showed that 4% of undiagnosed genomes were solved using SV calling.[109]

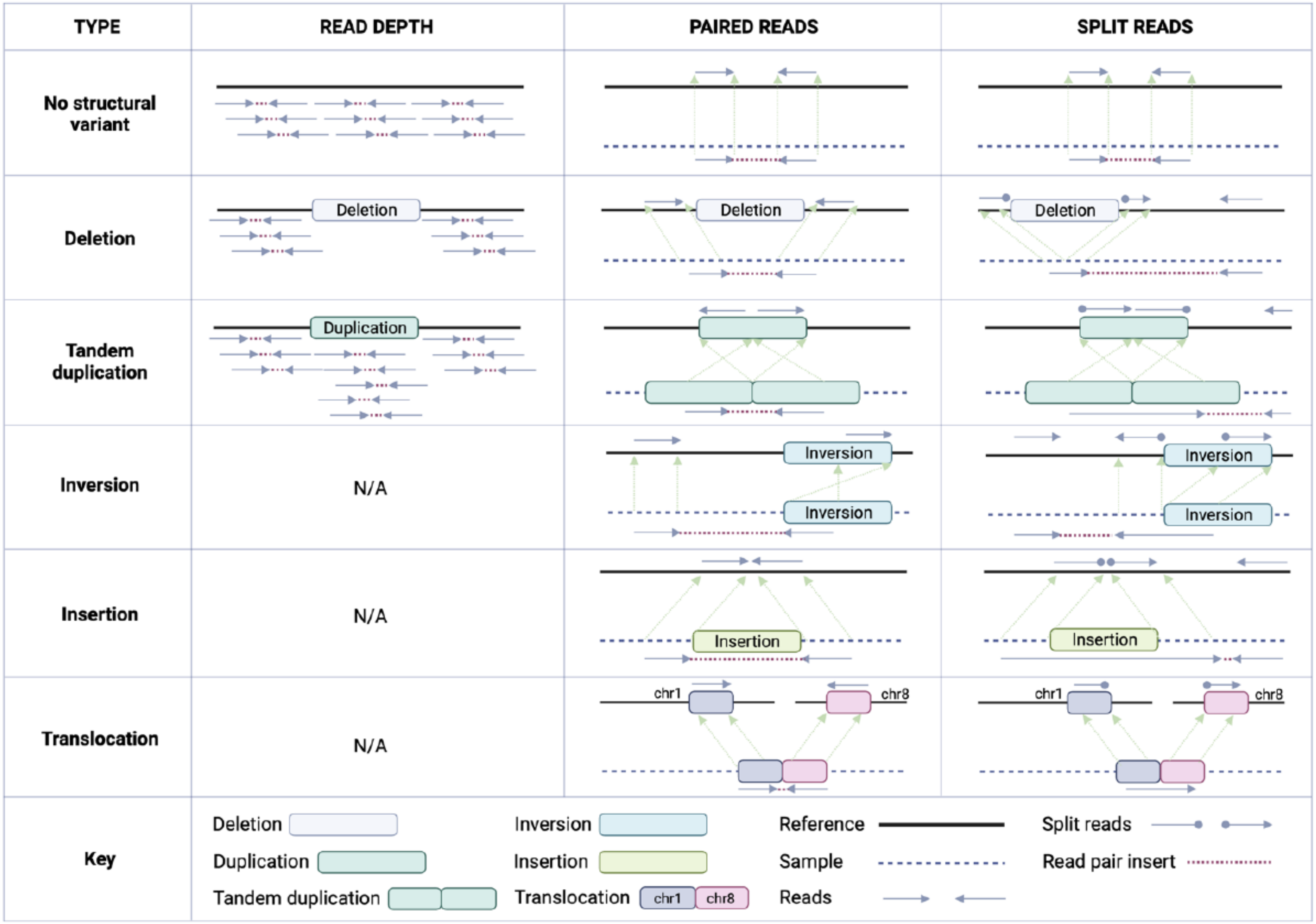
The most well studied SVs are unbalanced SVs or copy number variants (CNVs). Array comparative genomic hybridisation is the first-line diagnostic CNV test, however its low resolution and inability to detect balanced rearrangements mean that a large proportion of SVs are missed.[105] As technologies have advanced, the detection of SVs has moved away from cytogenetics towards NGS (**Table 1.6**). Methods to detect SVs from short-read WGS data has been made possible by an abundance of tools that exploit paired-end reads, read depth, and split reads (**Figure 1.2**). Theoretically, all SV classes can be detected at a finer scale than offered by traditional methods. However, these methods are limited, particularly from short-read data where mapping is notoriously inaccurate around repetitive regions, and most analytical pipelines do not support SV calling when their throughput is predominantly WES. Long-read sequencing (LRSeq) and optical mapping have emerged as solutions and have already revealed many times more SVs per genome when compared with short read sequencing (SRSeq); however, cost is still a substantial issue.[110] Linked-read technology, provided by 10X Genomics, integrates SRSeq with long range information. Long molecules of DNA are fragmented and labelled with a barcode; these fragments undergo SRSeq, producing a set of labelled reads that originate from the same DNA molecule. This technology is cheaper compared with LRSeq and facilitates simultaneous detection of small and large variants from a single library, however this technology is prone to the same limitations as both SRSeq and LRSeq.[111]

TABLE 1.6 | METHODS FOR THE DETECTION OF STRUCTURAL VARIANTS AND THEIR CORRESPONDING ADVANTAGES AND DISADVANTAGES

METHODS	ADVANTAGES	DISADVANTAGES
G BANDED KARYOTYPE	- Unbiased - Good for aneuploidy	- Limited resolution (5-7 megabases)
FISH	- Uses fixed interphase cells - Assesses balanced translocations - Detects mosaicism and tumour heterogeneity	- Requires multiple probes/assays to test multiple loci - Low resolution (100-200 kilobases)
ARRAY CGH	- Good for CNV and loss of heterozygosity - Better resolution than FISH - Low cost per data point	- Resolution still limited (50-100 kilobases) - Cannot detect balanced translocations - Cannot detect copy neutral CNVs - Low throughput
MLPA	- Detects small rearrangements up to 50bp - High throughput - Uses up to 40 probes	- Cannot detect copy neutral CNVs - Poor for mosaicism and tumour heterogeneity
SHORT READ SEQUENCING (SRSEQ)	- Can detect full range of genetic variation - High throughput - Resolution from 50bp - Greater per base accuracy than LRSeq - Deeper coverage	- Dependent on coverage - Susceptible to GC bias - Prone to errors particularly around low complexity regions - Requires WGS to assess all SV classes
LONG READ SEQUENCING (LRSEQ)	- Can detect full range of genetic variation - High throughput - Detects more SVs than SRSeq - More uniform coverage - Greater overall accuracy than SRSeq	- Expensive - Lower per base accuracy than SRSeq - Lower depth of coverage - Cannot resolve sequences larger than input DNA
LINKED-READ SEQUENCING	- Can detect full range of genetic variation - Constructs long range haplotypes - Low error rate - Lower cost compared with LRSeq	- Cannot resolve sequences larger than input DNA - Susceptible to GC bias - Reduced performance around small indel calling

[Bp – base pair; CGH – comparative genomic hybridisation; CNV – copy number variant; FISH – fluorescence in situ hybridisation; kbp – kilobase pair; LRSeq – long read sequencing; MLPA – multiplex ligation-dependent probe amplification; SRSeq – short read sequencing; SV – structural variant; WGS – whole genome sequencing].

FIGURE 1.2 | METHODS TO DETECT STRUCTURAL VARIATION USING NEXT GENERATION SEQUENCING DATA



***Read depth** is exploited to assess deletions and duplications. Reads aligned to the reference genome will show increased coverage over duplicated regions and decreased coverage in the presence of a deletion. **Paired reads** are sequenced from either end, leaving an insertion gap in the middle. <u>In a deletion</u> the read pairs align to the reference further apart than expected. Dotted green lines denote change in read position/orientation between sample and reference. <u>In a tandem duplication</u> the orientation of the read pair is reversed in the reference. <u>In an insertion</u> the read pairs are aligned closer together than expected i.e. their insert size is less than expected. <u>For inversions</u>, the paired reads align in the same direction (either in the forward or reverse direction). <u>For a translocation</u>, the reads pairs are mapped to different chromosomes. **Split reads** occur over breakpoints. <u>For a deletion</u>, reads across the breakpoint will be split in the reference. <u>For a tandem duplication</u>, reads spanning the concatenation of the duplication will be split where the insertion begins and ends. <u>For an insertion</u>, reads will split at the point where the insertion begins or ends. <u>For an inversion</u>, the right read pair is split at the point where the inversion begins. The remaining split read will change orientation and align at the reverse end of the inversion. <u>For a translocation</u>, reads spanning the join between translocated regions will be split and align with their independent chromosomes e.g. chromosome 1 (chr1) and chromosome 8 (chr8).*

1.3.11 Applying supporting evidence from literature

When a list of candidates remains following variant prioritisation, scrupulous assessment of the scientific literature is mandatory to evaluate biological significance. This is achieved by reading papers on a gene(s) of interest and accessing databases such as OMIM[112] (http://omim.org/), ClinVar[113] (https://www.ncbi.nlm.nih.gov/clinvar/), and DECIPHER[114] (https://decipher.sanger.ac.uk/), which catalogue genotype/phenotype relationships with degrees of supporting evidence. This part of analysis is perhaps the most time-consuming as it involves evaluating copious amounts of evidence. Often the only genotype/phenotype association is through a genome wide association study; yet rare pathogenic variants may not present the same phenotype.

There is also a risk that published papers make false associations between a gene and disease. This was apparent in historical papers that linked 'rare' (at the time) segregating variants as pathogenic for a particular phenotype based on public databases primarily representing Europeans. Some rare variants in European populations may be common in other ethnicities. Only when public repositories started to include a diverse range of ethnicities did these 'rare variants' prove to be common and not disease causing at all. For example, the *MYBPC3* (p.G278E) variant was classified as pathogenic for hypertrophic cardiomyopathy in African populations. The variant was later reclassified as benign when found to be common in the African population.[115] Additionally, as public repositories have increased in size, many presumed disease-causing variants have been deemed 'too common' to be pathogenic.[116] Removing these variants from the literature is impossible, but the data-sharing ethos of ClinVar ameliorates this issue by encouraging evidence-based submissions which may refute previous claims of pathogenicity. By leveraging knowledge from multiple submissions, clinical significance and assertion criteria, ClinVar can be exploited as an important tool for the accurate interpretation of variant pathogenicity.[117]

1.4 Current diagnostic approaches and challenges

1.4.1 Virtual gene panels and clinical diagnostics

For well-characterised rare genetic disorders, a rapid diagnosis can be made if a pathogenic variant is found in a gene already associated with that disease. This is the premise behind applying *in silico* (virtual) gene panels to sequencing tests such as exome or genome sequencing. Virtual gene panels focus analysis to putative clinically relevant genes to reduce the number of variants requiring assessment by clinical laboratories. The curation of which genes to include on gene panels has traditionally been subjective and variable. To standardise this process, Genomics England developed the publicly accessible platform, PanelApp.[118] This resource has a crowdsourcing review tool to allow each gene to be reviewed and commented on by international experts within the scientific community. PanelApp now provides all the gene panels that relate to genomic tests listed in the NHS National Genomic Test Directory.

If a patient's phenotype overlaps with the clinical features associated with a candidate disease gene in an applied panel, and a variant is found in that gene, then there is potential for diagnosis. However, not all variants found in clinically relevant genes are pathogenic. Gene panels remain problematic as they are outdated at the point that new discoveries are published; and it takes time for new genes to be added to panels upon expert consensus. Further, selection of which gene panel to apply to genomic data is subjective and little training is provided to those ordering tests. When gene panels are updated, this then necessitates iterative reanalysis of sequencing results, which whilst potentially boosting diagnostic rates by 10–15%[119,120], is seldom done due to constraints on workload and time.[121] Whilst gene panels are clearly important and reduce incidental findings, it still remains that ~50% of haploinsufficient disease are yet to be discovered.[122] Therefore, single use gene panels are rarely successful for the majority of exomes and genomes analysed.

1.4.2 Variants of uncertain significance

With millions of exomes and genomes being generated worldwide, the ability to generate data far outpaces the ability to interpret genomic results and return diagnoses.[1] For variants to reach a clinical threshold for diagnostic reporting, they need to be already established as disease-causing, predicted to truncate the protein product, or have robust evidence from established *in vivo* or *in vitro* studies. The American College of Genetics and Genomics-Association for Molecular Pathology (ACMG-AMP) guidelines have set out standardised variant interpretation guidelines that classify variants as: pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign or benign.[123] These guidelines are based on multiple lines of evidence, applied at different weightings that are combined to give the overall classification. Yet, despite a drive towards standardised guidelines, many diagnostic labs vary in their variant interpretation.[124,125] Likely pathogenic and pathogenic variants are considered likely to impact clinical management and are reported. However, most variants identified in sequencing results are VUSs lacking the functional evidence to determine pathogenicity. As more patients undergo exome and genome sequencing, the number of VUSs identified exponentially grows. Ultimately, the requirement for functional experiments required to validate VUSs is proving a major bottleneck and there is unmet demand for scalable, high-throughput functional assays to confirm pathogenicity and expedite the return of genetic diagnoses.

1.4.3 Traditional family-based approaches to rare genetic disorders

At present, most exome and genome analyses are conducted on a 'per family' basis, i.e. on a small number of related individuals, most commonly a trio (parents and child). Analysing next generation sequencing (NGS) data is labour intensive, sometimes taking several hours to compile a report, although this is variable dependent on laboratory.[126] The challenge of handling vast quantities of genomic data has improved with advancing methods; however, for each exome or genome sequenced (and depending on whether segregation analysis is available using family studies), there are anywhere from tens to thousands of plausible candidate variants.[9] Commonly, large numbers of variants are filtered out using allele frequency cut offs, and if virtual gene panels are applied, then this can dramatically reduce the number of variants assessed. If a variant is found in a known disease gene, a rapid diagnosis can often be made with rigorous variant curation against laboratory standardised guidelines.[123,125]

It is worth reiterating the difference between clinical testing and research-led sequencing studies. Clinical testing typically focuses on variants in established or known disease genes, whereas research studies have the scope to evaluate variants in genes of unknown function or not currently linked to disease. Unless patients undergoing diagnostic testing are additionally recruited into research studies, opportunities to evaluate variants in new disease genes are limited. That said, there is increasing involvement of diagnostic centres with research laboratories and affiliated universities; many families are now concomitantly offered diagnostic testing and recruitment to further research studies. For the Genomic Medicine Service in the UK, patients can consent to have their deidentified data deposited into a genomics research library. However, for the diagnostic labs capable of bridging the gap between clinical testing and research, many are ill-equipped to investigate the plethora of plausible disease candidates remaining after variant filtration and prioritisation. Theoretically, the only way to establish new diagnoses when novel genes are identified in a research setting is to conduct functional experiments on potential candidates. However, as most research laboratories do not have the finances, nor time or resources to support functionally validating many candidate variants without any guarantee that the selected variants are pathogenic. Few laboratories will invest resources into a particular variant for one patient without additional kindreds with overlapping phenotypes or prior published studies on the gene's function.

1.5 Utilisation of large cohorts

1.5.1 Large-scale programmes for novel gene discovery and diagnostics

Genomic sequencing has become increasingly affordable and possible. However, exome and genome sequencing are seldom first line investigations globally, with many healthcare systems and health insurance policies not covering the cost. This has perhaps inspired the creation of large-scale international sequencing programs, often with government funding, offering exome and/or genome sequencing to thousands of rare disease patients and their families with an aim to assess the long-term feasibility of such technologies in routine healthcare. These projects benefit from pooled resources and focus on diagnosing patients that were undiagnosed following conventional clinical testing, in addition to better elucidating the underlying mechanisms of Mendelian diseases. Such examples include the UK's 100,000 Genomes Project[3] and Deciphering Developmental Disorders study[127] as well as the USA's Centres for Mendelian Genomics, now delivered through the Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR) consortium.[128] These programs benefit from sequencing large numbers of patients with improved power to match patients with similar genotypes and phenotypes, both internally and externally through the Matchmaker Exchange.[129] Furthermore, most of these programs recruit patients for both clinical diagnostics and follow-on research, meaning that where possible, novel discoveries and variants of uncertain significance can be investigated further when a clinical diagnosis has not been made which has previously been a limitation of clinical diagnostic studies.[1]

1.5.2 The Genomic Medicine Service

With the cost of genome sequencing becoming ever competitive, genome sequencing is beginning to supersede exome sequencing in some institutes, including in the NHS. The UK government has pledged to create the most advanced genomic healthcare system globally to bolster the UK's position as a life sciences superpower. The UK's Genomic Medicine Service (GMS), which follows the now completed 100,000 Genomes Project, offers pioneering genome sequencing as a routine diagnostic test to NHS patients.[130] The GMS plans to sequencing half a million genomes by 2023/24 and help transform healthcare for maximum patient benefit.[131] However, one of the challenges in diagnosing patients with rare disease is the expanded scope of analysis and need to correlate results with patient phenotypes. With 3-4 million variants found in a genome; the GMS has adopted the use of virtual (PanelApp)[118] gene panels made available through the National Genomic Test Directory, to reduce noise and focus on the most salient regions of DNA. Further time and research are needed to fully realise the success of this approach longterm.

1.6 Novel gene discovery

For ~60% of rare disease patients who undergo clinical exome or genome sequencing, their sequencing report is non-diagnostic, even though for many, the causal variant is present but unrecognised in their sequencing results.[1,2,132] One of the biggest challenges in reaching a molecular diagnosis is the paucity of scientific knowledge into the biological function of all ~20,000 human genes. Therefore, diagnosing rare diseases is extremely challenging without a prior correlation between a clinical phenotype and causative gene. New gene disorders preclude detection when for every genome sequenced, millions of variants of uncertain significance reside in genes of unknown function.[133,134] Even the best computational methods available at present will typically overlook a gene of undetermined biological significance when analysing a patient's exome or genome. Therefore, new rare genetic diseases will be overlooked until further studies are undertaken, or new methods are developed to uplift novel gene discovery.

1.6.1 Why is novel Mendelian gene discovery important?

The significance of uplifting novel Mendelian gene discovery is not to be underestimated. Every new disease gene discovered goes towards ending the notorious 'diagnostic odyssey' of rare disease. This pertains to rare disease patients who move between specialties and undergo myriad diagnostic tests in search for a unifying genetic explanation.[135] For most, these often expensive evaluations only elucidate the clinical phenotype and seldom aid in diagnosis. In the UK, over a ten-year period, undiagnosed rare diseases have cost the National Health Service (NHS) in England an average of £13,064 ($18,279 USD) per patient, and in excess of £3.4 billion ($4.8 billion USD) in total.[136] In Australia, the cost per diagnosis using standard care is AU$27,050 ($21,241 USD) and in the USA the same cost basis was calculated at US$19,100.[137,138] Whilst these figures all showcase the cost burden of rare disease, it is ill advised to compare cost evaluations between countries due to differing healthcare systems.

Novel discoveries directly impact diagnostic potential. Diagnoses not only provide answers for patients and families but have far reaching clinical impact, including but not limited to guiding personalised treatments; offering patient support networks; collecting and gaining knowledge on disease trajectory and prognosis; enabling participation in research studies; informing reproductive choices; and impacting the health of relatives. Even when little can be done therapeutically following diagnosis, the importance of that diagnosis to patients and families should not be overlooked; when a cause is identified, this often alleviates guilt and blame felt by patients and families who believe a given rare disease is their fault.[139]

Novel gene discovery is critical in the research space to expand biological understanding of human genes and variation, and to identify therapeutic drug targets that may lead to successful and life-altering therapies.[140-142] Gene augmentation therapies have been developed for a number of conditions, for example: subretinal injection of adeno-associated virus vectors to deliver *RPE65* cDNA to treat Leber congenital amaurosis (Mendelian Inheritance in Man identifier (MIM): 204100);[143,144] and the Food and Drug Administration approved one-time intravenous administration of *SMN* cDNA to treat spinal muscular atrophy type 1 (MIM: 253300).[145,146] Small molecular therapies for cystic fibrosis (MIM: 602421) are well studied, and include Ivacaftor, which increases the time fraction that the CFTR channel remains open, and

Lumacaftor which increases the amount of CFTR that reaches the cell surface.[147,148] Development of antisense oligonucleotides are proving effective in pre-clinical and clinical studies to treat neurodegenerative diseases[149,150]; and it is hoped that identification of novel disease genes may guide further protein targets. In contrast to the development of new gene therapies, it is not uncommon for existing therapies to be repurposed when knowledge of a given gene and biological pathway is implicated in disease. For example, in 2011, autosomal recessive variants in *MTHFD1* (a gene involved in folate metabolism) were found to cause combined immunodeficiency and megaloblastic anaemia with or without hyperhomocysteinaemia (MIM:617780).[151] Simple folic acid has proven life-changing for patients with recessive mutations in *MTHFD1*.[152,153]

1.6.2 The prior decade of novel gene discovery

Since the advent of next generation sequencing technologies, there has been a stepwise acceleration in novel gene discovery leading to uplifted diagnostic rates for rare disease patients.[16,154] Between 2005 and 2009 there were ~170 novel discoveries per year. This is compared to ~240 per year between 2010 and 2014 when NGS became widely adopted.[154] In the history of disease-gene relationship discovery, NGS approaches are responsible for ~36% of all reported Mendelian disease genes. Their contribution to novel gene discoveries is accelerating, with 87% of new gene disorders now discovered using NGS approaches.[155] Novel discoveries are still progressing, although the pace of discovery appears to have reached a steady state that balances the time required to build international cohorts, undertake functional experiments, and publish findings.[16] Despite this, approximately 250 new genes are added to the literature annually and a recent review predicted that more than 6,000 Mendelian conditions remain to be discovered.[155] Therefore, with thousands of monogenic disease-gene relationships yet to be elucidated, there is clear evidence that the recognition of disease-causing variation in the exome is far from saturated.[16]

1.7 Strategies to uplift novel gene discovery

1.7.1 Collaborative, data-sharing approaches

Collaborative projects, data sharing, and building disease cohorts have proved invaluable in genomics. In 2010, *MLL2* (*KMT2D*) was discovered as the cause of Kabuki syndrome (MIM: 147920). Ten unrelated patients with the same characteristic clinical phenotype underwent exome sequencing. Seven of the ten individuals were found to have loss-of-function variants in *MLL2*, which led to its disease association.[156] Historically the approach of building a case-series of affected individuals has been a rate-limiting step, relying on local connections or collaborations built through conferences or publications. Given the rarity of monogenic disorders, it can take many years to accrue sufficiently sized cohorts with similar clinical features and genotypes. This method is therefore inefficient and inadequate to rapidly support novel gene discovery.[157]

In 2017, the directors' board of the American College of Medical Genetics and Genomics released a position statement on how genomic data sharing is critical to improve genetic healthcare.[158] With an ever more connected world, global efforts to share genotype and phenotype data have proved essential in the endeavour of novel gene discovery. Improved data governance, drives for open data-science, and advancing informatics methods have since led to the practice of genomic matchmaking, facilitating researchers and clinicians from across the globe to share phenotype/genotype data for accelerated discoveries.[157]

1.7.2 Role of data-sharing

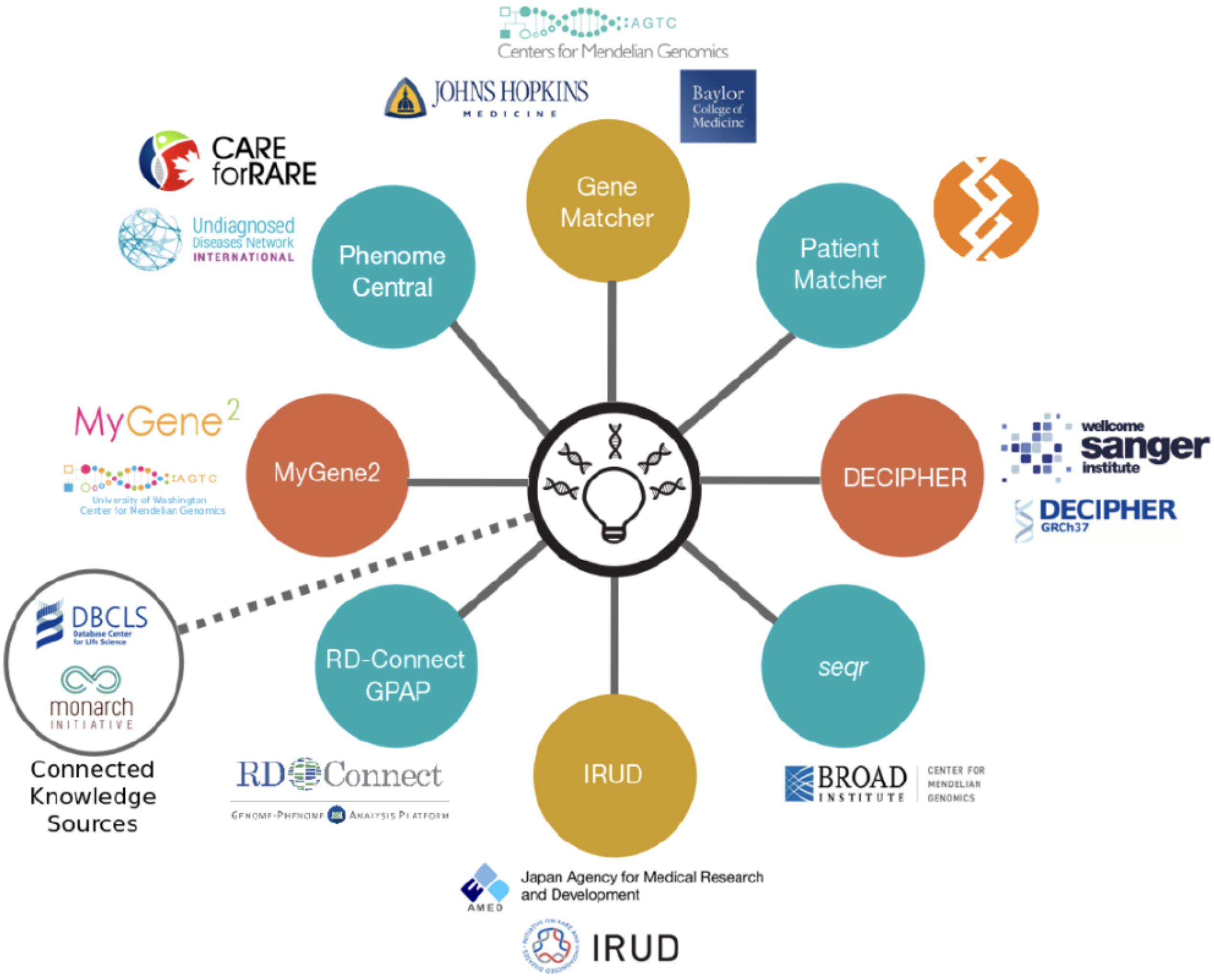
It is unquestionable that open science and data-sharing have been pivotal in uplifting diagnoses and advancing the field of genomic medicine. International collaborations are now commonplace in matching patients across the globe with specific genotypes, leading to high impact publications on novel gene discoveries.[129,157] Furthermore, the use of variant databases such as ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) have proven invaluable in providing the scientific community with a repository of variants, classified by pathogenicity, that can be applied to variant analysis for diagnostic interpretation.[113]

The success of open data science has been further driven by cloud computing. The presence of large datasets on cloud platforms can facilitate the access of desired data within a secure data-sharing platform. Examples include NHGRI's Genomic Data Science Analysis, Visualisation, and Informatics Lab (AnVIL) Space (https://anvilproject.org/data/) where rare disease data from the Centres for Mendelian Genomics along with data from additional projects such as 1000Genomes, Centres for Common Disease Genomics, and Genotype Tissue Expression (GTEx) can be accessed after application in the database of genotypes and phenotypes (dbGaP). Other trusted research environments include NHLBI's BioData Catalyst cloud platform with TOPMed data; Genomics England Research environment where 100,000 Genomes Project data are stored and accessed; and RD-Connect with rare disease genomic data from various European sources. In these trusted research environments, increasingly large amounts of data can be aggregated; researchers can bring tools directly to the data, share these analysis workflows, saving time, expense, and security risks of moving and maintaining local copies of large genomic datasets. However, data-sharing presents its challenges. There is still urgent need for an international code of conduct that provides clear, unified data-sharing rules across jurisdictions that comply with regional laws such as the European General Data Protection Regulation (GDPR) and the USA's Health insurance Portability and Accountability Act.[159]

1.7.3 Matchmaker Exchange

In 2015, the Matchmaker Exchange (MME) was launched, providing a systematic and robust approach to novel Mendelian gene discovery by facilitating a mechanism for matching patients across genomic centres, research laboratories, diagnostic laboratories, and physicians through a federated network (**Figure 1.3**).[129]

FIGURE 1.3 | THE MATCHMAKER EXCHANGE APPLICATION PROGRAMME INTERFACE AND ITS CONNECTED NODES


AGTC
Centers for Mendelian Genomics
JOHNS HOPKINS MEDICINE
Baylor College of Medicine
CARE forRARE
Undiagnosed Diseases Network INTERNATIONAL
Gene Matcher
Phenome Central
Patient Matcher
MyGene2
AGTC
University of Washington Center for Mendelian Genomics
MyGene2
DECIPHER
wellcome sanger institute
DECIPHER GRCh37
DBCLS Database Center for Life Science
monarch INITIATIVE
Connected Knowledge Sources
RD-Connect GPAP
seqr
IRUD
RD Connect
Genome-Phenome Analysis Platform
BROAD INSTITUTE
CENTER FOR MENDELIAN GENOMICS
AMED
Japan Agency for Medical Research and Development
IRUD

MME uses a federated network of nine connected nodes. Image taken with permission from https://www.matchmakerexchange.org/.

MME builds on the success of earlier genomic matchmaking platforms by connecting datasets through an application programming interface (API) enabling searches of multiple databases with a single query. The advantage of using a federated network enables individual submitters to maintain control and autonomy over their data and keep the content up-to-date, whilst ensuring compliance with their local and national data-sharing policies.[157] By identifying additional affected kindreds with overlapping phenotypes, the best candidate variants and genes can be targeted for functional validation. The MME API has been widely adopted by scientists and clinicians globally and has led to numerous international collaborations and publications. One such example is the discovery of a KMT2E-related neurodevelopmental disorder, O'Donnell-Luria-Rodan syndrome (MIM: 618512), following identification of 38 individuals from 36 families, of which 28 were ascertained using MME. This discovery goes beyond elucidating disease-gene aetiology and has identified a potential therapy already widely used in healthcare that could be evaluated for this syndrome.[160]

1.7.4 Patient-led approaches

It could be argued that no one is more invested in ending the diagnostic odyssey of rare disease than affected individuals and their families. In an era when patients are actively involved in research studies as participants[161], it is unsurprising that patients and caregivers are also invested in genomic matchmaking efforts.[162] Patients and families are beginning to take control of their own data and utilise open data-sharing and social media in an effort to discover new genetic disorders. Embedded within the MME API is a family facing platform called MyGene2 which gives patients and caregivers autonomy over their data, facilitating direct data-sharing when desired, while still enabling scientists and clinicians to access these shared anonymised data.[157]

Social networking sites such as Facebook, Twitter, and Instagram are also proving popular with patients/caregivers as a matchmaking resource.[163] In 2014, Matthew and Cristina Might harnessed the power of social media to identify additional cases of NGLY1 deficiency, leading to identification of a new gene disorder.[164,165] Following their son's diagnosis, the Might family explored options for conceiving a child unaffected by the same condition. Their son's diagnosis facilitated not only conception of a healthy sibling but a pathway for other affected families to conceive healthy children using preimplantation genetic testing or non-invasive prenatal diagnostic testing.[164] The Might family created a legacy for others to follow, having built a global community of families providing mutual support, in addition to facilitating research and international *NGLY1* meetings.[165]

Inspired by the success of the Might family, families across the globe have harnessed the networking potential of social media to match with other affected kindreds with similar phenotypes and genotypes. Indeed, social media additionally facilitated the identification of three children with variants of uncertain significance in *KDM1A*, leading to discovery of another novel gene disorder (MIM: 616728).[166] The success of such endeavours has now inspired the Undiagnosed Diseases Network (UDN), started at the National Institute for Health in 2008, with 11 additional clinical sites across the US, to use social media in a similar way. With appropriate consent, webpages are created for individual participants, showcasing the clinical phenotype, significant variants, and candidate genes. This approach has proven successful in identifying additional affected patients with variants in *NACC1* leading to the discovery of its associated phenotype.[163]

1.7.5 Phenotype-driven approaches

In recent years, novel gene discovery has shifted from phenotype-driven methods to genotype-driven approaches, i.e. taking genotype data and matching phenotypes to that genotype through matchmaking efforts, though both remain important.[155] Efforts to standardise phenotype terms through the Human Phenotype Ontology (HPO) database has aided comparative statistics using a universal library of agreed clinical terms involved in disease.[167] This has paved the way for computational phenotype analyses that can assess a candidate gene's relevance to phenotype data observed in patient(s). Several tools (**Table 1.7**) have been developed that estimate the similarity between HPO terms in an individual and those representing disease in a database. By incorporating phenotype ontology data across species, these tools are capable of prioritising candidate genes without known disease association.[168-170] Similar approaches have been commercialised, taking advantage of advanced artificial intelligence to identify and rank potential disease-causing variants following a multi-dimensional analysis; examples include Fabric Genomics (https://fabricgenomics.com/) and EMedGene (https://www.emedgene.com/). More experience and data are needed to understand the strengths and limitations of these tools.

TABLE 1.7 | FOUR PHENOTYPE-DRIVEN TOOLS FOR PRIORITISATION OF KNOWN AND NOVEL DISEASE GENES

Tool	Principle	Application	Access
Exomiser[134]	Uses random-walk analysis of protein-protein interaction networks, cross-species phenotype comparisons and a wide range of additional filters that consider prediction models, disease segregation and allele frequency.	Focused on identifying novel and known disease genes	http://www.sanger.ac.uk/science/tools/exomiser
eXtasy[171]	Prioritises non-synonymous variants predicted to be pathogenic using a fusion methodology that integrates multiple strategies in a phenotype-specific manner.	Focused on identifying candidates in novel and known disease genes.	http://extasy.esat.kuleuven.be/
Phevor[172]	Combines outputs of multiple biomedical ontologies and propagates patient phenotype information across and between ontologies for improved variant interpretation.	Focused on identifying candidates in novel and known disease genes.	http://weatherby.genetics.utah.edu/cgi-bin/Phevor/PhevorWeb.html
Phen-Gen[173]	Uses a systematic Bayesian framework which combines patient sequencing data with phenotype information for improved rare disease variant analysis of both coding and non-coding variation.	Focused on identifying candidates in novel and known disease genes.	http://phen-gen.org/

Bioinformatic phenotype-driven tools for the detection and prioritisation of variants in novel and known disease genes using the integration of patient genotype and phenotype data.

One of the challenges for novel gene discovery is the requirement for accurate, and deep phenotyping. Optimally, this should be collected longitudinally.[1] Whilst HPO terms do help to standardise the recording of phenotype information, and indeed are used universally in many databases including those connected through MME[129], they are often only collected at a point in time and may lack the 'full narrative' of the clinical history. This can be problematic when assessing new genotype-phenotype correlations, since for many neurodevelopmental disorders, phenotypes can significantly overlap. It can also be difficult to weight the severity or prominence of clinical features as the conversion to a list of terms tends to weight all the features similarly.

1.7.6 Functional validation using model organisms

Identifying the phenotypic effects of gene disruption may be possible using model organisms when there is enough conserved evolutionary function of the pathway/organ/system involving the gene of interest. Where the functional consequences of most human gene variants are yet to be established, model organism databases serve as a useful resource. Indeed, 58% of human genes have orthologs with disease-associated phenotypes reported in at least one model organism.[174] Although non-human models are not necessarily perfect proxies for human diseases, they can still serve as important biological models, particularly when data are aggregated across species. In recent years, publicly accessible databases have been made available for researchers to leverage the extensive body of genomic studies in model organisms. Examples include the Alliance portal[175] and MARRVEL[176].

1.7.7 Monarch Initiative

The Monarch Initiative (https://monarchinitiative.org) is an open science, collaborative project that aims to integrate phenotype-genotype data from a variety of species and sources.[174] Its user-friendly web portal promotes rapid assessment of phenotypes of orthologs in organisms and other species. Researchers can query genes, phenotypes, and diseases to identify candidate disease genes. Exomiser[134] and Genomiser[177] have utilised the Monarch Initiative in their gene prediction algorithms, which have led to diagnoses in participants in the UDN including the aforementioned discovery that the disruption of *STIM1* results in York platelet syndrome (MIM: 805070).[178]

1.7.8 Mouse knockout databases

Several mouse model organism databases exist including the Mouse Genome Database (MGD)[179,180], the Knockout Mouse Project (KOMP)[181], and the International Mouse Phenotyping Consortium (IMPC).[182,183] These projects are building comprehensive catalogues of mammalian gene function, genotype/phenotype associations, and detailed phenotype data from mouse knockouts of every protein-coding gene.[180,183] By 2019, the IMPC has fully or partially phenotyped 5,861 mouse genes, a third of which are non-viable.[183,184] Data from IMPC has aided the discovery of many novel Mendelian phenotypes.[185-187] That said, there is still much more to be gleaned from mouse data; of the >10,000 mouse genes linked to at least one non-lethal phenotype in a mutant strain in MGD, Bamshad *et al.* showed that human orthologs for 72% of those genes are yet to be associated with a Mendelian disorder, providing another rich data source for candidate genes awaiting discovery of the human Mendelian phenotype.[155]

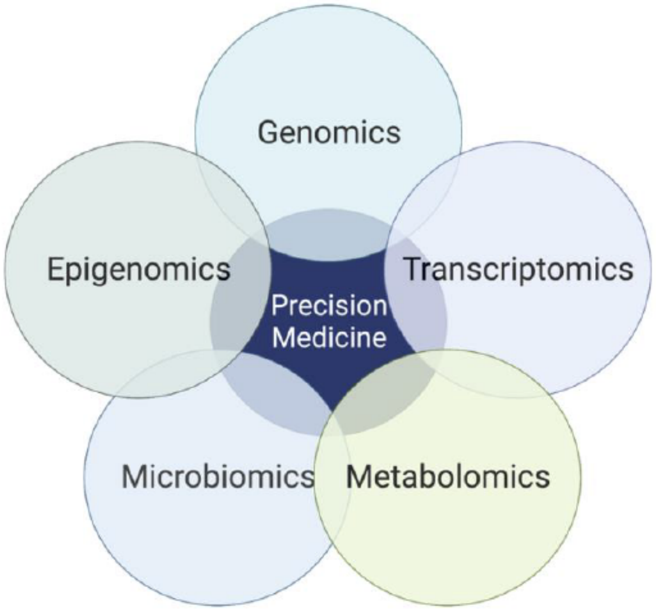
1.7.9 Incomplete penetrance and novel disease genes

Identifying novel disease genes can be challenged by incomplete penetrance, i.e. when a disease-causing variant does not always result in any clinical expression of the disease. If a novel candidate gene has been associated with a given phenotype, yet some or all alleles are incompletely penetrant, then it can be difficult to gather sufficient evidence for a new disease-gene association using traditional genetic evidence such as case observations and familial segregation. To mitigate this, larger cohorts that can support statistical association studies must be pursued. Furthermore, researchers are exploring how combinations of genomic variants such as oligogenic models or co-inherited protective alleles, environmental exposures, and mosaicism may impact the onset of Mendelian disorders.[188] One such approach is to specifically identify individuals that are resilient to rare disease, despite harbouring pathogenic variants.[90] Another area of interest is how *cis*-regulatory variation may modify the penetrance of coding variants.[189]

1.7.10 Integrating multi-omics data

Multiple omics technologies such as epigenomics, transcriptomics, metabolomics, microbiomics, and proteomics are being adopted as approaches in the effort to delineate the functional impact of genetic variation (**Figure 1.4**).[190] These integrative approaches can complement genomic data and aid in the validation and discovery of novel genes.

FIGURE 1.4 | A MULTI-OMICS APPROACH TO PRECISION MEDICINE



Schematic showing how the integration of multi-omics data is complementary and important for precision medicine.

The Genotype-Tissue Expression (GTEx) project (https://gtexportal.org/home/)[82] provides a public repository of tissue-specific gene expression and a multi-tissue reference for identifying variants associated with changes in gene expression, or expression quantitative trait loci (eQTL).[191] The GTEx consortium recently published results on their v8 release, providing insights into functional mechanisms and the architecture of genetic regulation.[192] The integration of transcriptome sequencing (RNA-seq) has led to improved diagnostic rates for Mendelian diseases[34,193], even when no strong candidate variants were identified from exome or genome data.[194] Expression outliers, altered splicing, and allelic imbalance in the transcriptome due to nonsense mediated decay can all be clues to candidate genes worth closer scrutiny in the exome or genome data.[195-197] Large scale transcriptome data can also be used in network analysis.[198] One caveat with RNA-seq is that splicing aberrations and differential gene expression is best assessed by sampling disease-relevant tissues. However, these may not always be clinically accessible e.g., brain tissue in neurodevelopmental disorders. Aicher *et al.*[199] showed that many splicing events in non-clinically accessible tissue are lowly expressed and poorly evaluated from more commonly accessible tissues such as skin and blood. The authors developed a tool (MAJIQ-CAT) which allows researchers to explore potentially accessible tissues that best represent splicing in genes of interest.[199] A recent preprint in 2021 describes an alternative approach that has advantages over expression-based methods. Minimum Required Sequencing Depth (MRSD) informs biosample selection (whole blood, lymphoblastic cell lines, and skeletal muscle) by estimating the minimum sequencing depth required from RNA-sequencing to achieve desired coverage across a given gene or gene panel. The authors reported high precision and their results suggest that lymphoblastic cell lines may be suitable for ~70% of established disease gene panels.[200]

The Encyclopaedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects have been instrumental in the generation of human reference epigenomes and epigenome maps, mainly from cell lines.[201-203] These data have successfully been used to conduct research on how the epigenome contributes to human development, environmental factors, and disease mechanisms.[201,203] More specifically, one of most commonly studied epigenetic phenomena, DNA methylation, is aiding diagnosis and gene discovery. Alterations in DNA methylation patterns are implicated in imprinting disorders and diseases of short tandem repeat (STR) expansions. The application of DNA methylation analyses has been successful in identifying molecular diagnoses in neurodevelopmental disorders where clinical microarray and other conventional genetic testing has been non-diagnostic.[204,205] In 2019, LaCroix *et al.*[206] investigated cases of Baratela-Scott syndrome (BSS) (MIM:615777) and identified hypermethylation of exon 1 of *XYLT1* associated with a GGC expansion and gene silencing. This not only confirmed BSS as a trinucleotide repeat expansion disorder but highlighted the relative prevalence of methylation abnormalities in disease pathogenesis of BSS. The hypermethylated allele accounted for 50% of the pathogenic alleles in their cohort, showcasing the importance of investigating epigenetic changes in disease cohorts with missing heritability.[206]

National biobanks such as the UK Biobank[207], United States All of Us Research Program[15], and Finland Biobank (FinnGen) provide opportunities to study genomic data and phenotype data alongside associated molecular markers from electronic medical records. Whilst their data are best studied in the context of complex disease, they are also important in rare disease by providing population level allele frequencies, biomarker results and phenotypic information for comparative analyses. Unlu *et al*. utilised Vanderbilt's biobank BioVU to identify a phenotypic profile that aided in the identification of a novel Mendelian syndrome CATIFA (cleft lip, cataract, tooth abnormality, intellectual disability, facial dysmorphism, attention-deficit hyperactivity disorder) that is due to loss-of-function of *RIC1* (MIM: 618761).[208]

The emerging application of metabolomics with exome/genome sequencing is helping to improve diagnostic rates in rare disease. Targeted and untargeted metabolomics are proving successful in validating variants of uncertain significance in inborn errors of metabolism.[209,210] It is hoped that with increasing research, metabolomics will continue to complement rich genomic data and aid in discovery of novel genes.

These aforementioned approaches often applied in combination have been pivotal in both clinical diagnostics and in identifying novel candidate disease genes. For example, a study in 2015 using epigenomics, comparative genomics, and genome editing identified a pathway for adipocyte thermogenesis regulation involving, *IRX3, IRX5,* and *ARID5B* in obesity[211], and in 2017, the complex I assembly factor, *TIMMDC1*, was established as a novel mitochondrial disease-gene by utilising genomic and transcriptomic sequencing.[193]

1.8 Constraint

Constraint concerns the observation that certain genes are more tolerant to mutation than others. This is of particular importance in rare genetic diseases, whereby many disease genes are constrained for mutation. With so many variants of uncertain significance identified in patients' sequencing data, methods capable of prioritising the best candidates for functional follow up have been an area of interest.

1.8.1 Constraint based approaches

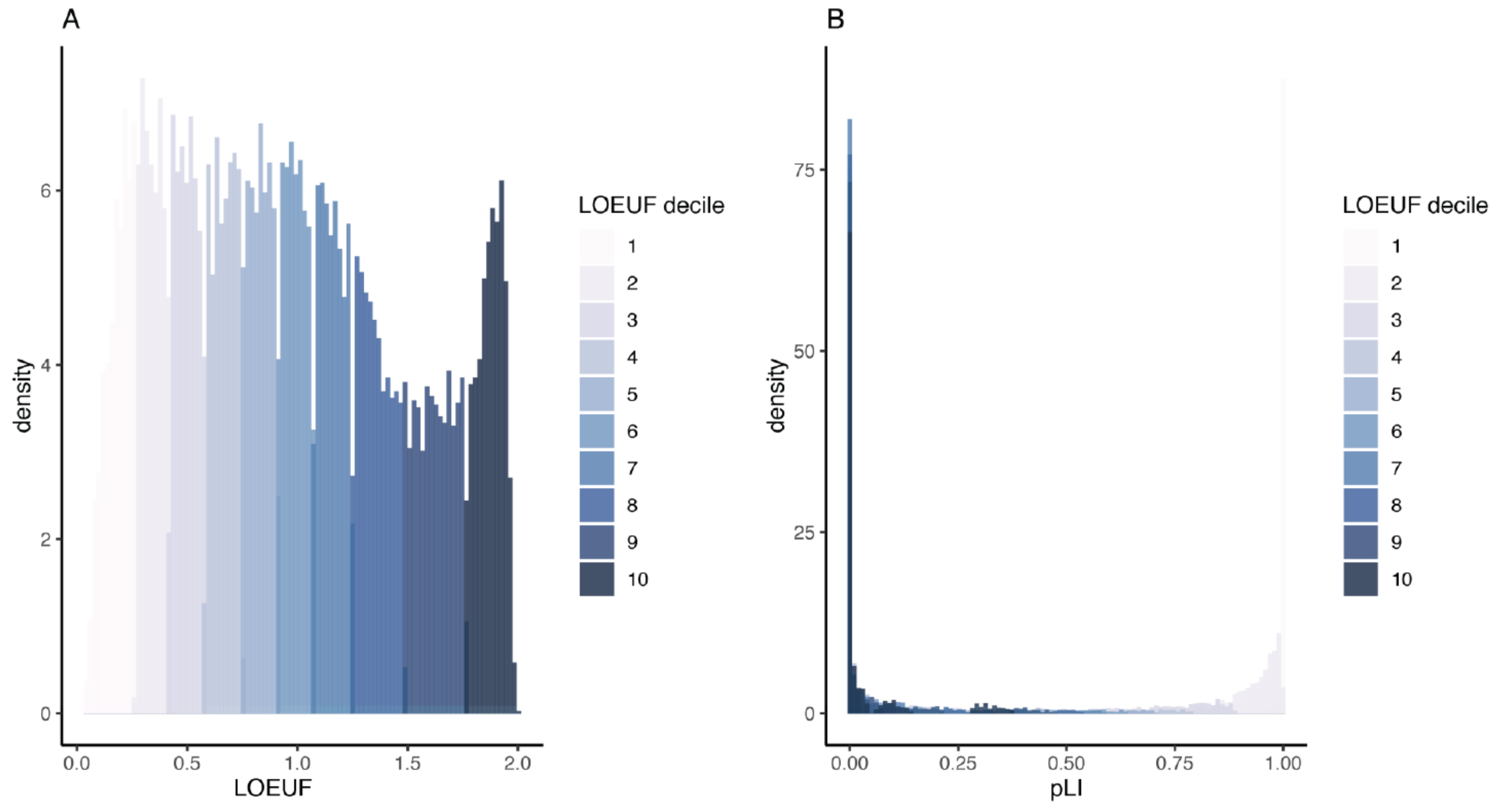
Given the mutation rate and the Earth's current population size, every variant compatible with life in a living human should theoretically be observed. Indeed, the aggregation of large population datasets has begun to reveal the spectrum of damaging variants across the human genome.[116,133] It is typical to observe approximately 100 loss-of-function (LoF) variants per genome with ~20 genes completely inactivated (knockouts) even in perfectly healthy individuals from the general population.[28] Population data can be utilised to evaluate the strength of natural selection at the gene level and to differentiate rare from common loss of function variants. As deleterious variants are purged from human populations through natural selection, there are opportunities to identify genes and regions that are constrained for variation compared to expected mutation rates, revealing which genes are most intolerant to inactivation of one (haploinsufficient) or both (knockout) copies.[133,212]

1.8.1.1 Loss-of-function constraint

In 2016, Lek *et al.*[213] defined a set of genes with high probability of intolerance to heterozygous predicted loss-of-function variation (pLI) modelled on ~60,000 exomes from the Exome Aggregation Consortium (ExAC) population database.[116] This pLI score can be used to identify candidate haploinsufficient disease genes constrained for loss-of-function in a dichotomous way; i.e. a gene is predicted to be haploinsufficient (pLI >0.9) or not (**Figure 1.5**).

In 2020, Karczewski *et al.*, refined the model and regenerated pLI scores utilising a larger dataset of ~141,000 exomes and genomes from the Genome Aggregation Database (gnomAD) (https://gnomad.broadinstitute.org).[133] The authors also developed the Loss-of-function Observed/Expected Upper-bound Fraction (LOEUF) score, a continuous metric which places >19,000 human genes on a spectrum of intolerance to knockout (**Figure 1.5**). Genes with the lowest LOEUF scores, i.e., the fewest predicted loss-of-function (pLoF) variants compared to an expectation, are the most constrained for loss-of-function, highlighting their potential biological essentiality. Both LOEUF and pLI were validated by comparison to several orthogonal indicators of constraint and shown to be accurate at discriminating haploinsufficient disease genes from autosomal recessive and polymorphic (unconstrained) genes.[133] A companion paper by Collins *et al.*[109] additionally showed that structural variants share the same pattern of constraint as LOEUF, and are responsible for about a quarter of all rare loss-of-function events per genome.

Figure 1.5 | Comparison of the distribution of pLI and LOEUF



*Window **A** shows a histogram of human genes across the LOEUF spectrum displaying a continuous pattern. Lower scores represent higher gene constraint (for loss-of-function). The histogram is coloured by LOEUF decile. Window **B** shows a histogram of human genes across the pLI spectrum. This spectrum is extremely dichotomous with many genes skewed towards either 0 (not constrained for loss-of-function) or 1 (constrained for loss-of-function). This can help to discriminate genes that are likely to cause disease through haploinsufficiency (pLI > 0.9). The dichotomous nature of pLI is by design, as initially the reference databases were too small to have adequate power to discern depletion for loss-of-function variation in small to medium length genes. The pLI distribution is coloured by LOEUF decile to show the overlap between scores. Higher pLI scores correlate with lower LOEUF scores as expected. The continuous nature of the LOEUF score provides more granular detail than pLI across the middle of the spectrum and can better stratify genes with moderate levels of constraint that may be implicated in recessive disease.*

As LOEUF identifies genes constrained for loss-of-function variation, there is an expectation that these genes would be enriched for dominant disease genes and to a lesser extent recessive disease genes. As of January 2021, 65% of genes in the lowest LOEUF decile are yet to have an Online Mendelian Inheritance in Man (OMIM) disease association (calculated using data from https://omim.org), highlighting thousands of high probability candidate disease genes awaiting discovery of the associated phenotypes.

Whilst a LoF variant in a LoF constrained gene is of particular interest in Mendelian disease, not all LoF variants called by NGS technologies are truly LoF. What may appear to be a LoF variant, may in fact escape nonsense mediated decay (NMD), through rescue mechanisms, such as in frame restoring splice sites or indels. Furthermore, LoF variants are commonly enriched for sequencing errors.[133,214] This means that careful prudence is necessary when interpreting LoF variants for diagnostic purposes.

1.8.1.2 Missense constraint

The majority of coding variants of uncertain clinical significance are missense variants, as reported in ClinVar, a public database where diagnostic laboratories and researchers share variant classifications (i.e. pathogenic, benign, uncertain significance).[215] Similar to methods for assessing loss-of-function constraint, methods to identify missense constraint have emerged by comparing the observed over expected number of missense variants modelled on population data.[29,30,116,216,217] However, missense constraint varies across a gene; for example, unstructured regions are often less constrained than important functional domains, which has necessitated the development of regional missense constraint models.[29,218] Furthermore, clustering patterns of pathogenic missense variants vary dependent on inheritance pattern. Turner *et al.*[219] showed that dominant missense variants cluster more than recessive variants. Therefore, testing for non-random clustering patterns may identify novel regions of interest across large sample sizes.[219] The application of these metrics has aided in the discovery of new gene disorders, including a *de novo* missense variant in a constrained region of *GABRA2* responsible for an early-onset epileptic encephalopathy (MIM: 618557).[220]

1.8.1.3 How constraint can help with diagnostics and novel gene discovery

Constraint metrics have become commonplace in the analysis and clinical reporting of variants. Gene level constraint is available, including missense constraint, pLI, and LOEUF scores in the gnomAD browser. Missense variants in missense constrained genes, and LoF variants in LoF constrained genes are typically of interest to clinical laboratories and can help with prioritising candidates for further review. Genes constrained for LoF, without a known disease association, are potential "yet to be discovered" disease genes and of interest in the research community; there is an expectation that LoF variants within these genes should manifest a phenotype.

1.8.1.4 Limitations of constraint metrics

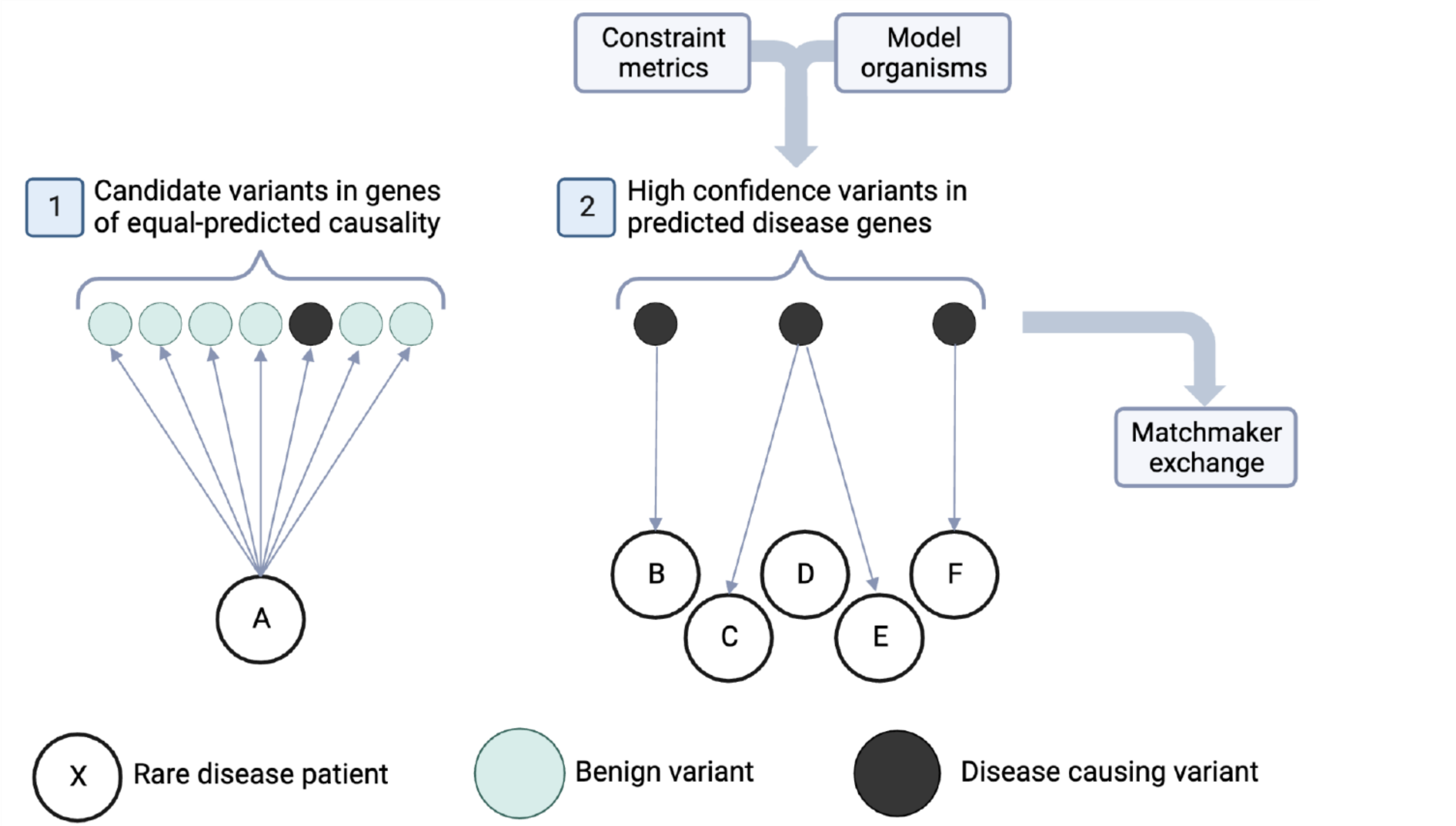
Caution is advised when interpreting constraint metrics such as pLI and LOEUF derived from population datasets; some pathogenic variants that would cause disease if present as germline variants may arise as somatic variants. Whilst many somatic variants are excluded from population databases using an allele balance between 20% and 80% for heterozygous variants, some may still evade this cut-off. Therefore, metrics based on presumed germline frequency, such as pLI and LOEUF, may be difficult to interpret in the context of known somatic mosaicism.[23]

1.9 Gene to patient approaches

Increasingly, government funding has invested in national sequencing projects for rare disease.[3,221] Despite 100,000 individuals with rare disease being sequenced in the UK as part of the Genomics England 100,000 Genomes Project (100KGP), the diagnostic rates are similar to those reported elsewhere in the literature.[3,222] However, the scale of such datasets welcomes opportunities for new approaches to novel gene discovery.

With increasing data available on genetic variation from a variety of sources including gene constraint, mouse models, phenotype-driven methods etc., there is scope to utilise the power of large cohort sizes for novel gene discovery. Instead of bringing a patient to a gene, there are opportunities, with large enough sample sizes, to be sufficiently powered to detect rare variation and bring candidate genes to large genomic datasets from patients (**Figure 1.6**). These “gene-to-patient” approaches can be applied to accelerate novel gene discovery and prioritise genes for functional studies.

FIGURE 1.6 | GENE-TO-PATIENT APPROACH FOR IMPROVED RARE DISEASE DIAGNOSTICS



*Scenario (**1**) shows a traditional patient-to-gene approach. Following variant analysis, rare disease patient A has several potential disease candidates, of which one (in black) is the disease-causing variant hidden within the sea of benign variation. Without prior knowledge that any of these variants are causative, the only way to test their pathogenicity is by expensive functional studies on genes of equal-predicted causality. In scenario (**2**), the approach is reversed. High-confidence disease-causing variants in genes identified by constraint metrics and model organism data can be matched to patients and compared to clinical phenotypes, circumventing the analytical noise precluding variant interpretation. In turn, this identifies the best candidates for follow-up and for data-sharing in the Matchmaker exchange. Variants/genes that match to more than one patient with the same or overlapping phenotypes can add credence to the method.*

1.10 Importance of novel gene discovery and improved diagnostic rates

1.10.1 Clinical impact of novel Mendelian conditions

Whilst novel gene discoveries widen the known functional repertoire of disease genes, the focus and drive are ultimately uplifting diagnosis rates and improving patient outcomes. There have been thousands of pivotal Mendelian discoveries throughout history, and each one is no more important than another, at least not for the families involved.

Since the discovery of the *CFTR* gene in 1989,[223] it is now possible to diagnose cystic fibrosis (MIM: 602421) rapidly, predict pancreatic functional status, and plan preventative care with modulator therapy.[224,225] In 2004, the discovery that hypermorphic or gain of function variants of *PCSK9* cause familial hypercholesterolemia type III (MIM: 603776)[226] has led to the successful development and Food & Drug Administration (FDA) approval of monoclonal antibodies against PCSK9, which is also used to treat non-familial forms of hypercholesterolemia.[227-230] As more collaborative, cohort-based studies have emerged in the NGS era, many candidate genes have been discovered that have directly impacted treatment and clinical outcomes. In one study on neurometabolic disorders, whole exome sequencing diagnosed 68% of patients and identified 11 novel candidate genes leading to a targeted intervention in 44% of patients.[231]

Diagnosing Mendelian disorders as a direct result of novel gene discovery not only impacts the primary patient involved but their families and caregivers. Families of children with rare genetic diseases are adversely impacted by lack of peer support groups and psychological support as well as delays in diagnosis.[232] Parents of children with rare disorders have called for better education, reduction in avoidable diagnostic delays, and early access to interventions and treatments.[233] For many, a genetic diagnosis can be life-changing, even in the absence of a therapeutic option.[234] Following diagnosis, quality of life is often improved by: participation in support groups that can provide longitudinal prognostic information, genetic counselling, and informed reproductive decisions with opportunities for pre-implantation genetic diagnosis or prenatal testing particularly in the case of inherited variants where there is a sizable recurrence risk.[235,236]

1.10.2 Financial impact of novel discoveries

Undiagnosed rare diseases are hugely expensive. A typical patient's diagnostic odyssey lasts an average of 8 years and costs a total of $5,000,000 throughout a patient's lifetime.[154] Two prospective Australian studies have shown that early exome sequencing is making significant head-way as a cost-saving diagnostic approach. Stark *et al.*[137] showed that integrating whole exome sequencing as a first-line test had an incremental cost saving per additional diagnosis of (converted to US dollars) $1,543 (95% CI: $92-$4,143). The cost per diagnosis was $4,248 (95% CI: $3,425-$5,588), $14,893 less than standard diagnostic care.[137] Tan *et al.* concluded that whole exome sequencing performed at initial presentation to tertiary care resulted in an incremental cost-saving of (converted to US dollars) $6,383 per additional diagnosis (95 CI: $3,045-$10,900) compared with standard diagnostic care.[237] However, cost-savings are only possible when sequencing can identify the causal variant. Therefore, every new genetic disorder identified, published, and shared in publicly available databases will have wide reaching diagnostic and cost-saving potential. Taking the average of costs saved per additional diagnosis from the two studies ($3,964) and extrapolating this on 100,000 patients could save an estimated $400 million US dollars.

1.11 Optimising diagnostic pipelines

1.11.1 Addressing the translational gap

Understanding of the genetic basis of rare disease is constantly changing with new genes and variation being linked to disease at a rapid pace. Given the direct application of these discoveries to the clinical diagnosis of rare disease in patients, guidance is needed for understanding what information is ready to be incorporated into clinical care and mechanisms are needed to quickly translate that information into medical practice. The Clinical Genome Resource (ClinGen) has developed a systematic framework for evaluating genetic and functional evidence for disease-gene relationships enabling their classification as definitive, strong, moderate limited, no human evidence, disputed or refuted with respect to their reported role in disease.[238] ClinGen supports Gene Curation Expert Panels that bring together international groups of disease and curation experts to evaluate gene-disease claims in their respective fields (https://clinicalgenome.org/affiliation). ClinGen's efforts are combined with other public and private gene curation efforts and are accessible within the Gene Curation Coalition database (thegencc.org). Currently, it is recommended that a gene-disease relationship reach moderate before it is included on predefined diagnostic gene panels for specific conditions.[239] However, when performing exome and genome approaches on individuals with rare disease, variation can be detected in genes that have not yet been linked to disease but may be strong candidates. Although practices vary between laboratories and countries, some professional standards recommend reporting these findings back to patients when there is a reasonable chance that new evidence may evolve over time to strengthen the gene-disease relationship, similar to the return of variants of uncertain significance in genes already linked to the patient's condition.[123,240] This approach also allows patients to be partners in solving the causes of rare disease.[241] It is hoped that such a framework will achieve global recognition and be universally adopted to ensure consistency in translating research findings into the clinic.

1.11.2 Looking to the future

It is estimated that by 2025, 60 million patients will have their genome sequenced in a research or healthcare setting.[242] While the sheer volume of data poses computational challenges, it also provides opportunities to learn more about the genetic architecture of health and disease. However, this necessitates improved methods for interpreting the spectrum of functional variation across all genes and particularly in the interpretation of non-coding variation, an area of investigation still in its infancy but beginning to make headway. Indeed, disruption of non-coding topological associated domains have been associated with limb malformations[105,243] and non-coding variants upstream of *PRDM13* and *CCNC* have been linked to North Carolina Macular Dystrophy.[244,245] Whilst efforts like the Atlas of Variant Effect Alliance are working towards achieving the mammoth goal of interpreting the impact of all genomic variation, there is still a long way to go.[246,247] It is expected that as data pours in across a variety of species and sources, more and more methods will adopt machine learning and deep learning techniques to find patterns and disease associations but the utility of these approaches are limited by the quality of the training data and other factors influencing data interpretation. For novel gene discovery, perhaps one of the most powerful resources would be to build a publicly available human knockout database that links naturally occurring null variants in genes and supportive functional evidence to shared human phenotype data. This is an exciting time for novel gene discovery - the end is by no means in sight.

Chapter 2 | Methods

This chapter describes some of the key methodology and tools applied in this thesis and aims to help signpost the reader to relevant sections in the 8 results chapters (**Chapters 3-10**).

2.1 Reference genome

The human reference genome was sequenced by 2003, taking 16 years to complete.[248] The reference genome represents a universal template to which all human sequences can be aligned and compared. Herein this section outlines the human reference sequence relevant to work undertaken in this doctoral thesis.

2.1.1 GRCh38

The Genome Reference Consortium Human Build 38 (GRCh38) is the assembly of the human genome released in 2013; it is accessible for download here: https://www.ncbi.nlm.nih.gov/grc/human. It was constructed from multiple donors and was sequenced base by base using Sanger sequencing. GRCh38 improves upon its predecessor, GRCh37, by providing a more complete reference sequence and altering 8000 nucleotides. Specifically, it uses more alternate configs (261 loci across 178 regions) to represent complex regions of the genome including the human leucocyte antigen (HLA) loci; corrects for sequencing artefacts previously represented in GRCh37; and has added centromeric sequence.[249]

2.2 Patient cohorts and data processing

2.2.1 100,000 Genomes Project dataset

The 100,000 Genomes Project, which started in 2014, was a government funded initiative, assigned to Genomics England (GEL) to deliver the sequencing of 100,000 whole genomes in the UK of 85,000 NHS patients and their families with rare diseases or cancer through 13 NHS Genome Medicine Centres (GMCs). GEL is a private company owned by the Department of Health and Social Care in the UK. Where possible, (parent/offspring) trios were recruited for rare disease to establish *de novo* variants, although more complex pedigrees existed if disease segregated through more family members. The project is now complete. The following subsections are relevant to results **Chapters 4-9**.

2.2.1.1 *Ethics, recruitment, and consent*

The 100,000 Genomes Project obtained ethical approval in 2015 (Research Ethics Committee (REC): 14/EE/1112; Integrated Research Application System (IRAS): 166046). Patient information sheets and consent forms are available at the Genomics England website: https://www.genomicsengland.co.uk/information-for-participants/participant-forms/.

Recruitment to the project was voluntary and involved informed consent by a trained healthcare professional for both diagnostic testing and research. For rare disease recruitment, eligibility involved suspicion of a rare disease (affecting < 1:2000) across a broad range of categories, that was likely monogenic or oligogenic in nature, and where patients had not been diagnosed after usual NHS care.[132] Where eligible, affected individuals and their close relatives were invited to join the study (usually an affected proband and their unaffected parents). Minors (under the age of 16) were consented by an authorised adult; although, once the patient turned 16, they had to provide their own consent to remain in the project. Patients of any age lacking capacity could have an authorised adult consent on their behalf. Individuals who had lost capacity after consenting for the project were removed from the study. Participants could opt out of secondary findings and carrier testing and withdraw from the project for any reason at any time. Recruitment for the project closed in 2018, although patients may still be offered whole genome sequencing on the NHS through genomic laboratory hubs.

2.2.1.2 *Data collection*

Blood, tissue, or saliva were collected from participants for genome sequencing. Phenotype data were provided by the treating clinician and recorded as HPO terms and later stored within LabKey, a data management system accessed behind the secure GEL firewall. Hospital episode data continue to be recorded throughout the patient's lifetime and is provided through NHS Digital.

2.2.1.3 *Accessing data securely and governance*

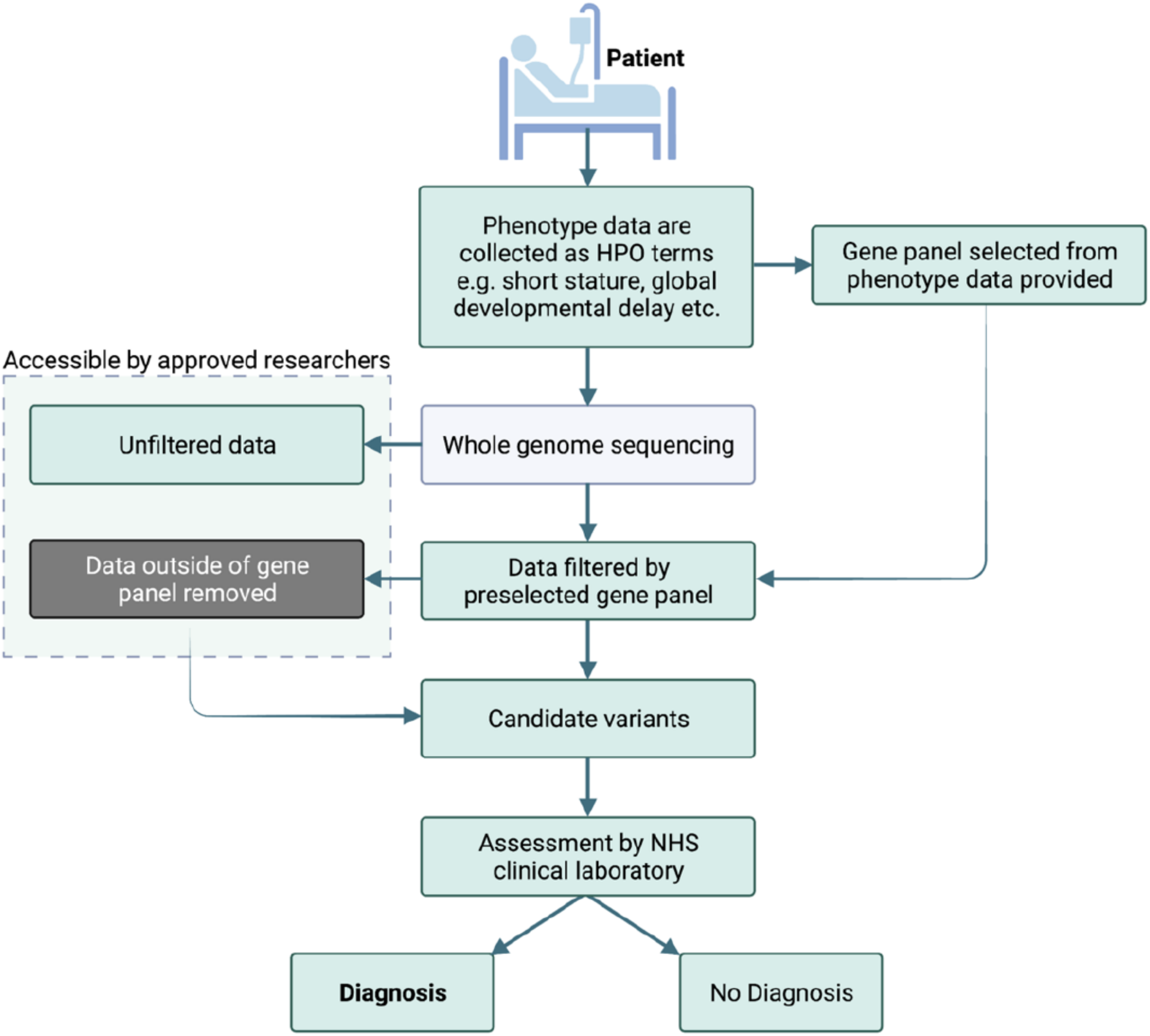
100,000 Genomes Project data are deidentified and accessible in a secure research environment (RE) to ensure that data from Genomics England is kept safely behind a firewall. Data from the project are typically released 2-3 times per year to reflect project updates and participant withdrawals; and to date there have been 16 releases, reflecting different numbers of participants in each release. There is software and a high-performance cluster built into the RE facilitating the analysis of data from the 100,000 Genomes Project. Technical specifics are outlined in **Methods 2.5.2**.

GEL ensures protection of participant data and is compliant with data protection legislation including GDPR. To ensure de-identified data, any information deemed "identifiable" such as name, address, and date of birth is removed and replaced with a unique identifier. Only aggregate phenotype or genotype data is allowed to be removed from the RE and this process is governed by an 'Airlock' Committee, which reviews requests. HPO terms or variants unique to less than 5 individuals within GEL are not allowed to be taken out of the RE. Variant level data associated with specific HPO terms (even if common) cannot be removed for any individual.

2.2.1.4 Data workflow

The fundamental workflow for the 100,000 Genomes Project is shown in **Figure 2.1.**

FIGURE 2.1 | 100,000 GENOMES PROJECT PATIENT TO DIAGNOSIS WORKFLOW


Patient
Phenotype data are collected as HPO terms e.g. short stature, global developmental delay etc.
Gene panel selected from phenotype data provided
Accessible by approved researchers
Unfiltered data
Whole genome sequencing
Data outside of gene panel removed
Data filtered by preselected gene panel
Candidate variants
Assessment by NHS clinical laboratory
Diagnosis
No Diagnosis

Patients recruited to the 100,000 Genomes Project have phenotyping data recorded as HPO terms. A gene panel is selected using PanelApp based on provided HPO terms. Whole genome sequencing is commenced. Where parental DNA is available, this is preferentially done as a (parent/offspring trio). Whole genome sequencing data are filtered by the pre-selected gene panel. Candidate variants in Tier 1 and Tier 2 genes are assessed by NHS accredited diagnostic laboratories. A report is generated and returned with a diagnosis, or no diagnosis. Data outside of the gene panel, not assessed by clinical laboratories are available for approved researchers to analyse. Any potential candidate variants (outside of the gene panel applied) can be referred to Genomics England for consideration of assessment by an NHS clinical laboratory. HPO – human phenotyping ontology; WGS – whole genome sequencing.

2.2.1.5 Informatics pipeline

All samples were sequenced with 150 base pair (bp) paired end reads in a single lane of the Illumina HiSeq 2500 instrument. Data were uniformly processed on the Illumina North Star V4 whole genome sequencing workflow (NSV4, V2.6.53.23). Samples were aligned to either GRCh37 or GRCh38. Most samples aligned to GRCh37 were later re-aligned to GRCh38.

BAM and VCF files that passed initial quality control (QC) checks were delivered by Illumina; germline samples required a sequencing quality of at least 30, and alignments had to cover at minimum 95% of the genome at 15x. *De novo* variant calling of single nucleotide variants and indels was performed using the Platypus variant caller. BAM and variant call files (VCF) are accessible for each participant in the RE.

GEL also make available an aggregate multi-sample VCF comprising 78,195 germline genomes. This dataset was generated by aggregating single sample gVCF files using gVCF genotyper. Genomic annotation was performed using Ensembl's Variant Effect Predictor (VEP) v85 and annotated against RefSeq. The multi-sample VCF is split into 1371 'chunks' of roughly equal size and includes the autosomes, sex chromosomes and mitochondrial positions.

2.2.1.6 Variant prioritisation and tiering

Participants' genomes were analysed in the context of their family structure using segregation analysis. PanelApp[118] virtual gene panels were applied to single nucleotide variants and indels as the primary filtering strategy, with removal of non-coding variants and common variants assuming a dominant disease with an allele frequency >0.001 in gnomAD, and an allele frequency >0.01 for presumed recessive models. Variants identified using the GEL pipeline were tiered to aid NHS diagnostic laboratories in their evaluation of primary findings (**Figure 2.2**). Protein coding variants not in the virtual gene panel applied were not expected to be assessed by NHS accredited clinical laboratories.

FIGURE 2.2 | GENOMICS ENGLAND VARIANT TIERING SYSTEM

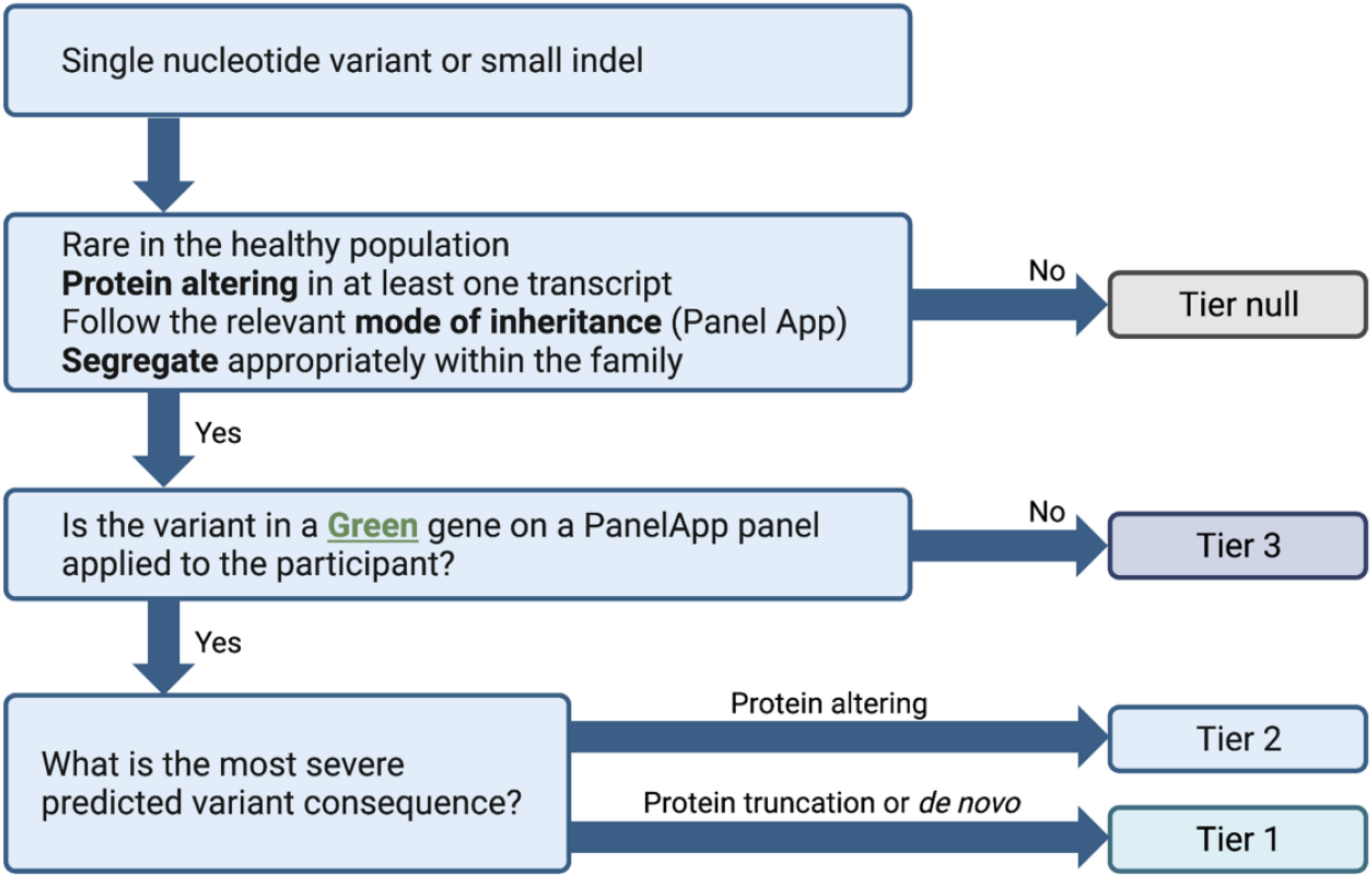

Single nucleotide variant or small indel
Rare in the healthy population
Protein altering in at least one transcript
Follow the relevant **mode of inheritance** (Panel App)
Segregate appropriately within the family
No
Tier null
Yes
Is the variant in a Green gene on a PanelApp panel applied to the participant?
No
Tier 3
Yes
What is the most severe predicted variant consequence?
Protein altering
Tier 2
Protein truncation or *de novo*
Tier 1

Figure adapted from https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38046769.
**Rare is defined by an allele frequency <0.001 for autosomal dominant and allele frequency <0.01 for autosomal recessive inheritance.*

2.2.1.7 *Returning results*

NHS accredited diagnostic laboratories review Tier 1 and Tier 2 variants and generate a clinical report. The outcome of these reports is available within LabKey in the 'GMC exit questionnaire' table.

2.3 Reference datasets

With increasing availability and affordability of next generation sequencing, efforts to aggregate population exome and genome data have become ever paramount. This section concerns reference data applied in **Chapters 3-10**.

2.3.1 gnomAD

The Genome Aggregation Database (gnomAD) is a freely accessible web-based resource which aggregates exome and genome data from a variety of international sequencing projects.[133] Over 195,000 individuals are included, and >140 principal investigators have contributed data from >60 studies. These data, to some extent, represent the 'general population', meaning that data are included from individuals with common, and complex diseases such as psychiatric disorders, cardiovascular diseases, and type 2 diabetes etc. However, gnomAD is specifically depleted for severe paediatric diseases, but some individuals with Mendelian disease may be included.[25] European participants are over-represented in gnomAD, with suboptimal representation of many communities, including Middle Eastern, Australasian, and African populations, inflating rare variants in these communities.

All the aggregated data have been processed and joint-called through the same BWA-Picard-GATK pipeline and *Hail* for data processing and analysis. Joint calling (on multiple samples) facilitates improved variant detection with greater sensitivity, particularly for low frequency variants. Rigorous quality control ensures the quality of the data and first and second-degree relatives are excluded to mitigate inflated allele frequencies. The gnomAD database (https://gnomad.broadinstitute.org) is openly available to the scientific community through a web browser which showcases variant and gene level statistics, including quality metrics and population allele frequencies across many ancestries. To maintain participant anonymity, data presented are such to align with the recommended guidance in the NIH data sharing policy (NOT-OD-03-032).[25]

Currently, there are two widely applied versions of gnomAD. The v2.1.1 dataset is aligned to (GRCh37/hg19) and annotated using VEP v85 using GENCODE v19 and includes 135,748 exome sequences and 15,708 whole genome sequences from unrelated individuals. The v3.1.2 dataset is aligned to (GRCh38) and annotated with VEP v95 using GENCODE v29 and spans 76,156 genomes. Full gnomAD access is available through Google Cloud Storage: gs://gcp-public-data--gnomad.

2.4 Matchmaker Exchange (MME)

MME (https://github.com/ga4gh/mme-apis) connects datasets containing both phenotype and genotype information through an Application Programming Interface (API) which enables searching of multiple databases with a single query and returning a list of profiles that match with the initial query.[157] Within Genomics England, MME is utilised by submitting genes to GeneMatcher[250], a node of the MME API. MME is utilised in **Chapters 4-6, 8-10**.

2.5 Bioinformatics and programming

The key bioinformatics and programming tools applied to the research presented are outlined in **Table 2.1**.

TABLE 2.1 | BIOINFORMATICS AND PROGRAMMING TOOLS AND SOFTWARE APPLIED

Tool	Description
R	A programming language and free software environment for statistical computing and graphics
Bedtools[251]	Flexible suite of tools for a wide range of genomic analysis tasks
VEP[252]	Variant annotation software that determines the effects of variants on genes, transcripts, protein sequences, and regulatory regions
LOFTEE[133]	Filters and flags erroneous or low confidence loss of function variants
OMIM API	Programmatic interface to query the OMIM database
LabKey	A software suite to analyse, integrate, and share biomedical data
IGV[253]	Interactive tool to visualise and explore genomic data

Abbreviations: API – application programme interface; IGV – Integrative genomics viewer; LOFTEE – loss of function transcript effect estimator; OMIM – Online Mendelian Inheritance in Man; VEP – variant effect predictor.

2.5.1 R packages

R is a programming language commonly utilised to manipulate and analyse data. R offers a range of packages to support data analysis. R has been used to process data in results chapters (**Chapters 3-4, 7-10**). The packages most utilised in this thesis are shown in **Table 2.2**.

TABLE 2.2 | MOST FREQUENTLY USED R PACKAGES

Package	URL
stringi	https://stringi.gagolewski.com/
magrittr	https://magrittr.tidyverse.org/
tidyverse	https://www.tidyverse.org/
viridis	https://github.com/sjmgarnier/viridis
patchwork	https://patchwork.data-imaginist.com/
lubridate	https://rdrr.io/cran/lubridate/man/lubridate-package.html
dplyr	https://dplyr.tidyverse.org/
ggplot2	https://ggplot2.tidyverse.org/
shiny	https://shiny.rstudio.com/
plotly	https://plotly.com/r/
Table1	https://rdrr.io/cran/table1/man/table1.html
UpSetR	https://github.com/hms-dbmi/UpSetR

2.5.2 GEL Research Environment

The GEL Research Environment (RE) is a secure virtual desktop consisting of two main platforms:

1) A Linux virtual desktop interface hosted by Inuvika
2) High performance cluster (HPC) – Helix.

The RE is accessible by login and provides access to software and tools, however it has limited memory and processing power. Intensive analyses require job submission to Helix, which has 50 nodes, operates on CentOS 7.6.1810 with 34 cores/node, with 1,700 job slots available. Each node can accommodate 22 batch jobs. For most Helix nodes, an IBM Load Sharing Facility platform is used to schedule jobs. Service modules available on the HPC are listed in **Table 2.3**. Commonly used environmental modules include BCFTools, BEDtools, BWA, CADD, FASTQC, GATK, Lumpy, PLINK, Pindel, SAMtools, STAR, Salmon, VEP, Picard, tabix and verifyBamID. A full list of Environmental Tools is available in **Supplementary Table S1**.

TABLE 2.3 | SERVICE MODULES AVAILABLE ON HELIX

Tool	Version	Module Load
EasyBuild	3.9.2	EasyBuild/3.9.3
TurboVNC	2.2.2	TurboVNC/2.2.2
autotools	-	autotools
charliecloud	0.9.7	charliecloud/0.9.7
clustershell	1.8.1	clustershell/1.8.1
cmake	3.14.3	cmake/3.14.3
gnu	5.4.0	gnu/5.4.0
gnu7	7.3.0	gnu7/7.3.0
gnu8	8.3.0	gnu8/8.3.0
hwloc	2.0.3	hwloc/2.0.3
llvm4	4.0.1	llvm4/4.0.1
llvm5	5.0.1	llvm5/5.0.1
papi	5.7.0	papi/5.7.0
paraver	4.8.1	paraver/4.8.1
pmix	2.2.2	pmix/2.2.2
prun	1.3	prun/1.3
rstudio-server	1.1.463	rstudio-server/1.1.463
rstudio_singularity	1	rstudio_singularity/1
singularity	3.2.1	singularity/3.2.1
valgrind	3.15.0	valgrind/3.15.0
websockify	0.8.0	websockify/0.8.0

As of 2023, the GEL research environment has transitioned to a Trusted Research Environment (TRE) on a cloud-based platform, supported by Amazon Web Services. The TRE hosts a linux virtual desktop and the Helix HPC.

2.5.3 Southampton High Performance Cluster – Iridis

The Southampton high performance cluster (Iridis) is one of the largest supercomputers in the UK. Access to Iridis is offered through application to researchers and post-graduates. Iridis facilitates running high intensity sequential jobs, or parallel jobs with distributed memory.

The latest generation of Iridis, Iridis5, boasts:

- 2.2 PB of storage
- 464 compute nodes each with 50 CPUs per node with 192 GB of memory
- Graphics cards including:
  - 20 Nvidia Tesla V100
  - 40 Nvidia HTX 1080Ti
- 4 high memory nodes each with 64 cores, 768 GB of RAM and 9 TB of local scratch space
- 2 data visualisation nodes with 22 usable cores, 384 GB of RAM and an Nvidia M60 GPU
- 3 login nodes with 40 cores and 384 GB of memory
- 20,000 processor cores provide 1,305 TFlops peak.

Iridis5 is accessible using the bash command line based on a UNIX stricture.

2.5.4 Broad Institute Cloud Computing

The Broad Institute no longer uses an HPC, but instead has collaborated with Google Cloud to form Google Genomics. Google Genomics provides a pay-per-use service to the Broad Institute, whereby data are stored using Google Buckets. The Translational Genomics Group at the Broad Institute utilises this service. The majority of compute used is from Kubernetes clusters. Compute usage for specific Broad related projects is shown in **Table 2.4**.

TABLE 2.4 | GOOGLE CLOUD SERVICES COMPUTE USAGE BY THE TRANSLATIONAL GENOMICS GROUP AT THE BROAD INSTITUTE

Project	Persistent vCPUs	Persistent RAM (GB)	# Of Persistent VMs	GCS Bucket Data Storage (TB)	Block Storage (VM Disks) (TB)	Other (Google) Cloud Services Utilised	Dynamic Compute (Dataproc) vCPU Hours (Nov 2021)	Average number of running Dataproc Worker nodes (assuming 8 vCPUs per worker)
gnomAD	50	224	6	343.3	21.9	Cloud Functions	130143	22
Seqr	64	402.5	15	87	19.3	CloudSQL (3 Postgres databases), Cloud Composer (1 environment), Cloud Functions, CloudBuild	18496	3
ClinGen	34	150.25	17	2	2.4	BigQuery, Cloudbuild, Confluent.Cloud kafka hosting	0	0
Totals	148	776.75	38	432.3	43.6		148639	

GB – gigabytes; GCS – Google Cloud Services; RAM – random access memory; TB – terabytes; vCPU – virtual central processing unit; VM – virtual machine.

2.6 Annotation tools

2.6.1 Constraint metrics

The availability of population datasets such as gnomAD has enabled research into variant level and gene constraint, by the comparison of expected rates of mutation compared with what is observed. Constraint metrics quantify this level of constraint and can help prioritise genes and variants in genomic analyses. This section lists the most applied LoF constraint metrics and software available to aid in genomic data interpretation, as showcased in **Chapters 3-7**.

2.6.1.1 pLI

The probability of LoF intolerance (pLI) is a constraint metric originally modelled on 60,000 exomes from the ExAC[22] database. The score is derived by comparing the number of observed pLoF variants (stop gained, frameshift, and essential splice site) in a population with the number expected under neutrality. The model assumes that the number of pLoF variants in a gene obeys a Poisson distribution with mean *λM,* whereby M is the expected number of pLoF variants in a sample under neutrality, and *λ* is the depletion in the number due to selection. Genes were categorised as followed:

- Neutral: λ_{Null}=1
- Recessive: λ_{Rec}=0.463
- Haploinsufficient: λ_{HI}=0.08.

Haploinsufficient and recessive *λ* values were derived by calculating the average reduction of pLoF variants in known recessive and haploinsufficient disease genes. The authors estimated the proportion of genes in each of the three categories above and obtained the maximum a posteriori probability that any given gene belonged to a given category. Scores >0.9 are considered to represent high probability haploinsufficient disease genes, extremely intolerant to LoF.

2.6.1.2 LOEUF

The Loss-of-function Observed/Expected Upper-bound Fraction (LOEUF) builds upon the pLI score. Modelled on 141,000 exomes and genomes from gnomAD, LOEUF assesses the degree of intolerance to pLoF variation for each gene using a continuous metric of the observed/expected ratio, and then applies an upper bound of a Poisson-derived confidence interval (at 90%) around the ratio. This upper bound is a conservative one directional metric, meaning that genes with low LOEUF scores <0.35 are depleted for pLoF variants, but genes with high LOEUF scores comprise genes without depletion and genes too small to estimate a precise observed/expected ratio.[133]

2.6.1.3 LOFTEE

The Loss-Of-Function Transcript Effect Estimator (LOFTEE) is a suite of tools (https://github.com/konradjk/loftee) that helps automate the removal of pLoF variants, likely to represent artefacts or false positives. LOFTEE applies stringent filtering criteria to remove pLoF variants likely to escape nonsense-mediated decay and is available as a VEP plugin. The LOFTEE V1.0 software has been applied in **Chapters 3, 4** and **7.** LOFTEE hard filtering and flags are shown in **Table 2.5**.

Table 2.5 | LOFTEE hard filtering criteria and optional flags

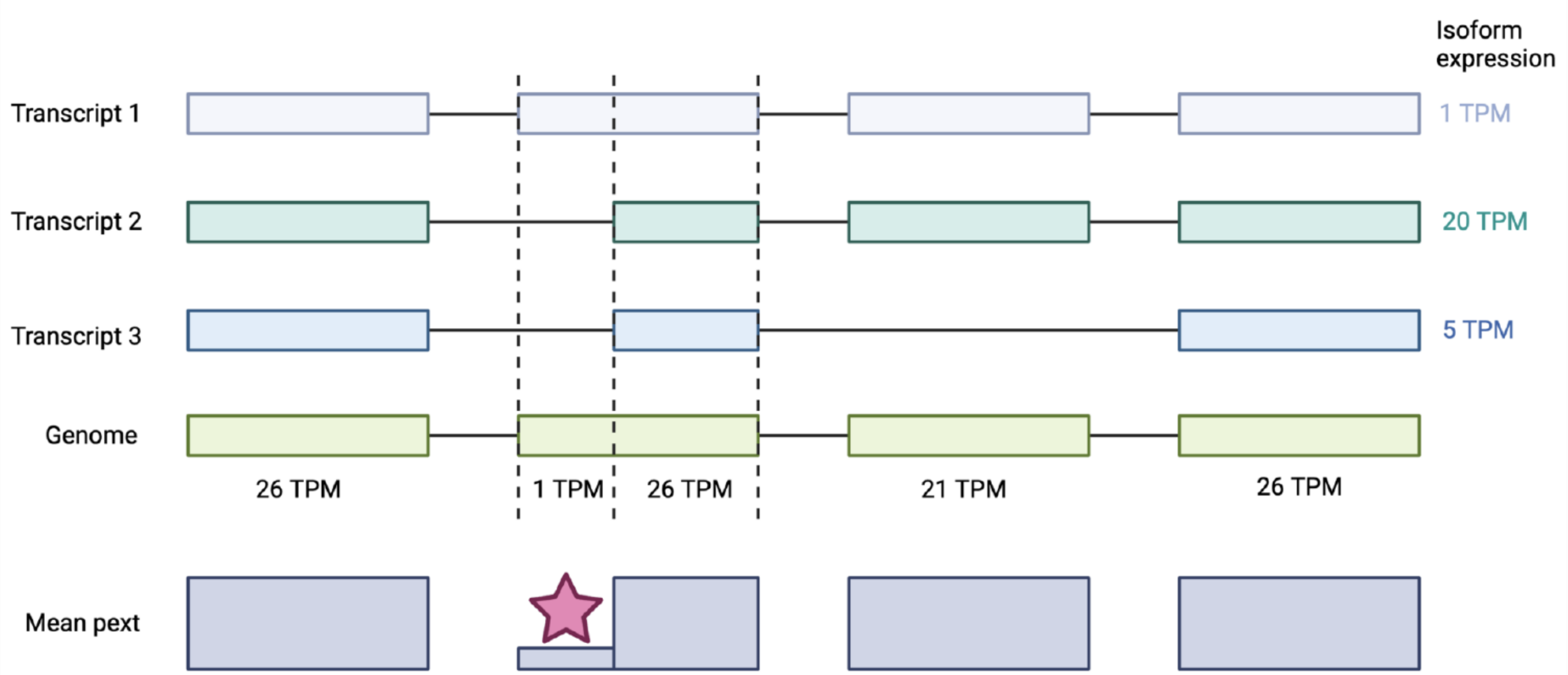
	Hard filters (variant removal)	Flags (non-stringent)
All variants	Variants whereby the loss of function allele is the ancestral state	N/A
	Variants in incomplete GENCODE transcripts	
Frameshift and stop gained variants	Variants in the last exon or within 50bp of the end of the transcript	Single exon genes
	Variants in an exon not contained within canonical splice sites	Exons devoid of the evolutionary protein-coding signature based on PhyloCSF
Splicing variants	Variants not predicted to affect a donor site	Variants in NAGNAG sites
	Splicing variants only disrupting untranslated regions	Intronic variants with a non-canonical splice site
	Variants not annotated by MaxEntScan	N/A
	Variants rescued by nearby in frame splice sites (within 15 base pairs)	
	Variants in small introns <15 base pairs	

Hard filters represent variants removed by LOFTEE. Flags are non-stringent and optional – these are highlighted but not removed by LOFTEE.

2.6.2 Proportion expressed across transcript (pext) score

In 2020, Cummings *et al.*[214] developed the proportion expressed across transcript (pext) score to aid in the interpretation of transcript-specific annotation. Pext is a tool that facilitates the rapid visualisation of isoform expression values across tissues using GTEx data. Pext can differentiate between weakly and highly evolutionarily conserved exons and thus serves as a surrogate for functional importance (**Figure 2.3**). Pext is applied in **Chapters 3- 7, 9-10**.

FIGURE 2.3 | SCHEMATIC OF HOW PEXT ENABLES VISUALISATION OF ISOFORM EXPRESSION ACROSS EXONIC REGIONS

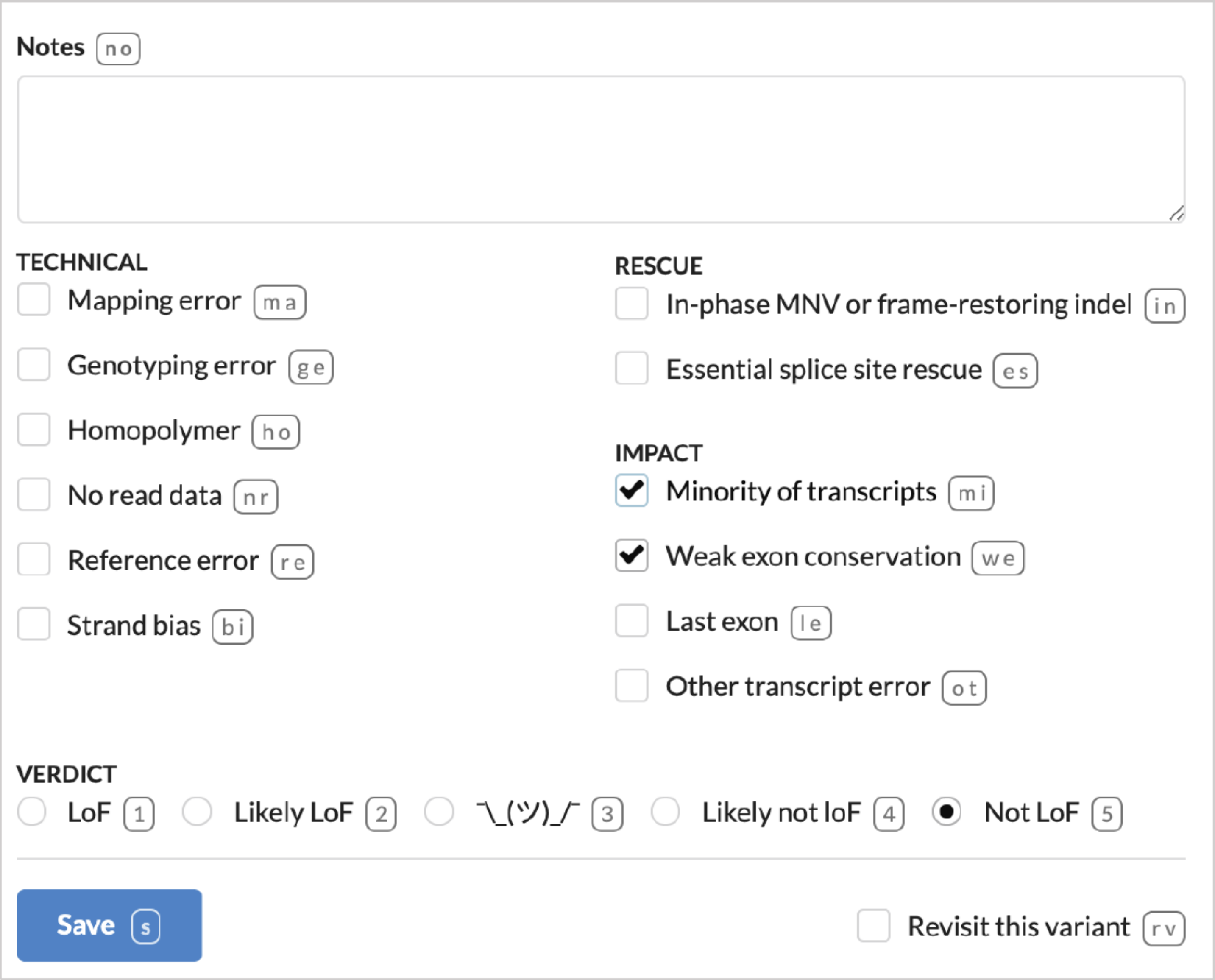


Pext values are integrated from the GTEx dataset by computing the median expression of a transcript for GTEx tissue samples. The expression of a given base is defined as the summed expression of all transcripts that touch that base. This is repeated for every GTEx tissue and then normalised by the whole gene expression in that tissue. Transcript 1 has low tissue expression compared with transcripts 2 and 3. The overhanging region of exon 2 (on transcript 1) has a base pext of 1 TPM, and therefore, this region is likely to have low conservation or be enriched for annotation errors. Therefore, a variant in the region of the pink star would be easily identified as a low priority candidate due to the relative drop in mean pext secondary to low tissue expression in GTEx across that region. The caveat to this rule is where transcript 1 is uniquely expressed in the tissue of interest, and therefore, expression values should be assessed against tissues of interest, rather than just the mean expression across all tissues. The integration of pext into the gnomAD browser allows the analyst to visualise both mean pext and tissue specific-pext. GTEx, genotype tissue expression; TPM, transcript per million.

2.6.3 Curation portal

A bespoke curation portal (https://github.com/macarthur-lab/variant-curation-portal) developed at the Broad Institute using *Hail* (https://hail.is) provides a platform to curate loss of function variants using a web application. The custom portal includes access to gnomAD variant pages with associated integrative genomics viewer (IGV) reads and the University of California, Santa Cruz (UCSC) browser for fast analysis and interpretation of pLoF variants identified in gnomAD. The portal includes a panel for rapid "tagging" of variants with error modes and flags, in addition to a free text box for comments, and the option to classify variants on a Likert scale from LoF to not LoF (**Figure 2.4**). The portal was applied in **Chapter 3**.

FIGURE 2.4 | CURATION PORTAL WEB INTERFACE FOR TAGGING AND CLASSIFYING VARIANTS



Notes no

TECHNICAL

□ Mapping error ma

□ Genotyping error ge

□ Homopolymer ho

□ No read data nr

□ Reference error re

□ Strand bias bi

RESCUE

□ In-phase MNV or frame-restoring indel in

□ Essential splice site rescue es

IMPACT

☑ Minority of transcripts mi

☑ Weak exon conservation we

□ Last exon le

□ Other transcript error ot

VERDICT

○ LoF 1 ○ Likely LoF 2 ○ ¯_(ツ)_/¯ 3 ○ Likely not loF 4 ● Not LoF 5

□ Revisit this variant rv

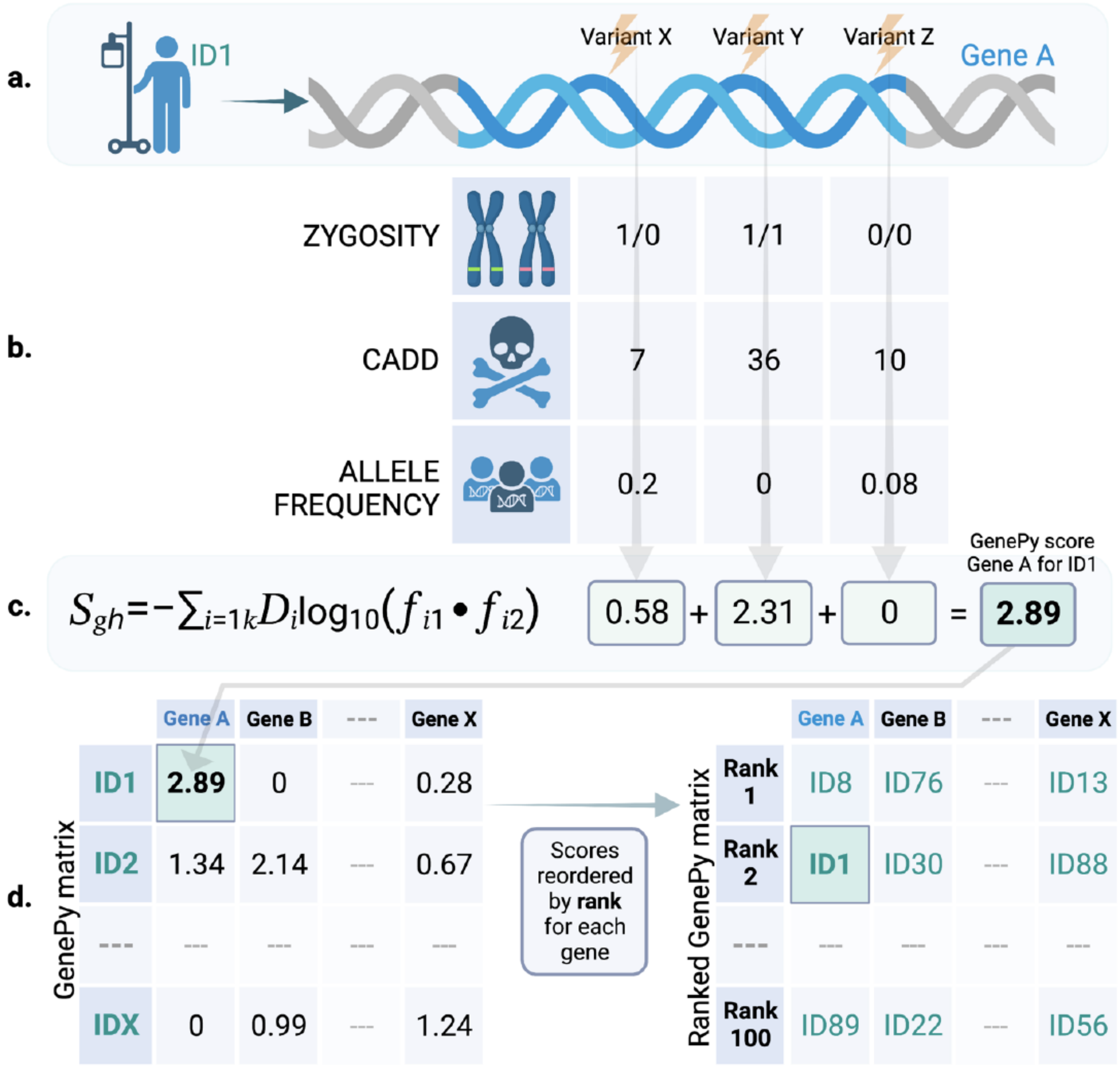
MNV – multi-nucleotide variant; LoF – loss of function.

2.7 Statistical methods

2.7.1 GenePy

At a high level, GenePy[81], is a dimensionality reduction algorithm developed at the University of Southampton. The software is open source, and the latest version is available here: https://github.com/UoS-HGIG/GenePy-1.4. GenePy rescales variant data into gene-level data for every individual in a given cohort. GenePy represented a genetic mixed model that integrates: 1) variant population frequency; 2) deleteriousness (reflecting pathogenicity, conservation, regulator effects); and 3) variant zygosity (i.e. the allele inherited from each parent). GenePy generates a single score per variant (and multiples the effects of two alleles at a single diploid locus) and then sums all the variant scores across a gene to generate a single score (**Figure 2.5**). Therefore, the contribution of all variation observed within a gene is modelled in an additive fashion, whereby the cumulative effect of pathogenicity from variants of varying effect sizes can be added to generate a single gene burden score, per gene, per individual.

FIGURE 2.5 | OVERVIEW OF HOW GENEPY WORKS


a.
ID1
Variant X
Variant Y
Variant Z
Gene A
b.
ZYGOSITY
1/0
1/1
0/0
CADD
7
36
10
ALLELE FREQUENCY
0.2
0
0.08
c.
$S_{gh}=-\sum_{i=1k} D_i \log_{10}(f_{i1} \bullet f_{i2})$
0.58 + 2.31 + 0 = 2.89
GenePy score Gene A for ID1
d.
GenePy matrix
Gene A | Gene B | --- | Gene X
ID1 | 2.89 | 0 | --- | 0.28
ID2	1.34	2.14	---	0.67
IDX	0	0.99	---	1.24
Scores reordered by rank for each gene
Ranked GenePy matrix
Gene A | Gene B | --- | Gene X
Rank 1 | ID8 | ID76 | --- | ID13
Rank 2	ID1	ID30	---	ID88
Rank 100	ID89	ID22	---	ID56

__a.__ Patient's DNA undergoes sequencing and subsequent processing to produce a file listing all variants identified in their data. __b.__ Each variant is individually annotated with biological information reflecting: zygosity i.e. the allele inherited from each parent; deleteriousness (D - can be user specified) and; frequency of the observed alleles (f) for in gnomAD. __c.__ These data are input into the GenePy algorithm for each variant and then summed across all variants observed within that gene for that individual. This step is run in parallel for all genes across all patients within the cohort. __d.__ The output is a matrix of all individuals by all genes. For certain applications, this matrix can be transposed such that for each gene, individuals are ordered by highest pathogenic variant loading.

For each individual sample *h* within a cohort $H = [h_1, h_2, \ldots, h_n]$, the loss of integrity of any given gene *g* in the RefGene database $G = [g_1, g_2, \ldots g_m]$ can be quantified as the sum of the effect of all (*k*) variants within its coding region observed in that sample, where each biallelic mutated locus (*i*) in a gene is weighted according to its predicted allele deleteriousness (D_i), zygosity and allelic frequency (f_i). The GenePy score S_{gh} for a given gene (*g*) in individual (*h*) is:

$$\boxed{S_{gh}=-\sum_{i=1k}D_i\log_{10}(f_{i1}\bullet f_{i2})}$$

At any one variant locus (*i*), both parental alleles are represented using f_{i1} and f_{i2} to embed the population frequency of allele$_1$ and allele$_2$ and, in doing so, model observed biological information on both frequency and zygosity. Any homozygous genotype therefore is simply the observed allele frequency squared whereas the product of each of the observed alleles is calculated for heterozygous genotypes. The latter can therefore accommodate variant sites with multiple alleles in addition to the typically encountered biallelic single nucleotide polymorphisms (SNPs). Hemizygotic variation from male X-chromosomes are treated as homozygotic. Where a variant may be novel to an individual or absent from reference databases, a lower frequency limit of 0.00001 is imposed. This lower limit is arbitrarily set to conservatively reflect the lowest frequency that can be observed in the largest current repository of human variation, gnomAD. The log function is applied to upweight the biological importance of rare variation.

2.8 Animal methods

2.8.1 Xenopus methods

These methods were performed by Dr Annie Godwin and Professor Matt Guille at the European Xenopus Resource Centre at the University of Portsmouth and are provided for reference.

2.8.1.1 Xenopus tropicalis husbandry

To generate all founder animals, female *X. tropicalis* were primed with 10 IU Human Chorionic Gonadotropin (Chorulon, Intervet) and boosted the next morning with 100IU. Each egg clutch was fertilized with cryopreserved *X. tropicalis* spermatozoa. All subsequent *ddx17* crispant generations occurred through natural mating. Male and female *X. tropicalis* were primed with 10 IU Human Chorionic Gonadotropin and boosted on the morning of mating with 100IU to obtain embryos. Following the boosting dose, animal pairs were placed in 16L buckets (containing 8L (50%) system water), under minimal lighting, at 24ºC - 25ºC and left to mate for 8-10 hours prior to embryo collection. All boosted and primed females were then placed in a temporary fill-and-dump tank comprising system water with additional salt (approximately 1.8 g per 6 L system water). The water was replaced every 24 hours, or sooner if visibly dirty. F_1 embryos were always obtained from genetically altered females with either wild-type males (F_1 heterozygous crispant offspring), transgenic males [Xtr.Tg(tubb2b:GFP)Amaya] RRID: EXRC_3001 males (F_1 heterozygous crispant offspring), or crispant mutant males (F_1 homozygous crispant offspring).

Xenopus embryos were cultured in 0.05X Marc's Modified Ringer's solution at 24 - 25ºC under 13-11 hour light dark cycles in a fill-and-dump set-up with 50% media changes every other day. Twice-daily health checks were also performed. At free-feeding stages, tadpoles were maintained in the fill-and-dump set-up, fed Sera Micron once daily with twice-daily 50% media changes. At stage NF50 (according to Nieuwkoop and Faber (NF))[254] tadpoles were re-housed within the recirculating MBK Ltd systems, under conditions equivalent to adult animals but with a lower water flow rate and 100% replacement of water every 10 days. All froglets were fed crushed Skretting Horizon 2.3 mm trout pellets twice daily (5 days per week) and once daily (2 days per week). Adult frogs were fed Skretting Horizon 2.3 mm trout pellets at least once daily 4 days per week. Juvenile frogs under 1 year of age were fed every day.

2.8.1.2 Generating and analysing crispant Xenopus tropicalis

The target regions within *ddx17* and *hdlbp* were identified in Xenbase, and the single-guide RNAs (sgRNAs) were designed using v10 of the *X. tropicalis* genome. Two single-stranded oligonucleotide templates for sgRNA synthesis were selected for high mutagenic activity, minimal predicted off-target events[255], and a high frameshift frequency.[256] Following the Taq-based method[257], single-stranded oligonucleotides (Invitrogen, UK) containing the T7 promoter were annealed and extended with the universal CRISPR oligonucleotide; this template was then transcribed with a T7 Megashortscript kit (Invitrogen, UK). The resulting sgRNAs were purified with SigmaSpin™ Sequencing Reaction Clean-Up columns (Sigma-

Aldrich), quantified with a NanoDrop 1,000 spectrophotometer (Thermo Fisher Scientific, Loughborough, UK), analysed by agarose gel electrophoresis, and stored as single-use aliquots at −80°C.

Across all founder animal experiments (unless otherwise stated), 600 pg of each sgRNA was co-injected with 2.6 ng Cas9 protein (Spy cas9 NLS, New England Biolabs) into a single-cell (unless otherwise stated) *X. tropicalis* embryos. Experiments analysing the effect of gene-editing on one side of the embryo included an additional 0.4 mg/mL Dextran Texas Red (ThermoFisher Scientific, D1863) tracer to the injection mixture. The efficiency of indel formation was assessed in genomic DNA from crispant embryos. Lysates were prepared from embryos collected at NF20, NF41-42 and from tail clips at NF48-50 by incubation at 56°C for 2 hours in 50 mM Tris, 1 mM EDTA (Sigma-Aldrich), 0.5% [v/v] Tween 20 (Fisher Scientific), 100 µg/mL Proteinase K (pH 8.5, Ambion - ThermoFisher). Proteinase K was inactivated for 15 minutes at 95ºC, and all samples were stored at -20ºC for up to 1 year. PCR amplification primers of the target region were designed with Primer3[258] software and BLASTed against v10 of the *X. tropicalis* genome (Xenbase). Template gDNA samples were amplified using GoTaq® G2 Green Master Mix (Promega Corporation) in a Thermal Cycler (VeritiPro™ Thermal Cycler, 96 well – ABI). Amplicons from the heterozygous *ddx17 and hdlbp* mating(s) were screened for mutants using the T7 endonuclease I (T7EI, New England Biolabs Ltd) mismatch detection assay prior Sanger confirmation. All remaining amplicons were visualised by agarose or polyacrylamide gel electrophoresis[259], column purified (SmartPure PCR Purification Kit, Eurogentec, Belgium) and Sanger sequenced (Azenta, UK).

2.8.1.3 *Xenopus wholemount in situ hybridisation*

To prepare cDNA from which to clone a *ddx17* or *hdlbp* cRNA probe, 10 wild-type *X. tropicalis* embryos were collected and frozen at -80ºC in a sterile 1.5mL microcentrifuge tube. Total RNA was extracted[260] and cDNA synthesised using the UltraScript 2.0 Reverse Transcriptase Kit (PCR Biosystems) using the manufacturer's instructions. The 3' target regions within *ddx17* and *hdlbp* were amplified with primers designed using Primer3. The resulting product was ligated into pGEM-T® easy vector system (Promega), transformed using One Shot™ TOP10 Chemically Competent *E. coli*, cultured, and purified using NucleoBond Xtra Midi kit for transfection-grade plasmid DNA (MACHEREY-NAGEL). Ten µg of the resulting plasmid DNA was linearised over 3 hours in a 50µL reaction. Reactions were activated by heat, and the linearised products were purified with SigmaSpin™ Sequencing Reaction Clean-Up columns (Sigma-Aldrich). A 50µL transcription reaction was set-up and incubated overnight at 37°C according to Broadbent and Read.[260] The next day, 10 units of RNase-free DNase I (Promega Corporation) was added to the reaction and incubated for 15 minutes at 37°C. The RNA probes were column purified (SigmaSpin™ Sequencing Reaction Clean-Up), analysed on an agarose gel, quantified, and stored in 500ng aliquots at -20°C following a 1:10 dilution in hybridisation buffer.

Terminally anesthetised embryos/tadpoles were fixed in MEMFA (0.1M MOPS pH7.4, 2mM EGTA, 1mM $MgSO_4$, 4% Stabilized Formaldehyde (Acros Organics)) for 1 hour at room temperature and dehydrated in 100% Methanol. For gene expression analyses, fixed embryos were rehydrated, bleached, acetylated, pre-hybridised and hybridised.[260] Probes were detected with a 1:2000 dilution of Anti-Digoxigenin-AP, Fab fragments Antibody (Sigma-Aldrich, in MAB blocking solution) and the colour reaction developed in Alkaline Phosphatase Buffer (AP: 0.1M Tris pH 9.0, 50mM $MgCl_2$, 0.1M NaCl, 0.1% Tween-20) containing a 1-in-4 dilution of BM Purple (Roche), over 3 hours. Staining was fixed in MEMFA for 30 minutes, embryos were dehydrated in Methanol and allowed to rest at -20°C for 48 hours before imaging.

2.8.1.4 *Phenotypic analysis of crispant tadpoles*

To ensure blinding, all data from assays performed on F_1 tadpoles were collected whilst the genotype of the tadpoles was unknown. For the F_0 animals, images were anonymised and then scored blindly by experienced *Xenopus* biologists at the European Xenopus Resource Centre. To identify any gross morphological differences between wild-type and crispant embryos/tadpoles, animals were anesthetised in 0.025% w/v neutralised tricaine solution, fixed and visually inspected with an AxioZoom V16 stereomicroscope (Zeiss, Jena, Germany) with fluorescence for visualising GFP-expressing animals.

Fixed embryos for whole-mount immunohistochemistry were stored dehydrated at -20°C for 24 hours prior to rehydration, bleaching and immunostaining.[260] Embryos were blocked in Donkey Serum, Anti-HNK-1/N-CAM mouse monoclonal (CD57, Sigma-Aldrich) was used as the primary antibody at a 1:200 dilution, Goat anti-Mouse IgM HRP (Invitrogen, ThermoFisher Scientific) was used as the secondary antibody at a 1:250

dilution and embryos were developed in Pierce^{TM} DAB Substrate Kit (ThermoFisher Scientific). Staining was fixed in MEMFA for 30 minutes with embryos dehydrated in Methanol and allowed to rest for 48 hours at -20^{o}C before imaging.

Following image analysis, genomic DNA was extracted from fixed tissue: after removal of the embryos from fixative, they were washed 3 times in 50 mM Tris, pH8, 5 mM EDTA, 50 mM NaCl, 0.5% SDS (250ul). The embryos were then placed in 250ul of this buffer containing 150ug Proteinase K and incubated at 55^{o}C for 3-6 hours. The tubes were mixed for 5 mins on a mixmate shaker at 1000 rpm and 5M NaCl (85 ul) was added. Following another 5 minutes on the shaker at 1000 rpm, the samples were centrifuged at 17000g for 7 minutes. The supernatant was carefully removed to a new tube, 200 ul of isopropanol and 10ug of glycogen were added, the tube mixed by inversion six times and the precipitate collected by centrifugation at 17000g for 5 minutes. After a 70% ethanol wash, the pellet was dried and resuspended in 50 ul water for analysis.

Quantification of *X. tropicalis* behaviour included the application of three assays: the Free Movement Pattern (FMP) Y-Maze, Light-dark transition assay, and *Xenopus* Locomotion assay. The FMP Y-Maze to assess working memory was conducted according to Ismail *et al.*[261] with the exception that our experiments were conducted in the Zantiks LT unit (Zantiks Ltd) in acrylic inserts containing 16 identical Y-Mazes (comprising: three 12(W) x 35(L) x 13(D) mm arms and a 12 x 12 x 12 mm central zone). For the Light-Dark transition assay, tadpoles were placed in standard 92 x 16 mm Petri-dishes (Sarstedt) filled with 0.05X MMR, containing dilute but equally distributed tadpole food mix. Filming was conducted using infra-red, enabling tracking of the distance moved by individual animals across two zones (an outer zone encompassing the perimeter of the dish and the remainder, central region of the petri-dish). Tadpoles were transferred into the Petri-dishes and placed into the Zantiks LT unit (Zantiks Ltd) for a pretrial time of 600 seconds with a 10-second auto-reference period for tracking. The lights were subsequently programmed to turn on and off in 5-minute time intervals over a 30-minute period. Data from each trial was output in three forms: distance travelled (mm) over time (1-second time bins), time spent (seconds) in each zone and an AVI video file. MK-801 ((+)-5-methyl-10,11-dihydroxy-5H-dibenzo(a,d)cyclohepten-5,10-imine]), was applied to examine the contribution of working memory across the trial with the concentration and delivery based on Cleal *et al.*[262] Diazepam (D0899, Sigma-Aldrich) was administered to examine the contribution of anxiety to the startle responses observed across the trial. Tadpoles were placed in 100 mL beakers obtaining 50 mL 0.05X MMR and either DMSO or 2 μM, 5 μM, 20 μM, 50 μM or 100 μM diazepam, for 1 hour prior to evaluation in the Zantiks unit. The lowest anxiolytic dose of diazepam, 5 μM, without sedative properties was selected for downstream work. For the *Xenopus* Locomotion assay, tadpoles were placed as described above in Petri-dishes, lights were maintained off and the distance moved by individual animals was tracked over a 10-minute trial period (following a 120-second pretrial period). The data from each trial was output in two forms: distance travelled (mm) over time (1-second time bins) and an AVI video file.

2.8.1.5 Xenopus experimental design and statistical analysis

All phenotyping experiments performed in founder animals were replicated in embryos from at least three different females. Sample size estimations were calculated *a priori* from pilot studies performed on 16 uninjected control tadpoles and 16 crispant tadpoles with the G*Power software package version 3.1.9.7. Statistical analysis did not account for embryos or tadpoles eliminated at specific humane endpoints (e.g. naturally occurring gastrulation defects) and all data were initially screened for anomalous points, those exceeding 1.5X the interquartile range, and these points were eliminated. Data obtained during the FMP Y-maze assay, was compared across groups with an analysis of covariance (ANCOVA) with significant effects assessed by Bonferroni's *post hoc* multiple comparison test correction. Continuous variables including distance travelled in the *Xenopus* locomotion assay and tadpole head size comparisons were compared between wild-type and crispant groups with the independent samples t-test. Grouped data are shown as the mean. Unless otherwise stated, statistical significance was as follows: *p < 0.05, **p < 0.01, ***p < 0.001 and ****p < 0.0001.

2.8.2 Mouse methods for ddx17

The following methods were curated and performed by Dr Julien Courchet and his laboratory at the University of Lyon and are provided for reference.

2.8.2.1 *Animals*

Mouse breeding and handling was performed according to experimental protocols approved by the CECCAPP Ethics committee (C2EA15) of the University of Lyon, and in accordance with the French and European legislation. For primary neuronal cultures, E15.5 pregnant female Swiss mice were used and purchased from Janvier Labs. For *in utero* electroporations, E15.5 timed-pregnant F_1 hybrid mice (129SV/J x C57BL/6JRj, hereafter termed 129B6-F_1 mice) were used. Females were maintained in a 12hr light/dark cycle and bred overnight with C57BL/6JRj males. Noon following breeding was considered as E0.5.

2.8.2.2 *DNA and plasmids*

Endotoxin-free plasmid DNA was obtained using the Macherey Nagel midi-prep kit according to the manufacturer's instructions. The following plasmids were used: the empty vector pCAG2, mVenus expressing vector pSCV2[263] and the control shRNA vector pLKO.1.[264] Two distinct shRNAs against DDX17 cloned in the pLKO.1 vector were used (Addgene plasmid # 10879; http://n2t.net/addgene:10879; RRID:Addgene_10879). Expression vectors for wild-type human DDX17 were generated by Vector Builder under a CAG promoter.

2.8.2.2 *Ex-vivo cortical electroporation and primary neuronal cultures*

The electroporation of dorsal telencephalic progenitors was performed by injecting plasmid DNA (1-2 µg-µL of endotoxin-free plasmid DNA) plus 0.5% Fast Green (Sigma; 1:20 ratio) into the lateral ventricles of the brain of E15.5 mouse embryos using a Picospritzer III microinjector (Harvard Apparatus).[265] Electroporations were performed on the whole head with gold-coated electrodes (GenePads 5 X 7mm, BTX) using an ECM 830 electroporator (BTX) and the following parameters: 5 pulses of 100 msec long pulses, 150 msec intervals, at 20V. Immediately after electroporation, cortices were dissected in Hank's buffered salt solution (HBSS) supplemented with HEPES (pH 7.4, 2.5mM), $CaCl_2$ (1mM, Sigma), $MgSO_4$ (1mM, Sigma), $NaHCO_3$ (4mM, Sigma), and D-glucose (30mM, Sigma), hereafter referred to as complete HBSS (cHBSS). Cortices were dissociated in cHBSS containing papain (Worthington, 20U/mg at least) for 15 min at 37°C, washed once in cHBSS containing DNase I (2.5mg/mL, Sigma) and then washed 3 times in cHBSS before being manipulated by trituration. Cells were then plated at 125 000 cells on 12mm glass coverslips coated with PDL and Laminin and cultured for 5 days in Neurobasal medium supplemented with B27 (1x), N2 (1x), Glutamax (2mM), and penicillin (10U/mL)-streptomycin (0.1mg/mL).

2.8.2.3 Immunostaining

Cells were fixed for 30 min at room temperature in 4% (w/v) paraformaldehyde in PBS 1X and washed 3 times in PBS 1X. Primary antibodies were incubated overnight at 4°C in permeabilisation buffer (PB: 0.1% Triton X-100, 0,5% BSA (Sigma), in PBS 1X). After 3 washes of 5 minutes in 1xPBS, coverslips were incubated with secondary antibodies for 1hr at room temperature in PB under agitation. Coverslips were mounted on slides with Fluoromount G (Invitrogen) and fixed with polish. The following antibodies were used: chicken anti-GFP (Rockland) (1:2000), rabbit anti-MAP2 (Invitrogen) (1:1000), mouse anti-SMI312 (biolegend) (1:1000). All secondary antibodies were Alexa-conjugated (Invitrogen) and used at 1:2000 dilution. Nuclear DNA was stained using Hoechst dye (1:5000).

2.8.2.4 In utero cortical electroporation

At E15.5, *in utero* cortical electroporation was performed on timed-pregnant 129B6-F_1 hybrid females as described by Meyer-Dilhet *et al.*[266] A mix containing 1.5 µg/µL endotoxin-free plasmid DNA (shRNA plasmid: pLKO.1 or pLKO-shDDX17 #1 or pLKO-shDDX17 #2 at 1 µg/µL ; mVenus coding plasmid pSCV2 at 0.5 µg/µL) plus 0.5% Fast Green (Sigma, 1:20 ratio) was injected into one lateral hemisphere of embryos using microcapillaries. Electroporation was performed using an ECM 830 electroporator (BTX) with tweezertrodes kit using 4 pulses of 100 msec long pulses, 500 msec intervals, at 42V to target cortical progenitors by placing the anode on the side of DNA injection.

2.8.2.5 Immunohistochemistry

Animals were sacrificed at day 21 postnatally (P21) by intracardiac perfusion of 5mL PBS 1X followed by 5mL of 4% paraformaldehyde (PFA, Electron Microscopy Sciences). Then post-fixation in 4% PFA overnight under agitation at 4°C, protected from light to prevent fluorescence loss. 80µm thick sections were performed using a Leica VT1000S vibratome. Slices were incubated overnight under agitation, at 4°C and protected from the light, with primary antibodies (chicken anti-GFP, Rockland) diluted at 1:2000 in permeabilisation buffer (PB). The following day, under agitation at room temperature, 3 washes in PBS 1X were performed for 10min, followed by incubation of slices in secondary antibody containing PB for 1hr (Goat anti-chicken antibody, Alexa 488, 1:2000, life technology). Nuclear DNA was stained using Hoechst dye (1:5000). Slices were mounted with 80µL of Fluoromount G (Invitrogen) and then fixed with polish.

2.8.2.6 Image acquisition

Confocal images were acquired in 1024x1024 mode with a Nikon Ti-E microscope equipped with the C2 laser scanning confocal microscope using the Nikon software NIS-Element (Nikon). The following objective lenses (Nikon) were used: x10 PlanApo; NA 0.45 for brain slices and x20 PlanApo VC; NA 0.75 for cell cultures. The parameters (gain and laser power) of the confocal detector were adapted to each experiment to maximise signal while avoiding saturated pixels and thus remain in the quantitative range. For *in vitro* experiments, representative neurons were isolated from the rest of the image using ImageJ. Contrast was enhanced and the background (autofluorescence of non-transfected neurons in culture) removed for better illustration of axon morphology.

2.8.2.7 Quantifications and statistical analyses

Statistical analyses were performed using Prism (GraphPad Software). Statistical tests and number of replicates are indicated in figure legends. For axonal morphogenesis experiments, large field acquisition of neuronal cultures was performed (typically a 4x4 tiling scan using a 20x objective). All neurons on the reconstituted image were quantified; axon length and number of branches were measured with the Nikon NIS Elements software. Axons shorter than 100µm and branches under 5µm were not counted. Quantifications were performed blind to experimental condition. Neuronal migration following *in utero* electroporations was quantified using imageJ.[267] TIF files were used to detect the x,y coordinates of the soma of electroporated neurons using thresholds. The boundaries of the cortex (ventricular zone and pial surface) were determined with the Path writer plugin from ImageJ. The coordinates for each detected neurons were converted as a percentage migration on a ventricular zone to pial surface axis using Excel. Neurons were quantified in bins corresponding to 10% migration (10 bins total). For axon projections, GFP intensity was measured in the ipsilateral and contralateral regions of the cortex.[264] Briefly, the ROI function of NIS AR element was used to measure signal intensity in the ipsilateral cortex. Average measurements were recorded of 5 ROIs in the layer 5 plexus, in the corpus callosum, and in a region without signal (background). After background subtraction, layer 5 signal was normalised to the corpus callosum signal, to normalise axon plexus to the intensity of electroporation. For the contralateral cortex, the intensity profile function of the NIS AR element was used to measure the fluorescence intensity of axonal projection on the contralateral side along the six layers of the cortex. Average signal in the corpus callosum was used to normalise GFP signal to the intensity of electroporation.

2.9 RNA-seq methods

The following methods were performed by Dr Cyril Bourgeois and his laboratory at the University of Lyon and are provided for reference.

2.9.1 Cell culture and transfection

Human SH-SY5Y neuroblastoma cells (ECACC #94030304) were grown and transfected with 60 nM of a mixture of 2 different siRNA against *DDX17* (Merck-Millipore), using Lipofectamine RNAiMax (ThermoFisher), as recommended by the manufacturer. Cells were harvested 48 hours after transfection.

2.9.2 Protein and RNA analysis

Protein extraction was carried out as previously described.[268] Total RNA were isolated using TriPure Isolation Reagent (Roche) and analysed by reverse transcription and qPCR.[269]

2.9.3 RNA-seq and Gene Ontology analyses

Directional RNA libraries were prepared from total RNA after removal of ribosomal RNA (lncRNA library, Novogene). High throughput sequencing of 150 bp paired-end reads was performed on an Illumina Novaseq 6000 platform (Novogene), generating at least 24 Gigabases of raw data per sample, for an average number of 75 million matched pairs of reads per sample. Raw reads were pre-processed using fastp[270] and mapped to the human reference genome (hg38) using HISAT2.[271] Mapped reads were filtered using SAMtools[272] and the number of reads per gene was counted using HTSeq.[273] Differential gene expression analysis was carried out using the DESeq2[274] package. Parameters for differential expression: P-value < 0.05, [log2(FC)] ≥ 0.50 and base mean ≥ 10. Gene ontology and gene-set enrichment analyses were performed using the ShinyGO 0.76.3 web interface.[275] 13886 gene IDs that presented an average normalised read count ≥ 100 in the control condition were used as a background list of genes,. The complete lists of enriched biological processes are provided in **Supplementary Dataset SD1**.

2.9.4 Reagents

2.9.4.1 siRNA (Merck-Millipore)

SiCtrl: CGUACGCGGAAUACUUCGA[dT][dT]

siDDX17-#1: UCAUGCAGAUUAGUUAGAA[dT][dT]

siDDX17-#2: CAGCAGACUUAAUUACAUU[dT][dT]

2.9.4.2 Antibodies

Anti-DDX17: rabbit polyclonal, #ab24601 (Abcam)

Anti-β-tubulin: rabbit polyclonal, #2146 (Cell Signalling Technology)

TABLE 2.6 | PRIMERS FOR QPCR

Gene	***Forward primer***	***Reverse primer***
GAPDH	CTATAAATTGAGCCCGCAGC	AGAAGATGCGGCTGACTGTC
RPH3A	CATCAACAGGGTGATTGCTCGA	CCTCATGTTTTCTAGGCGGTCC
PLXNA4	CTGGCTCAAGGTGAAGGACATC	CCAGGCCACAGAAGTTATCGTC
SDK2	CGCTACATTCTGGAGATGTCGG	TTTGCTGAACTGTCCTTTCCCC
MSX2	GCCTCGGTCAAGTCGGAAAATT	GGTCTTGTGTTTCCTCAGGGTG
CELF2	ATCAGTCTCAGACCATGGAGGG	CGCCTTTGCTCTTTGTCCTTCT
BTG2	CAGAGGCTTAAGGTCTTCAGCG	ACCAGTGGTGTTTGTAGTGCTC
PLXNA2	GTGTCAGAACAGCTCGTACCAG	AGGTCCTGAGGGTTGTCAATGA
IL11	GACCACAACCTGGATTCCCTG	CTCCAGGGTCTTCAGGGAAGA
NRP1	ACGAAACACATGGTGCAGGATT	ATCCGGGGGACTTTATCACTCC
CUX2	AATTGAACTCCGCCGGGAATTT	TACAGGAGCCACCATCTCTCTG
NEUROG2	CACAGGCCAAAGTCACAGCAA	CTCCAAGGTCTCGGATTTGACG
DRGX	TTGATGACGGGTTTCTGCGTAG	TTTCATGGCGAGCTCTTCTCTG
RET	GCCGTGAAGATGCTGAAAGAGA	TGCTTCAGGACGTTGAACTCTG
SEMA6A	GCCGATGTAGACACATGCAGAA	GCAGGAAGGGTTGAAGGCATTA
MAPT	GTTGGGGGACAGGAAAGATCAG	GTCTCCAATGCCTGCTTCTTCA
PBX1	TCTACCATACGGAGCTGGAGAA	CAGGAGATTCATCACGTGGGTG
LRP4	CAGATGGCAGCATGAGAACAGT	CAGTCCAATACATGTACCCGCC
SEMA6D	GGATCTGCCCTTCGCACAATAA	ATGGCATGAAGAAAGTGTGGCT
LRRK2	AGAAACGCTGGTCCAAATCCTG	CAAGACGATCAACAGAGGCACA
NONO	AAAGCTCTGGACAGATGCAGTG	TCATCTAACTGGTCCATGGGCT
RARB	TTGCTAAACGTCTGCCTGGTTT	TCGATTTAGGGTAAGGCCGTCT
UNC5D	GGCCAGCCATGCAGATATTCTT	GACCTTCAAACCTGAGCTCTCG
CBFA2T3	ACATCTGGAGGAAGGCTGAAGA	CTTTCTGCAGCTCCGACATGG
REEP1	TATTTGGCACCCTTTACCCTGC	TGTGAATGTCTCTGCTGTGGTG
TGFB3	CAGAGGATCGAGCTCTTCCAGA	ATTCTTGCCACCGATATAGCGC
MYCN	GACTGTAGCCATCCGAGGAC	CACAGTGACCACGTCGATTT
GPC3	TCCAGCCGAAGAAGGGAACTAA	ATTCCAGCAAAGGGTGTCGTTT
NANOS1	CGCTCTACACCACCCATATCCT	GAGAGCGGGCAGTACTTGATG

2.10 ACMG-AMP guidelines for the classification of variant pathogenicity

The American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) have developed guidance for the interpretation of sequencing data.[276] These guidelines comprise sets of rules to help classify variants into 5 categories: pathogenic; likely pathogenic; uncertain significance; likely benign; and benign.

Variants identified from sequencing studies (whether that is in a research or clinical setting) should be curated against ACMG-AMP guidelines. Variants that meet the threshold for ‘pathogenic’ or ‘likely pathogenic’ are deemed clinically reportable. The guidelines are a complex framework involving the synthesis of multiple levels of evidence in support or not in support of pathogenicity (**Table 2.7 and 2.8**).

Table 2.7 | Synthesising evidence of pathogenicity

		Benign criteria		Pathogenic criteria			
Strength of evidence		Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Odds of pathogenicity		-18.7	-2.08	2.08	4.33	18.7	350
Evidence category and corresponding ACMG-AMP codes	Population Data	BA1, BS1, BS2			PM2	PS4	
	Allelic Evidence & Co-segregation Data	BS4	BP2, BP5	PP1	PP1	PP1	
					PM3, PM6	PS2	
	Computational & Predictive data		BP1, BP3, BP4, BP7	PP2, PP3	PM1, PM4, PM6	PS1	PVS1
	Functional Data	BS3				PS3	
	Other		BP6	PP4, PP5			

Figure adapted from Strande et al.[277] *ACMG-AMP guidelines use evidence codes to classify sequence data by data type and strength. The first letter of the evidence code corresponds to whether the evidence supports a pathogenic (P) or benign (B) classification. The second letter corresponds to the relative strength: VS – very strong; S – strong; M – moderate; P – supporting. Odds of pathogenicity assume a prior probability of 0.1.*

TABLE 2.8 | EVIDENCE CODES USED IN THE ACMG-AMP GUIDELINES

			Rule
Pathogenic	Very strong	PVS1	Null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where loss-of-function is a known mechanism of disease
	Strong	PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change
		PS2	*De novo* (both maternity and paternity confirmed) in a patient with the disease and no family history
		PS3	Well-established *in vitro* or *in vivo* functional studies supportive of a damaging effect on the gene or gene product
		PS4	The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls
		PP1 (strong)	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease
	Moderate	PM1	Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation
		PM2	Absent from controls (or at extremely low frequency if recessive) in population databases
		PM3	For recessive disorders, detected *in trans* with a pathogenic variant
		PM4	Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants
		PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before
		PM6	Assumed *de novo*, but without confirmation of paternity and maternity
		PP1 (Moderate)	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease
	Supporting	PP1	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease
		PP2	Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease
		PP3	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)
		PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic aetiology
		PP5	Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation
Benign	Supporting	BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease
		BP2	Observed *in trans* with a pathogenic variant for a fully penetrant dominant gene/disorder or observed *in cis* with a pathogenic variant in any inheritance pattern
		BP3	In-frame deletions/insertions in a repetitive region without a known function
		BP4	Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)
		BP5	Variant found in a case with an alternate molecular basis for disease
		BP6	Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation
		BP7	A synonymous variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved
	Strong	BS1	Allele frequency is greater than expected for disorder
		BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age
		BS3	Well-established *in vitro* or *in vivo* functional studies show no damaging effect on protein function or splicing
		BS4	Lack of segregation in affected members of a family
		BA1	Allele frequency is >5% in population datasets

Once evidence is gathered, variants can be classified into their respective classification (**Table 2.9**).[276]

Table 2.9 | Combining criteria to classify variants

Pathogenic	Likely Pathogenic	Likely Benign	Benign
1. 1 Very Strong (PVS1) *AND* **a.** ≥1 Strong (PS1–PS4) *OR* **b.** ≥2 Moderate (PM1–PM6) *OR* **c.** 1 Moderate (PM1–PM6) and 1 Supporting (PP1–PP5) *OR* **d.** ≥2 Supporting (PP1–PP5) **2.** ≥2 Strong (PS1–PS4) *OR* **3.** 1 Strong (PS1–PS4) *AND* **a.** ≥3 Moderate (PM1–PM6) *OR* **b.** 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR* **c.** 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)	**1.** 1 Very Strong (PVS1) *AND* 1 Moderate (PM1–PM6) *OR* **2.** 1 Strong (PS1–PS4) *AND* 1–2 Moderate (PM1–PM6) *OR* **3.** 1 Strong (PS1–PS4) *AND* ≥2 Supporting (PP1–PP5) *OR* **4.** ≥3 Moderate (PM1–PM6) *OR* **5.** 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR* **6.** 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)	**1.** 1 Strong (BS1–BS4) and 1 Supporting (BP1–BP7) *OR* **2.** ≥2 Supporting (BP1–BP7)	**1.** 1 Stand-Alone (BA1) *OR* **2.** ≥2 Strong (BS1–BS4)

Rules for classifying variants. If other criteria are unmet or there is contradictory evidence, the variant should be classified as a variant of uncertain significance.

2.11 HDLBP binding methods

These methods were performed by Igor Minia and are made available for reference.

2.11.1 Protein expression and purification

Fragments of human HDLBP, comprising KH domains 5 - 9 were expressed in *E. coli* Rosetta DE3 strain. Three mutants of the same HDLBP fragment (single amino acid substitutions I471V and R677H) carrying an N-terminal GST-tag were also expressed in *E. coli*. The starter culture was inoculated from a frozen glycerol stock of bacterial cell cultures transformed with the aforementioned assembled proteins in 10 mL of LB media, containing 100 ug/mL ampicillin and incubated at 37 °C overnight with shaking. A total of 10 mL of each starter culture was rapidly centrifuged to pellet the culture. The pellet was then added into 8 L of LB containing 100 μg/mL ampicillin and incubated at 37 °C with shaking at 200 rpm. Every 30 minutes, the OD600 was checked until it reached 0.6 when all cultures were induced with 0.5 mM IPTG. The cultures were grown for another 3 hours. Bacterial cultures were harvested by centrifugation at 5000 g for 10 minutes. The supernatant was removed, and the cell pellet was washed with PBS, centrifuged for 10 minutes at 5000 g. The supernatant was then discarded and the cell pellet was flash frozen and stored at - 80 °C until further use. The pellet was thawed on ice for 15 min and resuspended in 10× volumes of the Bacterial Lysis Buffer (50 mM Tris-HCl, pH 7.8, 500 mM NaCl, 4 mM $MgCl_2$, 0.5 mM TCEP, 1mM PMSF). The cells were lysed by 8 freeze-and-thaw cycles and sonicated on ice to shear released DNA by ten cycles of 30 second bursts with 30 second cooling intervals (amplitude 90%). The resulting crude extracts were clarified by centrifugation at 48000 g for 45 min at 4 °C. The expressed GST-fusion protein was purified in a batch mode with Pierce Glutathione Magnetic Agarose Beads (Thermo Scientific) according to the manufacturer's instructions. The purified protein concentrations were determined with Nanodrop by diluting the protein sample in a storage buffer (500 mM sodium chloride, 50 mM sodium phosphate pH 7.5) and recording an absorbance spectrum from 340 – 220 nm. Absorbance at 280 nm was corrected by subtracting the background absorbance at 320 nm and converted to a concentration using a molar extinction coefficient of 53290 $M^{-1}cm^{-1}$. The purified proteins were flash-frozen with liquid nitrogen and stored at −80 °C until further use. The identity of the expressed HDLBP fragments were then further confirmed by Western blot and SDS-PAGE analysis.

2.11.2 Fluorescence anisotropy assay

RNA oligonucleotides were labelled with fluorescein in a two-step procedure as previously described.[278] Firstly, 1 nmol of RNA oligonucleotides was 5′-end thiophosphorylated overnight at 37 °C in 1× T4 polynucleotide kinase buffer supplemented with 0.5 mM ATP-γ-S, 5 mM DTT, and 10 units of T4 polynucleotide kinase followed by ethanol precipitation. Fluorescein was then added to the 5′ end of the RNA by incubating the RNA with 1mM fluorescein maleimide for 2 h at room temperature in the dark in 50 mM phosphate buffer pH 7.0 followed by ethanol precipitation. The fluorescence polarisation assay was performed as previously described[279] with minor modifications. Briefly, serial dilutions of human HDLBP fragment (KH domains 5 to 9) as well as the mutants of the same HDLBP fragment (single amino acid substitution I471V and R677H) with N-terminal GST-tag in 1× Binding buffer (20 mM Tris, pH 7.5, 60 mM KCl, 1 mM EDTA, 10% glycerol, 1 ng/µL tRNA, 1 ng/µLheparin, 0.4 U/µL RNasin, and 200 ng/µL BSA) were first added to the wells of a black 384-well flat-bottomed microplate (Corning® NBS™) followed by addition of the fluorescein-labeled TFRC1 RNA probe (auucccuuccuucaaucacacucaguuuccacc) or TFRC2 RNA probe (cacagcucuccuauugaaacuugcccagauguu) to a final concentration of 20 nM. Control samples included a no fluorescent RNA and zero protein concentration to establish background anisotropy and anisotropy of the free probe respectively. The final reactions were mixed by a microplate shaker, spun, and were incubated at room temperature for 30 min in the dark. The anisotropy values were measured and automatically calculated by the fluorescence polarization function of microplate reader SpectraMax iD5 (Molecular Devices), using the SoftMax® Pro7 software. The raw values were then analyzed and graphed in GraphPad Prism 9 software. The binding curves were fit using nonlinear model with 4 parameter logistic equation: where $\boldsymbol{a}$ and $\boldsymbol{d}$ describe the location of the upper and the lower asymptotes of the equation. These are the values that the anisotropy $\boldsymbol{y}$ approach as the log of the protein concentration $\boldsymbol{x}$ approach 0 and infinity. Parameter $\boldsymbol{c}$ is the inflection point between the two asymptotes and evaluated as the dissociation constant, that is the concentration of the protein at which half of the RNA probe is in the bound state. Parameter $\boldsymbol{b}$ is the slope at the transition point $\boldsymbol{c}$ and controls the rate of approach to the asymptotes.

Chapter 3 | Manual curation of pLoF variants in gnomAD

3.0 Contribution statement

This project was initiated during my time at the Broad Institute. Its aims were to manually curate predicted loss of function variants in gnomAD and identify potential false positive variants. Data were prefiltered by LOFTEE (as described in **Methods Chapter 2.6.1.3**) and provided as a tab separated file by K. Karczewski and N. Watts. A list of 61 haploinsufficient genes and pext scores (as described in **Methods Chapter 2.6.2**) was provided by B. Cummings. The curation portal used for manual curation was developed by N. Watts with input from myself and M. Singer-Berk (**Methods Chapter 2.6.3**). I developed heterozygous curation methods with input from A. O'Donnell-Luria and D. MacArthur. Methods were expanded for homozygous curation with input from myself and M. Singer-Berk. I curated all variants presented, with M. Singer-Berk being a second curator. E. England was a third curator. All other work presented is my own. Work from this project has been published as part of three publications in Nature: *Karczewski, K. J., Francioli, L. C. ,... SEABY, E. G., ... & MacArthur, D. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581, 434-443.* https://doi.org/10.1038/s41586-020-2308-7 (**Appendix Paper 3**); *Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., SEABY, E. G.,., ... & MacArthur, D. (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature, 581, 452-458.* https://doi.org/10.1038/s41586-020-2329-2 (**Appendix Paper 4**); and *Gudmundsson, S., Karczewski, K.J., Francioli, L.C., ... SEABY, E. G., ... & MacArthur, D. Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature (2021).* https://doi.org/10.1038/s41586-021-03758-y (**Appendix Paper 5**).[133,214,280]

3.1 Introduction

With an increasing number of individuals undergoing diagnostic next-generation sequencing, there is a clear need to accurately delineate pathogenic variants from benign variation. Predicted loss of function (pLoF) variants, including stop gained, essential splice, and frameshift variants, are typically rare and deleterious, with many known to cause severe Mendelian disease through gene inactivation. On the other hand, some genes are tolerant to complete inactivation without any severe phenotypic consequences. Understanding which genes are likely to cause disease is important for clinical diagnostics, as is identifying genes entirely tolerant to homozygous pLoF; indeed, identifying natural occurrences of homozygous knockouts can potentially help deprioritise genes in clinical variant analysis and identify potential "safe" drug targets. Therefore, accurate curation of pLoF variants can expand understanding into the biology and essentiality of genes.

However, not all pLoF variants identified by bioinformatics pipelines are truly LoF. Given the rarity and deleteriousness of most true LoF variants, it is expected that many variants annotated as pLoF by variant calling pipelines and included in population databases such as gnomAD[133] and TopMED[24], and representative of ostensibly healthy individuals, are enriched for mapping, genotyping and annotation errors. Automated approaches exist to try and identify these errors, notably LOFTEE and pext (as described in **Methods 2.6.1.3** and **Methods 2.6.2** respectively), yet the complexity underpinning the identification of false positive LoF variants renders these software approaches error prone. Thus, careful filtering and manual curation is considered the best practice to identify and remove erroneous LoF variants.[280]

3.1.1 What is the aim of this project?

This project aims to develop a robust framework to manually curate pLoF variants from the gnomAD database. Developed methods will apply criteria to flag variants that are enriched for technical and annotation errors, rescue events, and transcript errors and then score variants on a spectrum of likelihood of being a true LoF variant. These methods will be applied to two distinct subsets of gnomAD:

1) *Heterozygous pLoF variants filtered from a curated list of 61 known haploinsufficient genes. It is expected that many of these variants will be false positives due to the depletion of individuals with paediatric onset disease resulting from gene haploinsufficiency in gnomAD.*
2) *Homozygous pLoF variants, present in at least one individual in gnomAD, with intention to identify LoF tolerant genes that can be identified as potential therapeutic targets.*

3.2 Methods

3.2.1 Data acquisition

From the gnomAD v2.1.1 raw dataset, 345,458 pLoF variants passing the most stringent set of LOFTEE criteria (no filters or flags as described in **Methods Chapter 2.6.1.3**) were extracted.

3.2.1.1 pLoF variants in haploinsufficient genes

Haploinsufficient developmental delay genes were curated by the ClinGen Dosage Sensitivity Working Group.[281] To account for penetrance, only genes with more than 75% reported penetrance from a literature search were included, defined as conditions too severe to expect an individual could consent to participate in gnomAD without guardianship. Sixty-one genes were included, comprising 50 autosomal genes of high severity and high penetrance and 11 genes on chromosome X in which the phenotype was expected to be severe or lethal in males and moderate to severe in females. The majority (58/61; 95%) had a score of 3 with sufficient evidence for pathogenicity, whereas two genes (*CHAMP1*, *CTCF*) had a score of 2 (some evidence), and one gene (*RERE*) was unscored. The resulting gene list is available at gs://gnomad-public/papers/2019-tx-annotation/data/gene_lists/HI_genes_100417.tsv.

Essential splice acceptor/donor, stop gained, and frameshift variants were identified in the list of 61 haploinsufficient disease genes and extracted from gnomAD v2.1.1. Only pLoF variants that passed gnomAD random forest filtering were included.[133] A total of 401 heterozygous pLoF variants remained and were subject to manual curation. Of the 61 genes, 55/61 (90%) had at least one high quality pLoF variant in gnomAD. The pext score for each of these variants was then compared with the mean gene pext score across GTEx tissues to assess whether the pext could aid in the identification of false positive LoF variants.

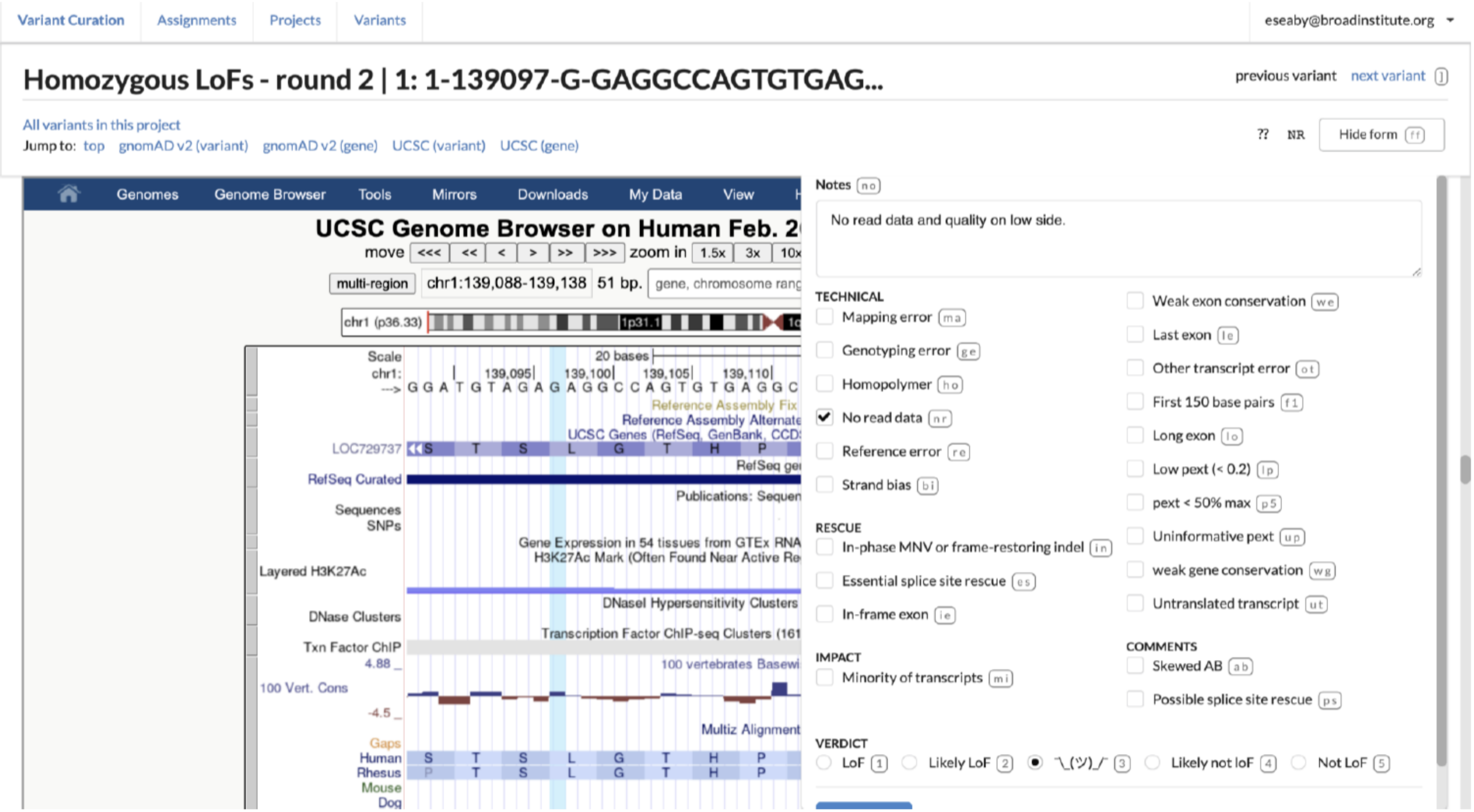
3.2.1.2 Homozygous pLoF variants across all genes

In gnomAD v2.1.1, 345,458 pLoF variants were identified. A total of 3,362 variants were extracted whereby at least one individual was homozygous for the pLoF observed, and the variants passed the random forest gnomAD filter. All 3,362 homozygous pLoF variants were subject to manual curation.

3.2.2 Manual curation

A total of 401 heterozygous pLoF variants in 61 haploinsufficient disease genes and 3,362 homozygous pLoF variants across all genes were manually curated using a custom build web portal (described in **Methods 2.6.3**). This portal included access to the gnomAD variant page, IGV reads for all individuals in gnomAD with the variant, and the UCSC browser (**Figure 3.1**). Two curators blindly assessed the same variants independently. If outcomes disagreed between curators, the variants were discussed until a consensus was reached. Where curators were unable to agree, a third independent curator validated the results.

FIGURE 3.1 | SCREENSHOT OF THE CURATION PORTAL



Homozygous LoFs - round 2 | 1: 1-139097-G-GAGGCCAGTGTGAG...

All variants in this project

Jump to: top gnomAD v2 (variant) gnomAD v2 (gene) UCSC (variant) UCSC (gene)

Notes [no]

No read data and quality on low side.

TECHNICAL

- □ Mapping error [ma]
- □ Genotyping error [ge]
- □ Homopolymer [ho]
- ☑ No read data [nr]
- □ Reference error [re]
- □ Strand bias [bi]

RESCUE

- □ In-phase MNV or frame-restoring indel [in]
- □ Essential splice site rescue [es]
- □ In-frame exon [ie]

IMPACT

- □ Minority of transcripts [mi]
- □ Weak exon conservation [we]
- □ Last exon [le]
- □ Other transcript error [ot]
- □ First 150 base pairs [f1]
- □ Long exon [lo]
- □ Low pext (< 0.2) [lp]
- □ pext < 50% max [p5]
- □ Uninformative pext [up]
- □ weak gene conservation [wg]
- □ Untranslated transcript [ut]

COMMENTS

- □ Skewed AB [ab]
- □ Possible splice site rescue [ps]

VERDICT

○ LoF [1] ○ Likely LoF [2] ● ¯_(ツ)_/¯ [3] ○ Likely not loF [4] ○ Not LoF [5]

Screenshot of the loss of function curation portal showing the variant of interest in UCSC (variant view). Tabs along the top allow the user to click between the gnomad variant view, gnomAD gene view, UCSC variant view and UCSC gene view. The curation form can be toggled on and off and keyboard shortcuts help the user classify the variants quickly.

Manual curation involved identifying error modes commonly found in pLoF variants (**Table 3.1**). Error modes were grouped into three main categories: ***technical errors***, ***rescue events***, and ***transcript errors***. These were subdivided into mapping error, strand bias, reference error, genotyping error, homopolymer sequence, in-frame multi-nucleotide variant or frame-restoring indel, essential splice site rescue, minority of transcripts, weak exon conservation, last exon, and other annotation error. Combinations of errors detected within these categories were applied to determine if a variant was likely to ablate gene function. After reviewing each variant for these three error modes, the variant was scored using a five-point likert scale: not LoF, likely not LoF, uncertain, likely LoF, and LoF (**Table 3.2**).

TABLE 3.1 | RULES FOR FLAGGING ERRORS IN MANUAL CURATION

	Technical errors						Rescue errors		Impact errors			
	Mapping error (M)	**Genotyping error (G)**	**Homopolymer (H)**	**No read data (N)**	**Reference Error**	**Strand Bias**	**MNV (P)**	**Essential splice site rescue (R)**	**Minority of transcripts (T)**	**Last Exon - (L)**	**Weak exon conservation (C)**	**Other transcript error (O)**
Rules	Tandem repeats (UCSC)*	Allele balance <80% for majority of HOM individuals	Repeat of >=7 base pairs (reference sequence)	No HOM read data available	Reference Sequence wrong (too small intron/MAF=1/gap in UCSC browser visible)	Variant falls on only forward or reverse strands, or there is a clear skew of strand in at least 1 HOM individual	In phase MNV (two SNV occurring in cis within the same codon) that changes nonsense variant to missense/synonymous variant	In frame splice site rescue within 21bp and validated by Alamut.	Variant falls on <=50% of Basic GENCODE transcripts (not including non-coding transcripts)	Variant falls in terminal coding exon or last 50bp of coding penultimate exon.	Entire/partial weak conservation of exon upon visual inspection in UCSC	Overprinting of exon (i.e. whichever transcript is using a different frame from main transcript) containing variant
	Human self-chained repeats (UCSC)*	GC rich region surrounding/including variant (approx. 50-150bp) except when unique/align scores are high	trinucleotide repeats >= 7	No HET or HOM read data available			Frame restoring indel					Re-initiation by downstream methionine in first coding exon for stop-gained variants only (also make note for these)
	Segmental dups (UCSC)*	GQ < 30 (Tag only if in majority of HOM individuals)					Insertion across the intron-exon boundary that maintains strong splicing predictions					Biologically relevant stop codon upstream of variant
	Complex variation in region e.g. multiple indels, SNVs	Low read depth < 15 (tag only if in majority of HOM individuals)										Biologically relevant start codon downstream of variant
		Low complexity sequence (by visualisation in IGV browser)										
Notes	Flag for mapping error if you have 3 of any combination of UCSC tracks covering the variant	Low complexity regions are defined as "poly-purine or poly-pyrimidine stretches, or regions of extremely high AT or GC content: (ensembl) and where you might see low read depth as a result *(especially in exonic sequences)*	Make sure to check multiple insertion bars if looking at an insertion to ensure there is no evidence of stuttering	For HET good AB is >=25%		Use this tag no matter where the strand bias occurs within the genomic sequence including the variant in question.	When looking at an indels, make sure you look at each homozygote with data to make sure the frame-restoring indel is in all cases and zoom out to see all potential sites for frame restoration. Check each insertion for stuttering.	Alamut needs to predict in frame rescue in majority of predictors or strong rescue with MaxEntScan. Strong rescue is defined by the rescue site having a score of at least 50% of the normal predicted splicing score for MaxEntScan.	For a variant in an exon with an unconserved overhang in a minority of transcripts you can assume the other xx will rescue and can use not likely LOF/not LOF. Check the gnomAD page for pext scores to look at expression relative to entire gene for cases where noncoding tx have high expression.	Do not tag when the variant results in disruption of more than 25% of coding sequence	View exon conservation on exon view in UCSC. **If entire gene is not conserved, do not use this tag.** For a variant in an exon with an **unconserved** overhang, if there is another tx, assume it will rescue and use not likely LOF/not LOF. Check gnomAD for pext scores to assess expression relative to entire gene where noncoding tx have high expression.	For overprinting: always tag but only affects function with stop gained. Overprinting transcripts can be identified by looking at conservation pattern of nucleotides for wobble effect. For stop/start codon upstream/downstream variant: use exon conservation to determine which start and stop sites are biologically relevant.

*A minimum of two tracks present

Abbreviations: AB – allele balance; bp – base pair; GQ – genotype quality score; HET – heterozygote; HOM – homozygote; IGV – integrative genomics viewer; LOF – loss of function; MAF – minor allele frequency; MNV – multinucleotide variant; SNV – single nucleotide variant; tx – transcript; UCSC – UCSC genome browser.

TABLE 3.2 | RULES FOR CLASSIFYING PLOF VARIANTS FROM MANUAL CURATION

	(1) LoF	(2) Likely LoF	(3) Uncertain	(4) Likely not LoF	(5) Not LoF
Criteria for scoring variants	*When any **one** of the following is seen in isolation*	*When any **one** of the following are seen in isolation or combinations of any slight genotyping/mapping errors with remaining error types within this category*	*Conflicting evidence/Ambiguous evidence*	*Potential re-initiation by downstream methionine within 1st coding exon for stop-gained variants in which the conservation pattern suggests the downstream MET could rescue from nonsense variation (downstream MET well conserved)*	*All HOM with AB<80%*
Rules for consideration	Absence of any evidence to refute a LOF consequence	Multiple genotyping errors (e.g. low DP, low GQ, GC rich all in combination)	No read data for HOM and HET + any change except splice variant**$**	Homopolymer repeat (indel variant only)	Frame restoring indel
	Mapping error	Multiple genotyping and mapping errors in combination (e.g. low DP and low AB with some repeat mapping)	Only HET data with low GQ **or** low AB **and** DP	In frame splice site rescue between 6 and 21bp and validated by Alamut (>/=80% 'raw' score of native site by MaxEntScan)	In phase multi-nucleotide variant (2 SNV occurring in cis within the same codon) that changes nonsense variant to missense/synonymous variant
	Genotyping error	Partial loss of exon conservation or low conservation of whole gene	Variant present in <=50% of coding GENCODE basic transcripts and have pext <= 0.80	Complex mapping issues	Complete splice site rescue within 6bp (gain/loss of </=2 amino acids) and validated by Alamut (>/=80% 'raw' score of native site by MaxEntScan)
	Strand bias	If >50% of the coding sequence was disrupted by a variant in the last exon, this was deemed probably LoF	In frame splice site within 6bp of natural splice site that is a NAGNAG site that is not predicted to splice by Alamut	Weak conservation of exon relative to the rest of the gene and with pext <= 0.8	Reference error
	No HOM data but good quality HET data for majority of individuals (high GQ **and** DP>15 **or** AB>80%)	Splice rescue within 21bp (not NAGNAG sites) but very weak signal as per Alamut (i.e., for rescue to be called, must be >/=80% 'raw' score of native sites by MaxEntScan)	Combination of multiple flags without complete confidence for not LOF classification	Biologically relevant stop site located upstream of variant (as identified by conservation)	Combination of multiple flags (multiple flags of indisputable evidence)
	Variant with minority tx **or** weak exon conservation (complete not partial conservation) with pext = 1 and no other tags (not including last exon and strand bias)	Homopolymer repeat with SNV of a different nucleotide than the homopolymer track	Re-initiation by downstream MET in first coding exon for stop gained variants in which first exon has not well conserved MET throughout	Biologically relevant start site located downstream of variant (variant in UTR)	
		Variant with minority tx **and/or** weak exon conservation tags (complete not partial conservation) with pext > 0.80 **and** any other tags excluding (strand bias/last exon)			

Notes	$ Splice variants with no read data should not be uncertain if they also have a good genotype quality (GQ) and one of 2: good depth (DP) or allele balance (AB) No need to check if splice variants result in exon skipping or nonsense mediated decay - treat as loss of function No extra consideration for variants with high minor allele frequency For each variant look at all read data available (both homozygous [HET] and heterozygous [HOM]) Abbreviations: AB – allele balance; bp – base pair; DP – depth; GQ = genotype quality; HET – heterozygote; HOM – homozygote; MET – methionine; LoF – loss of function; SNV – single nucleotide variant; tx – transcript; UTR – untranslated region.

Following application of flags and error modes, variants were classified on a 5-point Likert scale ranging from LoF to not LoF.

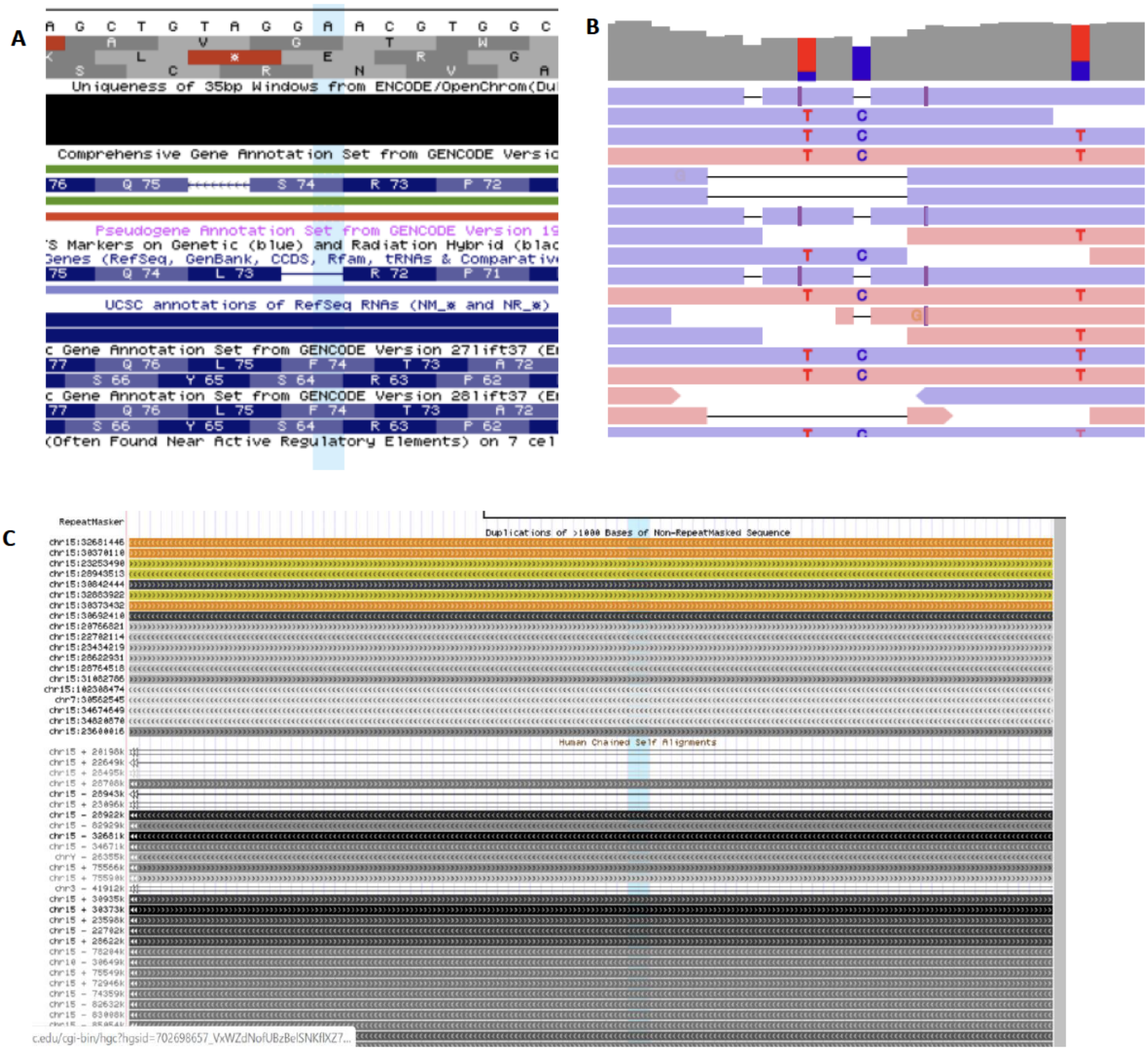
3.2.2.1 Technical errors

Technical errors include mapping errors and genotyping errors from sequencing issues; strand biases; misalignment of reads and repetitive regions that could be detected in IGV and the UCSC genome browser; and errors within the reference sequence.

Strand bias was flagged when a variant was skewed preferentially on the forward or reverse strand, or when the majority (>90%) of a given strand covered a region; this was often observed around intron–exon boundaries. Strand biases despite balanced coverage of the forward and reverse strands were weighted towards probably not LoF, whereas a strand bias due to skewed strand coverage was weighted alongside other genotyping errors. Reference errors were uncommon but identified by a small deletion in an exon, posing as a <5-base-pair intron (**Figure 3.2A**). These errors rendered the variant not LoF.

Most genotyping errors and strand biases in isolation were not deemed critical in deciding whether a variant was probably not LoF or not LoF, except for an allele balance ≤25%; this was to reduce identifying somatic variants arising through clonal haematopoiesis. Mapping errors were evident when reads around the variant harboured many other variants (**Figure 3.2B**), especially those with abnormal allele balances. UCSC tracks for large segmental duplications, self-chain alignments, and simple tandem repeats were utilised in determining mapping error status (**Figure 3.2C**). Genotyping errors were partially eliminated by upstream filtering for read depth, genotype quality <20, and skewed allele balance. Additional hallmarks for genotyping errors included homopolymer repeats (defined as an insertion or deletion within or directly neighbouring a sequence of seven or more of the same nucleotide or trinucleotide repeats), GC rich regions, and repetitive regions in which sequencing errors would be more common.

FIGURE 3.2 | TECHNICAL ERROR EXAMPLES IN THE UCSC BROWSER AND INTERACTIVE GENOMICS VIEWER

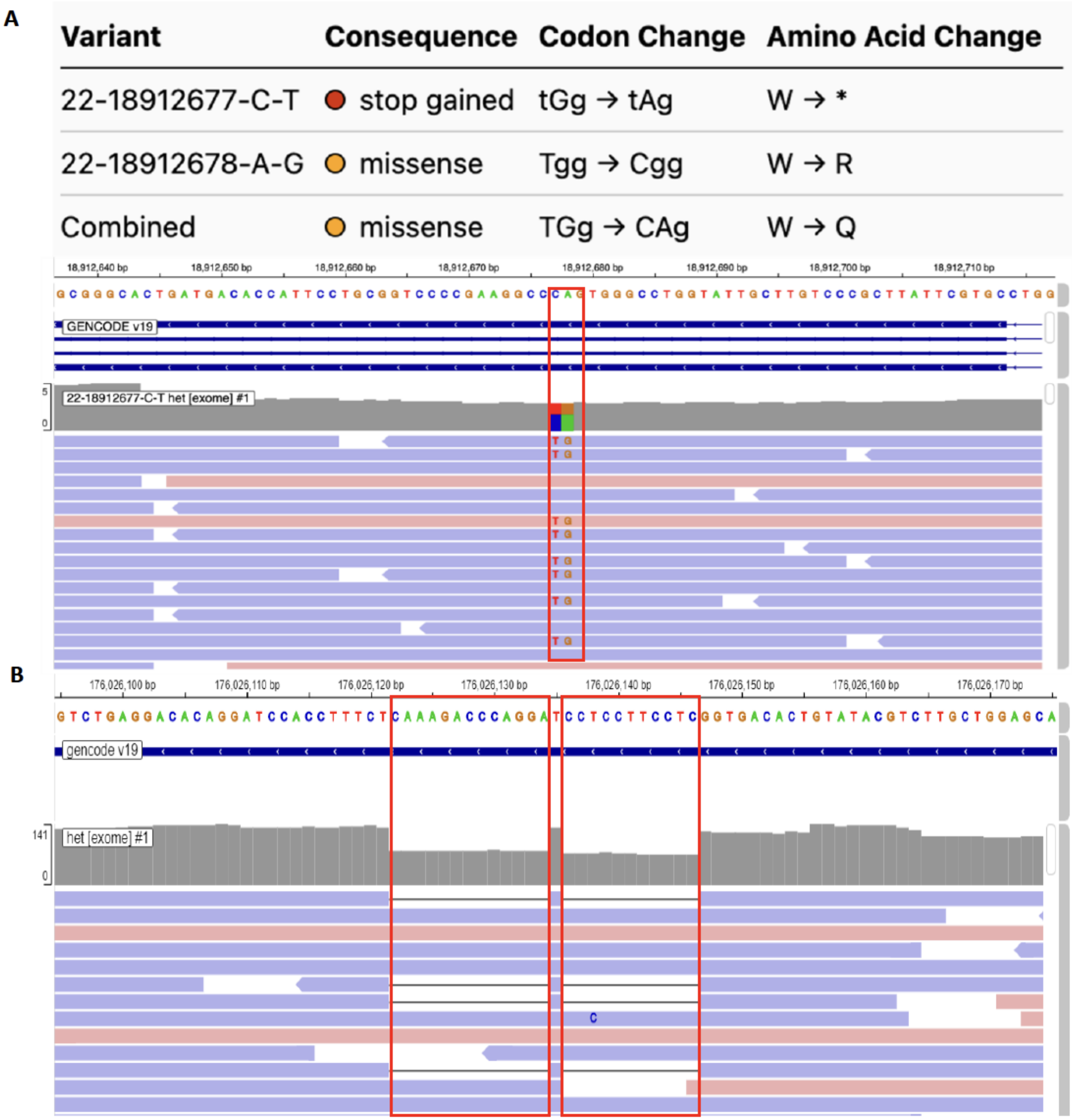


***(A)** Reference error in UCSC represented by a small one base pair deletion; variant identified by translucent blue line. Associated transcripts below deletion display overprinting. **(B)** Potential mapping error represented by polymorphic region with multiple deletions of varying sizes, insertions, and single nucleotide variants as shown in readviz IGV viewer. **(C)** Multiple repeat tracks with duplications of >1000 bases and human chained self-alignments shown in the UCSC browser.*

3.2.2.2 Rescue events

Rescue events include multi-nucleotide variants (MNVs), frame-restoring indels, and essential splice site rescues. MNVs were visualised in IGV and cross checked with codons from the UCSC browser; in frame MNVs that rescued stop codons were scored as not LoF (**Figure 3.3A**). Frame-restoring indels were verified by counting the length of the insertions and deletions to determine if the resulting variation disrupted the frame of the gene (**Figure 3.3B**). The window used to detect surrounding indels was approximately 80 base pairs in length. Essential splice site rescue occurs when an in frame alternative donor or acceptor site is present, which probably has a minimal effect on the transcript. Cryptic splice sites within 6 base pairs of the splice variant were considered a complete rescue, rendering the variant not LoF. Rescue sites >6 base pairs away but within ±21 base pairs were weighted with less confidence, scoring as probably not LoF. All potential splice site rescues were validated using Alamut v.2.11 (https://www.interactive-biosoftware.com/alamut-visual/). Splice sites were classified as rescues if the MaxEntScan score for the alternate sequence was ≥ 50% of the reference sequence score.

FIGURE 3.3 | RESCUE EVENT EXAMPLES



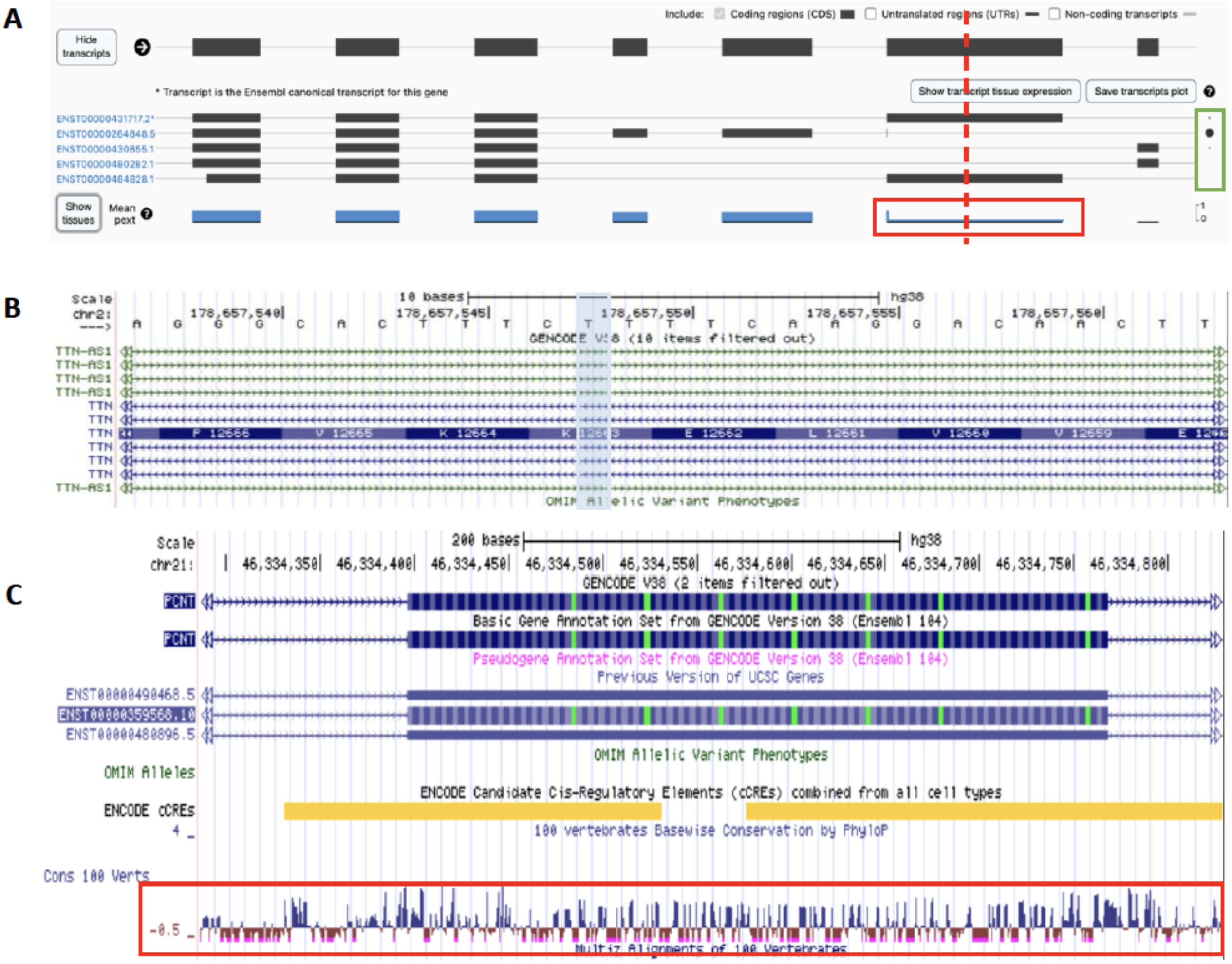
Variant	Consequence	Codon Change	Amino Acid Change
22-18912677-C-T	● stop gained	tGg → tAg	W → *
22-18912678-A-G	● missense	Tgg → Cgg	W → R
Combined	● missense	TGg → CAg	W → Q

*(A) Multinucleotide variant shown in IGV. Black box represents the codon. The first variant is called as a stop gained (C>T), however there is an adjacent missense variant in the same codon (A>G). The combination of the two variants in the codon is (CCA > CTG). The transcript is on the reverse strand, so the codon change should be the reverse complement (TGG > CAG), which results in a missense variant W -> Q. **(B)** Example of an inframe rescue. The first deletion shown within the blue box deletes 13 nucleotides. The second variant in orange deletes 11 nucleotides. The combination of the two deletions deletes 11 + 13 = 24 nucleotides which is in frame. Therefore, the protein will be truncated by 8 amino acids but will not result in loss of function.*

3.2.2.3 Transcript errors

Transcript errors were described as variants that occur in an exon found in a minority of transcripts for that gene, or that occur in a poorly conserved exon, or represent re-initiation events. The UCSC genome browser was used to detect all three situational errors. These were further assessed using pext (proportion expressed across transcripts) scores and only those with low overall expression relative to the gene were determined to be not LoF (**Figure 3.4A**). For an exon to be considered in a minority of transcripts, it had to be present in 50% or fewer of that gene's coding GENCODE v19 Basic transcripts (**Figure 3.4B**). These variants often affected poorly conserved exons, as determined by PhyloP,[65] PhyloCSF,[282] and visualisation in the UCSC browser (**Figure 3.4C**).[283] Exon conservation was determined by looking at the nucleotide base pair conservation based on PhyloP. Variants within the last coding exon, or within 50 base pairs of the penultimate coding exon were considered as uncertain LoF. If >50% of the coding sequence was disrupted by a variant in the last exon, this was deemed probably LoF. Re-initiation events were flagged when a methionine downstream of the variant in the first coding exon was predicted to restart transcription and were predicted to be probably not LoF. Variants occurring after a stop codon in the last coding exon were considered not LoF, particularly across the region of the exon or transcript in question.

FIGURE 3.4 | TRANSCRIPT ERROR EXAMPLES


A
Include: ☑ Coding regions (CDS) □ Untranslated regions (UTRs) □ Non-coding transcripts
Hide transcripts
* Transcript is the Ensembl canonical transcript for this gene
Show transcript tissue expression
Save transcripts plot
ENST00000431717.2*
ENST00000264848.5
ENST00000430855.1
ENST00000480282.1
ENST00000484828.1
Show tissues
Mean pext
B
Scale
chr2:
10 bases
hg38
178,657,540
178,657,545
178,657,550
178,657,555
178,657,560
GENCODE V38 (10 items filtered out)
TTN-AS1
TTN
P 12666
V 12665
K 12664
E 12662
L 12661
V 12660
V 12659
OMIM Allelic Variant Phenotypes
C
Scale
chr21:
200 bases
hg38
46,334,350
46,334,400
46,334,450
46,334,500
46,334,550
46,334,600
46,334,650
46,334,700
46,334,750
46,334,800
GENCODE V38 (2 items filtered out)
PCNT
Basic Gene Annotation Set from GENCODE Version 38 (Ensembl 104)
Pseudogene Annotation Set from GENCODE Version 38 (Ensembl 104)
Previous Version of UCSC Genes
ENST00000490468.5
ENST00000359568.10
ENST00000480896.5
OMIM Allelic Variant Phenotypes
OMIM Alleles
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
100 vertebrates Basewise Conservation by PhyloP
4
Cons 100 Verts
-0.5
Multiz Alignments of 100 Vertebrates

***(A)** Visualisation of a variant position across exons and transcripts as shown in the gnomAD browser. The top line shows all exons (dark grey boxes) across the gene. Five Ensembl transcripts are represented below with the green box showing the relative expression of each transcript across all tissues by size of the circle; the second transcript ENST00000264848.5 has the highest expression. The pext score is visualised at the bottom by the blue boxes. The 6th exon as shown in the red box has a relatively low pext score (determine by height of the box) when compared to the height of the first 5 exons. The most highly expressed transcript does not contain the full length of the sixth exon, which will influence the drop in pext score. Therefore, a variant (red dotted line) in the 6th exon has a low pext score and would only affect a minority of poorly expressed transcripts. **(B)** A variant (light blue strip) affects only one TTN transcript as visualised in the UCSC browser, representing a minority of transcripts. **(C)** Example of an exon in PCNT in the UCSC browser. The conservation track along the bottom (red box) based on PhyloP shows weak exon conservation.*

3.2.3 Classifying variants

For a variant to be considered as LoF, it had to have no rescue errors or other major error modes such as weak exon conservation or minority of transcripts. If a single minor error mode was noted for a variant, which include genotyping or mapping errors, it would be classified as likely LoF. In contrast, rescue errors were automatically classified as likely not LoF or not LoF. Multiple error modes (>= 3) resulted in a "not LoF" curation of the variant. Variants in which there was inconclusive evidence were curated as unknown (**Table 3.2**).

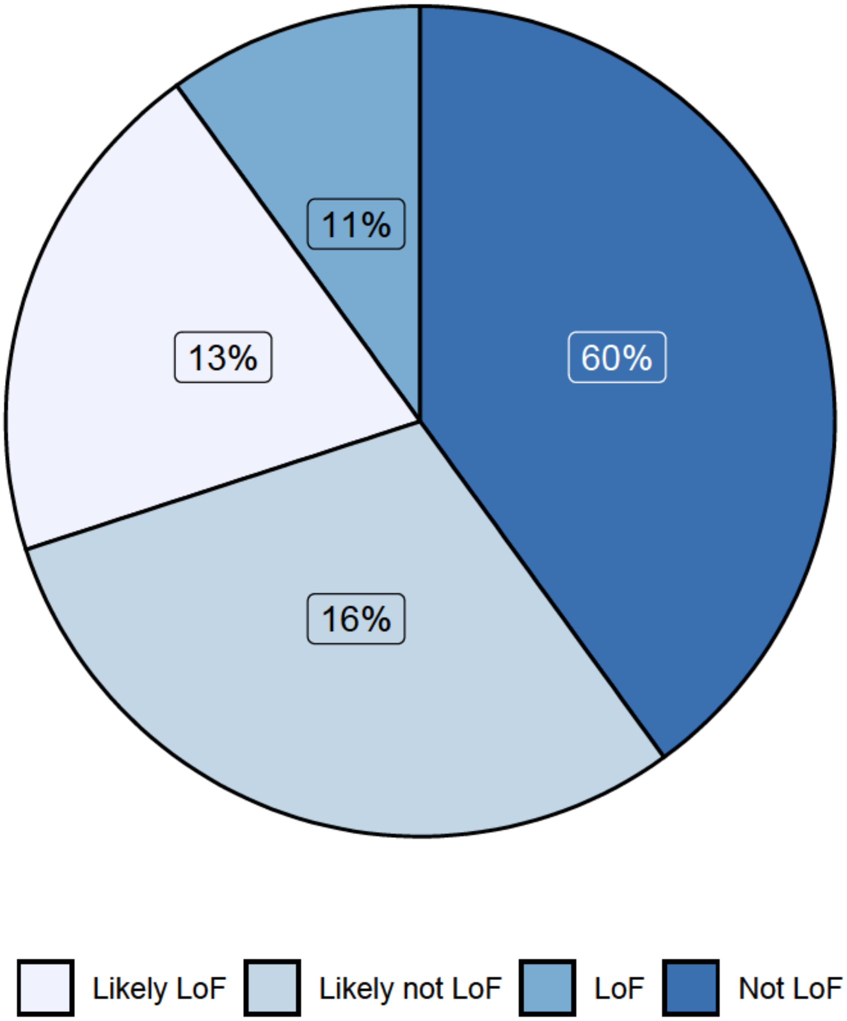
3.3 Results

Overall concordance between curators was >90% across projects. All discordant verdicts were resolved by consensus, either between the two curators or with help from a third curator.

3.3.1 Curation of 401 pLoF variants in 61 haploinsufficient disease genes

Four-hundred-and-one pLoF variants in the gnomAD dataset in 61 haploinsufficient severe developmental delay genes were manually curated. Flags were applied for any reason that the pLoF variant may not be a true LoF variant. Most variants 306/401 (76%) were classified as not LoF or likely not LoF (**Figure 3.5**). Full curation results are available in **Supplementary Dataset SD2**.

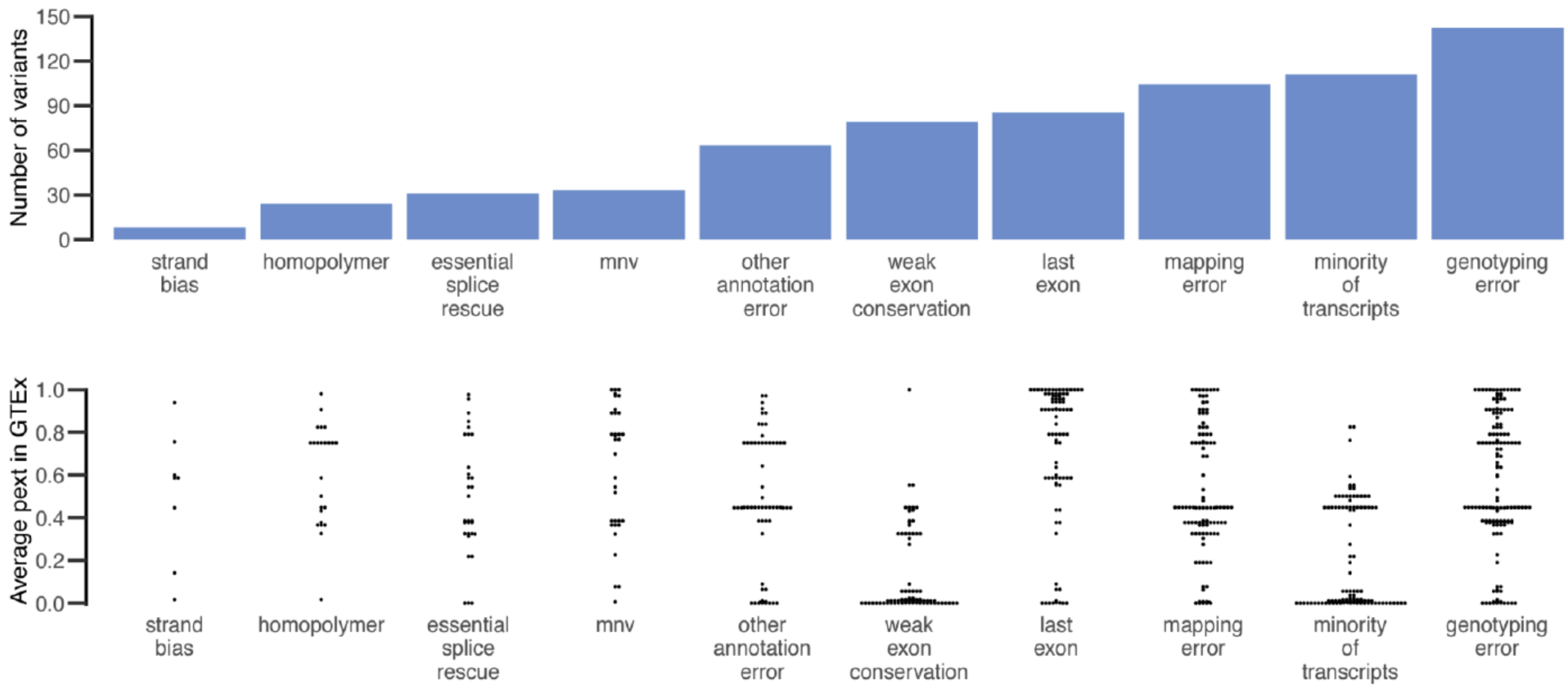
FIGURE 3.5 | MANUAL CURATION OF 401 PLOF VARIANTS IN 61 HAPLOINSUFFICIENT DEVELOPMENTAL DISEASE GENES



Disease distribution of curation verdicts for the 401 pLoF variants. Two-hundred-and-forty variants were classified as not LoF, 66 as likely not LoF, 52 as likely LoF and 43 and LoF.

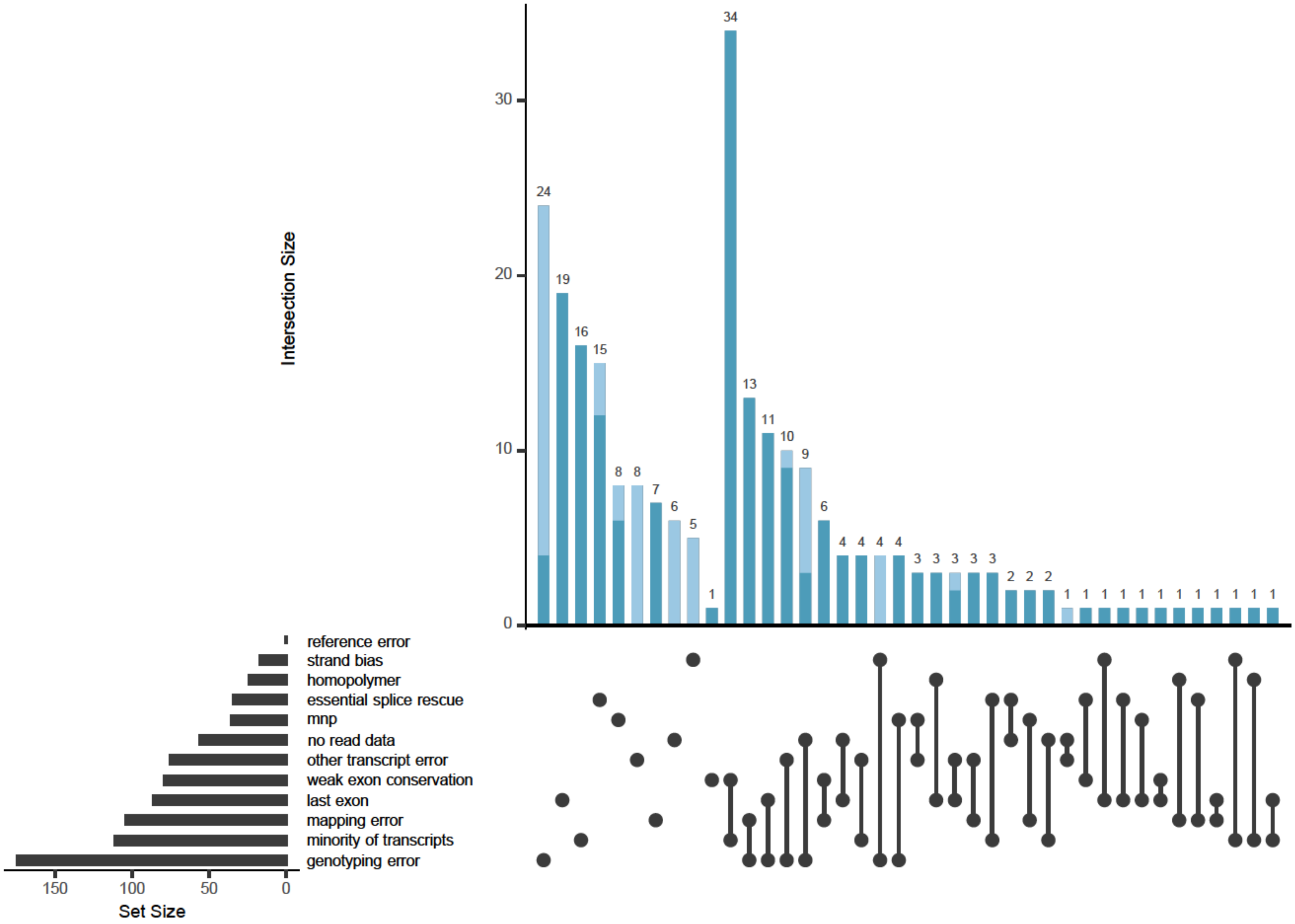
Minority of transcripts (whereby a variant disrupted less than 50% of GENCODE transcripts) and genotyping errors represented the greatest error modes for 401 pLoF variants in 61 haploinsufficient disease genes (**Figures 3.6 and 3.7**). Error modes compared with average pext scores across GTEx for each variant show pext scores are comparatively lower for minority of transcripts and weak exon conversation errors (**Figure 3.6**).

FIGURE 3.6 | MANUAL CURATION OF 401 PLOF VARIANTS IN 61 HAPLOINSUFFICIENT DEVELOPMENTAL DISEASE GENES



This figure was generated by Dr Beryl Cummings – using curation data analysed in this research chapter. Manual curation was performed according to error mode and pext score. Top, the frequency of each error mode present in the 306 variants classified as either LoF or likely not LoF. Transcript errors emerge as a major putative error mode in the annotation of these pLoF variants. Bottom, bee swarm plot shows the average pext score across GTEx tissues for each variant in the error categories. This shows that pext values are discriminately lower for variants that are annotated as possible transcript errors ($P = 4.1 \times 10^{-38}$, two-sided Wilcoxon test between transcript errors and other error modes).

FIGURE 3.7 | OVERLAP OF FLAGS APPLIED IN MANUAL CURATION TO 401 PLOF VARIANTS IN 61 HAPLOINSUFFICIENT DISEASE GENES

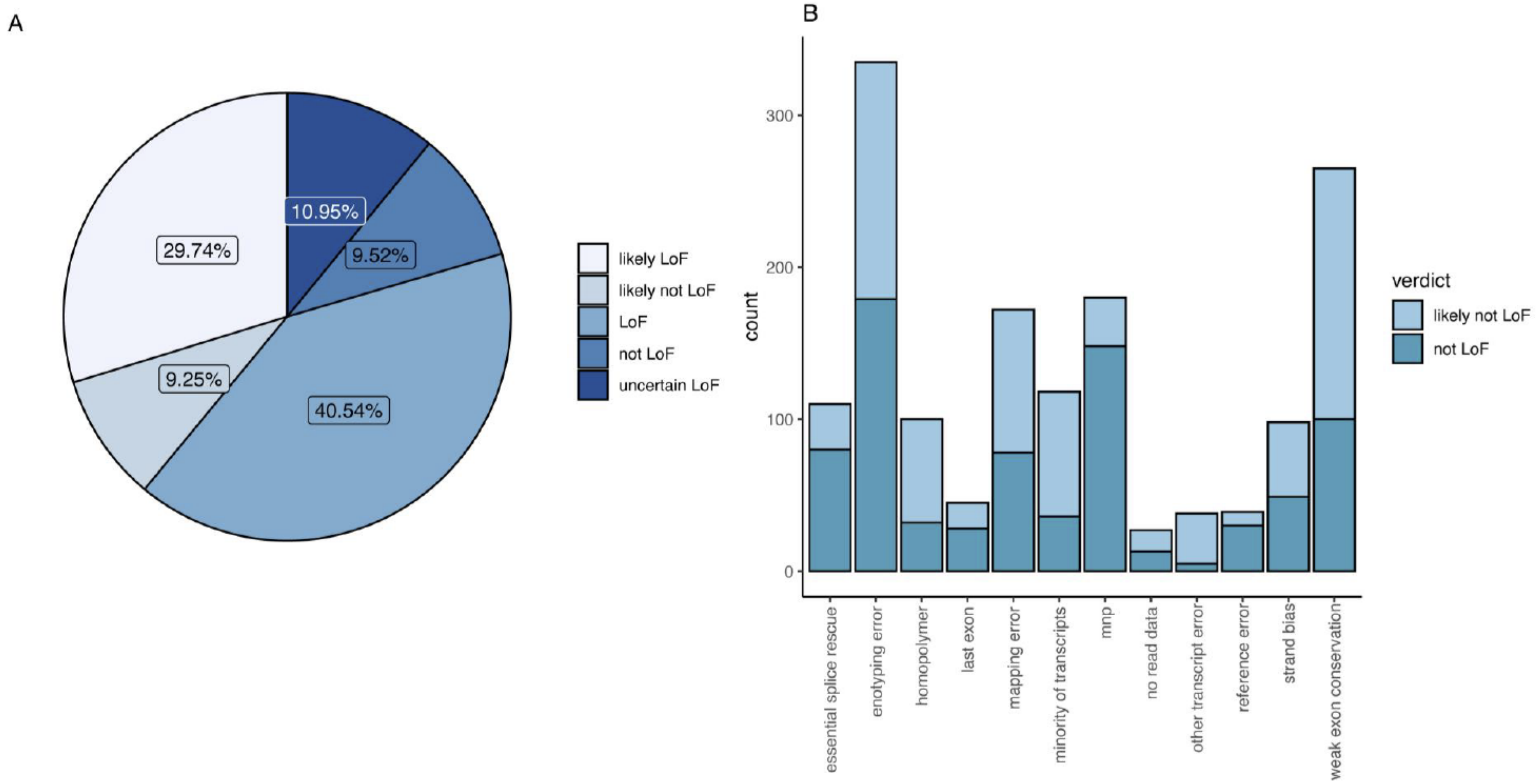


UpSet plot showing the overlap of flags applied to manual curation of 401 pLoF heterozygous variants. Light blue represents the proportion of variants classified as LoF or likely LoF. Teal represents variants classified as not LoF or likely not LoF. MNP – multinucleotide polymorphism.

3.3.2 Curation of 3,362 homozygous pLoF variants

The verdict of manual curation on 3,362 variants is shown in **Figure 3.8**. Curation resulted in removal of 1,141 (34%) variants, leaving 2,221 homozygous pLoF passing curation filters in a list of 1,815 genes where at least one individual was observed harbouring a homozygous knockout allele. This list likely excludes some genes with true LoF variants due to the strictness of curation and slight under-calling of homozygotes. Further, these results may also overestimate the effect of some pLoFs due to rescue mechanisms beyond the frame of reference (80 base pairs) applied in manual curation methods. However, the resultant list of 1,815 genes represents the best current estimate of confidently LoF-tolerant genes based on the gnomAD dataset. The full curation dataset is available in **Supplementary Dataset SD3**.

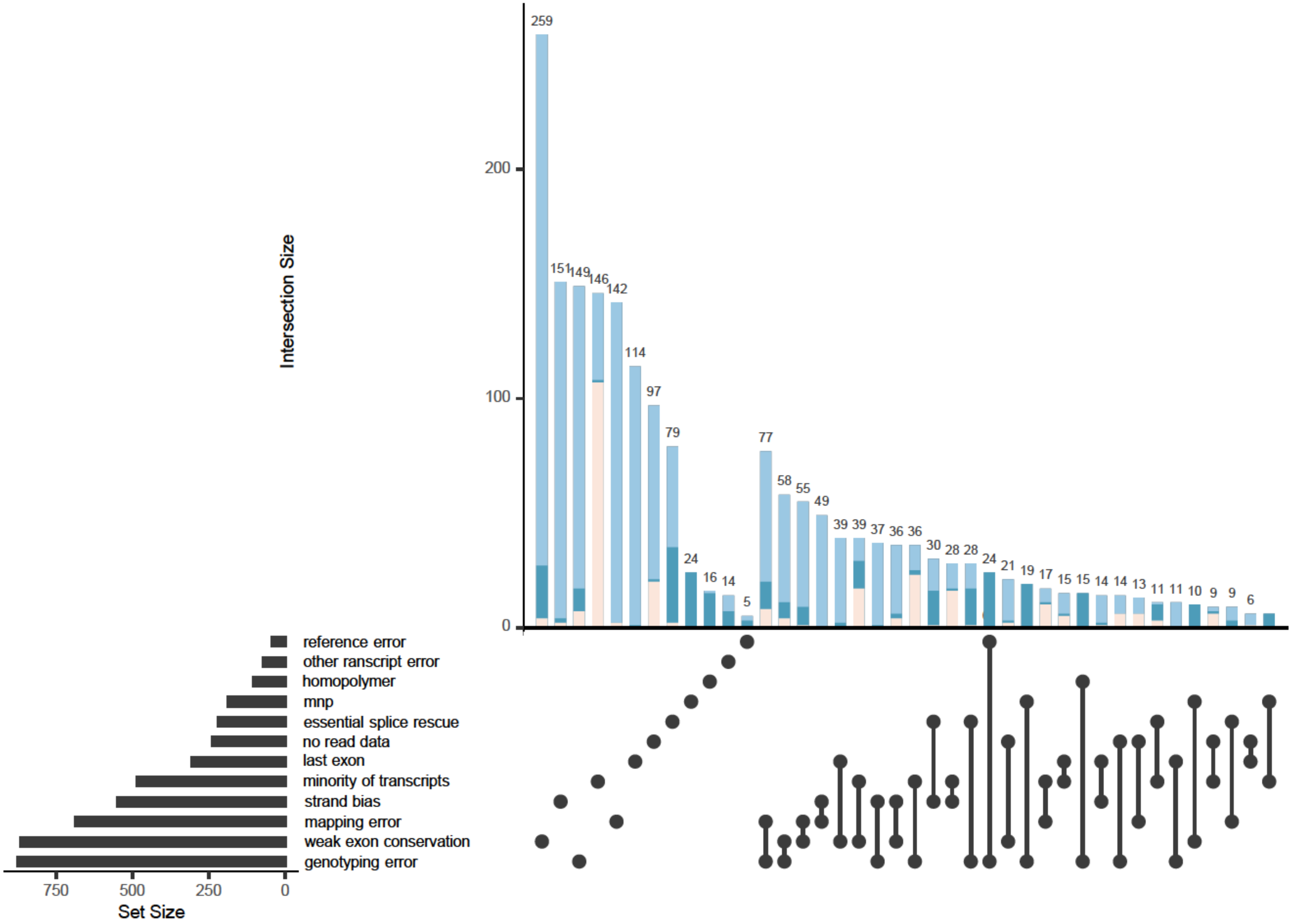
FIGURE 3.8 | LOSS-OF-FUNCTION VERDICT FROM MANUAL CURATION OF HOMOZYGOUS VARIANTS



*(**A**) Proportion of variants by verdict. (**B**) frequency of errors for removed variants. Mnp – multinucleotide polymorphism.*

The overlap of flags and outcomes are visualised in **Figure 3.9**. The most common overlapping feature was mapping and genotyping errors, however for the majority this did not result in the variant being classified as not LoF or likely not LoF.

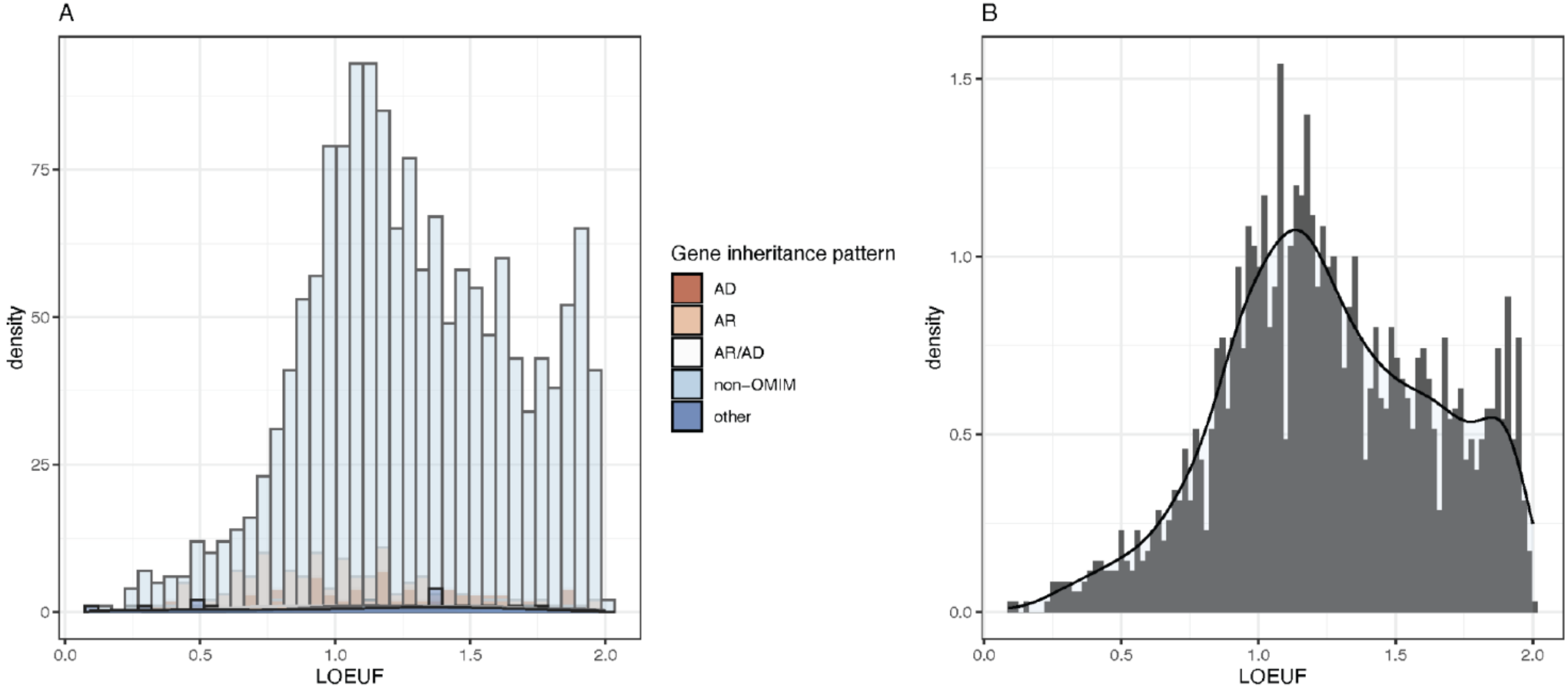
FIGURE 3.9 | OVERLAP OF FLAGS APPLIED TO MANUAL CURATION OF 3,362 PLOF HOMOZYGOUS VARIANTS



UpSet plot showing the overlap of flags applied to manual curation of 3,362 pLoF homozygous variants. Light blue represents the proportion of variants classified as LoF or likely LoF. Teal represents variants classified as not LoF or likely not LoF. Salmon represents variants that were classified as "uncertain". Mnp – multinucleotide polymorphism.

The final list of 1,815 genes tolerant of homozygous LoF variation (available in **Supplementary Table S2**) is depleted of essential genes (Fisher's exact test: OR = 0.10, p-value = 1.54 x 10^{-17}) obtained from https://github.com/macarthur-lab/gene_lists/tree/master/lists, with the majority being non-OMIM disease genes (**Figure 3.10A**). Further, the gene list is depleted of LoF constrained genes (p-value = 4.48 x 10^{-193}) using a LOEUF score <0.35 which represents genes considered to be intolerant to LoF, with many genes skewed towards the middle and end of the LOEUF spectrum (**Figure 3.10B**). Seventy-four genes had an OMIM autosomal dominant disease gene association, however the highest LOEUF score was 0.354 (**Supplementary Table S3**).

FIGURE 3.10 | 1,815 GENES TOLERANT OF HOMOZYGOUS LOF BY LOEUF SCORE AND OMIM DISEASE GENE STATUS



*(**A**) 1,815 genes tolerant of homozygous LoF across the LOEUF spectrum and separated by OMIM disease gene status. AD – autosomal dominant, AR – autosomal recessive, AR/AD – autosomal recessive and autosomal dominant, non-OMIM – non OMIM disease gene, other – anything else, including somatic mutations. (**B**) Density plot of 1,815 genes tolerant of homozygous LoF across the LOEUF spectrum.*

3.4 Discussion

3.4.1 Curation of 401 pLoF variants in 61 haploinsufficient disease genes

This project identified 401 high-quality pLoF variants in gnomAD v2.1.1 passing both sequencing and annotation quality filters in 61 haploinsufficient genes, whereby heterozygous pLoF variants are expected to cause severe developmental delay phenotypes with high penetrance. The gnomAD database comprises ostensibly healthy individuals, depleted of rare paediatric disease. Given the severity of expected phenotypes secondary to a heterozygous pLoF variant in anyone of the 61 haploinsufficient genes, and their low prevalence globally (ranging from 1 in 10,000 to less than 1 in 1,000,000), very few, if any true pLoF variants would be expected to be observed in gnomAD. Expectantly, 306/401 (76%) of these observed pLoF variants were artefacts, secondary to sequencing or annotation errors. Genotyping errors and variants being on the minority of transcripts were the most common error modes resulting in likely mis-annotation of pLoFs (**Figure 3.6**).

3.4.1.1 24% of pLoF variants in disease genes are predicted to cause LoF

Whilst manual curation removed 76% of variants in the 61 haploinsufficient disease genes, 24% of variants were curated as either LoF (11%) or likely LoF (13%), leaving 96 potentially "pathogenic" variants in individuals without evidence of severe paediatric disease. These variants could represent incomplete penetrance, although this is less likely as the genes were pre-selected as highly penetrant, but not impossible. It is possible that combinations of genomic variants or co-inherited protective alleles, environmental exposures, or mosaicism may impact the phenotypic consequences and penetrance of the 96 LoF variants.[188] Indeed, this is an active area of interest with research focused on identifying individuals resilient to rare disease despite harbouring pathogenic variants. It is hypothesised that *cis*-regulatory variation may modify the penetrance of coding variants in these cases.[90,189] Of course, another possibility is that the methods applied need refining and are still missing false positive LoF variants.

3.4.1.2 Transcript errors represent one of the most common error modes

Variants falling on low-confidence transcripts were identified as one of the most common error modes from manual curation. Transcript-specific annotation considers how a variant affects all annotated transcripts of that given gene. When comparing error modes identified from manual curation with the average pext score in GTEx, pext values were discriminately lower for variants that are annotated as possible transcript errors ($P = 4.1 \times 10^{-38}$, two-sided Wilcoxon test between transcript errors and other error modes). As a pext score can be calculated for each variant and transcript and exon conservation errors are common reasons to curate a variant as not LoF, the integration of pext into LoF curation could aid in the automated removal of false positive LoF variants.

3.4.2 Curation removes 34% of homozygous pLoF variants

Manual curation of 3,362 homozygous pLoF variants, observed in at least one individual in gnomAD, resulted in removal of 1141 (34%) of variants, with 2,221 high-confidence homozygous pLoF variants remaining in 1,815 unique genes. Fewer variants were removed than those in developmental disease genes, which was unsurprising as gnomAD is depleted for rare paediatric disease.

The most common error modes were genotyping errors, weak exon conservation, multinucleotide variants, and mapping errors. The final gene list of 1,815 genes is depleted of essential and disease genes, with most genes not having an OMIM-disease association. As expected, many genes fall within the higher end of the LOEUF spectrum suggesting these genes are tolerant to knockout. That said, there are a small number of genes (74 in total) with known autosomal dominant disease associations (**Supplementary Table S3**) although the highest LOEUF score was 0.354 suggesting that disease associations may represent gain of function mechanisms or transcript specific LoF. Indeed, the dominant disease gene with the highest LOEUF score was *TTN*. It is well documented that protein-truncating variants in *TTN* can cause dilated cardiomyopathy, yet LoF variants are present in healthy individuals, explained by different isoform expression.

3.4.3 Implications for ACMG-AMP guidelines

Current ACMG-AMP guidelines lack clear guidance on LoF curation. Many of the curation methods applied in this project are lacking from current guidance: ACMG-AMP do not enforce the use of read data for detecting artefacts. Methionine rescues are not discussed, only rescues in conjunction with initiation loss variants. ACMG-AMP do not explicitly mandate checking for exon conservation, nor the number or transcripts affected by a given variant. Allele balance, read depth, genotype quality, sequencing complexity, and strand bias are not discussed. There is no mention of reference errors or homopolymer repeats, nor multinucloeotide variants and frame restoring indels for nonsense and frameshift variants. There is no interpretation of cryptic splice sites/rescues, nor overprinting. This lack of curation of pLoF variants has important consequences for clinical labs interpreting pLoF variants. Seventy-six percent of pLoF variants identified in gnomAD in a haploinsufficient disease gene were errors. Many of these errors would pass current curation guidance by ACMG-AMP.

Working with colleagues at the Broad Institute, recommendations have now been provided on adjustments to the PVS1 criterion in the ACMG-AMP guidelines. These data are available in the following publication: *Singer-Berk, M., Gudmundsson, S., Baxter, S., SEABY, E. G., England, E., Wood, J. C., ... & O'Donnell-Luria, A. (2023). Advanced variant classification framework reduces the false positive rate of predicted loss-of-function variants in population sequencing data. The American Journal of Human Genetics.* (**Appendix Paper 6**).[284] This revised framework was applied to all high-confidence pLoF variants in 22 autosomal recessive disease genes from gnomAD v2.1.1. This revealed predicted LoF evasion of potential artefacts in 304/1113 (27.3%) of variants. The major reasons for LoF evasion were similar to data presented in this chapter, and included low pext scores, splice invasion, homopolymer repeats and presence in the last exon. Variants predicted to evade LoF were enriched for ClinVar benign variants and PVS1 was downgraded in 162/163 (99.4%) of predicted LoF evading variants assessed. In total 28/163 (17.2%) of variants were downgraded because of the revised framework. This translated to a change of variant classification with 20/28 (714%) of downgraded variants changing from likely pathogenic to a VUS.

3.4.4 Limitations

The methods applied to this chapter are not without limitations and potential criticism. Many of the rules applied are subjective, particularly pertaining to interpretation of weak exon conservation and technical errors by visualisation in the UCSC browser. This was perhaps evidenced by a 10% discordance between independent curators. Further, many of the rules developed were "best approximations", such as an arbitrary cut off ≥7 or greater (tri)nucleotide repeats being considered a homopolymer tract.

3.4.4.1 Gain of function

Whilst generating a list of 1,815 genes intolerant to homozygous knockout may prove useful as therapeutic drug targets, or even as genes predicted to not cause disease through total knockout, it is important to make clear that this curated gene list is not depleted for real disease-gene associations, particularly through gain-of-function (GoF). Sixty-four genes in the homozygous LoF tolerant list have known autosomal dominant associations. For example, *EHHADH* is a gene tolerant to homozygous knockout and has a LOEUF score of 0.97. Yet, published first-author work describes an heterozygous GoF variant in *EHHADH* causing idiopathic Fanconi syndrome.[285] This highlights that clinical interpretation of the "pathogenicity" of this gene list should be viewed with caution and interpreted only in the context of intolerance to LoF and not missense variation.

3.4.4.2 pext score

Whilst pext has shown promise in visually and computationally identifying weakly expressed transcripts, the metric is hard to interpret when pext scores are uniformly low across the gene (scores <0.2) due to overrepresentation of non-coding transcripts. One solution to this may be to calculate the delta pext score, which is the pext score as a percentage of the mean score to mitigate issues with both visual interpretation and absolute counts.

3.4.4.3 *Clonal haematopoiesis*

The gnomAD dataset attempts to exclude somatic variants by imposing an allele balance cut-off of 20-80% for heterozygous variants. Despite this, some somatic variants may remain. Indeed, a pathogenic LoF variant for Bohring-Opitz syndrome in *ASXL1* (p.Gly646Trpfs*12 is) present in 121 individuals in the gnomAD dataset. It is therefore possible that some pLoF variants may not be true loss-of-function variants because they represent somatic variants arising from clonal haematopoiesis. When assessing pLoF variants in known haploinsufficient disease genes, none of the genes selected were known to be associated with clonal haematopoiesis and neurodevelopmental disorders of autosomal dominant or X-linked inheritance, as defined by Brunet *et al.*[286] and Pich *et al.*[287] (**Table 3.3**). Similarly, none of these genes were present in **Supplementary Table S2**, which lists genes tolerant to homozygous knockout. In future, to reduce confounding by somatic mosaicism, it may be prudent to identify reported pathogenic/likely pathogenic variants present in gnomAD at allele frequencies higher than expected.

TABLE 3.3 | NEURODEVELOPMENTAL GENES ASSOCIATED WITH CLONAL HAEMATOPOIESIS

Associated genes	*AFF3, ARID2, ASXL1, BCOR, BRAF, BRCC3, CBL, CREBBP, CTCF, CUX1, DNMT3A, EZH2, FOXP1, GNB1, IDH2, KDM6A, KMT2C, KMT2D, KRAS, LZTR1, MYCN, NF1, NOTCH1, NRAS, PPM1D, PTPN11, PTPRD, RAD21, SETD2, SETDB1, SF3B1, SMC1A, SRSF2, STAG2, SUZ12, U2AF1*

All genes of autosomal dominant or X-linked inheritance.

3.5 Conclusions

Manual curation of pLoF variants identifies many errors missed by LOFTEE. These errors are more common in known haploinsufficient disease genes which is an important consideration when evaluating pLoF variants for clinical diagnostics, particularly when current ACMG-AMP guidelines lack guidance on LoF variant interpretation. However, manual curation is time-consuming and somewhat subjective, and methods may still require revising. Not all real 'to the best of our knowledge' LoF variants in disease genes manifest phenotypes and this may be explained by incomplete penetrance. On the other hand, genes tolerant to double knockout may still cause disease through gain of function.

3.5.1 Next steps

Manual curation identifies rules that could be incorporated into automated software, such as splice detection and transcript annotation. With increasingly rich datasets, supervised and unsupervised machine learning approaches may help identify factors or combinations of factors that may render a pLoF variant not real LoF, currently undetected by current methods. There is obvious need to improve the ACMG-AMP guidelines to incorporate many manual curation rules that would reclassify a “pathogenic” variant as benign. Lastly, there are opportunities to understand the mechanisms of potential incomplete penetrance for individuals in gnomAD harbouring putative neurodevelopmental pathogenic LoF variants.

Chapter 4 | A gene-to-patient approach accelerates novel disease gene discovery

4.0 Contribution statement

The concept and design of this study was developed by myself with input from my supervisory team. Patient data were collected and recorded by Genomics England (GEL). Data were accessible through my registered GEL project (RR359) titled: *Translational genomics: Optimising novel gene discovery for 100,000 rare disease patients*. The concept, methods, data processing, and analysis are all my own work with support from my supervisory team. The work presented herein is published as an Editor's Choice article in Genetics in Medicine: *SEABY, E.G., Smedley, D., Taylor Tavares, A.L., Brittain, H., van Jaarsveld, R.H., Baralle, D., Rehm, H.L., O'Donnell-Luria, A., Ennis, S. A gene-to-patient approaches uplifts novel disease gene discovery and identifies 18 putative novel disease genes. (2022). Genetics in Medicine* (**Appendix paper 7**). Additional data were published in a response to a letter to the editor: *Seaby, E. G., Baralle, D., Rehm, H. L., O'Donnell-Luria, A., & Ennis, S. (2022). Response to Ramos et al. Genetics in Medicine, 24(12), 2593-2594* (**Appendix paper 8**). The application of similar methods to 100,000 Genomes Project pilot data to identify novel candidates has been published in the 100,000 Genomes Project Pilot Report in the *New England Journal of Medicine: Smedley, D., Smith, K.R., … SEABY, E.G., … Caulfield, M. "The 100,000 Genomes Pilot on rare disease diagnosis in healthcare preliminary report." New England Journal of Medicine (2021)* (**Appendix Paper 9**).[132] One of the novel genes identified has been published as a phenotype case-series: *Janssen, B., van den Boogaard, M., Lichtenbelt, K., SEABY, E. G., et al. Loss-of-function variants in TAF4 are associated with a neurodevelopmental disorder. Human Mutation.* (**Appendix Paper 10**).[288]

4.1 Introduction

Next generation sequencing has revolutionised rare disease diagnostics; more patients than ever are receiving a molecular diagnosis for their rare genetic disorders. This has been driven by the ever-increasing rise in novel disease-gene discoveries, which is expanding the number of genes tested for in clinic.[155]

Making molecular genetic diagnoses is hugely important to patients and their families and can pave the way for therapeutic options, cascade testing, and family planning.[2]

However, most rare disease patients (up to 70% depending on clinical specialty) lack a definitive, molecular diagnosis.[1] Clinical genetic testing often involves application of a gene panel either as the ordered test or by the analysis strategy applied to exome and genome sequencing.[118] In the UK, the 100,000 Genomes Project only reported on variants in a pre-specified gene panel. Accredited clinical laboratories had no obligation to report on variants, including *de novo* variants outside of the panel applied.[132] Yet many patients harbour disease-causing variants not captured by a gene panel, or in genes yet to be associated with disease. Indeed, approximately 50% of genes thought to cause disease through haploinsufficiency are yet to be associated with a clinical phenotype.[122,289] Ergo, there is an unmet need for holistic and experimental approaches to identify novel disease genes and their associated phenotypes. These discoveries are critical for new genes to be added to diagnostic panels and for analytical approaches to uplift diagnostic rates.

4.1.1 Current barriers to novel gene discovery

Novel disease gene discovery is a protracted process that requires identifying multiple, unrelated patients with variants in the same gene affected with similar phenotypes. These 'discoveries' are then followed up with functional studies to provide evidence for gene causality.

Assessment of exome and genome data typically involves analysis of a small number of related individuals on a family-by-family basis. However, these analyses are time-consuming and resource intensive, often requiring commercial software and cross-checking public databases. Each family member has 3-4 million variants in their genome and ~30,000 variants in their genes. Assessing every potential disease-causing variant is simply impossible.[1] Whilst filtering techniques can restrict variant lists considerably, tens to thousands of variants of uncertain significance (VUS) typically remain with little to distinguish pathogenicity between them, particularly for genes of unknown function.[177] It is not possible to investigate all potential candidate variants since this necessitates intensive functional experiments on variants of ostensibly equal predicted causality, which is proving a major bottleneck. Researchers are reluctant to invest in expensive studies without persuasive evidence that a given candidate warrants pursuing; however, identifying which variants should be prioritised is challenged by the paucity of knowledge into the function of most human genes. Therefore, these VUSs end up as long lists of unreported variants present in a patient's sequencing results that no one has time to resolve or investigate further. In many cases, these lists will contain the causal variant and thus represent missed opportunities for molecular diagnosis.

4.1.2 The Matchmaker Exchange

One popular route to pursue candidate variants is through the Matchmaker Exchange (MME)[129] – please see **Chapter 1.7.3** for more detailed information. MME has proven successful in building case series of patients with shared phenotypes involving the same gene, which are later taken to publication.[157] However, this relies on knowing which gene candidates, of many, are best to submit to MME. Due to institutional restrictions on data sharing, it is not possible to query MME and return a list of genotypes and phenotypes for all submissions. Matches are only returned when two or more submitters match on the same gene or phenotype. Each match with another submitter requires electronic correspondence whereby both parties may choose to share variant and genotype specific data. Furthermore, there may be multiple matches per patient, making this method cumbersome and difficult to manage for large cohorts. Therefore, there are clear advantages to reduce the number of candidate variants for ongoing investigation.

4.1.3 100,000 Genomes Project

The 100,000 Genomes Project (**Methods 2.2.1**) was a government funded programme in the UK run by Genomics England (GEL) that has brought whole genome sequencing directly to patients with rare diseases through the National Health Service.[222] The project offered recruitment of patients for both clinical diagnostics and follow-on research, meaning that researchers can access anonymised data and further investigate variants of uncertain significance.[3,14]

4.1.4 Gene constraint

Mutation is random, giving rise to new variants of which most do not have biological impact; however, some variants have greater consequences and may help humans adapt and evolve, yet others may be harmful and cause disease. Natural selection purges deleterious variation from human populations as fewer individuals with damaging variants survive and reproduce. However, in large population databases, such as gnomAD, loss-of-function (LoF) variants are still observed because some genes are more tolerant than others to inactivation of one or even both gene alleles.[133] This principle can be exploited to identify genes with fewer LoF variants observed in population datasets compared with random and expected variant rates, signifying genes most intolerant to LoF.[109,133]

Karczewski *et al.*[133] developed the Loss-of-function Observed/Expected Upper-bound Fraction (LOEUF) score, which compared for each gene in gnomAD, the number of observed predicted LoF (pLoF) variants in 125,748 individuals compared with the number expected. LOEUF places >19,000 genes along a continuous spectrum of intolerance to gene inactivation, whereby low scores i.e., the fewest pLoF variants observed compared to expectation, are the most intolerant to LoF. Indeed, genes in the first LOEUF decile (equivalent to a score <0.2) have been validated as the most enriched for Online Mendelian Inheritance in (OMIM) haploinsufficient disease genes and show the greatest biological essentiality.[133] Yet, as of January 2021, 65% of genes in the lowest LOEUF decile are yet to have an OMIM disease association[2], leaving hundreds of undiscovered potential disease genes causing unrecognised phenotypes in patients.

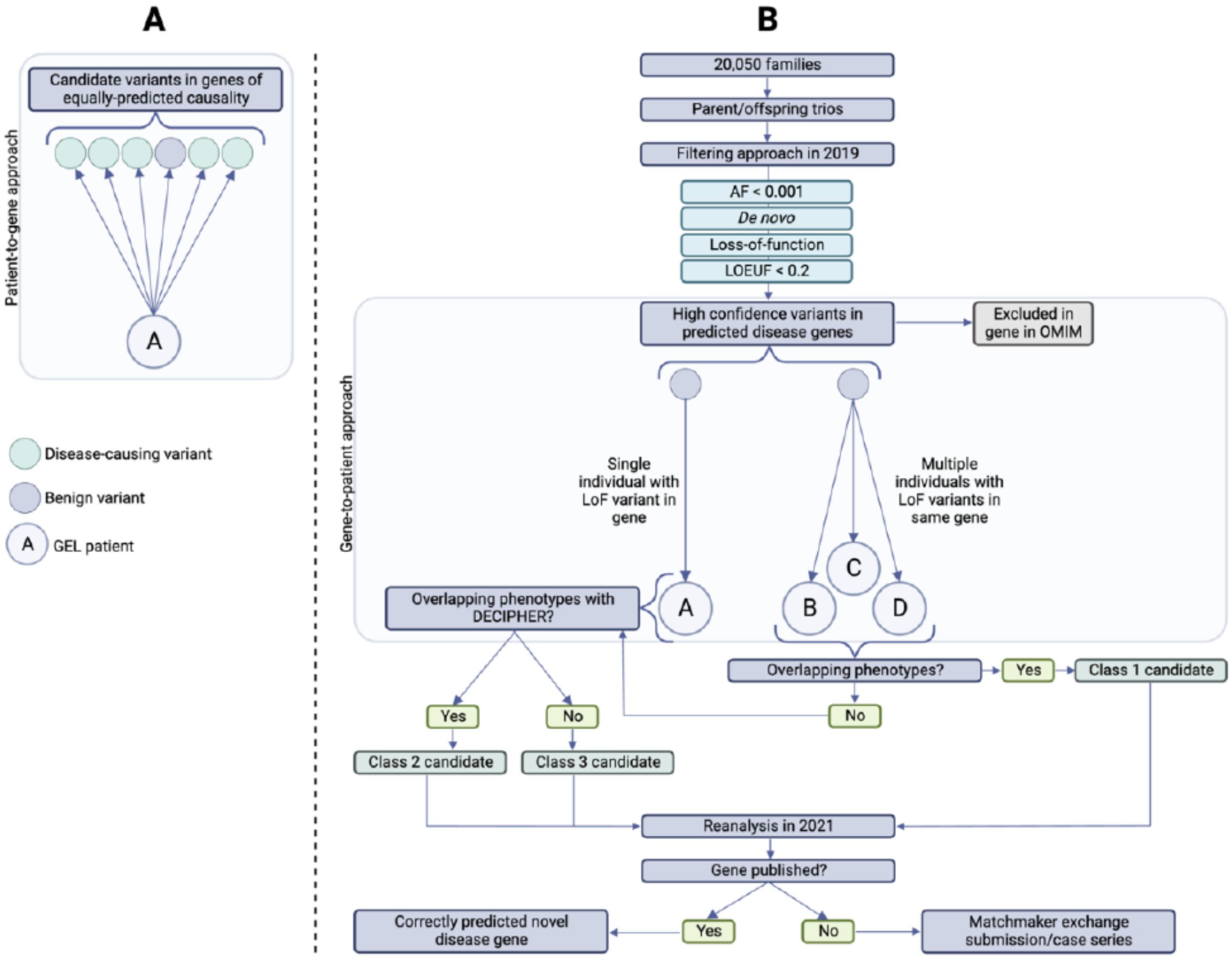
Whilst statistical methods exist to identify potential novel disease genes using excess *de novo* mutation analysis, such as DeNovoWEST[290] and DeNovolyzeR[212], these methods require huge cohorts of similar phenotypes, such as autism spectrum disorder. This study takes a non-statistical approach across a more heterogenous cohort and aims to uplift novel disease gene discovery by targeting pLoF variants (with the greatest pathogenic potential) in genes whereby inactivation of a single copy of the gene is highly probable to cause dominant disease. This method is applied to the 100,000 Genomes Project, which has brought genome sequencing directly to patients with rare diseases in the UK.[222] This method moves from a "patient-to-gene" approach to a "gene-to-patient" approach, whereby there is power to identify and assign rare putative pathogenic variation in predicted disease genes to patients, cohort-wide.

4.2 Methods

4.2.1 General methodological principle

This chapter proposes an objective filtering strategy that can be applied at scale. The LOEUF metric of intolerance to gene inactivation is applied to define a list of predicted haploinsufficient disease genes. Genes with a LOEUF score <0.2 (first decile) are selected, demonstrating the highest probability of representing autosomal dominant disease through haploinsufficiency.[133] By leveraging genomic and phenotypic data from rare disease trios in the 100,000 Genomes Project, an objective gene-to-patient approach is developed that filteres for rare, *de novo* pLoF variants in LoF constrained genes and matches these to rare disease patients. For this study, any variants in known OMIM disease genes (autosomal dominant or recessive) are excluded to focus exclusively on novel disease genes. Where more than one patient harbours a *de novo* pLoF variant in the same gene, this is defined as a *novel disease gene contender* and the patients are assessed for phenotype overlap. This approach reduces analytical noise to focus on the most likely novel disease genes (**Figure 4.1**) and identifies suitable candidates for functional validation.

FIGURE 4.1 | METHOD TO UPLIFT NOVEL DISEASE GENE DISCOVERY


A
B
Patient-to-gene approach
Candidate variants in genes of equally-predicted causality
A
Disease-causing variant
Benign variant
A GEL patient
20,050 families
Parent/offspring trios
Filtering approach in 2019
AF < 0.001
De novo
Loss-of-function
LOEUF < 0.2
Gene-to-patient approach
High confidence variants in predicted disease genes
Excluded in gene in OMIM
Single individual with LoF variant in gene
Multiple individuals with LoF variants in same gene
Overlapping phenotypes with DECIPHER?
A
B
C
D
Overlapping phenotypes?
Yes
Class 1 candidate
No
Yes
No
Class 2 candidate
Class 3 candidate
Reanalysis in 2021
Gene published?
Yes
No
Correctly predicted novel disease gene
Matchmaker exchange submission/case series

***(A)** A typical "patient-to-gene" approach, whereby patient A's exome or genome is analysed and multiple candidates remain of equally predicted causality. **(B)** A proposed "gene-to-patient" approach to identify novel disease genes, that challenges the widely adopted diagnostic analytical paradigm of exome and genome sequencing. 20,050 families with rare diseases were analysed from the 100,000 Genomes Project, with parent/offspring trios selected. In 2019 a filtering framework was applied across the entire cohort to extract de novo, putative LOF variants, with an allele frequency <0.001 in genes with a LOEUF score of <0.2. Variants were excluded from downstream analysis if the variant was in a known disease gene (as defined by OMIM). Predicted loss of function variants in novel genes constrained for loss of function, with a high probability of being disease-causing, were assigned to patients within GEL. Where multiple individuals in GEL had pLoF variants in the same candidate gene, overlapping phenotypes were assessed using recorded HPO terms. If shared phenotypes were present, the gene of interest was classified as Class 1. If no shared phenotypes were present between patients in GEL, their phenotypes were shared with high level phenotype data from DECIPHER. Overlapping features were classified as Class 2 and no overlap, Class 3. Where a single patient in GEL had a pLoF variant in a given candidate gene, their phenotype was also compared with DECIPHER and the gene was tiered accordingly. Data were reanalysed in 2021 to identify genes that had been added to OMIM or published since primary analysis in 2019. Unpublished genes are being taken forward as candidates.*

4.2.2 Data access

Access to the secure GEL research environment (RE) and high-performance cluster was obtained following information governance training and as a member of the Genomics England Clinical Interpretation Partnership (GeCIP): *Quantitative methods, machine learning, and functional genomics.* The research was approved with project ID: RR359 - *Translational genomics: Optimising novel gene discovery for 100,000 rare disease patients*. Access to the 100KGP dataset is restricted and only available as a registered GeCIP member in the Genomics England Research Environment. All data shared in this chapter were approved for export by Genomics England. The datasets and code supporting the current study, unavailable for export, are fully accessible within the Genomics England Research Environment in the shared directory: re_gecip/machine_learning/Ellie_Seaby/.

Access to the RE (Release V8), originally in 2019, provided access to an aggregate *vcf* file of 20,050 rare disease families called using the Illumina Starling pipeline and passing quality control parameters as previously described.[132] The majority of patients were children with neurodevelopmental disorders.[132]

Phenotype data for each patient were recorded by the referring clinician as a list of human phenotype ontology (HPO) terms.[167] The number of HPO terms varied considerably between patients with some individuals only having a single HPO term recorded. These data were stored within the RE in a LabKey data management system. The R LabKey package was used to extract HPO terms for each patient and merge these with genotype data.

4.2.3 Data filtering

Initial analysis was undertaken in October 2019. Full parent/offspring trios were selected for *de novo* analysis, reducing the number of available families from 20,050 to 13,949. Bespoke scripts utilising bcftools[291], VEP[252] and Exomiser[134] were developed to filter data. LOEUF scores were downloaded from gnomAD (http://gnomad.broadinstitute.org/downloads) and imported into the RE. Variants were excluded with an allele frequency (AF) >0.001 across all gnomAD populations and retained only if *de novo* and predicted loss-of-function (canonical splice site, frameshift, stop gain/nonsense, start loss, stop loss) on

RefSeq transcripts. Only variants in genes with a LOEUF score <0.2, representing the greatest LoF constraint, were retained. To account for potential false positive pLoF calls, LOFTEE v1.0 (https://github.com/konradjk/loftee) was applied, which removed low confidence variants such as those in the terminal exon. Variants remaining after LOFTEE filtering were deemed high confidence variants.

4.2.4 Merging genotype data with additional datasets

High confidence variants in putative disease genes that remained following the filtering approach in 2019 (AF < 0.001, *de novo*, pLoF, LOEUF <0.2) were classified as either found in a *known OMIM disease gene* (already associated with disease), or in a *non-OMIM disease gene* (not yet associated with disease); achieved by querying the OMIM application programme interface in October 2019. All novel disease gene contenders were compared with two mouse databases, the International Mouse Phenotyping Consortium (IMPC) database and Mouse Genome Informatics (MGI) database.[180,183]

4.2.5 Selecting high priority novel disease gene candidates

High confidence pLoF variants in novel disease gene contenders were selected as *candidate disease-causing variants*.

4.2.6 Phenotype overlap

Unrelated patients who shared a candidate pLoF variant in the same gene were assessed for phenotype overlap. This was achieved by computationally comparing HPO terms between individuals (using their coded identification number); a phenotype overlap occurred when any single HPO term matched exactly. Genes were prioritised as Class 1 candidates if more than one unrelated patient harboured a candidate disease-causing variant in the same gene and there was a phenotype overlap (**Table 4.1**). These novel disease gene contenders were further curated against the literature to ascertain if there were existing publications implicating any of the genes as disease-causing, prior to being indexed in OMIM.

For novel disease gene contenders with only one pLoF variant in the cohort (i.e., unique to one individual), high-level phenotypes were curated for each patient by manually upscaling their HPO terms to align with the terminology used in the publicly accessible DECIPHER (DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources) database (http://deciphergenomics.org). For example, hydrocephalus was upscaled to 'disorder of the nervous system' and atrial septal defect was coded as 'disorder of the cardiovascular system'.[127] High-level phenotypes of GEL patients were then compared with DECIPHER patients harbouring *de novo* variants (pLoF or missense) in the same gene. *De novo* missense variants in DECIPHER were included to increase the number of genes with an associated phenotype for comparison. When high-level phenotypes matched, these genes were classified as Class 2 candidates. In Class 3 candidate genes, phenotypes did not match, or no comparison was available (**Table 4.1**).

TABLE 4.1 | CLASSIFICATION OF NOVEL DISEASE GENE CONTENDERS

Gene Class	Classification rule
Class 1	pLoF variants identified in the same gene in two or more unrelated kindred in GEL and at least one HPO term exactly matched between affected individuals.
Class 2	A pLoF variant identified in a gene in one affected individual in GEL, whereby at least one high-level phenotype exactly overlapped between the GEL patient and an individual in DECIPHER with a *de novo* pLoF or missense variant in the same gene.
Class 3	No overlap in HPO terms between patients in GEL with pLoF variants in the same gene, or no overlap in high-level phenotypes between GEL patients and affected patients in DECIPHER with *de novo* pLoF or missense variants in the same gene. Or no available phenotype in DECIPHER for comparison.

4.2.7 Taking candidates forward

Permission was sought to submit genes to GeneMatcher[250] for Class 1 genes by filling in Clinician Contact Request forms within the RE. These forms were individually completed for all Class 1 candidates to obtain more detailed and current phenotype information from the patients' referring clinician, in addition to obtaining consent to share genotypes, phenotypes, and consent for publication with any matches made using the GeneMatcher node of MME. Where a successful match occurred and a case series was already underway, collaborations were sought with the patients' clinician to include their patient in the existing case series. Where no case series were established, new interest groups were initiated. This involved collecting phenotype data from collaborators and initiating collaboration with colleagues at the University of Portsmouth in conducting functional experiments in *Xenopus* on novel disease gene contenders.

4.2.8 Validation of method

To validate whether the method could correctly predict novel disease genes, novel disease gene contenders in 2019 were compared against an updated list of dominant OMIM disease genes from 2021, in addition to literature published between 2019 and 2021. If one of the predicted novel disease gene contenders from 2019 was added to OMIM or was published as a disease gene between 2019 and 2021, HPO terms of GEL patients were manually compared with the clinical phenotypes reported in the literature and/or OMIM to assess concordance. The method was considered to have correctly predicted a disease gene when any of the patients in GEL had significant overlapping features with the clinical presentation published for variants in the same gene and the GEL variant would meet, at minimum, likely pathogenic status by ACMG-AMP guidelines.[123,292] Any alternative diagnoses made by NHS accredited genetics laboratories between 2019 and 2021 were also assessed.

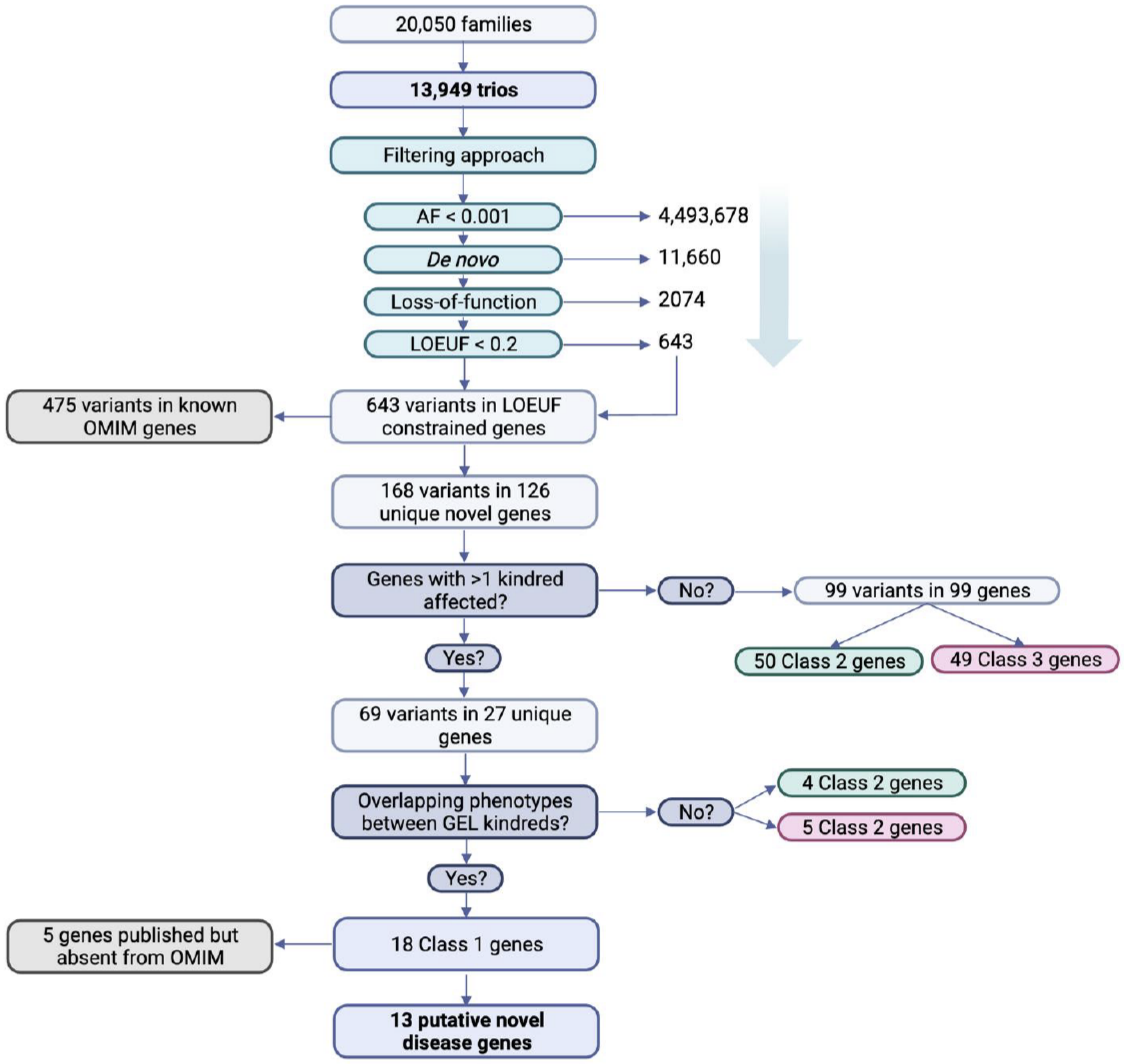
4.3 Results

Data from the 100,000 Genomes Project (13,949 trios, involving 41,847 individuals) revealed 643 rare (AF <0.001), *de novo* pLoF events filtered in 1,044 pLoF-constrained genes (**Figure 4.2**). 475 variants were in 148 known OMIM genes (as of October 2019) and 168 were in novel disease gene contenders (involving 126 unique genes). Of these, 27 genes had more than one GEL kindred affected and 18 had overlapping phenotypes, meeting Class 1 criteria (**Table 4.2**). Of these Class 1 genes, five were absent from OMIM but had been published in the literature (**Supplementary Table S4**). Six more of these genes have since been published as disease-causing genes with matching phenotypes to the GEL probands (**Table 4.2**).

Nine genes had more than one GEL kindred affected but the phenotypes between patients were non-overlapping meaning that there were no exact matches of HPO terms between patients; 4 genes met Class 2 criteria with high-level phenotypes overlapping with DECIPHER entries, and 5 genes met Class 3 criteria (**Supplementary Table S4**).

Ninety-nine variants in 99 unique genes were identified in 98 individuals (**Supplementary Table S5**). Of these, 50 genes were classified as Class 2 candidates meaning their high-level phenotypes overlapped with individuals in DECIPHER harbouring *de novo* pLoF or missense variants in the same gene. Forty-nine genes were Class 3 meaning no patients within GEL or DECIPHER had matching phenotypes involving the same gene.

FIGURE 4.2 | SUMMARY OF CLASS 1, 2 AND 3 RESULTS


20,050 families
13,949 trios
Filtering approach
AF < 0.001
4,493,678
De novo
11,660
Loss-of-function
2074
LOEUF < 0.2
643
475 variants in known OMIM genes
643 variants in LOEUF constrained genes
168 variants in 126 unique novel genes
Genes with >1 kindred affected?
No?
99 variants in 99 genes
50 Class 2 genes
49 Class 3 genes
Yes?
69 variants in 27 unique genes
Overlapping phenotypes between GEL kindreds?
No?
4 Class 2 genes
5 Class 2 genes
Yes?
5 genes published but absent from OMIM
18 Class 1 genes
13 putative novel disease genes

Summary of Class 1, Class 2, and Class 3 results. Six hundred and forty-three de novo predicted loss of function variants, with an allele frequency <0.001 were identified in genes with a LOEUF score <0.2 from 13,949 trios. Four hundred and seventy-five variants were in known disease (OMIM) genes. One hundred and sixty-eight variants remained in 126 unique novel candidate genes. There were 27 genes whereby more than one patient harboured a pLoF variant (69 variants in total).

Rapid and agnostic filtering of trios in GEL identified 18 Class 1 putative disease genes in 2019. Removing the five genes published in 2019 but absent from OMIM, left 13 putative novel disease genes of which seven have since been confirmed as disease-causing genes, matching the patient phenotypes that were identified up to two years prior to publication (**Table 4.2**).

TABLE 4.2 | 13 PUTATIVE NOVEL DISEASE GENES

Gene	Consequence	maxFreq (%)	Shared HPO terms across patients in GEL	Overlapping features between GEL patients and published literature	Publication status (June 2021)
HDLBP	frameshift frameshift start lost	0.000 0.000 0.000	Macrocephaly, intellectual disability, global developmental delay, delayed speech and language development, delayed fine and gross motor development, autism	N/A	Case series and functional studies underway
RIF1	frameshift frameshift	0.000 0.000	Delayed speech and language development, global developmental delay, delayed gross motor development, intellectual disability	N/A	Case series and functional studies underway
DDX17	frameshift splice acceptor	0.000 0.000	Horizontal nystagmus, global developmental delay, skeletal abnormalities	N/A	Case series and functional studies underway
CLASP1	stop gained stop lost	0.020 0.000	Delayed speech and language, global developmental delay, delayed gross motor development, intellectual disability	N/A	Case series underway
ANKRD12	splice acceptor frameshift frameshift	0.000 0.001 0.000	Intellectual disability, global developmental delay	N/A	Case series and functional studies underway
CASZ1	frameshift frameshift	0.000 0.000	Intellectual disability	N/A	Case series underway
TAF4	frameshift stop gained splice donor	0.000 0.000 0.008	Seizures, spasticity, brain atrophy, cerebellar signs	Intellectual disability, abnormal behaviour, abnormal brain MRI	https://onlinelibrary.wiley.com/doi/full/10.1002/humu.24444
ZNF292	frameshift frameshift	0.001 0.000	Global developmental delay, facial shape abnormalities, and intellectual disability	Intellectual disability, Global developmental delay, delayed speech, microcephaly, skeletal abnormalities, seizures, dysmorphic features, abnormal face shape	https://www.nature.com/articles/s41436-019-0693-9
SETD1A	stop gained frameshift	0.000 0.000	Global developmental delay, intellectual disability	Delayed speech and language development, intellectual disability, seizures, global developmental delay, dysmorphic facial features, hypotonia	https://link.springer.com/article/10.1007/s12264-019-00400-w
ANKRD17	stop gained frameshift	0.000 0.000	Delayed speech and language developmental, delayed gross motor development, intellectual disability	Intellectual disability, delayed speech and language development, dysmorphic features	https://www.sciencedirect.com/science/article/abs/pii/S0002929721001385
USP7	stop gained stop gained	0.000 0.007	Global developmental delay and abnormal facial shape	Intellectual disability, seizures, hypotonia, global developmental delay, facial shape deformation, feeding difficulties	https://www.nature.com/articles/s41436-019-0433-1
TANC2	stop gained frameshift	0.000 0.000	Intellectual disability	Intellectual disability, global developmental delay, behavioural abnormalities, autism, impaired speech development, seizures, delayed motor development	https://www.nature.com/articles/s41467-019-12435-8
SPEN	stop gained stop gained	0.000 0.000	Intellectual disability	Developmental delay/intellectual disability, autism spectrum disorder, behavioural abnormalities, dysmorphic features, and obesity/increased BMI	https://www.sciencedirect.com/science/article/abs/pii/S000292972100015X

A table of 13 putative novel disease genes identified from analysis in 2019. Shared phenotypes between patients involving pLoF variants in the same gene are listed. Grey cells highlight candidates that have since been published in June 2021. For these, shared phenotypes between patients in GEL and patients included in publications by 2021 are recorded. Abbreviations: BMI – body mass index; GEL – Genomics England; HPO – Human phenotype ontology; MaxFreq – Maximum allele frequency in gnomAD v2.1.1 and 1000 Genomes phase 3 data; N/A – not available for comparison.

4.3.1 Investigating and validating putative disease genes

By 2023, 24/126 (19%) of the novel disease gene contenders identified were published by independent groups. Class 1 candidates were the highest predictors of disease genes with 12/18 (67%) having been functionally validated and published confirming their status as new disease genes. Of the Class 2 and Class 3 genes occurring in unique individuals, 2/50 (4%) and 10/49 (20%) have been published with evidence of causality, respectively. Of the remaining six Class 1 genes yet to be validated, case series/and or functional experiments are underway. By 2023, 20 patients had likely pathogenic or pathogenic variants independently identified in alternative known disease genes by GEL diagnostic laboratories. In total, 126 novel disease gene contenders were identified.

4.4 Discussion

An objective filtering strategy was rapidly applied across a large cohort and identified 18 high-confidence putative novel disease genes of which 12/18 (67%) have since been validated through functional experiments and confirmed as disease-causing. Additionally, further 108 novel disease gene contenders were identified.

In total, 24/126 (19%) of the genes identified in the study have been validated as disease-causing and diagnoses are being returned to patients who would otherwise have a negative genome report. This was achieved by a targeted gene-to-patient approach applied to the 100,000 Genomes Project with the power to detect very rare, pLoF variation in genes most intolerant to LoF. However, only in time will the full specificity as well as sensitivity of the approach be determined.

4.4.1 Class 1 genes and internal matches in GEL

Since initial analysis, 12/18 (67%) Class 1 genes have undergone functional validation and been published by independent groups confirming their status as novel disease-gene discoveries and it is anticipated that this number to increase over time. Class 1 genes outperformed Classes 2 and 3 (Fisher's exact test <0.0001) likely due to the greater specificity and granularity of phenotypes available for internal matching within GEL.

There were nine genes whereby unrelated patients in GEL had pLoF variants in the same gene, yet no patient shared the same HPO term. However, three of these genes have since been published and the published phenotypes overlap with the GEL patients (**Supplementary Table S4**). This may be explained by variability in HPO terms reported in GEL; some patients had many HPO terms recorded, yet others had only one or two. In **Table 4.2**, patients with pLoF variants in *SPEN* and *TANC2* only overlapped by one HPO term (intellectual disability). Yet, when the disease phenotype was further delineated in published case series for both genes, many more features observed in the GEL patients were consistent with the reported phenotypic spectrum. This highlights the need for longitudinal and deep phenotyping data in automated gene discovery studies.

Due to the automated process of exact HPO term matching between GEL patients, there was potential to miss overlapping phenotypes recorded with subtly different nomenclature, e.g. one patient with intellectual disability (HP:0001249) would not match another patient with mild intellectual disability (HP:0001256).

4.4.2 Class 2 and 3 genes

More Class 3 genes 10/50 (20%), were published as novel causal genes by 2021 than Class 2 genes 2/49 (4%); Fisher's exact test 0.027. This may be due to small sample sizes but could reflect a weakness in Class 2 and 3 classification (**Table 4.1**). Comparing high-level phenotypes is potentially problematic as it lacks the granularity required to assess clinical overlap. Furthermore, high-level phenotypes of GEL patients were compared with patients in DECIPHER harbouring *de novo* missense variants, which are considerably more common and less likely to be pathogenic, increasing the possibility of false disease-phenotype associations. Additionally, it is possible that some patients in the cohort were also in DECIPHER, however due to data anonymity this could not be verified.

4.4.3 Lessons learnt

Class 2 and Class 3 genes may be better assessed through MME. Sharing more detailed phenotype data would provide the granularity to assess true clinical overlap. In GEL, this step involved contacting the patient's clinician for permission to share data with matches through MME; and this process was not always successful. As MME involves manual correspondence between peers, this cannot be easily automated, highlighting the advantages of internal phenotype matching within the same cohort. MME has been fully utilised for novel gene candidates following contact with the referring clinician, however presenting these results is not possible due to restrictions on data sharing.

4.4.4 Novel gene discovery remains time consuming

Whilst the method presented is rapid at identifying highly promising novel candidate genes, there remains persistent time requirements to validate any results through case series and functional experiments; however, the strength of the method is in rapidly identifying which VUSs to pursue and therefore shortening the process of discovery. Six Class 1 genes have been identified and case series are underway in all. Functional studies have started in four Class 1 genes (**Table 4.2**) and two genes are presented in this thesis (**Chapters 5** and **6**). Therefore, this method provides the opportunity to identify salient candidates for follow-on studies, providing patients with the opportunity to have their "most damaging" VUS investigated, when typically, no candidates would have been pursued.

4.4.5 Clinical implications

The discovery of novel disease genes facilitates both anterograde and retrograde diagnoses for patients. Clinical gene panels are constantly updated as new genes are discovered. Although not commonplace, there are increasing drives to re-analyse existing exome and genome data considering novel discoveries.[293]

Making diagnoses is extremely important for patients and their families and helps to end the notorious "diagnostic odyssey" of rare genetic disease that burdens so many patients. Whilst for many patients, a molecular diagnosis does not offer an immediate therapy, diagnoses are still highly valuable and can link patients and families with support networks, provide knowledge on disease prognosis, facilitate participation in follow-on research studies, and inform reproductive choices.[1] However, for a subset of patients, a diagnosis does confer access to life-altering therapies including repurposed drugs, gene augmentation therapies, small molecule therapies, and antisense oligonucleotides.

This method extends beyond novel gene discovery and is being utilised to rapidly identify pathogenic variants in known disease genes. These data are presented in **Chapter 7**.

4.4.6 Considerations and limitations

4.4.6.1 Requires trios with predominantly neurodevelopmental phenotypes

Trios were utilised for *de novo* analysis, meaning that families without trio data were excluded; this raises a legitimate concern regarding genetic test inequity across the globe. It is important to caution on the interpretability of pLoF variants in constrained genes without available segregation data. It has been calculated that if trio data were unavailable in the 100,000 Genomes Project, >2100 high priority pLoF variants would need curating, of which many would be present in an unaffected parent. In fact, in the 100,000 Genomes Project, only 6% of pLoF variants passing QC in novel genes constrained for LoF were *de novo*.[294] Therefore, where possible, trio data still remains the most powerful tool to increase confidence in pLoF variant pathogenicity. For circumstances where segregation is unavailable, more caution in the interpretation of pLoF variants is recommended.

The 100,000 Genomes Project is enriched for patients with rare neurodevelopmental disorders and therefore there is risk when comparing patient phenotypes within a cohort enriched for similar phenotypes.[132] Caution was applied when defining ‘phenotype overlap’ as any two patients exactly matching on one HPO term, however due to the variability in number of HPO terms reported in GEL, this maximised sensitivity of Class 1 genes and enabled the correct prediction of *SPEN* and *TANC2* as novel disease genes.

4.4.6.2 Statistical rigor

The large number of neurodevelopmental disorders in the dataset caused by many heterogenous genes precludes the reliability of statistical methods to confirm/refute novel disease gene contenders, although with ongoing genome sequencing in the UK this will likely be overcome. Further, there was specific focus on pLoF variants only, meaning that there is insufficient power even with 3 *de novo* variants per gene (the maximum observed for class 1 candidates) to reach statistical significance using a case/control Fisher's test and multiple test correction.[212] Instead there is reliance on the established approach of identifying overlapping phenotypes to further prioritise the best candidates for functional validation.

4.4.6.3 Allele frequency

A liberal AF of <0.001 was used, yet the highest variant frequency observed was 0.0002. The presence of these rare, *de novo* variants within gnomAD could represent recurrent *de novo* variation.[295] While a more restrictive AF would increase confidence of pathogenicity, pathogenic disease variants can be present in population databases due to incomplete penetrance, effects of cis-regulatory variation, and adult-onset disease.[280] Nevertheless, this studied cohort is probably depleted for adult-onset diseases as early-onset conditions are more likely to have complete trios.

4.4.6.4 Prioritising haploinsufficiency

This method is enriched for haploinsufficient disease genes, and biallelic observations were not prioritised in this analysis.[133] Using a LOEUF score <0.2 (top decile), enabled selection of the genes most highly constrained for LoF, although the expectation was that these would be associated with dominant inheritance, meaning the approach is not enriched for autosomal recessive novel gene discovery. Several genes in the top decile may be embryonically lethal, although observation of *de novo* LoF variants in these genes is not expected in this cohort. With higher LOEUF thresholds it is likely that further haploinsufficient disease genes and even more recessive disease genes will be found, but at the expense of increased noise.[133]

4.4.6.5 Classification of pLoF variants

Start loss and stop loss were included within the category of LoF variants, however these variants often do not constitute true LoF and show selection signatures more similar to missense variants.[116] Six start/stop loss variants were observed, and therefore potential misclassification of these variants is not expected to have substantively impacted the analysis. The current analysis strategy also misses other LoF variants e.g. untranslated region variants, extended splice site, and structural variants (SV). Research into these potential LoF disrupting variants using tools such as UTRannotator[296], spliceAI[64] may further expand the disease gene candidate list.

4.4.6.6 *False positive pLoF variants*

Not all pLoF variants truly cause LoF; many are enriched for technical, rescue, and impact errors (see **Chapter 3**).[133,280] Whilst *in silico* tools can identify some of these errors, manual curation is the most effective method to identify potential false positives.[280] However, this process is extremely time-consuming and not yet standardised; therefore there is risk that false positive LoF variants were included in the analysis.[280] It is expected that these false positives are more likely to be variants with higher allele frequencies in gnomAD or in Class 2 and 3 genes whereby detailed phenotype data cannot be assessed for overlap. Indeed, 15 of the pLoF variants in Class 2 and 3 genes were in individuals who had an alternative disease-causing variant (**Supplementary Table S5**). Whilst this does not rule out the potential for a second diagnosis, which occurs up to 5% of the time[122], it does raise the possibility of a variant without functional impact.

4.5 Conclusion

Utilising a large cohort and adopting a highly efficient gene-based approach can accelerate novel gene discovery and target the most appropriate variants and genes for functional validation. This can uplift diagnostic rates and add new disease genes to clinical gene panels.

As rare disease cohorts continue to increase, there is increasing demand to automate analyses and reduce the burden of variants requiring analysis by clinical scientists. With increasing study sizes, this method should be better powered to detect rare disease-causing variation shared across individuals but necessitates real-time comparison to previously generated large datasets if the approach is to be used in routine diagnostics. Assessing phenotype overlap is an important methodological step and, with drives towards data sharing, there is opportunity to securely access data and apply automated phenotype matching within and across cohorts using trusted research environments, such as the NHGRI's Genomic Data Science Analysis, Visualisation, and Informatics Lab (AnVIL) space.[297]

It is anticipated that this method can be applied by other researchers to their own cohorts; however, it is important to emphasise need for trio analyses, and encourage prudence when determining what constitutes LoF. This work demonstrates that gene-based approaches can successfully identify novel disease genes, and with larger rare disease cohorts it is hoped that more discoveries will be identified for the benefit of patients, their families, and the wider scientific community.

Chapter 5 | Seaby-Ennis Syndrome

5.0 Contribution statement

This chapter builds upon the results of **Chapter 4**, whereby *DDX17* was identified as a novel disease gene candidate. I curated an international case series of 11 patients, obtaining consent and detailed phenotype data, and coordinated an effort bringing together local and international collaborators to functionally validate the gene as disease-causing. *Xenopus* modelling was performed by A. Godwin, T. Fletcher, and M. Guille at the University of Portsmouth. Mouse modelling was undertaken by J. Courchet's laboratory in Lyon. RNA-seq analysis was performed by C. Bourgeois' laboratory in Lyon. All other work is my own. This work is available as a first-author preprint and is in review at the *American Journal of Human Genetics*: *SEABY, E. G., Godwin, A., Clerc, V., … Ennis, S. (2023). Monoallelic de novo variants in DDX17 cause a novel neurodevelopmental disorder. medRxiv, 2023-09.*

5.1 Introduction

RNA helicases have essential biochemical roles in all aspects of RNA metabolism, including unwinding and annealing RNA molecules and remodelling ribonucleoprotein complexes.[298,299] They share twelve conserved motifs including the signature DEAD motif compromising the amino acid sequence Asp-Glu-Ala-Asp. DDX17, also known as DEAD box protein 17, (and its close homolog DDX5), are highly energy dependent DEAD-box RNA helicases involved in diverse cellular processes, notably gene expression, biogenesis of miRNAs via their interaction with the Drosha/DGCR8 complex, and the regulation of cell fate switches and biological transitions.[300,301] They are coregulators of several transcription factors including MYOD, a master regulator of muscle differentiation and SMAD proteins, which mediate transforming growth factor beta induced epithelial-to-mesenchymal transition.[301,302] Additionally, they are components of the spliceosome and regulate alternative splicing.

DDX17, located on chromosome 22q13.2, has been shown to be involved in the control of Repressor Element 1-silencing transcription factor (REST) related processes that are critical during the early phases of neuronal differentiation.[303] Through its association with REST, DDX17 promotes its binding to the promoter of certain REST-targeted genes and coregulates the transcriptional repression activity of REST. DDX17 and the REST complex are downregulated during neuroblastoma cell differentiation, affecting activation of neuronal genes. Furthermore, DDX17 and DDX5 regulate the expression of multiple proneural microRNAs which target the REST complex during neurogenesis, implicating DDX17 in neuronal gene repression.[303] In 2022, Suthapot *et al.*[304] focused on characterising chromatin occupancy of DDX17 and DDX5 in hPSCs NTERA2 and their neuronal derivates. They showed that the expression of both helicases is abundant throughout neural differentiation of the hPSCs NTERA2, preferentially localised within the nucleus and that they occupy chromatin genome-wide at regions associated with genes related to neurogenesis. Both DDX17 and DDX5 are mutually required for controlling transcriptional expression of these neurogenesis-associated genes but are not important for maintenance of the stem cell state of hPSCs. In contrast, they are critical for early neural differentiation of hPSCs, possibly due to their role in the upregulation of key neurogenic transcription factors such as SOX1, SOX21, SOX2, ASCL1, NEUROG2 and PAX6. Critically, DDX17 and DDX5 are important for differentiation of hPSCs towards NESTIN and TUBB3 positive cells, which represent neural progenitors and mature neurons, respectively. However, those studies used a DDX17 and DDX5 co-depletion approach to address the function of these factors in neurogenesis, and information regarding the specific contribution of each helicase to this process is lacking.

To date, *DDX17* has no disease-gene relationship. The *DDX17* gene is highly constrained for loss-of-function (LoF), i.e. fewer LoF variants in *DDX17* are observed in population datasets than would be expected under a null mutational hypothesis (37.7 expected, 1 observed, pLI=1.0 in gnomAD). Genes can be quantified by constraint to LoF using the Loss-of-function Observed/Expected Upper-bound Fraction (LOEUF) score, which places genes along a continuous spectrum of intolerance to haploinsufficiency.[26] Genes highly constrained for LoF, represented by low LOEUF scores, are highly associated with known haploinsufficient disease genes.[26,305] However, the majority of genes in the lowest LOEUF decile are not yet associated with a disease phenotype but may be expected to cause disease if mutated through LoF.[306] *DDX17* has a LOEUF score of 0.13, suggesting that haploinsufficiency of the gene is not tolerated. Considering its role in neuronal differentiation, muscle differentiation and alternative splicing, one might expect that *DDX17* is an essential gene in neurodevelopment and may present such a phenotype in humans. Therefore, identification of patients with LoF variants in *DDX17* may help characterize a new gene-disease relationship.

5.2 Methods

5.2.1 Study subjects and data acquisition

Access to the 100,000 Genomes Project data was obtained through membership of a Genomics England Clinical Interpretation Partnership, with approved project RR359: *Translational genomics: Optimising novel gene discovery for 100,000 rare disease patients*. Deidentified whole genome sequencing and phenotype data (stored as human phenotype ontology terms) were accessible in the Genomics England Research Environment.

Additional study subjects were identified through GeneMatcher.[307] In total, 11 patients consented to participate in the study. Parents and legal guardians of all affected individuals provided written consent for the publication of their results alongside genetic and clinical information. Guardians of patients 5, 6, 7 and 10 explicitly consented to have photographs published.

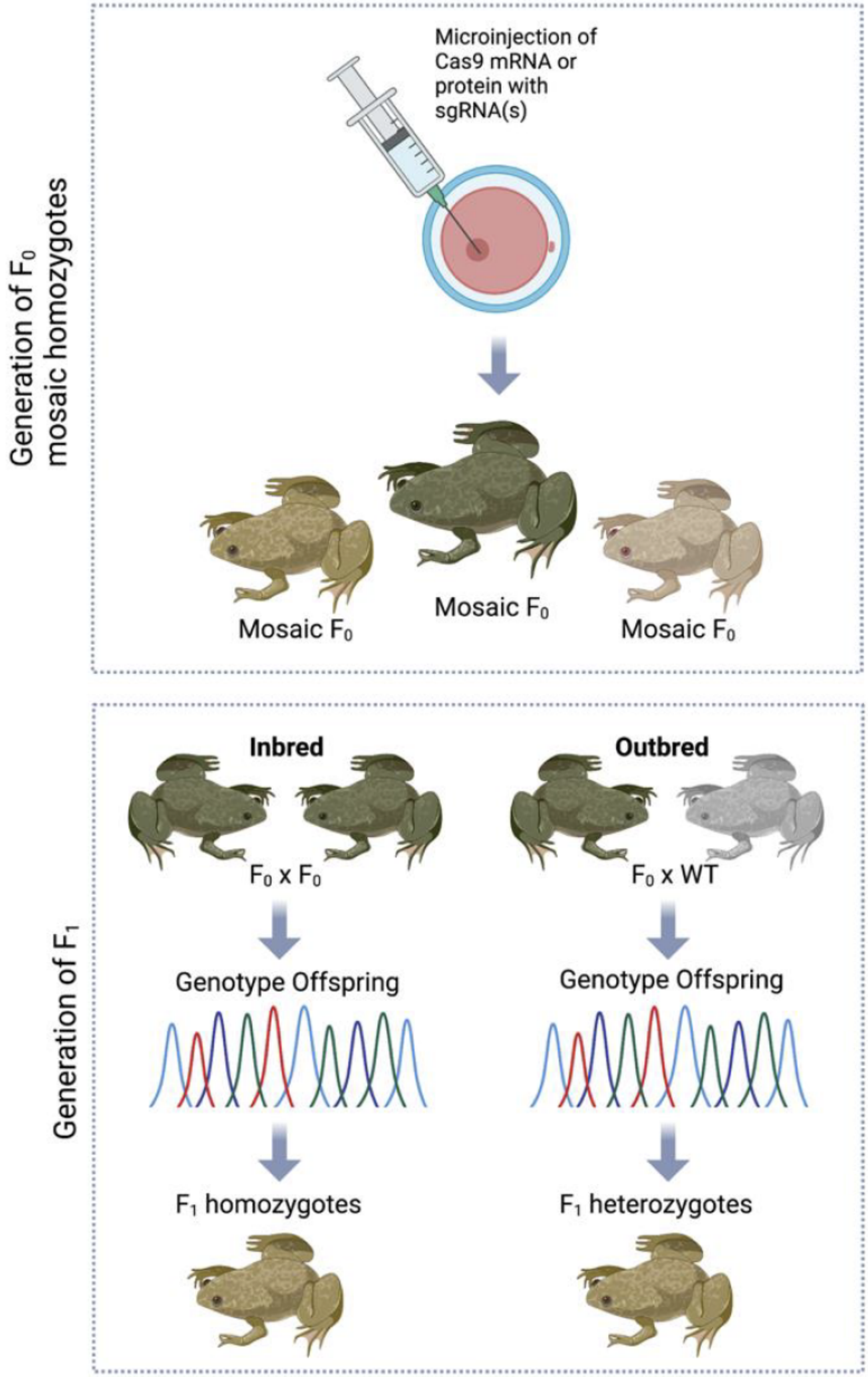
5.2.2 Sequencing and data analysis

All patients, except for patient 3, had trio exome sequencing performed. Data processing and variant filtering and prioritisation were carried out by in house pipelines at respective host centres. Patient 3 had trio whole genome sequencing undertaken as part of the 100,000 Genomes Project[132] and their data was filtered using the DeNovoLOEUF filtering strategy (see **Chapter 7**).[305]

5.2.3 *Xenopus* methods

Adult Nigerian strain *Xenopus tropicalis* were housed within the European *Xenopus* Resource Centre (EXRC; https://xenopusresource.org), University of Portsmouth, in recirculating MBK Ltd systems maintained at 24ºC - 27ºC (13-11-hour light-dark cycle) with 10% daily water changes. All *Xenopus* work was completed in accordance with the Home Office Code of Practice under PP4353452 following ethical approval from the University of Portsmouth's Animal Welfare and Ethical Review Body. Detailed methods on the generation of *X. tropicalis* founder animals and ddx17 crispants; wholemount *in situ* hybridisation; phenotypic analysis; experimental design and statistical analysis are available in **Methods 2.8.1.** In short, homozygous and heterozygous ddx17 crispants were generated (**Figure 5.1**). *Ddx17* crispants were assessed for morphological abnormalities. They were further subjected to three behavioural assays (Free Movement Pattern (FMP) Y-Maze, Light-dark transition, and *Xenopus* Locomotion) to test neurodevelopmental phenotypes. Finally, frog brains were examined for neuronal outgrowth.

FIGURE 5.1 | GENERATION OF F_0 AND F_1 *XENOPUS* MODELS



Generation of F_0 mosaic crispant homozygotes using microinjection of Cas9 mRNA or protein with sgRNA(s). F_1 generations occur through natural mating, either inbred (F_0 crispant x F_0 crispant) or outbred (F_0 crispant x wild-type (WT)). F_1 animals are genotyped to identify whether they are F_1 crispant homozygotes or F_1 crispant heterozygotes.

5.2.4 RNA-seq methods

Detailed methods are available in **Methods 2.9**. In summary, human SH-SY5Y neuroblastoma cells (ECACC #94030304) were grown and transfected with 60 nM of a mixture of 2 different siRNA against *DDX17* (Merck-Millipore, see sequences in **Supplementary Table S6**). Protein extraction was carried out as previously described[268] and total RNA were isolated. Directional RNA libraries were prepared from total RNA after removal of ribosomal RNA (lncRNA library, Novogene). High throughput sequencing of 150 bp paired-end reads was carried out on an Illumina Novaseq 6000 platform (Novogene), generating an average number of 75 million matched pairs of reads per sample. Raw reads were pre-processed and mapped reads were filtered using SAMtools[272] and the number of reads per gene was counted using HTSeq.[273] Differential gene analysis was carried out with the DESeq2[274] package. Parameters for differential expression: P-value < 0.05, [log2(FC)] ≥ 0.50 and base mean ≥ 10. Gene ontology and gene-set enrichment analyses were carried out using the ShinyGO 0.76.3 web interface.[275]

5.2.5 Mouse methods

Mouse breeding and handling was performed according to experimental protocols approved by the CECCAPP Ethics committee (C2EA15) of the University of Lyon, and in accordance with the French and European legislation. Detailed methods on *ex vivo* cortical electroporation and primary neuronal cultures; immunostaining; *in utero* cortical electroporation; immunohistochemistry; image acquisition; and quantifications and statistical analyses are available in **Chapter 2.8.2**.

5.3 Results

The DeNovoLOEUF filtering strategy was applied, as previously described[305] and detailed in **Chapter 7**, to 13,494 parent/offspring trios in the 100,00 Genomes Project, focusing on genes with a LOEUF score <0.2 with no prior disease gene association. One individual harbouring a heterozygous pLoF variant in *DDX17* was identified. Using the GeneMatcher platform, a further 10 patients with *de novo* variants in *DDX17* were identified all presenting with neurodevelopmental phenotypes. Representatives for these participants were then invited to join our research study and referring clinicians were asked to complete a standardised phenotype table (**Supplementary Dataset SD4**). The summary of the phenotypic features of the 11 patients (from 11 independent families) harbouring *de novo* heterozygous variants in *DDX17* are provided in **Table 5.1** and further detailed in **Supplementary Dataset SD4**. All variants were absent from gnomAD[26] v2.1.1 and v.3.1.2.

TABLE 5.1 | CORE PHENOTYPIC FEATURES OF COHORT WITH HETEROZYGOUS DE NOVO VARIANTS IN *DDX17*

	Predicted Loss of Function					Missense					
Patient	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
Age (at last visit)	3y 8m	13y	13y 8m	17y 6m	17y 6m	4y	7y 0m	15y	16y	17y 3m	23y
Sex	F	M	M	M	M	M	M	F	M	F	M
Variant coordinates (GRCh38)	22:38499457_38499458del	22:38494928C>CAT	22:38495939T>C	22:38494973TC>T	22:38493717delT	22:38493736T>C	22:38495016C>T	22:38498504G>A	22:38495900T>C	22:38498463T>C	22:38506032G>T
Variant consequence	Arg161Glyfs*7	c.997_998dupAT	c.739-2A>G	Arg318Hisfs*36	Asn462Metfs*16	Gln454Arg	Arg304His	Pro203Leu	Gln259Arg	Thr217Ala	Ala69Asp
Macrocephaly	N	N	N	Y	At birth	N	N	N	Y	N	N
Dysmorphic facial features	Y	Y	N	Y	Y	Y	N	Y	N	Y	N
Age walking independently	24m	18m	16m	14m	14m	>2y	18m	2y	14m	27m	unknown
Intellectual disability	mild-moderate	N	mild-moderate	mild	mild	unknown	N	moderate	mild	N	moderate

ADHD	N	N	Y	Y	Y	unknown	N	N	Y	unknown	Y
Autism spectrum disorder	N	Y	Y	N	N	N	N	N	Y	Y	N
Delayed speech and language development	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Global developmental delay	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y
Neurodevelopmental delay	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y
Gross motor delay	Y	Y	N	N	N	Y	Y	Y	Y	Y	N
Fine motor delay	Y	N	Y	N	Y	Y	Y	Y	Y	Y	Y
Stereotypy	N	Y	N	N	N	Y	Y	N	N	Y	N
Generalised hypotonia	Y	Y	N	N	N	Y	N	Y	N	Y	N

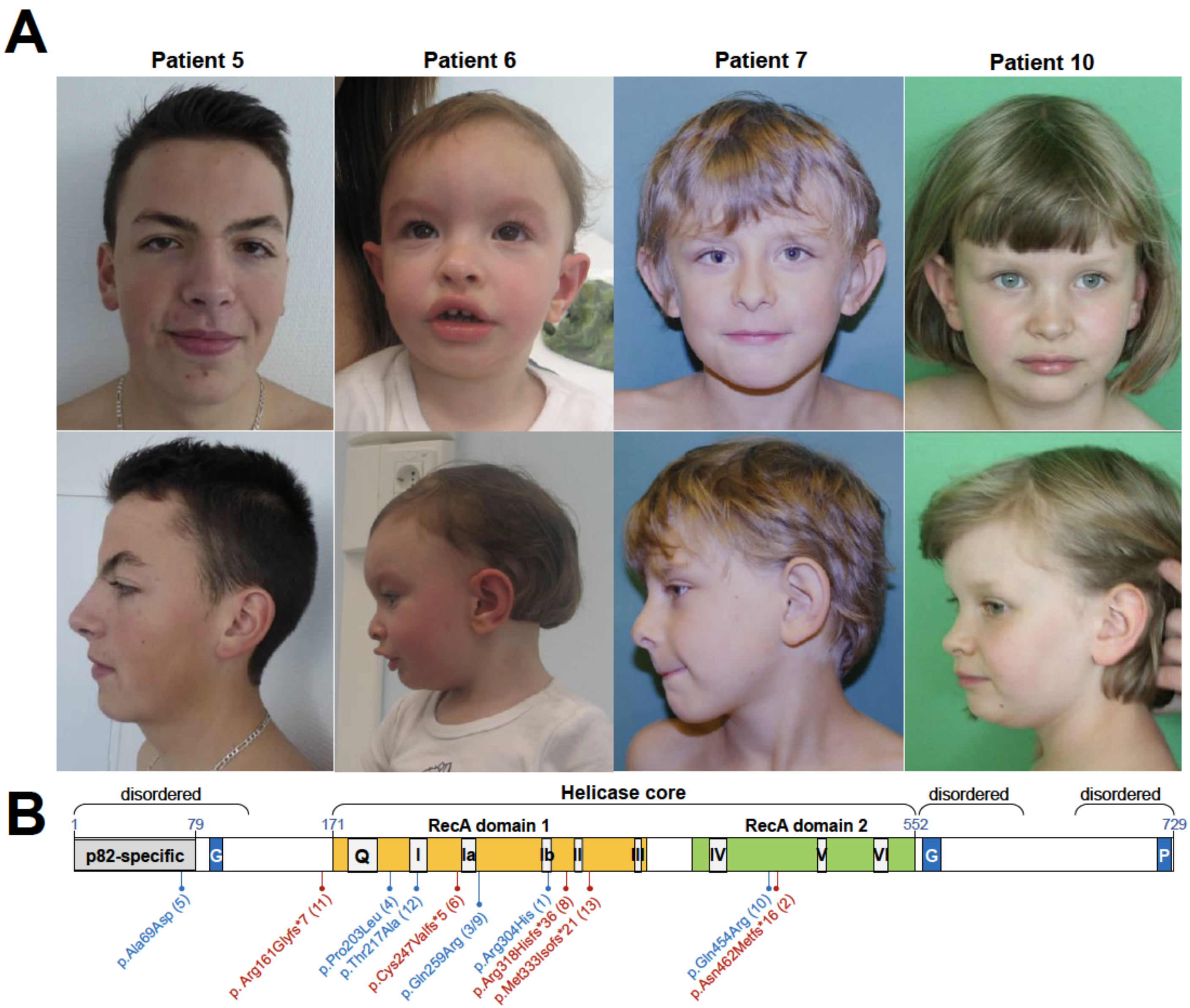
ADHD - attention deficit hyperactivity disorder; F – female; M – male; m – months; N – no; Y – yes; y – years. Blue cells denote presence of a feature. Grey cells represent unavailable data.

The cohort comprises 8 males and 3 females, all of whom are alive and have a median age of 15y at the latest available follow up. The median age of walking was 18 months. Intellectual disability, ranging from moderate to mild (IQ 56-83), is prevalent in 7/10 (70%) of patients. Sixty-four percent (7/11) of the cohort have dysmorphic facial features. Overlapping facial dysmorphology between patients includes: telecanthus; synophrys; upslanting palpebral fissures; depressed nasal bridge; posteriorly rotated ears; high arched eyebrows; epicanthus; telecanthus; frontal bossing; micrognathia and strabismus. Three patients have 5th finger clinodactyly, two have clubfeet, and three have 2,3 toe syndactyly (**Supplementary Dataset SD4**). Fifty-six percent (5/9) have attention deficit hyperactivity disorder (ADHD) and 4/11 (36%) have features of autism. Ninety-one percent (10/11) have delayed speech and language development, 9/11 (82%) have global developmental delay, and 11/11 (100%) have neurodevelopmental delay. Gross motor delay is prevalent in 8/11 (73%), and 9/11 (82%) have fine motor delay. Thirty-six percent (4/11) have stereotypy and 5/11 (45%) have generalised hypotonia. Patient 9 (height 178cm (Z=0.40); weight, 55.7kg (Z= -0.89)) and patient 4 (height 165cm (7th percentile); weight, 51.5kg (4th percentile)) have signs of macrocephaly (**Supplementary Dataset SD4**) with Z-scores of 3.97 and 3.23, respectively. Patient 5 had signs of macrocephaly at birth, which normalised through infancy. Eight participants had brain MRI scans of which 4 patients showed abnormalities including: left lateral compartment greater than right; monolateral temporal cortical dysplasia; asymmetry of the cerebral cortex and right sided nonspecific demyelination; generalised brain demyelination, and periventricular white matter hyperintensities (**Supplementary Dataset SD4**). No other obvious asymmetry was observed.

5.3.1 Molecular genetic findings

Following variant filtering and prioritisation of exome and genome data, no likely pathogenic or pathogenic variants as curated using American College of Medical Genetics and the Association for Molecular Pathology (ACMG-AMP) guidelines[276] were identified that fully explained the patients' phenotypes. All participating centres identified *de novo* variants of uncertain significance in *DDX17* that were of sufficient interest to submit to GeneMatcher. Five variants were pLoF and 6 variants were missense (**Figure 5.2**).

FIGURE 5.2 | *DDX17* PATIENT VARIANTS AND PHOTOGRAPHS



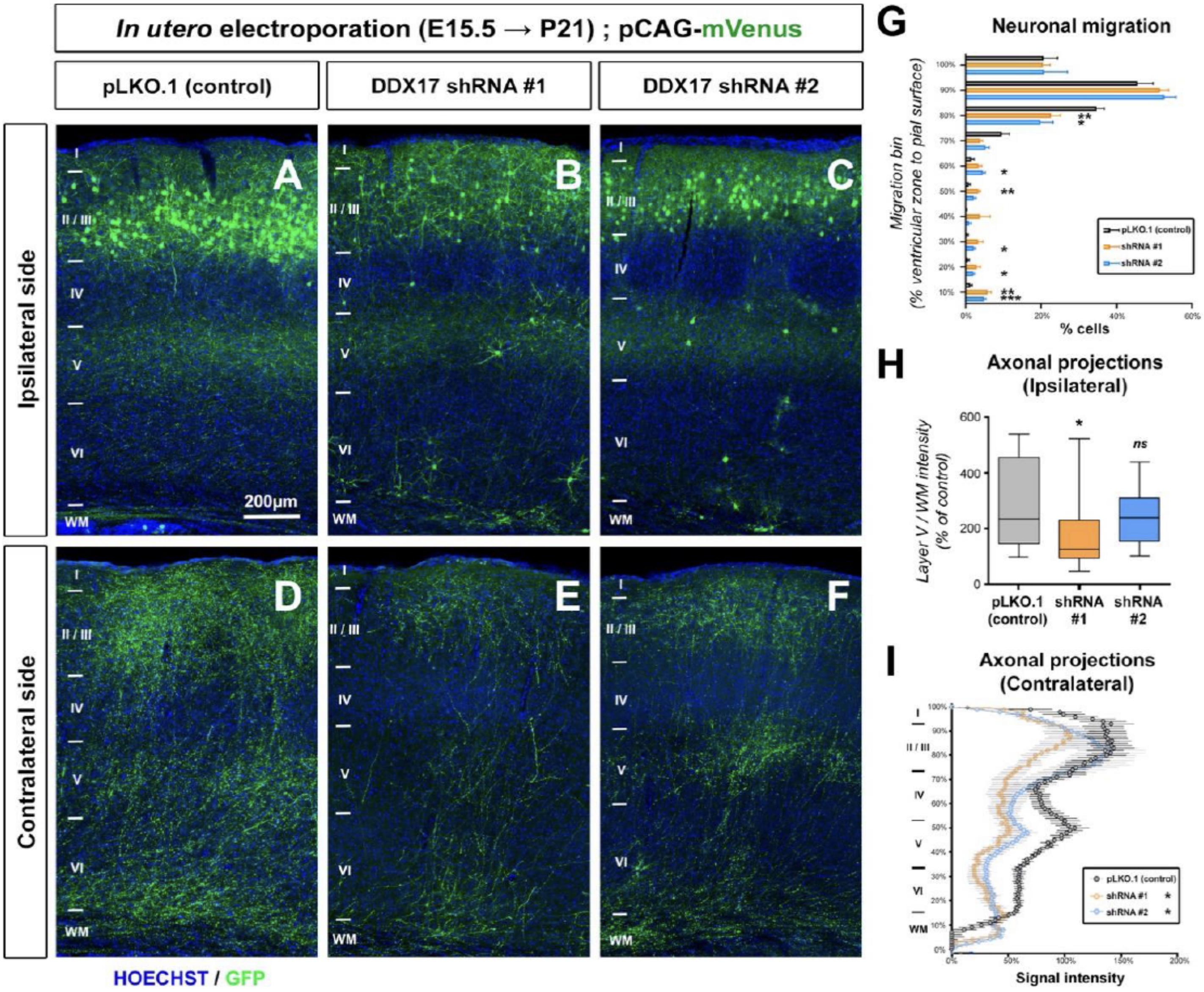
***A**. Photographs of patients 1, 2, 10 and 12. Shared features reported between these patients include posteriorly rotated ears, arched eyebrows, telecanthus, and depressed nasal ridge. **B**. Gene ideogram whereby variants in blue are missense and variants in red are loss-of-function.*

Four patients in the cohort had additional variants of uncertain significance reported (**Supplementary Dataset SD4**). Patient 7 had compound heterozygous pathogenic variants in *ACADM* associated with medium chain fatty acid dehydrogenase deficiency; NM_000016.5(ACADM):c.799G>A,p.(Gly267Arg) and NM_000016.5(ACADM):c.985A>G,p.(Lys329Glu). Patient 3 had a *de novo* Yq11.21-qter deletion and *de novo* Ypter-p11.3 duplication for which the significance is unknown. Patient 6 had a VUS in *HCFC1* associated with methylmalonic aciduria and homocysteinaemia; NM_005334.3(HCFC1):c.4418C>T.p.(Thr1407Met) and patient 10 harboured a maternally inherited 16q23.3 (81,477,800 – 81,552,781) VUS.

5.3.2 DDX17 supports cortical neuron development in the mouse

The identification of several variants in *DDX17* associated with neurodevelopmental features prompted assessment of DDX17 reduction on cortical development in animal models. Julien Courchet and his team performed *In Utero* Cortical Electroporations (IUCE) in the mouse using two distinct shRNA plasmids targeting *Ddx17* (shDDX17 #1 or shDDX17 #2). Electroporations were performed at embryonic day (E)15.5, the developmental stage at which progenitors give rise to callosal-projecting pyramidal neurons. By P21, in control conditions, all electroporated neurons (visualised by mVenus fluorescence) reached the superficial layers of the cortex (layers II/III) (**Fig 5.3A**). In contrast, upon knockdown of *Ddx17,* defects in neuronal migration were observed (**Fig 5.3B-C**) and after quantification, a statistically significant fraction of neurons did not reach the most superficial cortical layers in conditions electroporated with shRNA plasmids (**Fig 5.3G**). Despite this, neuronal polarisation and axon formation was not impaired. In control conditions, axons of layer II/III neurons progress through the corpus callosum to reach the contralateral hemisphere, and branch extensively on ipsilateral layer V (**Fig 5.3A**), as well as contralateral layers II/III (**Fig 5.3D**). In shRNA-electroporated animals, a trend toward a decrease of axon density in the ipsilateral side was observed (**Fig 4a.2B-C,** quantified in **Fig 5.3H**), and especially a strong reduction of contralateral axon density (**Fig 5.3E-F**, quantified in **Fig 5.3I**). There was no difference in axon density in the white matter (WM), indicating that the same proportion of axons reached the contralateral hemisphere regardless of *Ddx17* expression. These results demonstrate that DDX17 is required for cortical development in the developing mouse brain.

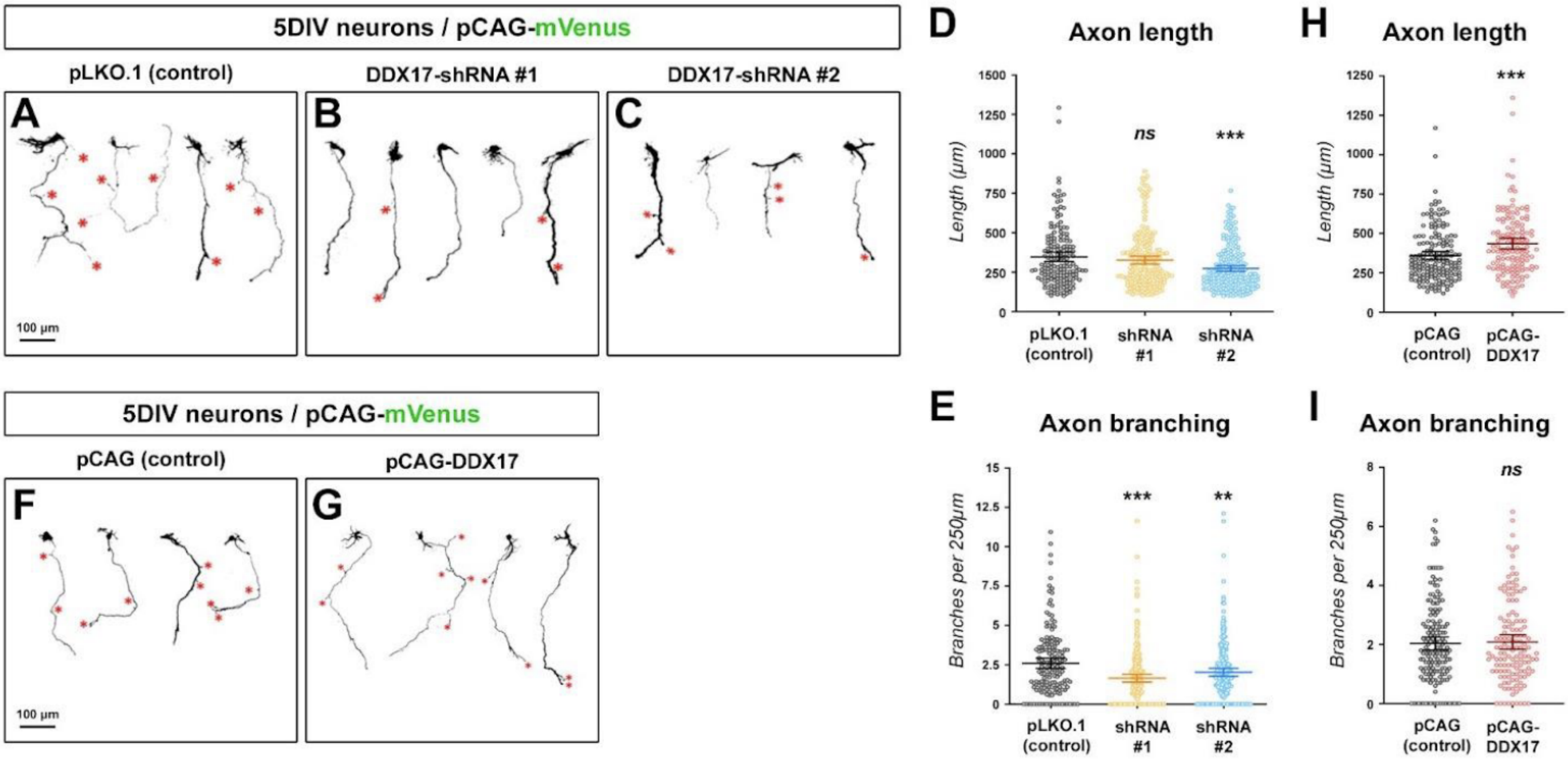
FIGURE 5.3 | KNOCKDOWN OF DDX17 DECREASES CORTICAL AXON COMPLEXITY IN THE MOUSE *IN VIVO*



*Results produced by J. Courchet. **(A-F)** Histochemistry of the ipsilateral or contralateral side of mice at P21 following in uteroelectroporation with pLKO **(A,D)** or after loss of function of DDX17 (**B-C** and **E-F**) and the fluorescent protein mVenus. (**G**) Quantification of neuronal migration defects upon knockdown of DDX17. Soma position was quantified on a ventricular zone to pial surface axis. Each bin represents 10% migration. Data: average + SEM, N = 6 sections out of 3 animals (2 sections per animal). Analysis: Two-way ANOVA with multiple comparisons. *p<0.05, **p<0.01, ***p<0.001. **(H)** Quantification of normalised mVenus fluorescence in layer V of the ipsilateral cortex (min, max, median, 25th, and 75th percentile). NpLKO=13, NshDDX17-1=18, NshDDX17-2=25. Analysis: One-way ANOVA with Dunn's multiple comparisons. ns:p>0.05, *p<0.05. (I) Quantification of normalised mVenus fluorescence along a radial axis in the contralateral cortex (Average ±SEM) (H) in control condition pLKO) or after knockdown of DDX17. NpLKO=13, NshDDX17-1=18, NshDDX17-2=25. Analysis: Two-way ANOVA. *p<0.05.*

Because the wiring of the brain results from sequential biological processes, defects in the early steps (e.g. neurogenesis or neuronal migration) can lead to alterations in the later biological processes such as axon development. To ensure that the axonal phenotypes observed *in vivo* do not result from abnormal neuronal migration, *in vitro* neuronal cultures were used to quantitatively assess axonal development at a single cell resolution. Julien Courchet's team performed *Ex Vivo* Cortical Electroporation (EVCE) at E15.5 to target neuronal progenitors in the dorsal telencephalon and cultured neurons for 5 days *in vitro* (5 DIV). In both shRNA conditions, axonal development was impaired compared to the control condition (pLKO.1) (**Fig 5.4A-C**). The inhibition of DDX17 expression decreased axon length and reduced collateral branch formation (**Fig 5.4D-E**). Two independent shRNA plasmids produced markedly similar phenotypes, indicating that this phenotype is unlikely to be an off-target effect of the shRNAs. To confirm this result, overexpressing human DDX17 by EVCE was tested. Following electroporation, an increased axon length compared to the control condition was observed, mirroring the effect of DDX17 knockdown (**Fig 5.4F-G**, quantified in **Fig 5.4H**). Interestingly, more collaterals were counted per neuron, this increase was due to the increase in axon length and axon branching did not differ from the control condition when normalised per axon length. Overall, these results demonstrate that DDX17 is important for axon development in mouse cortical neurons.

FIGURE 5.4 | DDX17 IS NECESSARY AND SUFFICIENT FOR AXON DEVELOPMENT

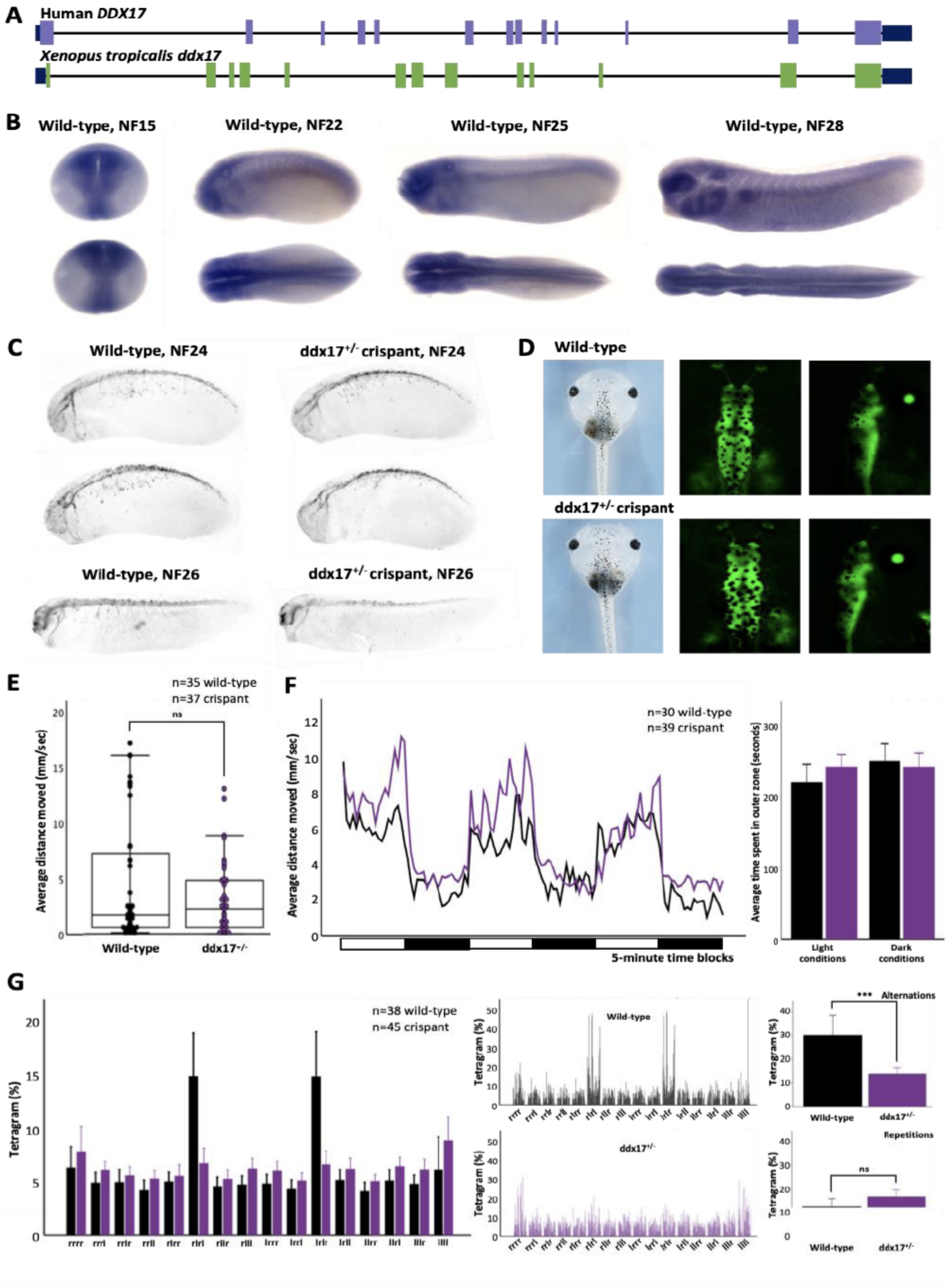


*Results produced by J. Courchet. **(A-C)** Representative images of mVenus expressing cortical neurons (5 DIV) in control condition (pLKO.1) or after loss of function (DDX17-shRNA #1 and DDX17-shRNA #2). Red star (*) point to collateral branches of the axon. **(D-E)** Quantification of axon length and number of collateral branches of 5 DIV neurons in the indicated conditions. Bars represent the average and 95% CI. Statistical tests: Kruskal-Wallis test with Dunn's post-test (each condition compared to control condition). **(F-G)** Representative images of mVenus expressing cortical neurons (5 DIV) in control conditions, or upon overexpression of DDX17. Red star (*) points to branch/collateral position. **(H-I)** Quantification of axon length and number of collateral branches of 5 DIV neurons in the indicated conditions. Bars represent the average and 95% CI. Statistical tests: Kruskal-Wallis test with Dunn's post-test. (D-E) $N_{(pLKO.1)}$=169 , $N_{(shDDX17\ \#1)}$=205, $N_{(shDDX17\ \#2)}$=228. (H-I) $N_{(pCAG)}$=168, $N_{(pCAG\text{-}DDX17)}$=134. ns: p>0.05, **: p<0.01, ***: p<0.001*

5.3.3 *ddx17* crispants have reduced axon outgrowth and working memory

To test the effect of loss-of-function *ddx17* variants in an intact animal, crispant *X. tropicalis* models were used; the exon structure of the human and *Xenopus* genes are similar (**Fig 5.5A**) and the proteins produced have 68% amino acid identity (**Fig S2A**). The expression pattern of *ddx17* mRNA has not previously been reported in *Xenopus* and *in situ* hybridisation shows it to be expressed most highly in neural tissues including the migratory neural crest, brain, eye, and otic vesicle (see the purple staining in **Fig 5.5B**).

FIGURE 5.5 | HETEROZYGOUS *DDX17 X. TROPICALIS* CRISPANTS APPEAR MORPHOLOGICALLY NORMAL BUT SHOW REDUCED AXON OUTGROWTH AND HAVE A WORKING MEMORY DEFICIT



Results produced by A. Godwin. ***A.*** *Xenopus tropicalis and Homo sapiens have the same exon-intron structure.* ***B.*** *A developmental series of wild-type X. tropicalis were fixed and underwent in situ hybridisation with a probe specific for ddx17, the blue stain shows where this gene is expressed. The highest levels of ddx17 mRNA are in neural tissues although it is detectable more widely.* ***C.*** *Control and crispant embryos were fixed at the stages shown and stained for neuron bodies and axons using HNK1 monoclonal antibody. The extension of axons ventrally from the neural tube is reduced in crispants at stage 24 (4/4 embryos) although growth does continue (see stage 26), scoring was blind and prior to genotyping.* ***D.*** *Brightfield microscopy showed no clear distinctions between control and crispant tadpoles across a range of stages and, when the neural tissue was labelled transgenically this too failed to reveal any gross-morphological distinctions.* ***E.*** *Tadpoles at stage NF42, similar to those shown in D, underwent automated movement analysis in a Zantiks MWP unit. In all cases the analysis was performed blind (with genotyping subsequent to measurements) the black data points represent wild-type animals with purple showing crispant data.* ***F.*** *The control and crispant animals move similar distances over time but crispants show a greater initial startle response than controls. Interestingly all tadpoles spend the majority of each 300 second block in the most proximal arena, closest to the edge of the dish (F).* ***G.*** *The main change caused by heterozygous ddx17 LOF becomes clear when working memory was tested in the free movement pattern Y-maze; the crispants have lost the alternating search pattern shown by all vertebrates.*

First, a mosaic homozygous crispant knockout of *ddx17* was generated using two non-overlapping CRISPR/Cas9 gene-editing complexes designed to disrupt exon 7 and these were mated to produce F_1 heterozygous and compound heterozygous tadpoles (**Fig S1A**). The introduction of indels into the *ddx17* locus was tested by Sanger sequencing of the target region and demonstrated strong penetrance of indels in the founder animals (**Fig S1B**), producing a *ddx17* mosaic knockout (*Xtr.ddx17*$^{\text{em1EXRC}}$). The phenotypes produced by each sgRNA were indistinguishable, showing they were not due to off target effects. The heterozygous offspring produced by outcrossing mosaic founders were first identified using the T7E1 mismatch detection assay (**Fig S1C**) with the genotype, confirmed by Sanger sequencing (**Fig S1D)**, revealing frequently occurring deletions of 5bp, 22bp, 131bp and 136bp (**Fig S1E**), resulting in a frameshift leading to protein truncation within exon 7 (*Xtr.ddx17*$^{\text{em2EXRC}}$). Inbreeding of the mosaic founders produced homozygous or compound heterozygous tadpoles as shown by Sanger sequencing (**Fig S1F**) (*Xtr.ddx17*$^{\text{em3EXRC}}$). The homozygous tadpoles bearing larger deletions (131bp or 136bp) in one allele (**Fig S1G**), were found at later stages to have a series of further, progressive deletions as demonstrated by polyacrylamide gel electrophoresis (**Fig S1H**).

Phenotypically, more than half of the founder crispant embryos showed evidence of gastrulation defects (**Fig S3A**). The remaining animals showed no gross morphological defects (**Fig S3B,C**). The gross craniofacial morphology of *ddx17* crispants was tested by injecting CRISPR/Cas9 complexes into one-cell of a dividing two-cell embryo. This results in the effects of the protein truncation being concentrated on one side of the embryo along the left-right axis, with the other side acting as an internal control; no altered morphology was observed in the crispants (**Fig S3C**). A decrease in head size was observed prior to free-feeding stages (average 5.93 mm^2 for control tadpoles and 4.89 mm^2 for crispant tadpoles (NF41, n = 16), t(30)=3.85, $p<0.001$). This significant decrease in head size was no longer apparent at later stages of development (average 15.9 mm^2 for control tadpoles and 16.5 mm^2 for crispant tadpoles (NF48, n = 16), t(30)=-5.05, p=0.617, **Fig S3D**).

The phenotype of patients with *ddx17* variants include neurodevelopmental deficits and these can now be modelled in *X. tropicalis*. Here, in addition to the Free-Movement Pattern (FMP) Y-maze that is already validated in *X. tropicalis,*[261] another assay has been adapted to assess working memory in *Xenopus*. In this assay, animals are exposed to 5-minute alternating periods of light and dark; their initial response is to move (startle) as the lights go off, but they quickly learn that the dark period does not signal danger and reduce their response to it. This attenuated response to the changing environmental conditions demonstrates a second measure of short-term working memory in the tadpoles. Importantly this behaviour can be probed by pharmacological agents which disrupt working memory and attenuate anxiolytic responses. Following treatment with the NMDA receptor antagonist MK-801, tadpoles can be observed to startle repeatedly in response to the light-dark transition period. The startle response is also abolished following administration of 5 µM diazepam (**Fig S4A**). Once established, these assays were applied to test the neurodevelopmental phenotypes of the *ddx17* crispants.

Comparing the movement of control tadpoles and F_0 *ddx17* crispant animals at NF48 showed that crispant animals move less over the 10-minute trial period (average 4.0 mm/sec for control tadpoles and 2.7 mm/sec for crispant tadpoles (n = 48, t(82)=-2.6, p=0.012), **Fig S3F**). A significant reduction in locomotion was additionally observed in a second assay investigating the tadpoles' response to 5-minute light-dark transition periods (average 3.9 mm/sec for control tadpoles and 2.3 mm/sec for crispant tadpoles (n = 24, t(46)=3.4, p=0.001), **Fig S3G**). Unusually, these mosaic crispant tadpoles were not observed to startle at these light-dark transitions. In the FMP Y-maze assay that tests working memory, founder crispant tadpoles performed fewer alternations in their search patterns than controls (average alternations 22.8% for control tadpoles and 19.5% for crispant tadpoles (n = 20, F(1, 33) = 0.366, p = 0.550)) and more repetitions (average repetitions 9% for control tadpoles and 12.8% for crispant tadpoles (n = 20, F(1, 33) = 3.272, p = 0.080)) (**Fig S3H**).

Since there was clear evidence of movement and neural defects in the F_0 crispants, together with reduced axon length observed in electroporated mouse brains deficient in DDX17, neuron outgrowth was examined in *Xenopus* using an anti-HNK-1 antibody. Embryos injected with control (*tyr*)[308] and *ddx17* gene-editing complexes restricted to one side of the embryo reveal reduced axon outgrowth in *ddx17* crispants on the ipsilateral side (compare the injected and uninjected lateral views in **Fig S4E**). Similarly, in the heterozygous *ddx17* model axon outgrowth is visibly reduced compared with controls at NF stage 24 with evidence of continued axon outgrowth at later stages (NF26 onwards (**Fig 5.5C & Fig S4B**)).

In comparison to the founder animals, F_1 crispant tadpoles bearing non-mosaic heterozygous or compound heterozygous/homozygous indels in *ddx17* show no gross morphological or developmental abnormalities including in gastrulation (**Figs 5.5D, S4C & S5A**), hatching rate and head size (**Fig S5B**). Gross structural differences between control and crispants were not seen in the forebrain, midbrain or hindbrain regions of heterozygous *ddx17* crispant animals, even when bred in a [Xtr.Tg(tubb2b:GFP)Amaya] RRID: EXRC_3001 background (**Fig S4C**) to make differences more obvious. Post-hatching, all animals moved away from tactile stimuli (to the head) and at later stages adopted a normal, head down filter-feeding posture with the ability to navigate freely within their respective environments. Further, no abnormalities in locomotive behaviour consistent with descriptors of seizure activity in tadpoles were observed.[309,310] The locomotive activity of wild-type tadpoles, heterozygous or homozygous, non-mosaic crispants in *ddx17* was indistinguishable at NF48 when tracked across a 10-minute trial period (**Fig 5.5E & S5C**). Similarly, comparative average locomotive activity in the light-dark assay was not significantly different from controls in either non-mosaic crispant group, with all tadpoles noted to spend the majority of each of the 5-minute blocks at the edge of the dish.

Unlike wild-type tadpoles however, both non-mosaic crispant tadpole models were noted to respond with increased locomotion to multiple dark-light transition periods, although startling was more pronounced in the homozygous model (**Fig 5.5F & S5D**). Since this indicated reduced memory, a well characterised working memory assay was used, the 1 hour free-movement pattern Y-maze, to compare wild-type and homozygous or heterozygous *ddx17* crispant tadpoles. Although not significant, homozygous *ddx17* tadpoles performed fewer alternations (average alternations 22.6% for control tadpoles and 15.2% for homozygous *ddx17* tadpoles ($n = 49$, $F(1, 95) = 3.761$, $p = 0.055$)) but significantly more repetitions ($n = 49$, $F(1, 95) = 10.983$, $p = 0.001$) than controls (**Fig S5E**). This increase in repetitions taken together with the unattenuated startle response suggests increased levels of anxiety in the homozygous *ddx17* tadpole group.[311] In comparison, the heterozygous *ddx17* model demonstrated significantly fewer alternations (F(1,

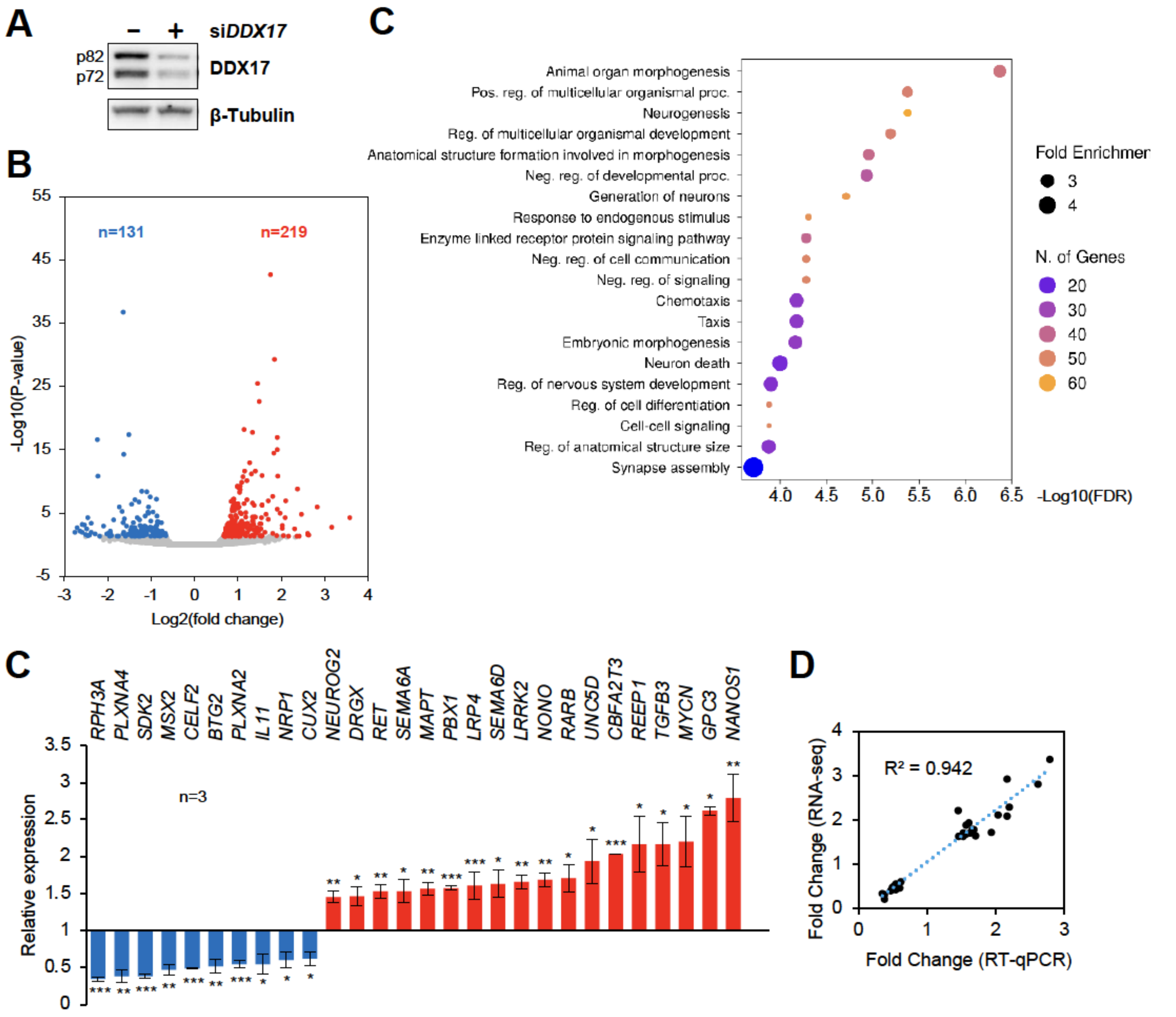
80) = 14.25, p < 0.001) without a significant increase in repetitions (F(1, 80) = 3.29, p = 0.074) when compared to wild-type tadpoles (**Fig 5.5G**). When considered alongside the response observed in the light-dark assay, this shows a clear deficit in the short-term working memory.

Overall, the crispant *ddx17* models show very similar phenotypes, with evidence of reduced axon outgrowth and working memory consistent between them. The mosaic founder animals however move less than the wild-type tadpoles unlike the other crispant models and have a transiently reduced head size.

5.3.4 RNA-seq analysis

Finally, to gain insight into the possible functions and target genes of DDX17 in a human cellular context, transcriptomic analysis of neuroblastoma SH-SY5Y cells was performed whereby the expression of the *DDX17* gene was knocked-down using a mixture of 2 different siRNAs (**Fig 5.6A**). 350 genes differentially expressed genes were identified in *DDX17*-KD cells compared to control cells (**Fig 5.6B** and **Supplementary Dataset SD5**). The functions of this set of genes were significantly associated to developmental processes, in particular the development and functions of the nervous system (**Fig 5.6C, S6,** and **Supplementary Dataset SD1**). For instance, the expression of several development-associated transcription factors (*MSX2*, *TBX3*, *GATA3*, *FOSL1*, *NEUROG2*, *SMAD6* and *SMAD9*, *SOX13*, *DRGX*, *RARB*, *MYCN…*) was deregulated upon *DDX17* KD. Of note, looking at the molecular functions associated to those genes also revealed the presence of a significant number of *trans*-membrane receptors (31 genes) and receptor ligands (15 genes) (**Supplementary Dataset SD5**), including several receptors/ligands associated with axon guidance (*DCC*, *EFNB2*, *PLXNA2* and *PLXNA4*, *SEMA6A* and *SEMA6D*, *RET*, *ROBO2*, *UNC5D…*).

FIGURE 5.6 | DDX17 CONTROLS THE EXPRESSION OF GENES INVOLVED IN NERVOUS SYSTEM DEVELOPMENT



*Results produced by C. Bourgeois. **A.** Western-blot showing the siRNA-mediated depletion of DDX17 protein in SH-SY5Y cells. B. Volcano plot showing the genes that are impacted by DDX17 KD in SH-SY5Y cells, as predicted from the RNA-seq analysis. Significantly altered genes (downregulated in blue and upregulated in red) were identified as described in the Methods section.*

***C.** Gene ontology analysis using ShinyGO for the genes impacted by DDX17 KD. Only the top 20 of the GO enriched biological processes are shown (see Supplementary File A for the full list of enriched terms).*

***D.** Validation of the effect of DDX17 knockdown on the steady-state expression of a selection of genes. RT-qPCR data were first normalised to GAPDH mRNA level in each condition, and the normalised mRNA level of each gene in the DDX17 knockdown condition was then normalised to the control condition, set to 1. Data are expressed as the mean value ± S.E.M. of independent experiments (n = 3). Unpaired Student's t-test (*P-val < 0.05; **P-val < 0.01; ***P-val < 0.001).*

***E.** Correlation between the measured fold change of expression (x-axis) and the corresponding predicted fold change value (y-axis) for the 28 genes shown in panel E.*

Subsets of genes of which steady-state expression was negatively (131 genes) or positively (219 genes) altered by *DDX17* KD were separately analysed (**Supplementary files SD1 & SD5**). This analysis showed again a significant occurrence of GO terms associated with neurogenesis in both groups of genes, but it also underlined a link between downregulated genes and body morphogenesis, while the group of upregulated genes was associated more specifically to cell signalling pathways (**Supplementary Fig S7**). To validate this computational analysis, 28 genes were selected from the two subgroups of mis-regulated genes and measured their mRNA level by RT-qPCR assays in mock-depleted and DDX17-depleted SH-SY5Y cells. Consistently, the expression of each tested gene was altered as predicted from the RNA-seq, with a strong combined correlation score (R^2=0.942) (**Figs 5.6D** and **5.6E**).

Collectively, these results indicate that the *DDX17* gene is involved in several processes during the development of vertebrates, in particular in the development of the nervous system, most likely because they control the expression of subsets of genes with key functions in the neural system.

5.4 Discussion

This is the first study to describe *de novo* heterozygous *DDX17* variants associated with features describing a novel neurodevelopmental disorder, Seaby-Ennis Syndrome. Mouse and frog animal models provide strong evidence that DDX17 plays important roles in the developing nervous system. More specifically, *Ddx17* knockdown impaired neuronal migration and axon development in the brain of newborn mice and reduced axon outgrowth and branching in primary cortical neurons *in vitro*. In agreement with these results, crispant tadpole *ddx17* models, including a heterozygous F_1 model, also presented a reduced axon outgrowth phenotype. Crispant tadpoles also had clear functional neuronal defects. Since the region and developmental state of the central nervous system in the mice and tadpoles in which this effect has been noted are distinct, it suggests that *DDX17* has an important role in widely distributed neurodevelopmental processes. The conservation of function across evolutionarily distant species such as mice and frogs strongly support that the role of *DDX17* is conserved in humans too. These *in vivo* results are therefore consistent with the hypothesis that heterozygous loss-of-function *DDX17* variants identified in patients induce a significant alteration of the function of the protein during neuronal development, resulting in the observed phenotype.

The Seaby-Ennis syndrome phenotype associated with *de novo* variants in *DDX17* is consistent with a neurodevelopmental disorder, typified by mild-moderate intellectual disability, delayed speech and language development and global developmental delay. Sixty-four percent (7/11) of the cohort have dysmorphic facial features, although for many this is subtle. Overlapping dysmorphology between patients includes: synophrys; upslanting palpebral fissures; depressed nasal bridge; posteriorly rotated ears; high arched eyebrows; epicanthus; telecanthus; and strabismus. Some patients have gross and fine motor delay, generalised hypotonia, sterotypy, and evidence of autism spectrum disorder.

There were no substantial differences in phenotype severity between patients harbouring missense variants versus loss-of-function variants, suggesting DDX17 haploinsufficiency causes the observed phenotype. Since all missense variants but one fall within the helicase domain (**Fig 5.2**), it suggests that they impact the structure and/or activity of this domain in a way that deeply alters the overall function of DDX17, similarly to loss-of-function variants. Preliminary molecular modelling analyses did not reveal any significant modification of the DDX17 structure in which patient missense mutations were introduced (data not shown). However, this analysis was based on the only known 3D structure of human DDX17, which is limited to the helicase core domain[312]. Recently it has been shown that the two disordered and flexible flanking domains also strongly affect the helicase activity of Dbp2, the yeast DDX17 ortholog.[313] It is thus currently impossible to accurately predict the impact of variants on DDX17 function, without taking into account the interactions between the structured and unstructured regions of the protein.

We, and others[303,304] have shown previously a role of DDX17 in the retinoic acid induced differentiation of neuroblastoma SH-SY5Y and pluripotent embryonal NTERA2 cells, respectively. However, this effect was most evident when DDX17 knockdown was combined with the concomitant depletion of its paralog DDX5. This work now demonstrates that the downregulation of DDX17 alone is sufficient to alter neuronal development, both *in vivo* and *in vitro*. Both DDX17 and DDX5 have largely redundant functions, which probably explains why their joint depletion has such a strong effect compared to single protein depletion. Interestingly, two distinct shRNAs targeting DDX17 alter neuronal migration in the mouse cortex. Although shRNA-based strategies are prone to off-target disruption of neuronal migration in the mouse cortex[314], the migration phenotype is compatible with a previous observation that DDX17 controls the activity of the Repressor Element 1-silencing transcription factor (REST) complex during neurogenesis[303] and that the REST/CoREST complex regulates neuronal migration.[315] Future studies using genetic knockout models will demonstrate the specificity of the migration phenotype. Furthermore, this work reports that DDX17 plays a role in axon morphogenesis that is independent of its function in neuronal migration.

Xenopus frogs have been used as pioneer model organisms since the mid-twentieth century, mainly in discovery research[316]. Gene editing was found to be exceptionally effective in them and their application as tools for studying disease has increased.[317] *X. tropicalis* are diploid tetrapods with very few gene duplications. Their genome structure has high levels of synteny with humans[318] and the initial determination that 80% of human disease genes have orthologues in this species is now thought to be an underestimate.[319] Research has shown them to be highly suited to testing the links between a variant of uncertain significance and human disease phenotypes.[261,320] This can often be achieved without breeding the animals due to the efficiency of CRISPR/Cas resulting in very low levels of mosaicism in founders. Hence, they represent a rapid and cost-effective assay for gene-disease associations, filling an important gap between the mouse and zebrafish models. Here, mosaic founders, and both heterozygous and compound heterozygous/homozygous F_1 non-mosaic models, have been used to test the effect of a truncation in ddx17. Herein, the use of *Xenopus* in disease modelling has been expanded through application of a behavioural assay to tadpoles (the light dark-transition assay) that has been used previously in other models. The measurements of both working memory and anxiety with small molecule inhibitors was then confirmed. Additionally, mosaic homozygous F_0 crispants were directly compared with non-mosaic F_1 homozygous and F_1 heterozygous animals. There were stronger phenotypic effects in mosaic founder animals. This is a known phenomenon, which may be associated with a failure to activate compensatory mechanisms in mosaic animals, including crispants in another aquatic model, zebrafish (reviewed by Rouf *et al.*[321]). This suggests that *Xenopus* behave like zebrafish in this respect.

These data offer some limited insights into the mechanism whereby DDX17 variants affecting its function relate to the disease phenotype. Since DDX17 is known to regulate gene expression at multiple levels, the different pathological features associated with *DDX17* mutations likely result from the altered expression of some of its target genes and transcripts. Indeed, the transcriptomic analysis showed that 350 genes may be impacted, a large proportion of which are important for development and morphogenesis, and most particularly for neurogenesis. This includes several key transcription factors (*NEUROG2*, *RARB*, *MYCN…*), the deregulation of which could have direct and indirect effects on many other genes during embryonic development. Furthermore, the DDX17-dependent regulation of several genes coding for trans-membrane receptors and ligands associated with axon guidance is also of particular significance, considering the altered axonal development observed upon DDX17 knockdown in mice and tadpoles, and the neurological phenotype observed in patients. Whilst further work is needed, the goal of this study is to establish DDX17 as a novel neurodevelopmental disease gene and enable identification of more patients to further elucidate the genotype-phenotype relationship.

5.5 Conclusion

Eleven patients with neurodevelopmental phenotypes harbouring monoallelic *de novo* variants in *DDX17* were identified, describing a new disorder, Seaby-Ennis Syndrome. Functional experiments (*in vitro* and *in vivo*) show that *DDX17* is important in neurodevelopmental processes, in keeping with the observed human phenotype. *Ddx17* knockdown of newborn mice showed impaired axon outgrowth, and reduced axon outgrowth and branching was observed in primary cortical neurons *in vitro*. The axon outgrowth phenotype was replicated in crispant *ddx17* tadpoles, including in a heterozygous (F_1) model. Crispant tadpoles had clear functional neural defects and showed an impaired neurobehavioral phenotype. Transcriptomic analysis further supports the role of *DDX17* in neurodevelopmental processes, particularly neurogenesis. These results strongly support that monoallelic variants in *DDX17* cause a neurodevelopmental phenotype.

Chapter 6 | *De novo* variants in *HDLBP* are associated with a neurodevelopmental phenotype

6.0 Contribution statement

This chapter contains unpublished results and builds upon the work of **Chapter 4**, whereby *HDLBP* was identified as a candidate novel disease gene. I have curated an international case series of 7 patients, obtaining consent and detailed phenotype data, and have coordinated an effort bringing together local and international collaborators to functionally validate the gene as disease-causing. *Xenopus* modelling was performed by A. Godwin, T. Fletcher, and M. Guille at the University of Portsmouth. RNA-binding experiments were undertaken by I. Minia and M. Landthaler in Germany. All other work is my own with support from my supervisory team.

6.1 Introduction

HDLBP is a protein first described in 1987 by Graham *et al.*[305] Its name originates from the observation that it binds to high density lipoprotein. It is ubiquitously expressed, with high expression in secretory tissue such as the pancreas, thyroid, and gastrointestinal tract. HDLBP comprises 15 hnRNP homology (KH) domains, more than any other human KH-domain protein. The KH domain is highly conserved across diverse organisms such as Bacteria, Archaea and Eukaryotes, suggesting it arose early in evolution.[306] KH domains predominantly interact with mRNAs and comprise a core motif folded into a βααβ unit.[306]

HDLBP has been shown to contribute to biological processes including protein aggregation and translation. It localises to both the cytosol and endoplasmic reticulum (ER).[307]The precise function of HDLBP is not well established however recent work by Zinnall *et al.* shows that HDLBP directly interacts with more than 80% of ER localised mRNAs,[307] and that absence of HDLBP results in decreased translational efficiency of HDLBP-target mRNAs, and impaired protein synthesis and secretion in model cell lines. Other functions of HDLBP include RNA transport, chromosome segregation, genomic stability, DNA damage repair, and heterochromatin-mediated gene silencing.[308,309]

HDLBP is a gene highly constrained for loss-of-function (LoF). Fewer LoF variants are observed in healthy human populations than would be expected under a null mutational hypothesis.[286] This suggests that *HDLBP* is intolerant to mutation. Genes constrained for LoF are highly associated with haploinsufficient disease genes[287], suggesting that *HDLBP* may present a novel disease gene. In 2009, Felder *et al*.[310] described a patient with autism and brachymetaphalangy meeting criteria for 2q37 deletion syndrome (Albright Hereditary Osteodystrophy-like syndrome or Brachydactyly-Mental Retardation syndrome, MIM: 600430). Genes within the 2q37.3 region include *FARP2, GPC1, PASK, KIF1A* and *HDLBP*, all of which are involved in neuronal differentiation and skeletal processes. Felder *et al*. used lymphoblastoid cell lines derived from their patient and his family and performed RNA expression analyses. Their results showed that *FARP2, PASK* and *HDLBP* were significantly downregulated in the patient's cell line, supporting that *HDLBP* contributed to the patient's neurodevelopmental phenotype.[310] Additionally, research by Banday *et al.* has shown a role for *HDLBP* in DNA double-strand break repair, whereby it is hypothesised that haploinsufficiency of HDLBP may contribute to autism spectrum disorder (ASD).[309] This hypothesis is further supported by Satterstrom *et al*.[311], who performed large-scale exome sequencing on 11,986 patients with ASD and 23,598 controls. They applied an enhanced analytical framework, integrating *de novo* and case-control rare variation to identify risk genes for ASD. *HDLBP* was one of the genes implicated in both the functional development and neurobiology of ASD.

Despite the implication that *HDLBP* may be involved in autism spectrum disorder, the gene is not currently a recognised disease gene in the clinical genetic community. Targeted studies are required to unambiguously determine any pathogenic consequences of variants in this loss-of-function constrained gene.

6.2 Methods

6.2.1 Accessing patient data

Through membership of a Genomics England Clinical Interpretation Partnership, access was obtained to deidentified whole genome sequencing and phenotype data stored in the Genomics England Research

Environment. The DeNovoLOEUF method was applied to identify patients with variants in genes (without a known disease association) constrained for LoF (see **Chapter 4** results). Two patients were identified with predicted LoF variants in *HDLBP*. Additional study subjects were identified through GeneMatcher.[307] In total, 7 patients consented to participate in the study. Parents and legal guardians of all affected individuals provided written consent for the publication of their results alongside genetic and clinical information.

6.2.2 *Xenopus* methods

Adult Nigerian strain *Xenopus tropicalis* were housed within the European *Xenopus* Resource Centre (EXRC; https://xenopusresource.org), University of Portsmouth, in recirculating MBK Ltd systems maintained at 24ºC - 27ºC (13-11-hour light-dark cycle) with 10% daily water changes. All *Xenopus* work was completed in accordance with the Home Office Code of Practice under PP4353452 following ethical approval from the University of Portsmouth's Animal Welfare and Ethical Review Body. Detailed *Xenopus* methods are available in **Methods 2.8.1.** In summary, CRISPR/Cas9 sgRNAs were used to target exon 6 of *hdlbp* to generate mosaic homozygous *hdlbp* crispants. Two experiments were performed. In experiment 1, a range of concentrations of each sgRNA was injected at the one-cell stage; indel penetrance and phenotype were assessed in the resulting embryos. In experiment 2, 300pg of gene-editing components was injected into one-of-two-cell embryos and the resultant phenotype was assessed. F_0 crispants were assessed for morphological abnormalities and they had their survival and locomotion assessed.

6.2.3 RNA-binding methods

RNA-binding methods are summarised in **Methods 2.11**. Briefly, human wild type HDLBP protein (KH domains 5 to 9), and two mutant proteins (KH domain 5 to 9) containing patient-specific missense variants I471V (patient 4) and R677H (patient 7), were expressed in *E. coli*. The protein fragments were then purified. The wild type and mutant HDLBP fragments were incubated with fluorescein-labelled TFRC1 and TFRC2 RNA probes. A fluorescence polarisation assay was performed to measure the binding capacity of the HDLBP proteins to the two RNA probes. Binding curves were fitted using a non-linear logistic regression model and the binding capacity was measured.

6.3 Results

Seven patients with monoallelic variants in *HDLBP* were identified. This step involved contacting clinicians from across the globe using the GeneMatcher platform and providing collaborators with a comprehensive phenotype template to capture salient phenotyping data, which was later harmonised and summarised. Research consent was then obtained for each of the seven participants to facilitate publication of phenotype and genotype data.

Six of the patient variants were confirmed *de novo*, and one was observed in a patient from a child-mother duo whereby the variant was absent in the maternal sample. All seven patients have a neurodevelopmental phenotype (see **Table 6.1** and **Supplementary Dataset SD6**). All variants are absent from gnomAD[26] v2.1.1 and v.3.1.2.

Table 6.1 | Patient variants and phenotype

	Loss-of-function			Missense			
Patient	1	2	3	4*	5	6	7*
Age (at last visit)	10y 8m	2y 5m	5y	Unknown	13y	7y	3y 6m
Sex	M	F	M	F	F	F	M
Variant	p.Gly748ArgfsTer20	p.Lys672ArgfsTer4	c.1373-1G>A	p.Ile471Val	p.Arg839His	p.Gln392Glu	p.Arg677His
Inheritance	*De novo*	*De novo*	*De novo*	Absent from maternal sample and missing paternal sample	*De novo*	*De novo*	*De novo*
Conductive hearing impairment	N	N	Y	N	Y	N	N
Dysmorphic facial features	N	Y	Y	Y	N	N	N
Age walking independently	12m	18m	19m	13m	Unable to walk without orthoses	13m	18m
Intellectual disability	Moderate to severe	Mild	N	Borderline	N	N	Borderline
ADHD	N	Suspicion	N	Y	N	Y	N
Autism spectrum disorder	Y	N	N	Y	N	Not yet assessed	Y
Delayed speech and language development	Y	Y	Y	Y	N	Y	Y
Global developmental delay	Y	Mild	Mild	Mild	N	Y	Borderline
Gross motor delay	Y	Mild	Mild	Y	Y	Y	N
Fine motor delay	Y	Mild	Mild	Borderline	N	Y	N
Muscular hypotonia	N	N	Y	N	Y	N	N

*F – Female, M – male, m – months, N – No, Y – Yes, y – years. *Variants modelled in RNA-binding studies. Grey cell – unknown data.*

The cohort comprises 3 males and 4 females. 4/7 (57%) have intellectual disability ranging from borderline to moderate-severe. Forty-three percent (3/7) patients have ADHD and 3/7 (47%) have confirmed ASD. Eighty-six percent (6/7) patients have delayed speech and language development, and the same 6/7 (86%) patients have global developmental delay. Seventy-one percent (5/7) have fine motor delay, 6/7 (71%) have gross motor delay and 2/7 (29%) have muscular hypotonia. Twenty-nine percent (2/7) have conductive hearing impairment. Forty-three percent (3/7) have non-overlapping dysmorphic facial features. These include malar hypoplasia, long philtrum, high arched palate, short ears with over-folding of the helical rim, short palpebral fissures, dental hypoplasia, and a laryngeal cleft.

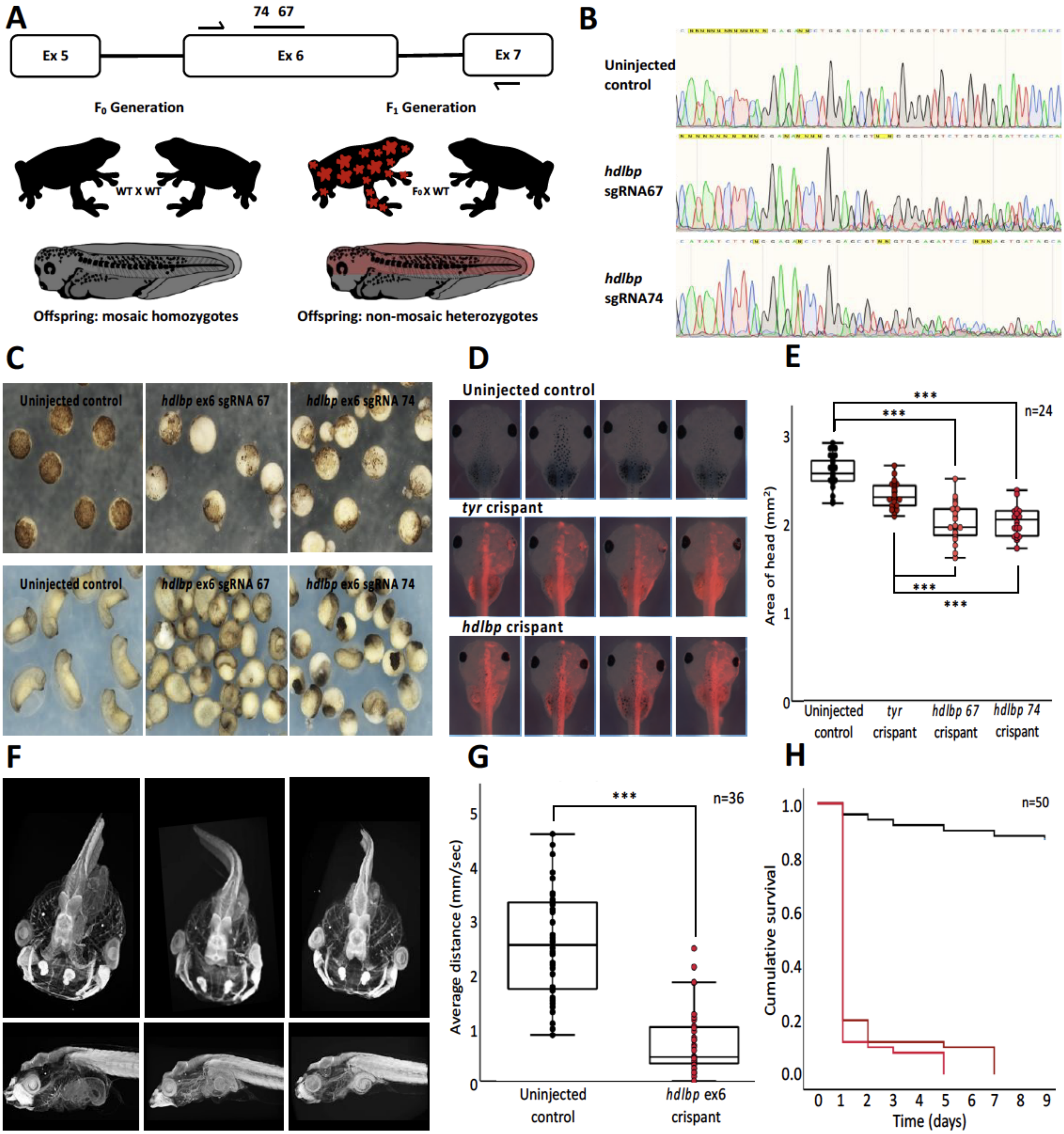
6.3.1 Molecular genetic findings

All variants in *HDLBP* are variants of uncertain significance (VUSs). Six variants are confirmed *de novo*, and one variant is from a mother-daughter duo, whereby the variant is absent from the mother, but the father's DNA is unavailable. Three variants are predicted LoF and 4 variants are missense. Patient 3 has a paternally inherited likely pathogenic variant in FOXP1:NM_001244815.1:c.47_57del:p.Thr16IlefsTer62. Patient 4 has an essential splice site variant in FBN2: c.3847+1G>T (absent from mother but status in father unknown), and a maternally inherited variant in DSG2:c.1912G>A; p.Gly638Arg. Patient 7 has a *de novo* VUS deletion 1:145382174-145831406, containing no known GenCC[322] disease genes.

6.3.2 *Xenopus* results

Non-overlapping sgRNAs targeting exon 6 of *HDLBP* demonstrate clear indels, confirming successful experiments (see **Figure 6.1B**). Mosaic F_0 animals show significant gastrulation defects; the white patches in the early embryo are consistent with extensive cell death that has been reported as apoptotic. The small 'caps' of pigmented cells are consistent with failures in epiboly, although maternal expression (from Xenbase[323]; https://xenbase.org) is very low, so this is not unexpected. Only 20% of F_0 tadpoles survived past day 1 (**Figure 6.1H**). F_0 survivors have small heads as can be seen from the 1-of-2 cell injections, and this observation is distinct from both the injected (tyr) controls and uninjected (wild type) controls (**Figure 6.1D & E**). MicroCT images of highly mosaic surviving F_0s show that the main anatomical structures are present, but that the anterior end of the head appears compressed in the dorsal-ventral plane. F_0 animals have significantly reduced movement compared to wild type (**Figure 6.1G**).

FIGURE 6.1 | SURVIVING F_0 HDLBP CRISPANTS SHOW MORPHOLOGICAL AND MOTOR ABNORMALITIES

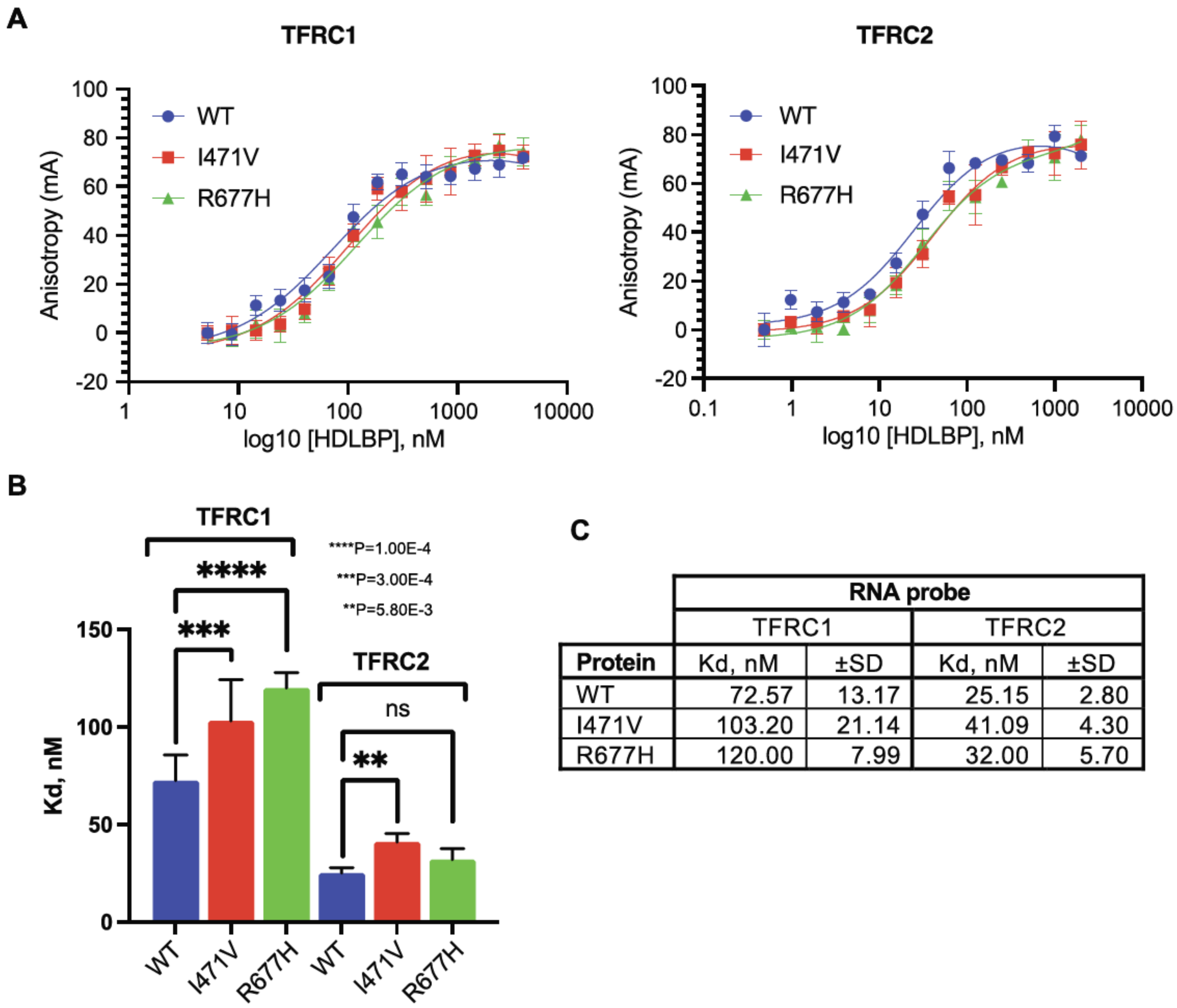


A. *Two guide sgRNAs target exon 6 of HDLBP to produce mosaic F_0 homozygotes (experiments complete) and an F_1 generation of non-mosaic heterozygotes (experiments ongoing).* ***B.*** *Both guide RNAs (sgRNA67 and sgRNA74) show successful deletions as evidenced on the sequencing traces.* ***C.*** *Hdlbp crispants (67 and 74) show severe gastrulation defects.* ***D.*** *Injected hdlbp crispants (at the one-of-two-cell stage) have small heads and are distinct from both tyr crispants (injected controls) and the uninjected controls. Dextran staining (red) shows which side of the animal was injected.* ***E.*** *The head size of hdlbp crispants is reduced compared to injected and uninjected controls.* ***F.*** *MicroCT images of very mosaic surviving F_0s (at stage 45; day 4) show an unusual downward projection of the head (inferior images).* ***G.*** *Hdlbp crispants are significantly less motile than uninjected controls.* ***H.*** *80% of hdlbp 67 crispants (red) and 74 crispants (burgundy) die by day 1. The remainder die by days 5-7.*

6.3.3 RNA-binding results

To assess the effect of patient variants on the RNA binding properties of HDLBP, wild type fragment HDLBP (KH domains 5-9) and two mutant proteins (I471V [patient 4] and R667H [patient 7]) were expressed and purified. A fluorescent anisotropy assay was used to determine the apparent dissociation constant (K_d) of the respective proteins with TFRC1 and TFRC2 RNA probes, which were selected based on previously published PAR-CLIP data.[324] The *in vitro* binding assay showed that wild type HDLBP protein fragment interacted with both RNA probes with a binding affinity around 73 nM for TFRC1 and 25 nM for TFRC2. The two mutant proteins had noticeably reduced binding to both RNA probes (**Figure 6.2**). Both I471V and R677H mutants showed a statistically significant reduction in binding to TFRC1 ($p = 3.0 \times 10^{-4}$ and $p = 1.0 \times 10^{-4}$ respectively). Only the I471V mutant showed a statistically significant reduction in binding to TFRC2 ($p = 5.8 \times 10^{-3}$).

FIGURE 6.2 | RNA BINDING AFFINITY OF HDLBP (KH5-9) AND ITS MUTANTS TO TFRC1 AND TFRC2



	RNA probe			
	TFRC1		TFRC2	
Protein	Kd, nM	±SD	Kd, nM	±SD
WT	72.57	13.17	25.15	2.80
I471V	103.20	21.14	41.09	4.30
R677H	120.00	7.99	32.00	5.70

*Apparent dissociation constants of recombinant GST-tagged HDLBP protein fragments (KH5 KH9 - domains corresponds to protein fragment B in Zinnall et al.) were determined by fluorescence anisotropy binding assays. FAM labelled RNA probes TFRC1 and TFRC2 were incubated with GST HDLBP fragments (either wild type or containing single amino acid substitution I471V [patient 4] and R677H [patient 7]). Anisotropy was measured (**A**) and Kd determined from the binding curves (**B**). The summary for three independent Kd values is shown in the table (**C**).*

6.4 Discussion

To the best of knowledge, this is the first description of *de novo* monoallelic variants in *HDLBP* associated with neurodevelopmental disorder in humans. The most commonly shared features include global developmental delay, gross motor delay, and speech and language delay. Three patients have evidence of ASD, however another three patients are likely to be too young for formal assessment. There was no clear difference in phenotype between patients harbouring predicted loss-of-function variants compared with missense variants.

Three patients had additional variants in other genes. Patient 3 has a likely pathogenic variant in *FOXP1*; however the variant is paternally inherited, and the father is unaffected, therefore its role in the patient's phenotype is unclear. Patient 4 has a maternally inherited variant in *DSG2*, which is associated with arrhythmogenic right ventricular cardiomyopathy. A screening cardiac MRI showed no evidence of the cardiac phenotype expected from a variant in *DSG2*. Patient 7 has a *de novo* deletion in chromosome 1:145382174-145831406, which contains no known disease-causing genes and therefore its significance remains uncertain but is not predicted to be pathogenic.

Data from *Xenopus* studies show that HDLBP is crucial in early development. This was evidenced by a high burden of homozygous indels in *Xenopus tropicalis* causing embryonic lethality. Indeed, the majority of F_0 tadpoles did not survive past day 1 with many displaying extensive gastrulation defects (**Figure 6.1**). Data from the International Mouse Phenotyping Consortium (mousephenotype.org) show that complete *hdlbp* mouse knockouts are non-viable with no pups surviving. This is perhaps unsurprising given how constrained the *HDLBP* gene is for loss-of-function in gnomAD (LOEUF = 0.15 and pLI = 1).

To assess morphology in crispant *Xenopus*, sgRNAs were injected into one-of-two cell embryos. This allows one half of the embryo to develop normally i.e. to be wild-type and the other half to be mutant; this is because the first cell division in *Xenopus* is left-right.[325] The surviving *hdlbp* crispants had significantly smaller heads than wild type, with some subtle asymmetry affecting the injected side of the animal. An injected *tyr* control behaved similarly to the (uninjected) wild type control (**Figure 6.1D**). The small head size may be secondary to craniofacial defects and gross-motor deficits. But to better elucidate the cause, quantitative measurements of the head pre- and post- feeding are needed. MicroCT images of surviving F_0 tadpoles were generally unremarkable. However there may be subtle differences seen in the organisation of the branchial arches on the injected side, as well as what appears to be the architecture of the hind brain. Crispant tadpoles displayed an unusual downward projection of the head, of which the cause is unclear. This may represent a reduced forebrain, but many more tadpoles would need imaging to be certain. Crispant tadpoles moved significantly less than wild type when their distance was recorded in a single well of a 6-well plate over a 5-minute period. Initially at stage 41, the tadpoles ate normally but by stage 43, the health and motility of the tadpoles rapidly declined. It is unclear whether this resulted from an inability to appropriately filter feed as a sequela of a systemic movement problem, or whether another cause affected the feeding ability, which subsequently impaired locomotion. Due to the motor deficits observed in the tadpoles and their poor overall survival, further behavioural assays were not performed.

Xenopus experiments clearly support the hypothesis that *HDLBP* is an important gene in development. Homozygous *hdlbp* crispants display a severe phenotype, particularly affecting overall survival, locomotion, and head size. These results recapitulate a homozygous model, whereby the phenotype may be vastly different from that presenting in an heterozygous state, as is observed in the patients presented in this case series. It is likely that a homozygous knockout of *HDLBP* in humans is embryonic lethal. That said, there is some homogeny in phenotypes observed between the homozygous *Xenopus* crispants and the humans with heterozygous missense and pLoF variants. Six of seven patients have motor delay. No patients have signs of microcephaly, however the small heads seen in *Xenopus* crispants may reflect abnormal brain development, for which most patients in this case series have some degree of intellectual disability or neurodevelopmental deficit. However, to better model the human phenotype, an F_1 generation of heterozygous *Xenopus* crispants is needed and these experiments are in progress.

HDLBP is an essential RNA-binding protein. RNA-binding experiments of two mutant HDLBP proteins to two RNA probes show statistically significant reduced RNA-binding when compared with wild type. It was not possible to perform the RNA-binding assays directly on patient samples due to lack of access to tissue. However, the evidence presented further supports the observation that variants in *HDLBP* impair protein function and may lead to functional deficits. Further work is needed to elucidate how these observations mechanistically correlate with observed phenotypes in humans.

6.5 Conclusion

De novo monoallelic variants in *HDLBP* are associated with a neurodevelopmental phenotype in humans. Functional experiments in *Xenopus* show that *HDLBP* is a crucial gene in the developing tadpole and that homozygous knockout is mostly lethal. Surviving (very mosaic) homozygotes have small heads, move significantly less than wild type and show morphological abnormalities which may implicate impaired brain development. RNA-binding studies assessing the binding capacity of mutant HDLBP (using patient 4 and patient 7's variants) to two RNA probes, demonstrate statistically significant impaired binding capacity compared to wild type; this supports the pathogenicity of variants identified in the cohort. Functional studies are ongoing and an F_1 heterozygous *Xenopus hdlbp* crispant population has been generated, whereby further experiments are underway. It is hoped that results from these experiments will further corroborate the hypothesis that *HDLBP* is a haploinsufficient novel disease gene.

Chapter 7 | Improving diagnostic rates in the 100,000 Genomes Project

7.0 Contribution statement

The concept and design of this study was developed by myself with input from my supervisory team. I wrote and performed all bioinformatics analysis in the secure Genomics England Research Environment. The resultant software, DeNovoLOEUF, is all my own work and is available in GitHub (https://github.com/lecb/DeNovoLOEUF). My bioinformatics analysis identified 62 variants, believed to be missed diagnoses. These were curated by Dr Simon Thomas in an NHS diagnostic laboratory to ACMG guidelines. This work is published in Human Genetics: *Seaby, E. G., Thomas, N. S., Webb, A., Brittain, H., Taylor Tavares, A. L., Baralle, D., ... & Ennis, S. (2023). Targeting de novo loss-of-function variants in constrained disease genes improves diagnostic rates in the 100,000 Genomes Project. Human Genetics, 142(3), 351-362* (**Appendix Paper 12**).

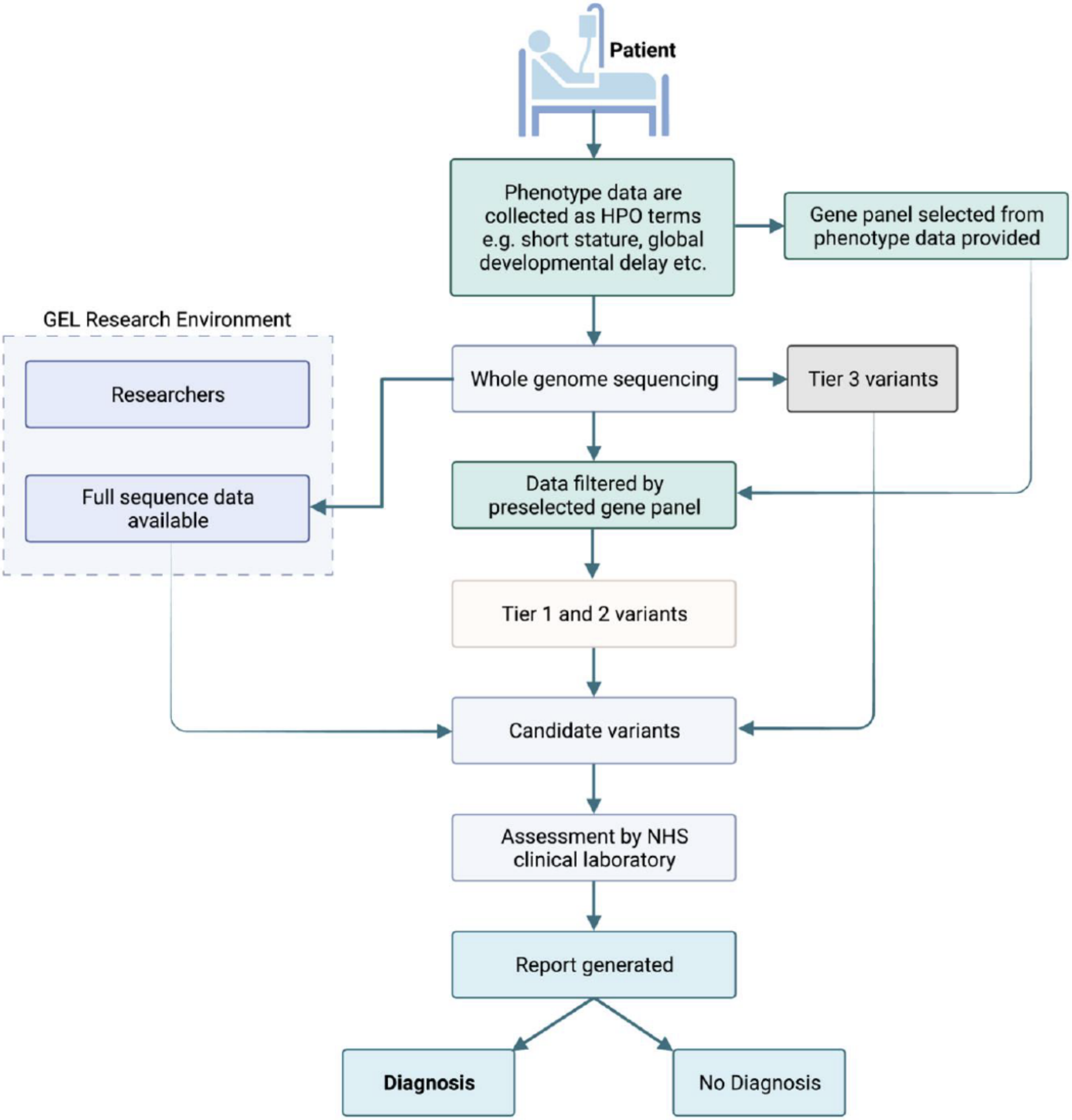
7.1 Introduction

With transformative advances in genomic medicine, there has been an exponential rise in the number of individuals undergoing exome and genome sequencing. A shift towards large-scale international sequencing programs is improving affordability and accessibility of such sequencing for diagnostic purposes, where conventional clinical tests have failed to yield a diagnosis.[2,16] The 100,000 Genomes Project (100KGP), was a research project embedded within the UK National Health Service and the precursor to offering whole genome sequencing (WGS) as a clinical test.[3,132,222] This pioneering project benefited from sequencing vast patient numbers with rare genetic diseases with improved power to identify multiple patients with overlapping phenotypes and genotypes, however the number of cases that required clinical assessment for diagnostic reporting versus resources available created a significant bottleneck.

Diagnostic rates for the 100KGP were similar to the international average for rare diseases.[326] The flagship 100KGP paper showed that an estimated diagnostic uplift from 15% to 20% could be achieved beyond prior testing but that the time and level of additional resources required to analyse the full genome was beyond routine diagnostic testing.[132] As a result, the project adopted the use of predefined gene panels (Genomics England PanelApp)[118] to target sequencing analysis to the most relevant genes selected from the Human Phenotype Ontology (HPO)[167] terms provided by referring clinician (**Figure 7.1**).[132] Whilst this approach restricted the number of variants assessed and improved the efficiency of the variant curation process applied to each patient's genome for diagnostic reporting, it risked missing variants in genes outside of the gene panel applied. At the latter stages of the project, many NHS England clinical laboratories additionally reviewed all *de novo* variants and top Exomiser[134] results, although this was never mandatory. This learning has also informed guidance for the evaluation of genome sequencing, which is now available as a clinical test in the NHS through the Genomic Medicine Service.

Accurate phenotyping is essential for gene panel selection, yet there is huge variability in the phenotypes reported by clinicians. For some cases in the 100KGP, only a single HPO term was reported. As WGS becomes more widespread, appropriate matching of HPO terms with optimal panel(s) may be less error prone for experienced geneticists but will represent a challenge for the wider community of clinicians expected to routinely refer patients. Furthermore, whilst the use of HPO terms aids in the standardisation of reporting phenotype data, it represents a cross-sectional timepoint analysis without resource for reanalysis. Ultimately, using HPO terms and relying on predefined gene panel options lacks the full clinical narrative and challenges gene panel selection. Therefore, there is need to expand genome analysis beyond gene panels to enable a more agnostic and comprehensive genome analysis, yet this needs to be balanced with the number of variants that require manual assessment for diagnostic reporting and the risk of incidental findings. Targeting variants with high pathogenic potential across the entire exome provides an opportunity to rapidly identify diagnostic variants and uplift diagnostic rates. Genotype-driven analysis approaches are complementary to the phenotype-drive approach currently utilised by 100KGP.

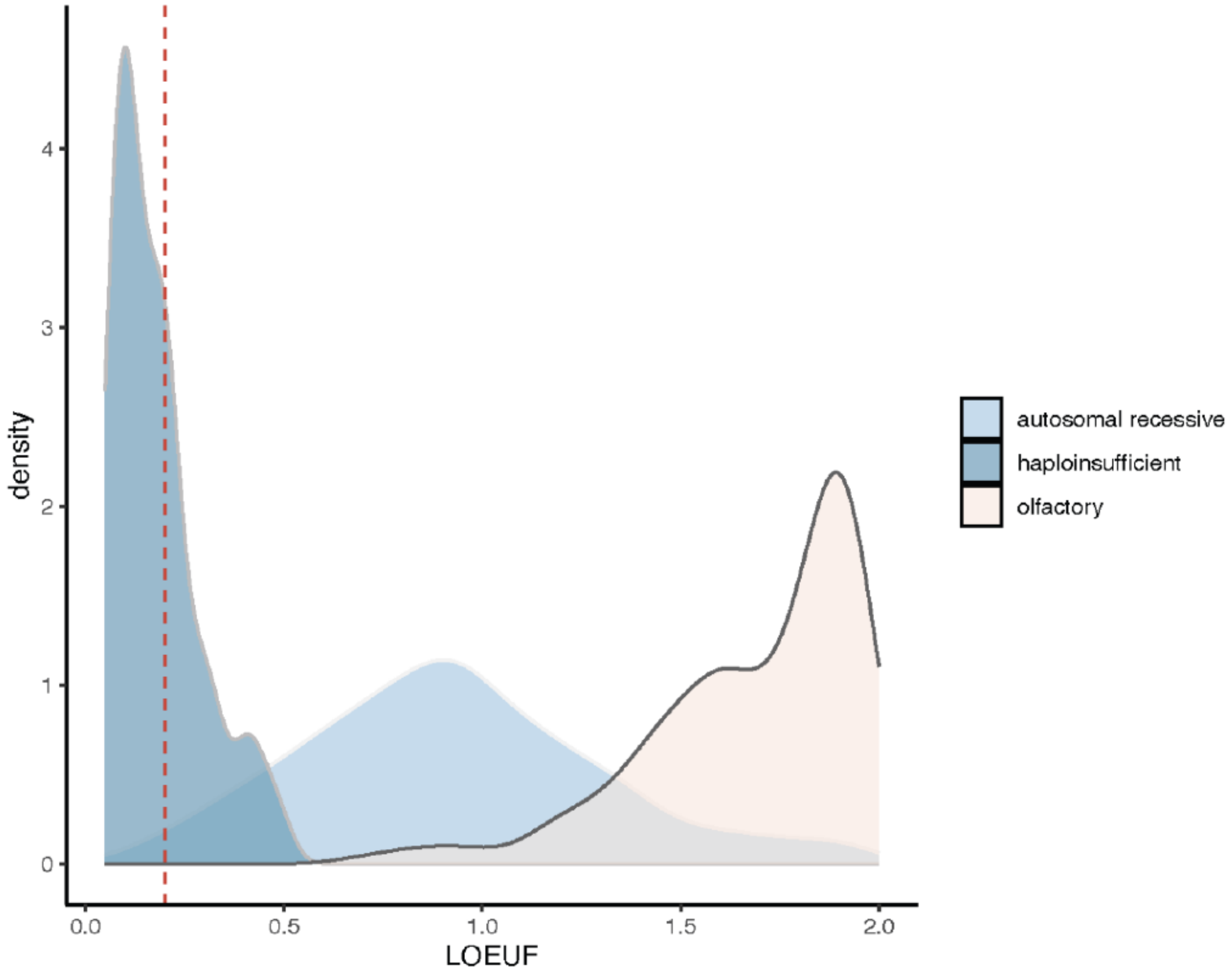
FIGURE 7.1 | GENOMICS ENGLAND 100,000 GENOMES PROJECT WORKFLOW


Patient
Phenotype data are collected as HPO terms e.g. short stature, global developmental delay etc.
Gene panel selected from phenotype data provided
GEL Research Environment
Researchers
Full sequence data available
Whole genome sequencing
Tier 3 variants
Data filtered by preselected gene panel
Tier 1 and 2 variants
Candidate variants
Assessment by NHS clinical laboratory
Report generated
Diagnosis
No Diagnosis

Phenotype data were collected from patients and recorded as HPO terms. These terms informed the virtual gene panel(s) applied for data analysis. This contrasts with the new UK Genome Medicine Service, where the clinician selects the panel(s) applied. In 100KGP, the patient underwent whole genome sequencing (WGS), and their sequencing data were filtered using the pre-selected gene panel(s). Data were also filtered by allele frequency and variant segregation (*) and these variants were classified into Tier 1 and Tier 2 variants as previously described.[132] *Candidate variants within the gene panel were identified and assessed by an NHS accredited diagnostic laboratory and a report was generated and returned to the patient. Diagnostic laboratories were under no obligation to review variants outside of the virtual panel(s) applied, however rare variants outside of the predefined panel(s) were available for analysis as Tier 3 variants. Variants outside of the gene panel(s), including full raw sequencing data, remain accessible to approved researchers for interrogation. Potential candidate variants identified through this route can be reported via Genomics England for potential return for local review by clinical laboratories.*

Across humans, some genes are extremely depleted or constrained for variation predicted to result in loss-of-function (LoF).[28] That is to say there is negative selection against the loss or inactivation of one allele. By comparing the observed over the expected rate of predicted loss-of-function variants in large population databases, it is now possible to compute the degree of constraint a given gene has for inactivation.[133,327] The loss-of-function observed over expected upper bound fraction, or LOEUF score, is a metric that places each gene on a continuous scale of loss-of-function constraint. Low scores are highly correlated with disease genes and gene essentiality, with the first LOEUF decile (<0.2) being enriched for haploinsufficient disease genes (**Figure 7.2**) and the greatest burden of OMIM disease entries.[133] Loss of function variants in extremely LoF constrained genes are therefore prime targets for potential diagnoses.

FIGURE 7.2 | LOEUF SCORE COMPARED WITH HAPLOINSUFFICIENT, AUTOSOMAL RECESSIVE AND OLFACTORY GENE LISTS



Density histogram of the LOEUF score compared with haploinsufficient, autosomal recessive, and olfactory gene lists, publicly available from: https://github.com/broadinstitute/gnomad_lof. Haploinsufficient genes are enriched for low LOEUF scores, with a natural cut off at 0.2 (dotted line). Autosomal recessive genes sit in the middle of the distribution and olfactory genes show tolerance to loss-of-function with high LOEUF scores.

This project aimed to utilise the sequencing and phenotype data generated through 100KGP and apply a transferable and rapid filtering method, called DeNovoLOEUF, that screens for putative pathogenic variants. This gene agnostic method targets variants with the highest diagnostic yield in rare disease patients and enables clinical curators to focus on the most important findings, regardless of the gene panel applied; therefore improving efficiency for cases where a diagnosis could be rapidly identified.

7.2 Materials and Methods

7.2.1 Data access

Permission to access to the secure GEL research environment (RE) and high-performance cluster (HPC) was obtained following information governance training and with membership of a Genomics England Clinical Interpretation Partnership (GeCIP): *Quantitative methods, machine learning, and functional genomics*. The following project was approved (RR359 - *Translational genomics: Optimising novel gene discovery for 100,000 rare disease patients*) which permitted access to 100KGP sequencing and phenotype data (Release V8). This included an aggregate *.gvcf* file comprising 13,949 rare disease trios with *de novo* variants, called using the Illumina Platypus pipeline.[132]

Access to the 100KGP dataset is restricted and only available as a registered GeCIP member in the Genomics England Research Environment. All data shared in this chapter were approved for export by Genomics England. The datasets and code supporting the current study, unavailable for export, are fully accessible within the Genomics England Research Environment in the shared directory: re_gecip/machine_learning/Ellie_Seaby/.

7.2.2 Phenotype data

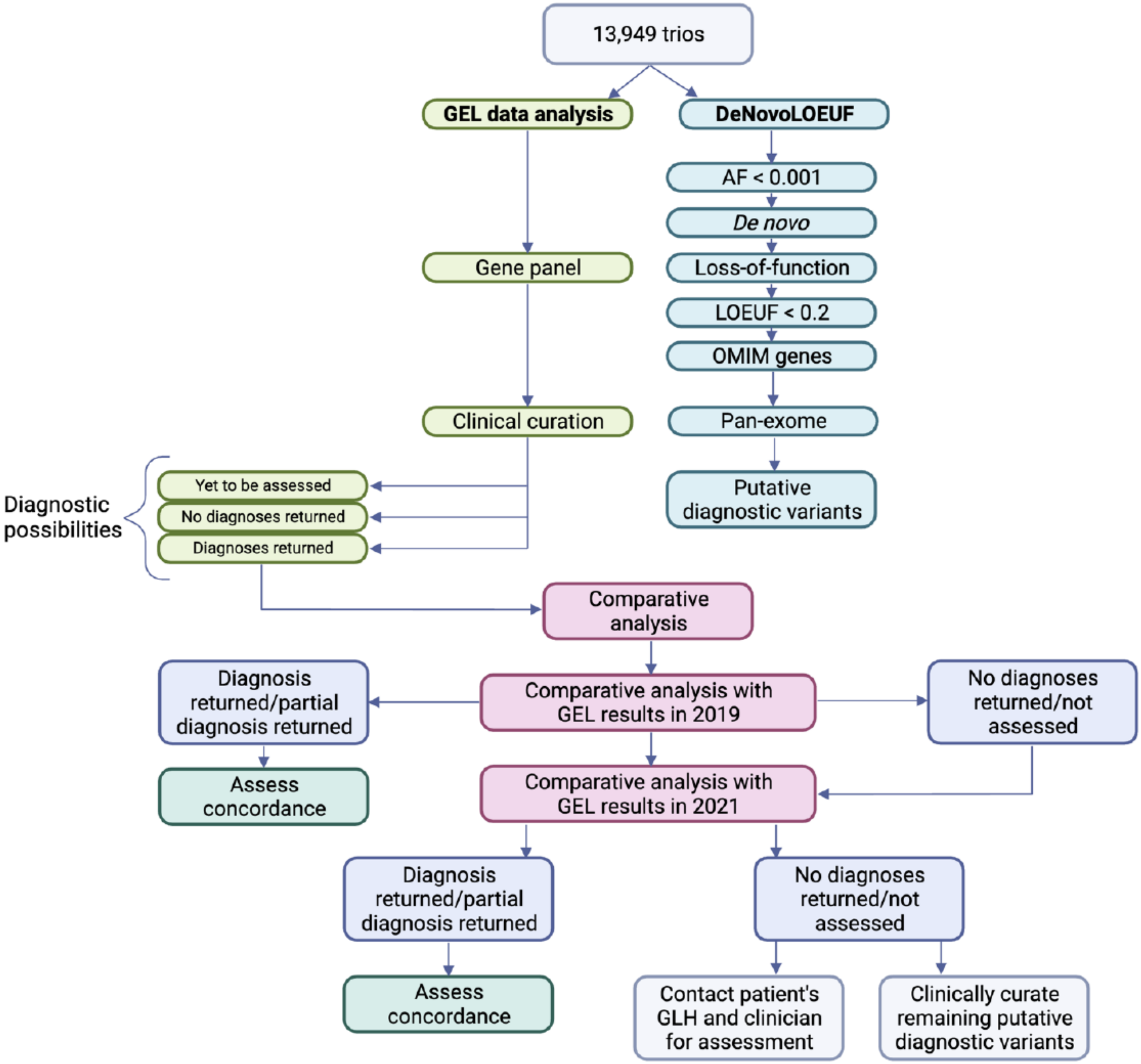
Referring clinicians recorded phenotype data as categorical HPO terms. These were accessible in the RE by querying HPO terms stored in mysql tables in a LabKey data management system. Gene panels were selected by GEL based on the phenotype terms provided. A summary of high-level phenotypes of the patient population is available in **Supplementary Table S7**.

7.2.3 Data analysis

Data analysis was first performed in Autumn 2019 (**Figure 7.3**). Bespoke scripts were developed to query the aggregate *.gvcf* file. Only variants that passed Illumina QC as previously described were selected.[132] The DeNovoLOEUF filtering strategy was then applied: Firstly, *de novo* variants with an allele frequency <0.001 in gnomAD v2.1.1 (all populations) were selected. This variant list was further restricted to predicted loss of function (pLoF) variants including nonsense, frameshift and essential (canonical +/- 2 base-pairs) splice site variants. LOEUF constraint gene scores, downloaded from the gnomAD browser (https://gnomad.broadinstitute.org), were imported into the research environment. Rare, *de novo*, pLoF variants were then further restricted to genes with a LOEUF score of <0.2 (n=1044), approximately equivalent to the first LOEUF decile, representing a list of genes most highly constrained for loss-of-function and predicted to cause disease through haploinsufficiency, as outlined in the flagship gnomAD paper and visualised in **Figure 7.2**.[133] Only variants in LOEUF constrained genes with known disease gene associations in the OMIM[328] database (n=335), first accessed and downloaded as a flat *.txt* file in October 2019, were kept for further analysis. Variants in autosomal recessive disease genes were excluded; this left 293 genes. Variants in novel disease genes were considered beyond the scope of this disease-gene focused assessment and are presented in **Chapter 4**.[329] LOFTEE v1.0[133] was applied to flag variants as potential false positives. Variants in the terminal exon were not excluded. Variants remaining following DeNovoLOEUF filtering steps were considered *putative diagnostic variants*.

Clinical outcome data pertaining to diagnostic reports and individual specific phenotype information were extracted querying Labkey using the RLabKey package (v2.9.0) in R (v4.0.3). These phenotype data were computationally merged with filtered putative pathogenic variants for each patient. The diagnostic report status for each patient, which included any returned pathogenic variants, was extracted computationally by querying the ‘GMC exit questionnaire’ table in LabKey.

FIGURE 7.3 | SUMMARY OF METHODS


13,949 trios
GEL data analysis
DeNovoLOEUF
AF < 0.001
De novo
Loss-of-function
LOEUF < 0.2
OMIM genes
Pan-exome
Putative diagnostic variants
Gene panel
Clinical curation
Diagnostic possibilities
Yet to be assessed
No diagnoses returned
Diagnoses returned
Comparative analysis
Diagnosis returned/partial diagnosis returned
Comparative analysis with GEL results in 2019
No diagnoses returned/not assessed
Assess concordance
Comparative analysis with GEL results in 2021
Diagnosis returned/partial diagnosis returned
No diagnoses returned/not assessed
Assess concordance
Contact patient's GLH and clinician for assessment
Clinically curate remaining putative diagnostic variants

Putative diagnostic variants identified by the filtering approach were compared with the diagnostic reports for the same patients. Following comparative analysis in 2021, if a negative report had been issued, or the case was still under review, the patient's Genomics Laboratory Hub and referring clinician were contacted and variants of interest were shared. If no response was received, the remaining variants were clinically curated by Dr Simon Thomas in the Wessex Regional Genomics Laboratory as per ACMG-AMP guidelines. AF – allele frequency; GEL – Genomics England; GLH – Genomics Laboratory Hub; LOEUF – Lower Observed/Expected Upper-bound Fraction; OMIM – Online Mendelian Inheritance in Man

7.2.4 *Comparative analysis*

Putative diagnostic variants were compared with the diagnoses returned to patients recruited to 100KGP at two time points (October 2019 and April 2021) to assess concordance between the DeNovoLOEUF filtering method and the analysis strategy by GEL.

At the first time point, putative diagnostic variants extracted using DeNovoLOEUF were compared against variants declared in the Genome Medicine Centre exit questionnaire of the RE as being returned to patients in their diagnostic report. This was to assess the positive predictive value of the method. It was expected that many patients would not have had a diagnostic report returned in 2019, i.e. their case status was "yet to be assessed". The comparative analysis was therefore repeated at the second time point (18 months later in April 2021), to assess whether the method correctly predicted additional diagnoses determined over time as the proportion of closed 100KGP cases increased.

Cases that were not assessed or reported as negative (i.e. no diagnosis identified) by 2021, were re-curated by NHS Clinical Scientists to standardise curation of any novel diagnoses not originally detected through the 100KGP. This was achieved in two ways. Firstly, the patient's Genomic Laboratory Hub (GLH), previously known as the Genome Medicine Centre, and referring clinician were contacted to discuss the variant(s) found. Both the GLH and clinician were asked whether the variant was already known about and/or had been returned as a diagnosis. Communication with the GLH and referring clinicians often prompted local multidisciplinary team meetings followed by diagnostic laboratory confirmation of the variant. Secondly, if the variants were unknown to the GLHs, or no response was received from the centres contacted, Dr Simon Thomas (NHS Clinical Scientist) in the Wessex Regional Genetics Laboratory, an established GEL diagnostic reporting centre, curated the remaining variants alongside the patients' phenotypes as per the ACMG-AMP guidelines.[124] These curation results then determined how many of the remaining putative diagnostics variants would meet a partial or full diagnosis.

7.2.4 Testing the method on non-trio data

In June 2022, an additional 6,101 families with complex family structures, whereby *de novo* analysis was not possible, were assessed. In these families, rare (AF <0.001), pLoF variants, in OMIM disease genes with a LOEUF score <0.2 were extracted. This filtering mimicked the DeNovoLOEUF strategy, aside for removing the *de novo* filter. Only variants present in affected individuals were retained. As before, any putative diagnostic variants were compared with the reported variants in the patient's GMC exit questionnaire.

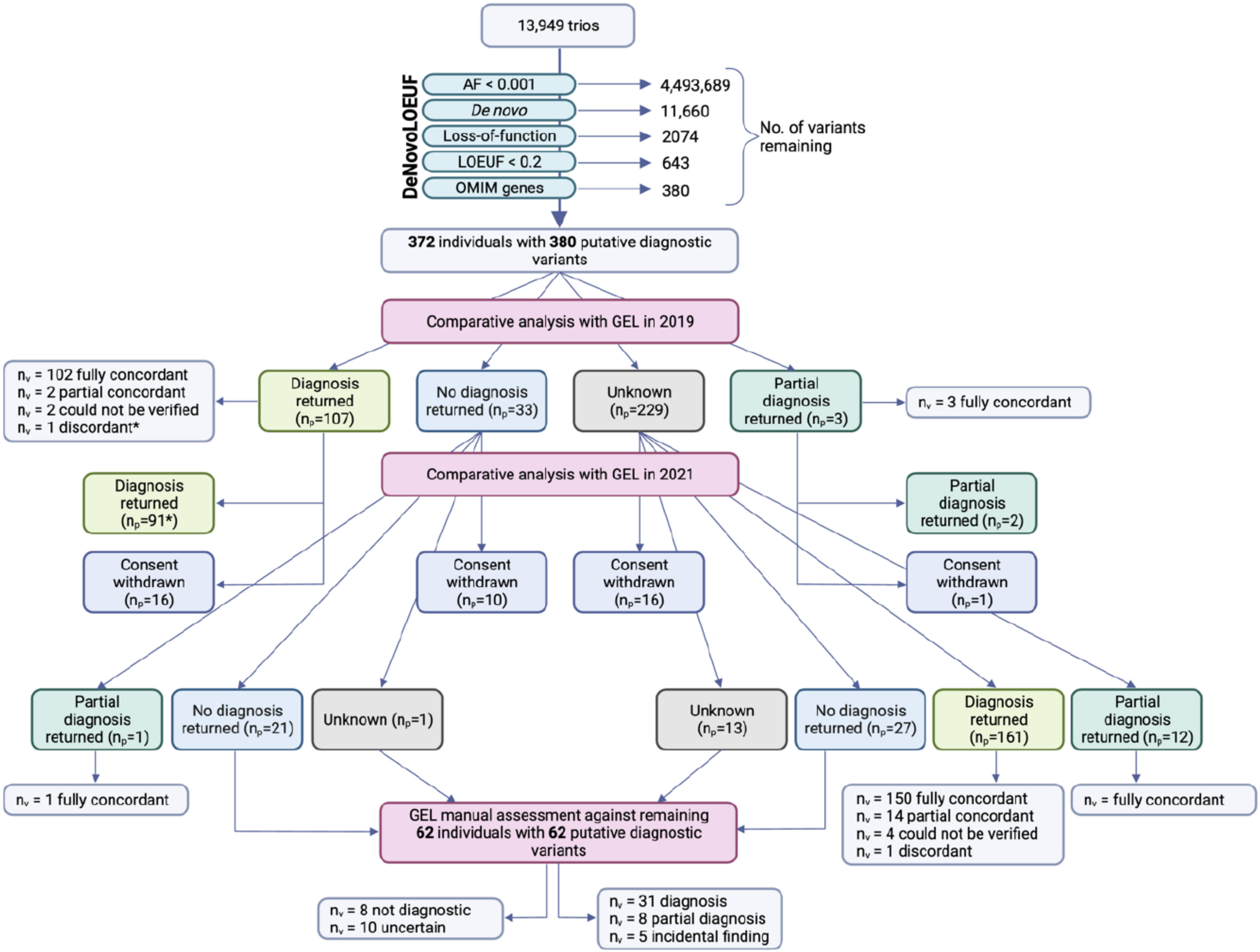
7.2.5 Iterative re-analysis

In August 2022, the DeNovoLOEUF method was applied to newly discovered disease genes (classified as 'definitive' or 'strong' in GenCC[322]) with a LOEUF score <0.2, published between 2019 and 2022 to assess for possible diagnostic uplift. These variants were then curated by a Dr Simon Thomas in an NHS diagnostic laboratory.

7.3 Results

A total of 380 putative diagnostic variants were identified by DeNovoLOEUF in 372 patients. Of these variants, 339/380 (89%) were in the Exomiser top ranked results. There were more variants than patients due to some individuals harbouring more than one *de novo* variant in the same gene ($n_{\text{patients}} = 2$) or having more than one *de novo* variant in two different genes ($n_{\text{patients}} = 6$). The patients with two *de novo* variants in the same gene were explored further and these variants did not represent a complex structural event. Results stratified by time-point assessment are shown in **Figure 7.4**.

Figure 7.4 | Summary of results using DeNovoLOEUF on 100,000 Genomes Project Patients


13,949 trios
DeNovoLOEUF
AF < 0.001
De novo
Loss-of-function
LOEUF < 0.2
OMIM genes
4,493,689
11,660
2074
643
380
No. of variants remaining
372 individuals with **380** putative diagnostic variants
Comparative analysis with GEL in 2019
n_v = 102 fully concordant
n_v = 2 partial concordant
n_v = 2 could not be verified
n_v = 1 discordant*
Diagnosis returned (n_p=107)
No diagnosis returned (n_p=33)
Unknown (n_p=229)
Partial diagnosis returned (n_p=3)
n_v = 3 fully concordant
Comparative analysis with GEL in 2021
Diagnosis returned (n_p=91*)
Partial diagnosis returned (n_p=2)
Consent withdrawn (n_p=16)
Consent withdrawn (n_p=10)
Consent withdrawn (n_p=16)
Consent withdrawn (n_p=1)
Partial diagnosis returned (n_p=1)
No diagnosis returned (n_p=21)
Unknown (n_p=1)
Unknown (n_p=13)
No diagnosis returned (n_p=27)
Diagnosis returned (n_p=161)
Partial diagnosis returned (n_p=12)
n_v = 1 fully concordant
GEL manual assessment against remaining **62** individuals with **62** putative diagnostic variants
n_v = 150 fully concordant
n_v = 14 partial concordant
n_v = 4 could not be verified
n_v = 1 discordant
n_v = fully concordant
n_v = 8 not diagnostic
n_v = 10 uncertain
n_v = 31 diagnosis
n_v = 8 partial diagnosis
n_v = 5 incidental finding

In October 2019, 380 predicted loss of function variants in 372 individuals were identified in known OMIM disease genes using DeNovoLOEUF. At the time, 29% (107/372) of patients had a single diagnostic variant returned and 33 patients had their case closed as 'negative'. Ninety-five percent (102/107) of variants identified by DeNovoLOEUF were entirely concordant with the formal GEL diagnosis returned. In two patients where a pathogenic variant was correctly identified, a second variant in a different gene also contributed to the diagnosis (partial concordant). Two variants were excluded from comparative analysis as it was not possible to verify the returned diagnosis in the GEL research environment's exit questionnaire. One variant was discordant (*) between the variant identified by DeNovoLOEUF and the reported outcome data from the 100KGP; however, this variant was subsequently confirmed as pathogenic by 2021. Three patients had a partial diagnosis returned, meaning a single variant was returned to the patient but that it did not fully explain the phenotype; all three were fully concordant with DeNovoLOEUF. 232 cases were 'unknown' meaning that no formal report had been returned to the patient. In summary, at the first time point, the method correctly detected 99% (107/108) of reported diagnostic variants. For the comparative analysis in 2021, 43 variants were excluded from downstream analysis due to patients being withdrawn from 100KGP. A further 150 variants were fully concordant with our method, 14 were partially concordant, 13 variants were concordant with a returned partial diagnosis, and 1 variant was discordant. Following assessment of the remaining 65 variants in 65 individuals, 31 cases were considered diagnostic, 8 cases partially diagnostic, and 5 cases were incidental findings. Eight variants did not explain the phenotype, and 10 cases were uncertain meaning that there was insufficient clinical information to determine causality. n_p = number of patients; n_v = number of variants.

7.3.1 Comparative analysis of method in 2021

In April 2021, 284/380 (75%) of all variants initially identified were confirmed as either fully diagnostic or partially diagnostic, including 17 patients who had diagnostic reports returned in 2019 prior to being withdrawn from the study. A single discordant variant identified in 2019 (**Figure 7.4**) was reclassified as a pathogenic variant by 2021 and returned to the patient as diagnostic. Twenty-six patients harbouring 26 variants identified in 2019 were unavailable for further assessment in 2021 and classified as "consent withdrawn" due being removed from the trusted research environment as re-consent was not established for child participants reaching adulthood. Only one variant identified did not match the variant returned to the patient, meaning GEL had returned an alternative diagnosis, however the variant identified by DeNovoLOEUF is a known pathogenic variant in ClinVar. This patient had only one HPO term recorded "cystinosis", which was consistent with the biallelic variants reported by GEL. The clinician responsible for this patient has been contacted to gain further clinical information and establish whether the putative diagnostic variant identified may be a missed additional diagnosis. Six variants could not be assessed for concordance as the variant returned by 100KGP was not identifiable in the GEL research environment. Sixty-two variants in 62 individuals identified using DeNovoLOEUF remained unresolved in 2021 and required further scrutiny. All 62 variants were in Tier 3 of the GEL tiering system as they were *de novo* loss-of-function variants but were not in the original gene panel(s) applied.

7.3.2 Assessment of remaining 62 variants in 62 unique individuals

Contact was established with seventeen (27%) of the remaining 62 patient's GLHs and referring clinicians. Following this connection, all 17 patients had their DeNovoLOEUF variants confirmed as disease-causing following independent validation in their local NHS laboratories. Contact was not established with the remaining 45 patients' referring clinicians and/or GLH and therefore the outstanding 45 variants were manually curated by two clinical scientists working in an NHS accredited genomic diagnostic laboratory (**Supplementary Dataset SD7**). Of these variants, 14 were designated diagnostic, 8 were partially diagnostic, and 5 were identified as incidental findings. Eight variants were considered not diagnostic, and

10 cases were uncertain with insufficient clinical information to confirm causality for the patient's phenotype (**Table 7.1**).

Table 7.1 | Classification of remaining 62 variants

Diagnosis (n=31)	Partial diagnosis (n=8)	Not diagnostic (n=8)	Uncertain (n=10)	Incidental Finding (n=5)
Confirmed following GLH contact (n=17)	Confirmed following manual curation (n=8)	Poor phenotype fit (n=3)	Pathogenic by ACMG-AMP guidelines (n=5)	Confirmed following manual curation (n=5)
Confirmed following manual curation (n=14)		Intronic on MANE Select transcript (n=2)	VUS by ACMG-AMP guidelines (n=5)	
		XLR gene and patient female (n=1)		
		Artefactual variant (n=1)		
		Splice rescue (n=1)		

Half (5/10) of the uncertain variants were classified as pathogenic by ACMG-AMP guidelines[123] but there was insufficient clinical information to confirm causality for the patient's phenotype; there was an average of 4 HPO terms per 'uncertain' case compared with 18 HPO terms for patients with a diagnosis. Of the 8 non-diagnostic cases, one variant was in an X-linked recessive (XLR) gene and the patient was female. A further variant passed GEL quality control filtering in 2019 but upon more detailed inspection was artefactual. One variant was within 6 base pairs of an alternative splice site, predicting a full splice rescue. Three variants were a poor phenotypic fit. Two variants were intronic on the Matched Annotation from NCBI and EBML-EBI (MANE)[330] Select transcript and not present in an exon of a MANE Plus Clinical alternative transcript harbouring known pathogenic variants.

7.3.3 Summary of results

In summary, 324/332 (98%) of the variants identified through DeNovoLOEUF filtering (excluding incidental findings, variants for withdrawn participants, or in patients where there was inadequate phenotype or genotype reporting in GEL) were classified as diagnostic or partially diagnostic (**Table 7.2**).

TABLE 7.2 | TABULATED SUMMARY OF RESULTS

GEL diagnoses returned	***2019 comparison***	***2021 comparison***	***Following curation/GLH***
Fully concordant	102	253	284
Partially concordant	5	32	40
Couldn’t be verified	2	6	16
Discordant	1 [discarded in 2021]	1	8
Incidental	0	0	5
PPV	107/108 = 99%	285/286 = 99%	324/332 = **98%**

Summary of results comparing variants identified by DeNovoLOEUF compared with diagnostic reports returned to patients, in addition to variant curation. The 2021 comparison is a cumulative comparison of the 2021 and 2019 comparisons. Full concordance is where a single variant identified was confirmed as the pathogenic variant. Partial concordance denotes the case where the variant identified was pathogenic but did not explain the full phenotype, or a second variant in a different gene was returned to the patient. Variants that could not be verified were those where it was not possible to see which variant had been returned to the patient as diagnostic, or there was not enough clinical information to determine causality. These 16 variants, plus the 5 incidental findings were not included in the PPV calculations. GLH – Genomics Laboratory Hub; PPV – positive predictive value.

7.3.4 DeNovoLOEUF on non-trio data

Filtering for rare, pLoF variants in known OMIM disease genes with a LOEUF score <0.2 on 6,101 families with complex family structures, revealed a further 776 putative diagnostic variants in 757 individuals. 270/757 (36%) of individuals had diagnoses returned that were fully concordant with DeNovoLOEUF. The number of individuals per pedigree was significantly different between concordant and discordant cases (Wilcoxon test, $p=6.02^{-16}$), with discordant patients having a median pedigree structure of one individual (singleton).

7.3.5 Proportion of disease-causing variants detected by DeNovoLOEUF

The total number of *de novo* pLoF variants detected by GEL and captured by DeNovoLOEUF was explored. A total of 2,074 *de novo* pLoF were identified in GEL (**Figure 7.4**). Of these, 480/2074 (23%) were confirmed diagnoses. Of these diagnoses, 380/480 (79.2%) were in LOEUF constrained genes (score <0.2). Different LOEUF cut offs were tested, including a score between 0.2 – 0.4 and between 0.4 - 0.6. The positive predictive values were 69% and 45% respectively.

7.3.6 Iterative re-analysis

Re-running LOEUF on newly discovered genes between 2019 and 2022 that were not present in the original OMIM gene list, identified 13 new *de novo* pLoF variants. Of these, 12/13 (92%) have been confirmed as diagnostic (**Supplementary Dataset SD8**).

7.4 Discussion

This chapter describe a fast, unbiased filtering strategy, DeNovoLOEUF, to identify potential pathogenic variants with high positive predictive value and specificity. Human gene LOEUF scores are available in **Supplementary Table S8**. Unlike the approach adopted in the 100KGP with panel-based tiering, this genotype-driven DeNovoLOEUF method is agnostic to phenotype and independent of gene panels which often change over time and require optimal panel selection. Indeed, 39 diagnostic or partially diagnostic variants (in 39 known disease-associated genes) were detected that had been missed by standard 100KGP diagnostic protocols due to the gene in which the causal variant was identified not being included on the gene panel selected. Ninety percent (35/39) of these genes were included on different gene panels, meaning they were recognised "PanelApp disease genes" but the gene panel selection was suboptimal. One of the issues with PanelApp is in selecting the 'correct' panel based on the HPO terms provided. For example, the gene *CLCN5* is a disease gene for Dent disease 1 (MIM: 300009), a recognised renal tubulopathy. This gene is on the following PanelApp gene panels: "nephrocalcinosis", "unexplained kidney failure in young people", "skeletal dysplasia" and "unexplained paediatric onset end stage renal disease". However, it is not on the "tubulopathies" panel, which is perhaps the most appropriate panel selection. For most cases where GEL missed the diagnosis, the panel selected was not inappropriate *per se*, but did not encompass the exact panel required, again highlighting why 'agnostic to phenotype' approaches are critical to ensure increased diagnostic yield, particularly in the UK where gene panels are the selected analytical method.

In view of the recognition of diagnoses outside of the selected gene panel(s), some NHS accredited laboratories adopted a policy, when reviewing results from the 100KGP available within the 100K results portal, to assess *de novo* variation and Exomiser top-ranked results to uplift the resultant diagnoses. This approach to reporting has been informative for the strategy within subsequent large-scale WGS endeavours. When including these results, 26% of all causal variants returned by NHS labs in the 100KGP were not in the initial gene panel applied, exemplifying the issue with gene panel analysis strategies.[326] However, re-analyses involve re-visiting sequencing data and clinical cases, something which could be mitigated by screening and prioritising highly putative diagnostic variants in the first instance. Using DeNovoLOEUF as a screening strategy would have immediately identified 321 pathogenic variants, saving considerable time and money. On average, DeNovoLOEUF added only one extra variant for assessment in ~3% of all rare disease probands (0.023 variants per person).

This method is rapid, having identified 172 variants in 2019 that could not be efficiently returned to the patients' clinical teams as the processes for returning research results to the clinic were not supported at scale. As a result of collaboration with colleagues at Genomics England, a form that enables the submission of multiple potential diagnoses for different participants via a single submission within the RE, is now in place. DeNovoLOEUF utilises an effective screening approach to detect highly penetrant putative diagnostic variants across a large cohort. It should however be noted that whilst 285/333 (86%) of all the variants identified were fully diagnostic, 40/333 (12%) were a partial diagnosis, meaning that the variant was considered pathogenic but did not fully explain the phenotype. Additionally, following manual curation there were 10 variants whereby it was not possible to confirm whether the variant explained the phenotype; all these patients lacked sufficient clinical data to determine causality even though five of the variants were pathogenic by ACMG-AMP guidelines. These patients had a median of 4 HPO terms compared with 18 HPO terms for patients with a diagnosis. This highlights some of the challenges with phenotyping in a large-scale national sequencing project and that using HPO terms are sometimes insufficient to make a diagnosis and post-analysis communication with the clinical care team is a critical component of molecular diagnosis as emphasised by ACMG clinical practice guidelines.[331] Genomics England is actively supporting improvements at the clinical-research interface to enable collaborations between researchers and clinicians and in patient phenotyping by the provision of Hospital Episode Statistics data within the RE as a longitudinal record of participants' phenotypes.

With ever increasing application of genome sequencing and a drive to sequence a further 5 million genomes in the UK, there is clear demand to find efficient analytic strategies. This method shows promise as a suitable adjunct to the current protocols to identify causal variation in 100KGP, the NHS Genomic Medicine Service, or other similar international initiatives. DeNovoLOEUF is capable of prioritising putative pathogenic variants for diagnostic laboratories, saving time and resources. Furthermore, one of the draw backs of applying gene panels is that many are already outdated at the point of use, with new genes being consistently added to the literature. This method can be easily applied iteratively for re-analysis as new genes are discovered and added to ClinVar[113], HGMDPro, or GenCC[322], before being indexed in OMIM. However, at the time of method development, OMIM represented the best available repository of disease genes with a standardised method for curating genotype-phenotype relationships. However

DeNovoLOEUF was re-run in 2022 on new disease genes (added to GenCC) after the initial analysis in 2019; this identified a further 13 variants of which 12 have been confirmed as diagnostic.

Whilst the DeNovoLOEUF method has a high PPV, nine variants (1 from the 2021 analysis and 8 from curation analysis) were discordant with the diagnosis returned. One variant was in an X-linked recessive gene and the patient was female. A further variant passed quality control filtering but was artefactual when the reads were directly visualised. Four variants were in a disease gene inconsistent with the patient's phenotype and there was an average of 3 HPO terms per patient. One variant was within 6 base pairs of a full splice rescue, rendering it not truly LoF. Two variants were pLoF on a non-canonical transcript that was poorly expressed across disease-relevant tissues using the pext score based on GTEx,[87] and intronic on the MANE Select transcript. Whilst one option would be to limit the method to variants on the MANE Select and MANE Plus Clinical transcripts, this is potentially problematic as not all genes have been curated to define additional transcripts to be included in the MANE Plus Clinical resource.[1]

7.4.1 Limitations and opportunities

Whilst this genotype-first method diagnosed patients missed by the initial 100KGP diagnostic strategy, it does not replace the importance of a phenotype-driven approach. It would be foolhardy not to look at all variants in a gene with a close phenotype match to the patient. This method is best applied as a complementary screening strategy and will not diagnose most patients, especially those with variants in non-constrained genes, or with pathogenic missense or extended splice site variants or with inherited variants. It also does not negate the need for variant curation, since some pLoF variants may not result in LoF and there is yet to be an automated method capable of replicating full manual curation.[280]

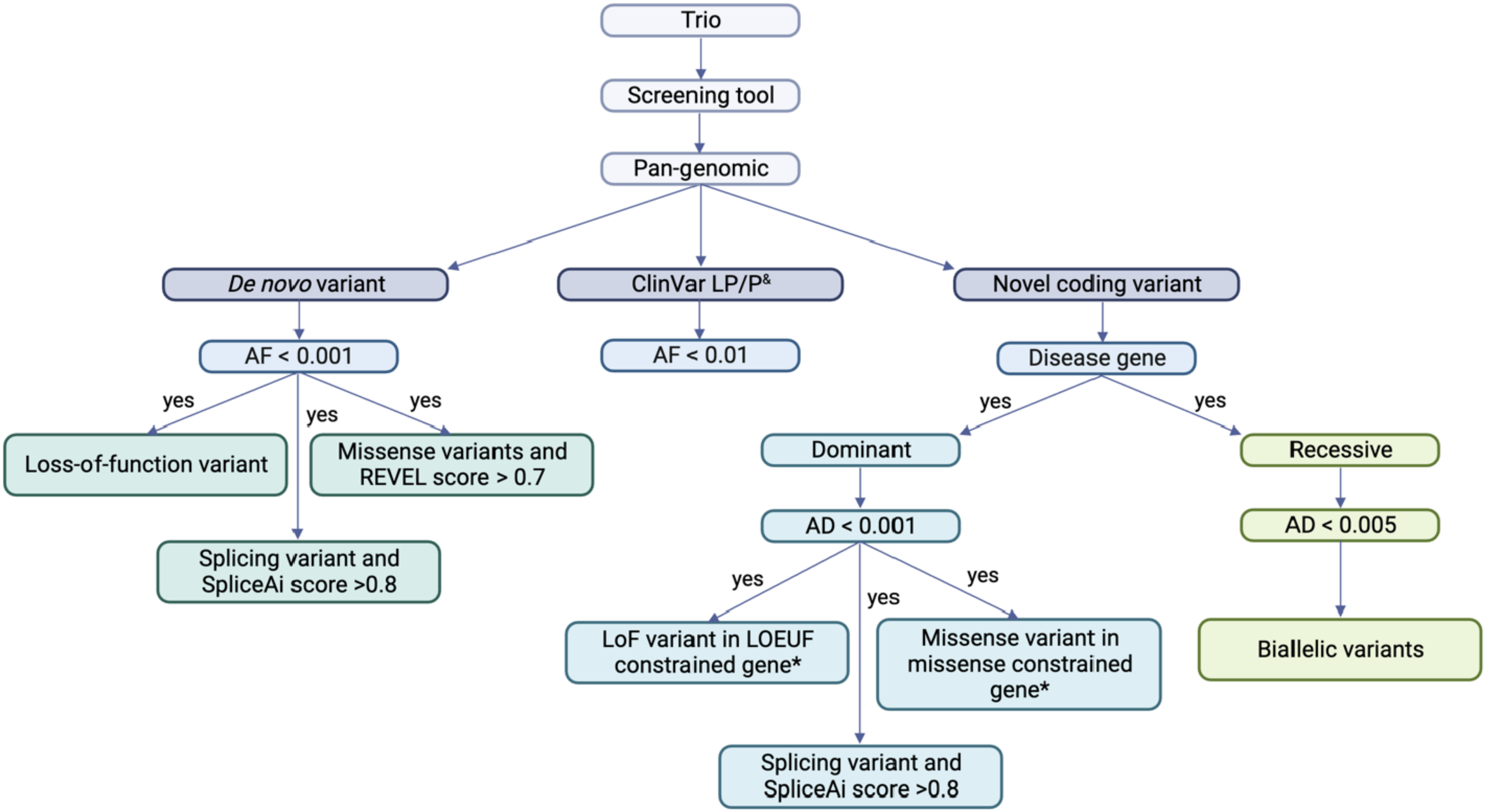
This screening tool selects *de novo* variants, meaning that potential pathogenic variants were excluded in patients without trio data and in diseases whereby disease segregation may be expected e.g. cardiac or immune disease. When applying the same filtering strategy, minus the *de novo* filter, to complex family structures or singletons, an additional 270 diagnoses were identified. This yielded a PPV of 36% vs 98% for trios. Used prospectively, this means there is an increased probability of type 1 errors, although some of the unsubstantiated variants may represent real diagnoses. Unsurprisingly, for non-trio cases where returned diagnoses were discordant with DeNovoLOEUF, the median family size was 1, with 631/678 (93%) of cases being proband-only (singletons).

Genes in the first LOEUF decile were selected to increase the specificity of the method. As shown in the flagship gnomAD paper, a LOEUF score of <0.2 is the most highly enriched for haploinsufficient disease genes. In total across all genes, there were 2074 *de novo* pLoF variants called in Genomics England of which 480/2074 (23%) were diagnoses. Of these, 79.2% were in LOEUF constrained genes (score <0.2). Expanding the approach to *de novo* pLoF variants in disease genes using a higher LOEUF threshold will inevitably increase the diagnostic yield, however this must be balanced with increased noise, an increased number of variants for review, and significantly reduced specificity. When expanding the analysis to a LOEUF score between 0.2-0.4 the positive predictive value reduced to 69%, and with a LOEUF score between 0.4 and 0.6 this reduced again to 45%. This approach could be adopted for downstream analyses.

DeNovoLOEUF also risks identifying incidental findings including those in the ACMG secondary findings v3 list, which are a consequence of WGS and a trade-off for increased diagnostic yield. Applying a LOEUF cut off <0.2 identified 10 genes in the ACMG list, although these genes could be manually removed from DeNovoLOEUF if preferred, or if the patient has not consented for secondary findings.[332]

Whilst DeNovoLOEUF has a high PPV, there are opportunities to refine the method to increase sensitivity at the expense of specificity. A revised screening pipeline is proposed to not only identify *de novo* variants in LOEUF constrained genes, but also screen for all known rare ClinVar pathogenic or likely pathogenic variants regardless of the gene panel applied (**Figure 7.5**). The suggestion is to place *de novo* variants, ClinVar variants and novel coding variants into a new tier for assessment by NHS clinical laboratories to complement the current tiering system.[132] Adding this additional review approach, along with a phenotype-driven variant analysis, is consistent with recently released best practices in genome analysis released from the Medical Genome Initiative.[333] From reviewing 20 genomes, it is estimates that this approach would yield 3-9 potential additional variants per trio, using a LOEUF cut off <0.35, a missense constraint z-score >3,[334] and likely pathogenic/pathogenic ClinVar entries with 2 stars or above. The aim is to achieve a higher diagnostic yield per number of variants assessed by diagnostic labs, whereby the most salient variants are prioritised first.

FIGURE 7.5 | PROPOSED GENOTYPE-FIRST SCREENING STRATEGY



A possible refined screening strategy. All de novo variants with an allele frequency (AF) < 0.001 would be filtered, and loss of function variants would be prioritised, in addition to missense variants with a REVEL score >0.7, and splicing variants with a SpliceAI score > 0.2. All known ClinVar likely pathogenic (LP) and pathogenic (P) variants would be reviewed independent of zygosity and prioritised with an AF < 0.01. Novel coding variants (absent from population databases) in a known OMIM disease gene would be extracted and divided by disease mechanism. For dominant genes, variants are filtered with an AF of < 0.001. Retained variants would be prioritised if they were a loss of function (LoF) variant in a LOEUF constrained gene; a missense variant in a missense constrained gene, using a z-score >3 using the statistical method described by Samocha et al.[212]*; or a splice site variant with a SpliceAI score >0.2. For recessive genes, variants are filtered with an AF < 0.005. Any biallelic (phased) variants are retained.*

**Denotes user-specified cut offs for LOEUF and z-score missense constraint which can be tailored (or removed) as required.*

&Suggest using ClinVar 2 and above.*

7.5 Conclusion

This chapter presents a targeted screening tool, DeNovoLOEUF, that can be applied at scale to rapidly identify putative pathogenic variants with a 98% positive predictive value. The method complements current family-based analyses and can add value by identifying diagnostic variants missed by filtering strategies that adopt predefined disease-targeted gene panels. A total of 39 pathogenic variants were identified that were missed by the initial 100KGP variant prioritisation strategy. With 5 million more genomes being sequenced on the NHS, and many other international sequencing studies underway, this method alongside the new GEL initiative to report on Exomiser top ranked variants can help rapidly and effectively improve diagnostic efficiency and uplift diagnostic rates for the benefit of rare disease patients and their families.

Chapter 8 | Diagnosing cardiomyopathy phenotypes in the 100,000 Genomes Project

8.0 Contribution statement

The concept, methods and analyses applied to this chapter were carried out by myself under the supervision of my supervisory team.

8.1 Introduction

Cardiomyopathies (CM) are myocardial disorders defined by an abnormal structure and function of the heart in the absence of coronary artery disease or an alternative haemodynamic cause.[335] There are five common types of CM including dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), arrhythmogenic right ventricular cardiomyopathy (ARVCM), left ventricular noncompaction cardiomyopathy (LVNCM) and restrictive cardiomyopathy. Commonly, these are categorised by cardiac imaging, notably echocardiogram. All five subtypes are associated with monogenic disease (**Supplementary Dataset SD9**). Cardiomyopathies may also arise from systemic disorders such as haemochromatosis, amyloidosis, drug toxicity, and hypertension; as part of complex monogenic syndromes such as Noonan syndrome; and as a direct result of neuromuscular disease e.g. Duchenne Muscular Dystrophy.

8.1.1 Hypertrophic cardiomyopathy (HCM)

HCM is genetic disorder, which in the absence of a haemodynamic cause, is characterised by left ventricular (LV) hypertrophy and a non-dilated left ventricle with preserved or increased ejection fraction. Commonly, hypertrophy is asymmetrical with the basal interventricular septum most severely affected with histological evidence of myocyte hypertrophy and interstitial fibrosis. Left ventricular diastolic dysfunction is often observed. For many patients, HCM is well tolerated and managed, however HCM is also associated with sudden cardiac death in adolescents and young adults, usually through ventricular tachycardia.[336]

It is estimated that a diagnosis is found in 60% of familial cases with HCM, but in <50% of overall cases.[337] Most of the genes implicated in HCM are sarcomeric, including *MYH7* and *MYBPC3,* which make up >50%

of cases.[336] Variants in *MYH7* are predominantly missense, whereas the majority in *MYBPC3* are loss-of-function. The sarcomere is the fundamental motor unit of the cardiomyocyte and is composed of two principal components: the thick filament (myosin) and the thin filament (actin). Muscle contraction results from the integration between myosin and actin. For HCM, most pathogenic missense variants are believed to be dominant negative, where the mutant protein is incorporated into the sarcomere, but where its interaction with the wild-type protein disrupts normal sarcomeric assembly and function.[338]

8.1.2 Dilated cardiomyopathy (DCM)

DCM is defined by an enlarged left ventricle with variable degrees of systolic dysfunction, which is often progressive. DCM predisposes to heart failure and is associated with increased risk of arrhythmia. It can be a monogenic disease or secondary to disease processes such as valvular disease, infection, inflammation, toxins, and hypertension.[339] For familial DCM, the diagnostic rate is estimated to be 25-40%.[340] Secondary DCM is still likely to be influenced by genetic factors augmenting disease risk.[335] The majority of genetic DCM is inherited autosomal dominantly, although autosomal recessive, X-linked recessive and mitochondrial inheritance also occur.[335] Monogenic forms of DCM are most often caused by protein truncating variants in titin (*TTN)* (~20-25% of cases) and *LMNA* which encodes lamins A and C (5-8% of cases).[83,335,341] Titin is the largest protein in the human body and an important determinant of the elastic, contractile and signalling properties of cardiac and skeletal muscle. The function of TTN is heavily influenced by the alternative splicing of the *TTN* gene. In DCM it is unclear whether truncated proteins lead to dominant negative effects or act via haploinsufficiency.[342]

8.1.3 Arrhythmogenic right ventricular cardiomyopathy (ARVCM)

ARVCM is a disease of the myocardium caused by atrophy and cell death. It has a prevalence of between 1:1000 to 1:5000 and is characterised by progressive fibrofatty replacement starting and the epicardium and extending transmurally, interfering with electrical impulse conduction and causing ventricular arrhythmias.[343] Approximately 50% of ARVCM is thought to be familial, caused by heterozygous variants in genes encoding desmosomal proteins, such as *PKP2, DSG2, DSC2, JUP* and *DSP*. ARVCM tends to present after puberty, usually between the second and fourth decade of life.[344,345] Desmosomes are essential in maintaining the structural integrity of the ventricular myocardium. Mutated desmosomal proteins are believed to result in detachment of cardiac myocytes by loss of cellular adhesions, leading to cell death and substitution of fibrofatty adipose tissue.[346] The molecular diagnostic yield for ARVCM is estimated to be 50%.[340]

8.1.4 Left ventricular noncompaction cardiomyopathy (LVNCM)

LVNCM is a rare heterogenous disorder of the myocardium with a failure of compaction thought to be due to arrest of endomyocardial morphogenesis. It has an incidence of 5:10,000. LVNCM is characterised by prominent trabeculations, intratrabecular recesses, and a left ventricular myocardium with compacted and non-compacted layers.[347] LVNCM shows variable expressivity and can present at any age. It can range from asymptomatic disease, detected incidentally on echocardiography, to fulminant end-stage heart failure.[348] Patients with LVNCM are at increased risk of ventricular tachyarrhythmias and sudden cardiac death. Several genes have been implicated in LNVCM, including *ACTC1, MIB1, MYBPC3, MYH7* and *TPM1* and it is estimated that 20-30% of cases are genetic in nature.[349] It is believed that genes involved in LVNCM may share a final common pathway, however the underlying molecular mechanisms are unknown.[350] The diagnostic yield from genetic testing in LVNC varies from 9-41% dependent on patient selection and genes assessed.[351]

8.1.5 100,000 Genomes Project

The 100,000 Genomes Project (100KGP) was a government funded scheme in the UK, run by Genomics England (GEL) offering genome sequencing to ~100,000 patients and their families with rare diseases, cancers, and infectious diseases. The diagnostic workflow for the 100KGP is described elsewhere in **Figure 2.1, Methods 2.2.1.4**. The 100KGP provides a database of rich genotype and phenotype data; for CMs, this offers the potential for diagnostic discovery where conventional clinical testing has been unsuccessful in determining a molecular cause.

Inclusion criteria for CMs in the 100KGP was initially strict, requiring prior testing of specific genes (see https://files.genomicsengland.co.uk/forms/Rare-Disease-Eligibility-Criteria.pdf). In addition, for paediatric cases, at least another family member had to be affected.

This chapter aims to assess the diagnostic efficacy of cardiomyopathies in the 100KGP, which represent heritable and non-neurodevelopmental phenotypes. The chapter specifically explores the diagnostic rate achieved by 100KGP for affected participants and their relatives recruited under different subtypes of cardiomyopathy. Further, this chapter explores whether application of a cardiomyopathy 'super panel', inclusive of all genes related to cardiomyopathies in PanelApp can achieve a diagnostic uplift and identify missed diagnoses. Predicted loss-of-function variants in genes without a known disease-gene relationship are also explored for patients with cardiomyopathy phenotypes. Pathogenic variants in cardiomyopathy genes that were returned to patients *without* a cardiomyopathy phenotype are also discussed.

8.2 Methods

8.2.1 Data access

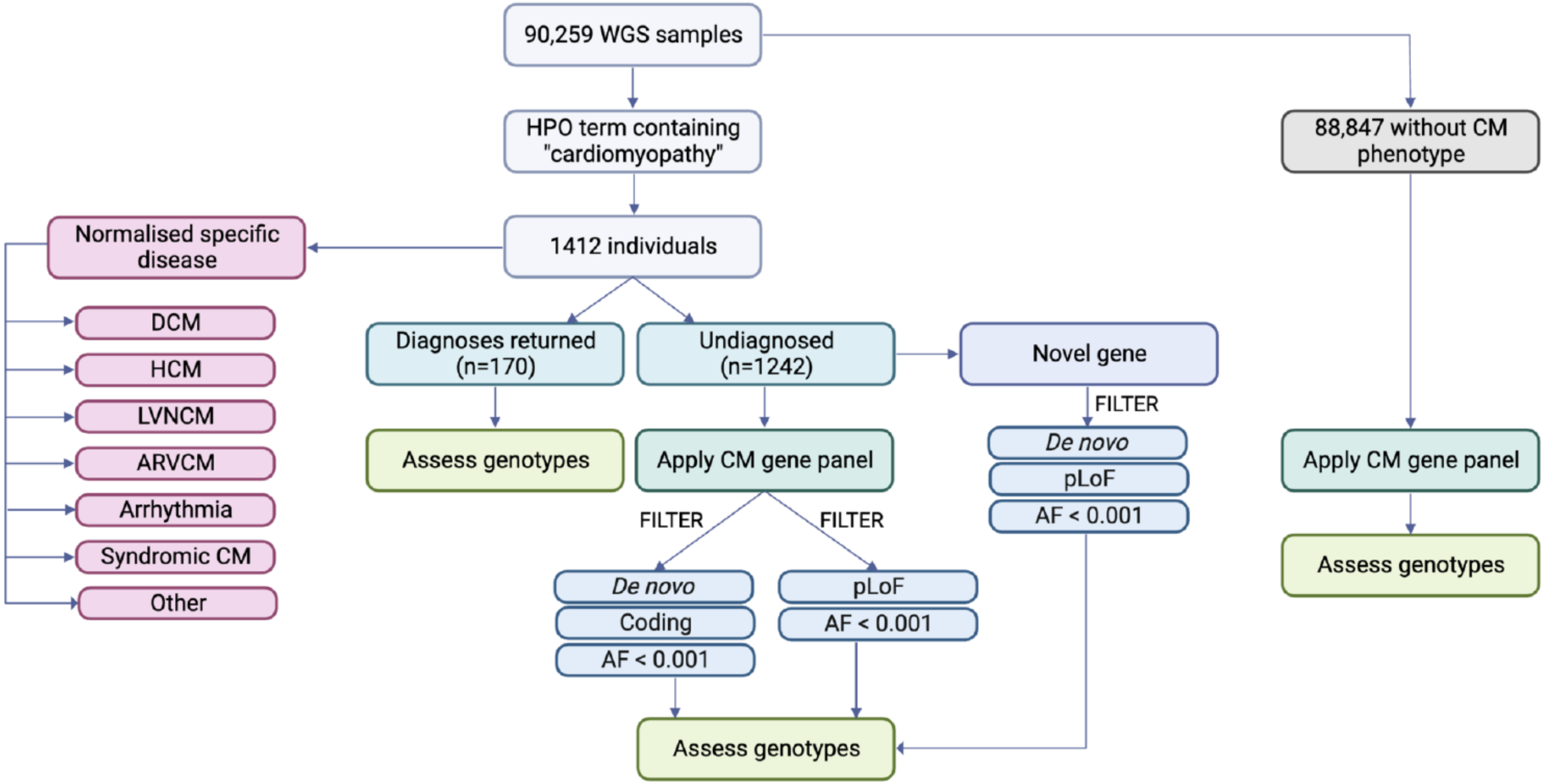
Access to the Genomics England Research Environment (RE) and high-performance cluster was obtained through membership of a Genomics England Clinical Interpretation Partnership: *Quantitative methods, machine learning, and functional genomics*. This provided access to whole genome sequencing and phenotype data for 90,259 individuals (release version 14, January 2022) sequenced as part of the 100,000 Genomes Project. Variants returned to patients by the 100KGP are available in the 'gmc exit questionnaire' Labkey table in the RE.

Access to the 100KGP dataset is restricted and only available as a registered GeCIP member in the Genomics England Research Environment. All data shared in this chapter were approved for export by Genomics England. The datasets and code supporting the current study, unavailable for export, are fully accessible within the Genomics England Research Environment in the shared directory: re_gecip/machine_learning/Ellie_Seaby/.

8.2.2 Identifying patients with cardiomyopathies

Phenotype data in the RE are stored as human phenotype ontology (HPO)[167] terms submitted by the referring clinician and stored in mysql databases within a LabKey data management system. Participants (probands or relatives) were defined as having a CM phenotype if the term "cardiomyopathy" was contained within the free text of any one of their HPO terms e.g. dilated cardiomyopathy or hypertrophic cardiomyopathy etc. Each participant was then subcategorised by cardiomyopathy subtype based on their 'normalised specific disease'. This is a high-level category that participants within GEL were recruited under and provided the basis for their gene panel selection by Genomic England. Normalised specific disease categories pertaining to cardiomyopathies included: ARVCM, DCM, HCM, and LVNCM. However, some patients with a CM phenotype were recruited under a non-CM normalised specific disease category. These cases were manually reviewed and grouped as either 'syndromic CM', 'Arrhythmia' or 'Other'. 'Syndromic CM' was defined as a syndrome or rare disease whereby cardiomyopathy is a known associated feature, e.g. Noonan syndrome or Duchenne Muscular Dystrophy. 'Arrhythmia' was defined by any known arrythmia e.g. LongQT syndrome, Brugada etc. Finally, 'Other' was defined as participants recruited with specific disorders ostensibly unrelated to CMs, such as chronic kidney disease, hearing loss, thoracic aortic aneurysm disease etc (**Figure 8.1**).

FIGURE 8.1 | METHODOLOGICAL WORKFLOW

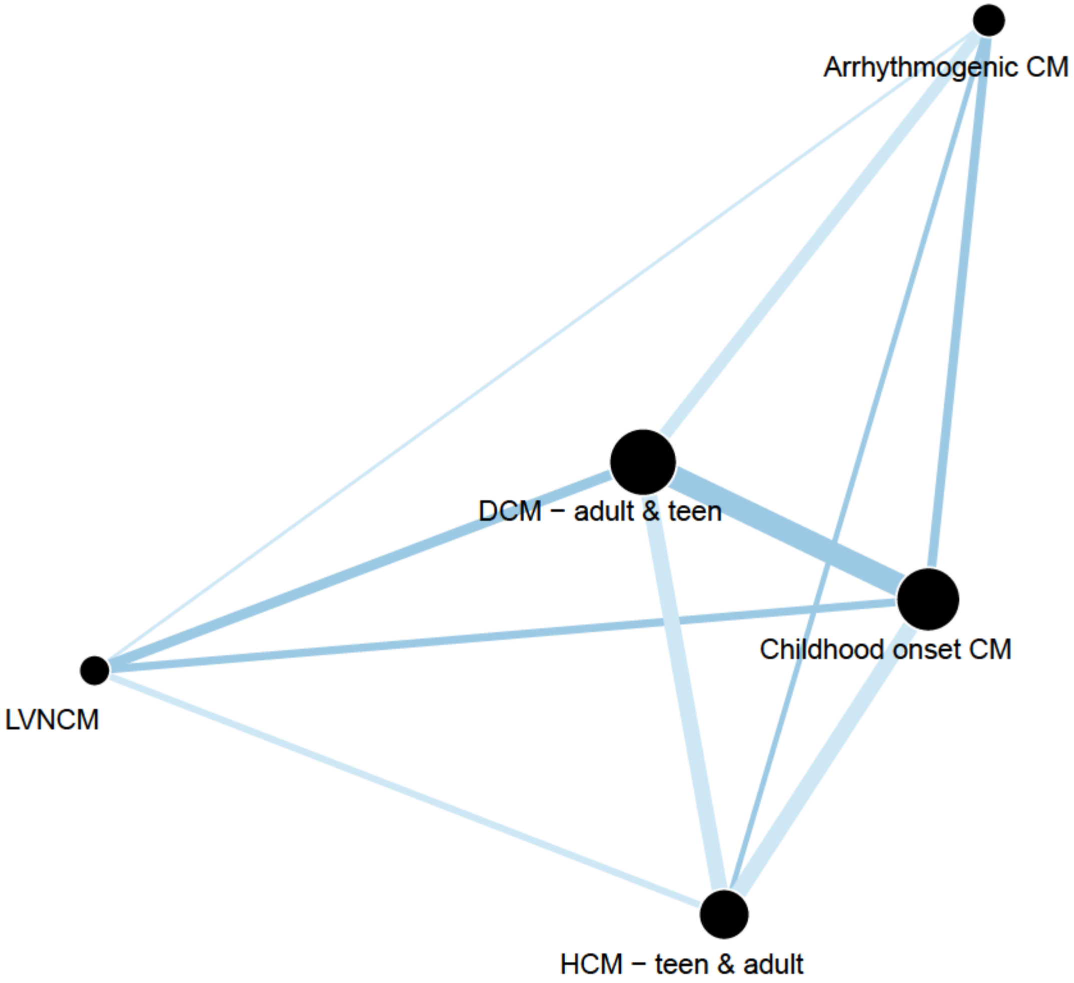


90,259 whole genome sequencing (WGS) samples and their associated phenotype data were accessed in the Genomics England Research Environment. 1412 participants were identified as having "cardiomyopathy" contained within the free text of their HPO terms; these participants were recruited under a range of normalised specific diseases i.e. recruitment categories. Of these, 170 had a molecular diagnosis returned by the 100KGP. Genotypes were assessed and data were collected on genes involved. 1242 participants with a CM phenotype had a negative report returned through 100KGP. For these, a CM super gene panel (as defined in 8.2.3) was applied followed by application of two different filtering strategies: 1) any de novo coding variant with an allele frequency (AF) in gnomAD v2.1.1 <0.001; and 2) any predicted loss-of-function (pLoF) variant with an allele frequency in gnomAD < 0.001. These genotypes were manually interrogated for missed diagnoses, considering allele frequency, segregation data, in silico metrics and suitability of gene involved. Separate to application of a CM-gene panel, 1242 undiagnosed cases were filtered for novel genes (not yet associated with disease i.e. absent from OMIM). Variants were retained if de novo, coding and with AF <0.001 in gnomAD, or the variant was pLoF with an AF <0.001 in gnomAD. For the 88,847 participants without a CM phenotype, a CM super panel was applied and variants were assessed against ACMG criteria for CM diagnoses.

8.2.3 Identifying diagnosed individuals with cardiomyopathies

To identify diagnosed individuals with CMs, all participants with a CM phenotype had their gmc exit questionnaire assessed for reported ‘likely pathogenic’ or ‘pathogenic’ variants (in any gene). Summary statistics were collected on the genes involved for each subtype of CM in addition to the diagnostic rates. This was followed by assessment of participants without a cardiomyopathy phenotype that had a pathogenic variant retuned by 100KGP but in a cardiomyopathy super panel, comprising a permissive list of CM-related genes (**Table 8.1 and Supplementary Dataset SD9**). The cardiomyopathy super panel included any ‘green’ gene from the following PanelApp[118] gene panels: *Arrhythmogenic cardiomyopathy, Cardiomyopathies - including childhood onset, Dilated cardiomyopathy - adult and teen, Dilated Cardiomyopathy and conduction defects, Hypertrophic cardiomyopathy - teen and adult, Left Ventricular Noncompaction Cardiomyopathy*. The overlap between the PanelApp gene panels is shown in **Figure 8.2**.

FIGURE 8.2 | CARDIOMYOPATHY SUPER PANEL- OVERLAP BETWEEN PANELAPP GENE SETS



Schematic representation of the overlap between PanelApp gene panels containing cardiomyopathy related genes. The CM super panel comprises all six PanelApp cardiomyopathy gene panels. Node size corresponds with size of panel. Edge weight corresponds with gene overlap between panels. Abbreviations: CM – cardiomyopathies; DCM – dilated cardiomyopathies; HCM – hypertrophic cardiomyopathies; LVNCM – left ventricular non-compaction cardiomyopathies.

Table 8.1 | Cardiomyopathy super panel – collation of PanelApp gene panels containing cardiomyopathy related genes

PANELAPP	GENES
ARRHYTHMOGENIC CARDIOMYOPATHY	*DES, ANK2, CAVIN4, CDH2, CTNNA3, DSC2, DSG2, DSP, FLNC, JUP, LDB3, LMNA, PKP2, PLN, RBM20, RYR2, SCN5A, TGFB3, TMEM43, TTN*
CARDIOMYOPATHIES - INCLUDING CHILDHOOD ONSET	*AARS2, ABCC9, ACAD9, ACADVL, ACTA1, ACTC1, ACTN2, AGK, AGL, ALMS1, ALPK3, ANK2, ANKRD1, APOPT1, ARSB, ATP5D, ATPAF2, B3GAT3, BAG3, BCS1L, BRAF, BTK, CACNA1C, CBL, CDH2, COA5, COA6, COA7, COX10, COX14, COX15, COX20, COX6A1, COX6B1, COX7B, CPS1, CPT2, CRYAB, CSRP3, CTF1, CYC1, DES, DHCR7, DMD, DNAJC19, DOLK, DSC2, DSG2, DSP, DTNA, EMD, EPG5, ETFA, ETFB, ETFDH, EYA4, FAH, FASTKD2, FHL1, FHOD3, FKRP, FKTN, FLII, FLNC, FNIP1, FOXRED1, GAA, GALNS, GATA6, GBE1, GLA, GLB1, GLRA1, GNS, GSN, GUSB, HADHA, HADHB, HCN4, HFE, HGSNAT, HRAS, IDH2, IDS, IDUA, ILK, JPH2, JUP, KIF20A, KRAS, LAMA4, LAMP2, LDB3, LMNA, LRPPRC, LYRM7, LZTR1, MAP2K1, MAP2K2, MCM10, MIB1, MLYCD, MMACHC, MRAS, MRPL44, MT-TI, MUT, MYBPC3, MYH6, MYH7, MYL2, MYL3, MYLK3, MYPN, NAA15, NAGLU, NDUFA1, NDUFA10, NDUFA11, NDUFA2, NDUFA4, NDUFA6, NDUFA9, NDUFAF1, NDUFAF2, NDUFAF3, NDUFAF4, NDUFAF5, NDUFAF6, NDUFAF8, NDUFB11, NDUFB3, NDUFB8, NDUFS1, NDUFS2, NDUFS3, NDUFS4, NDUFS6, NDUFS7, NDUFS8, NDUFV1, NDUFV2, NEBL, NEXN, NF1, NKX2-5, NONO, NRAP, NRAS, NUBPL, PCCA, PCCB, PDLIM3, PET100, PKP2, PLD1, PLN, PNPLA2, PPA2, PPCS, PPP1CB, PPP1R13L, PRKAG2, PTPN11, RAF1, RASA2, RBM20, RHBDF1, RIT1, RNF220, RPL3L, RYR2, SCN5A, SCO1, SCO2, SDHA, SDHAF1, SDHD, SGCD, SGSH, SHMT2, SHOC2, SLC22A5, SLC25A20, SLC25A4, SLC30A5, SOS1, SOS2, SPEG, SPRED1, SPRED2, SURF1, TAB2, TACO1, TAZ, TCAP, TGFB3, TMEM126B, TMEM43, TMEM70, TMPO, TNNC1, TNNI3, TNNI3K, TNNT2, TOR1AIP1, TPM1, TSFM, TTC19, TTN, TTR, UQCC2, UQCRB, VCL, NA*
DILATED CARDIOMYOPATHY - ADULT AND TEEN	*ABCC9, ACTC1, ACTN2, ANK2, ANKRD1, BAG3, CDH2, CRYAB, CSRP3, DES, DMD, DOLK, DSC2, DSG2, DSP, EMD, EYA4, FHOD3, FKRP, FKTN, FLII, FLNC, GATA6, GATAD1, JUP, LAMP2, LDB3, LMNA, MYBPC3, MYH6, MYH7, MYLK3, MYPN, NEXN, NKX2-5, NRAP, PKP2, PLN, PRDM16, RBM20, RHBDF1, RPL3L, RYR2, SCN5A, SGCD, SLC6A6, SPEG, TBX20, TBX5, TCAP, TMEM43, TNNC1, TNNI3, TNNI3K, TNNT2, TPM1, TTN, VCL*
DILATED CARDIOMYOPATHY AND CONDUCTION DEFECTS	*ABCC9, ACTA1, ACTC1, ACTN2, ALMS1, ANKRD1, BAG3, CAVIN4, CRYAB, CSRP3, CTF1, DES, DMD, DMPK, DNAJC19, DOLK, DSC2, DSG2, DSP, EMD, EPG5, EYA4, FHL1, FHL2, FKTN, FLNC, GATAD1, GLA, HAMP, HFE, HFE2, IDH2, ILK, JUP, LAMA4, LAMP2, LDB3, LMNA, MPO, MYBPC3, MYH6, MYH7, MYL2, MYL3, MYPN, NEBL, NEXN, NKX2-5, NPPA, PDLIM3, PKP2, PLN, PPP1R13L, PRDM16, PRKAG2, PSEN1, PSEN2, RAB3GAP2, RAF1, RBM20, RYR2, SCN1B, SCN5A, SDHA, SGCB, SGCD, SLC40A1, SPEG, SYNE1, SYNE2, TAZ, TBX20, TCAP, TFR2, TMEM43, TMPO, TNNC1, TNNI3, TNNT2, TPM1, TTN, TTR, TXNRD2, VCL, XK*
HYPERTROPHIC CARDIOMYOPATHY - TEEN AND ADULT	*ACADVL, ACTA1, ACTC1, ACTN2, AGL, ALPK3, ANKRD1, ATAD3A, ATP5E, BRAF, CACNA1C, CALR3, CASQ2, CAV3, COA5, CRYAB, CSRP3, DES, FHL1, FHOD3, FLNC, FOXRED1, FXN, GAA, GLA, GLB1, GUSB, GYG1, HRAS, JPH2, KCNQ1, KLF10, LAMP2, LDB3, LMNA, LZTR1, MAP2K1, MAP2K2, MRPL3, MT-TI, MT-TL1, MYBPC3, MYH6, MYH7, MYL2, MYL3, MYLK2, MYO6, MYOM1, MYOZ2, MYPN, NEXN, NRAS, PDLIM3, PLN, PRKAG2, PTPN11, RAF1, SCO2, SHOC2, SLC25A3, SLC25A4, SOS1, TCAP, TMEM70, TNNC1, TNNI3, TNNT2, TPM1, TRIM63, TSFM, TTN, TTR, VCL*
LEFT VENTRICULAR NONCOMPACTION CARDIOMYOPATHY	*ACTC1, CASQ2, DNAJC19, DTNA, LDB3, LMNA, MIB1, MYBPC3, MYH7, MYPN, PRDM16, SDHA, TAZ, TNNI3, TNNT2, TPM1*

8.2.4 Identifying potential missed diagnoses

To identify potential missed diagnoses, a CM super panel (**Table 8.1**) was applied to participants with a cardiomyopathy phenotype that did not have a diagnosis returned by GEL. Following application of the super panel, variants were filtered for *de novo* coding variants with a maximum allele frequency (AF) of 0.001 in gnomAD v.2.1.1.[352] The *de novo* filter was then removed, and a new filter was applied, selecting all rare (AF < 0.001), predicted loss of function (pLoF) variants (essential splice site, stop gained, frameshift) in the super panel. Variants passing filtration were assessed alongside the patient's HPO terms and family structures to identify potential missed diagnoses. For patients with high priority variants that may represent diagnoses, a 'Clinician Contact Request Form' was submitted through the GEL Research Environment to enable direct contact with the patient's clinician. Potential diagnostic variants identified by users of the RE can be returned to GEL through the Diagnostic Discovery Pathway. All variants of interest were submitted to this pathway and cross-checked against the 'Diagnostic Discovery' LabKey table, which holds a record of variants discussed at the Diagnostic Discovery Committee meetings.

8.2.5 Identifying potential novel genes

To identify any potential novel gene candidates for patients with a CM phenotype, rare (AF < 0.001), *de novo* pLoF variants in novel genes i.e. not listed as having a disease gene relationship in OMIM were scrutinised by looking at gene constraint (e.g. LOEUF in gnomAD), tissue expression in GTEx, and subject to a wider literature review.

8.3 Results

Of a total of 90,259 participants in the 100KGP, 1412 (1.6%) had a cardiomyopathy-related phenotype. Of these participants, there was a median of 3.5 HPO terms recorded per individual. The majority (1052/1412 [74.5%]) of participants were proband-only, 252/1412 (17.8%) were duos, 87/1412 (6.1%) were trios, and 21/1412 (1.5%) were more complex family structures. Many more children were enrolled as trios, 48% vs 5% of adults. Sixty-nine percent of adults were proband-only.

Fifty percent of the cardiomyopathy patients were recruited under HCM followed by 30.5% under DCM, 7.2% under ARVCM, 3.9% under LVNCM, 4.7% under syndromic CM and 3.1% under other (**Table 8.2**).

TABLE 8.2 | SUMMARY OF PATIENTS WITH A CARDIOMYOPATHY PHENOTYPE AND THEIR DIAGNOSES

		PATIENTS WITH CARDIOMYOPATHY PHENOTYPE (n=1412)							
		DCM	HCM	ARVCM	LVNCM	Syndromic CM	Arrythmias	Other	
	No. of pts (% of total)	431 (30.5%)	712 **(50.4%)**	102 (**7.2%**)	55 (**3.9%**)	67 (**4.7%**)	9 (**0.6%**)	36 (**2.5%**)	
	No. pts diagnosed (**% of total diagnosed**)	57 (**13.2%**)	76 (10.7%)	3 (**2.9%**)	6 (**10.9%**)	19 (**28.3%**)	4 (**44.4%**)	5 (**13.9%**)	**In CM super panel?**
	MYBPC3	0	34	0	0	1	0	1	CM gene
GENES	*SMAD3*	0	0	0	0	0	0	1	No
	LDLR	0	2	0	0	0	0	1	No
	MYH7	3	21	0	2	2	1	1	CM gene
	PKD2	0	0	0	0	0	0	1	No
	KCNH2	0	0	0	0	0	1	0	No
	SCN5A	0	0	0	0	0	1	0	CM gene
	SCN6A	0	0	0	0	0	1	0	No
	TTN	27	2	0	0	2	0	0	CM gene
	LZTR1	0	1	0	0	2	0	0	CM gene
	DSP	7	0	0	0	1	0	0	CM gene
	FLNC	4	0	0	0	1	0	0	CM gene
	ALMS1	0	0	0	0	1	0	0	CM gene
	COL1A2	0	0	0	0	1	0	0	No
	FKRP	0	0	0	0	1	0	0	CM gene
	MFN2	0	0	0	0	1	0	0	No
	NDUFA4	0	0	0	0	1	0	0	CM gene
	NONO	0	0	0	0	1	0	0	CM gene
	POU3F3	0	0	0	0	1	0	0	No
	SGCG	0	0	0	0	1	0	0	No
	SLC20A2	0	0	0	0	1	0	0	No
	TAZ	0	0	0	0	1	0	0	CM gene
	NKX2-5	0	0	0	2	0	0	0	CM gene
	TNNT2	1	1	0	1	0	0	0	CM gene
	ACTN2	0	0	0	1	0	0	0	CM gene
	PKP2	0	1	1	0	0	0	0	CM gene
	DSG2	0	0	1	0	0	0	0	CM gene
	PLN	0	0	1	0	0	0	0	CM gene
	TNNT1	0	4	0	0	0	0	0	No
	MYL2	0	2	0	0	0	0	0	CM gene
	TPM1	0	2	0	0	0	0	0	CM gene
	DES	0	1	0	0	0	0	0	CM gene
	PRKAG2	0	1	0	0	0	0	0	CM gene
	PTPN11	0	1	0	0	0	0	0	CM gene
	RAF1	0	1	0	0	0	0	0	CM gene
	TNNC1	0	1	0	0	0	0	0	CM gene
	TNNI3	0	1	0	0	0	0	0	CM gene
	BAG3	5	0	0	0	0	0	0	CM gene
	RBM20	5	0	0	0	0	0	0	CM gene
	LMNA	3	0	0	0	0	0	0	CM gene
	PLD1	1	0	0	0	0	0	0	CM gene
	PPP1R13L	1	0	0	0	0	0	0	CM gene
	Total count	57	76	3	6	19	4	5	

Table showing a summary of 1412 participants with a cardiomyopathy phenotype, split by cardiomyopathy subtype, and the number of cases (and percentage) where a diagnosis was made. For each case with a diagnosis, the gene implicated was recorded and counted for each cardiomyopathy subtype. The final column describes whether a diagnostic gene was in the CM super panel. The rows "No. of pts", and "No. of pts diagnosed" are coloured as a heat map, whereby high values are red and lower values are green. For example, for the no. of patients, 712 (50.4%) were HCM (in red) which represented the highest count for each CM subtype. For the columns "DCM" to "other", the columns (across all gene names) are also coloured as a heat map, with high values being red and low values green. For example, for DCM, the most diagnosed gene was TTN (red, n=27).

8.3.1 Diagnostic rate

Of 1412 participants with a cardiomyopathy phenotype, 170/1412 (12%) had a confirmed diagnosis in a total of 42 unique genes. Sixteen diagnoses in 11 of these genes were not in the CM super panel as defined by PanelApp (**Table 8.1**). The diagnostic yield for DCM was 13.2%, 10.7% for HCM, 10.9% for LVNCM, 2.9% for ARVCM, 28.3% for syndromic CM, 13.9% for other, and 44.4% for participants with an arrhythmia, although this group was small (**Table 8.2**).

TTN was the most common gene affected in DCM (27/57 [47%] of diagnosed cases), followed by *DSP* (7/57 [12.3%]), *BAG3* (5/57 [8.8%]), and *RBM20* (5/57 [8.8%]) (**Table 8.2**). *MYBPC3* was the most common diagnosed gene for HCM (37/76 [48.7%]), followed by *MYH7* (21/76 [27.6%]). Variants in *PKP2, DSG2* and *PLN* were molecular diagnoses for 3 participants with ARVCM. Four genes were mutated in 6 participants diagnosed with LVNCM, *MYH7* (2/6 [33.3%]), *NKX2-5* (2/6 [33.3%]), *TNNT2* (1/6 [16.7%]), *ACTN2* (1/6 [16.7%]). In the syndromic CM cases, five diagnoses were not on the CM super panel (*DSC2, MFN2, POU3F3, SGCG, SLC20A2*), ostensibly unrelated to CM, meaning that the aetiology of the CM remains unresolved. Two participants with syndromic CM had molecular diagnoses in *DSP* and *MYBPC3,* genes that cause isolated cardiomyopathy, suggesting their CM was caused by a variant in a gene entirely unrelated to their primary syndrome. In the 'other' category, one participant with polycystic kidney disease had a variant in *MYH7* and one participant with thoracic aortic aneurysm disease had a variant in *MYCPC3*; both genes are known to cause isolated cardiomyopathies. Three participants with arrythmias had variants in *KCNH2, SCN5A, SCN6A*.

8.3.2 Missed diagnosis in undiagnosed participants with a CM phenotype

For 87/1242 (7.0%) of the undiagnosed participants recruited as trios with a CM phenotype, a *de novo* screen was performed for rare coding variants in the CM super panel. Three *de novo* variants were identified (**Table 8.3**). Following contact with the patients' clinicians, the variants in *RYR2* and *NKX2-5* have now independently been confirmed pathogenic by ACMG-AMP guidelines. The variant in *TTN* is currently being functionally validated by an independent lab.

TABLE 8.3 | POTENTIAL MISSED *DE NOVO* DIAGNOSES IN CM-RELATED GENES

Normalised specific disease	*Sex*	*Gene*	*Variant*	*Functional Class*	*ACMG*	*Max AF*	*Status*	*Assembly*
Arrhythmogenic Right Ventricular Cardiomyopathy	F	RYR2	1:237919589:A:G	missense	pathogenic	0	independent diagnosis	b37
Left Ventricular Noncompaction Cardiomyopathy	M	NKX2-5	5:173232976:G:T	missense	pathogenic	0	independent diagnosis	b38
Dilated Cardiomyopathy	F	TTN	2:178741570:C:T	missense	VUS	0	functional work ongoing	b38

The second filtering strategy involved removing the *de novo* filter and assessing all 1242 undiagnosed individuals with a CM phenotype and looking for rare pLoF variants in the CM super panel. This approach yielded a total of 196 variants. After clinical review of the variants, 36/196 (18.3%) remained as plausible candidates supported by phased data (where available) and phenotype data consistent with the gene harbouring a variant (**Supplementary Dataset SD9**). Of these 36 variants, three had already been independently confirmed by an NHS accredited diagnostic lab as causal (through personal communication with referring clinicians); three had been reported by GEL as variants of uncertain significance (as reported in the gmc exit questionnaire); and a further 30 have been submitted to Genomics England for validation through their diagnostic discovery pathway.

8.3.3 Novel gene candidates

To identify potential novel gene candidates across all cardiomyopathy phenotype subgroups, rare *de novo,* pLoF variants in novel genes were extracted. One variant in *CYFIP1* was identified (**Table 8.4**). The gene is highly constrained for loss-of-function (LOEUF = 0.3), and is ubiquitously expressed across tissues, including the heart. Upon further review of the literature, the gene was found to be associated with autism and schizophrenia.[353]

TABLE 8.4 | *DE NOVO* PLOF VARIANTS IN NOVEL GENES FOR PATIENTS WITH A CM-PHENOTYPE

Normalised specific disease	***Sex***	***Gene***	***Variant***	***Functional class***	***Max AF***	***De novo***	***Genome Assembly***
Hypertrophic cardiomyopathy	*Female*	*CYFIP1*	15:22933819:G:T	Stop gained	0.0006	Y	b38

8.3.4 Diagnosing patients *without* a CM-phenotype with variants in non-CM-related genes

There were 182 participants without a listed CM phenotype that had a pathogenic variant returned by 100KGP in the CM super panel (**Supplementary Table SD10**). Of these, 33/182 (18.1%) had a pathogenic variant in *NF1*, 27/182 (14.8%) in *PTPN11* and 10/182 (5.5%) in *ALMS1*.

8.4 Discussion

8.4.1 The spectrum of diagnoses for participants with a CM phenotype

In the 100KGP, 1412/90,259 (1.6%) of participants had “cardiomyopathy” mentioned in any one of their HPO terms. Despite 100KGP preferring trio genome sequencing, 1052/1412 (74.5%) of participants were sequenced as singletons. Only 87/1412 (6.1%) were full trios enabling *de novo* analysis, of which the vast majority were children (48% of all children with CMs were sequenced as trios). The high rate of proband-only cases is perhaps unsurprising as many cardiomyopathies present in adult life, whereby availability of parents for sequencing may be limited due to increasing age. A small number (4.7%) of participants with CM phenotypes had CM as part of another syndrome, and this was much higher for children than for adults.

In total, 170/1412 (12%) of participants with a CM phenotype had a diagnosis returned, of which 16/170 (9.4%) were in genes not on the CM super panel. Twenty-five percent of diagnoses were not on the gene panel selected by 100KGP. The diagnostic rate was highest for patients with syndromic CM (28.4%) versus 13.2% for DCM, 10.7% for HCM, and 10.9% for LVNCM and 3/102 (2.9%) for patients with ARVCM. The molecular diagnostic rate for those recruited under CM subtypes was well below reported rates in the literature[337,340,351]; ARVCM demonstrated a particularly low genetic diagnostic rate, with only 2.9% of participants diagnosed versus an expected 50%. The low diagnostic rate may, in part, be because of the strict inclusion criteria for CM patients in 100KGP. Initially, CM participants required negative conventional testing i.e. a negative gene panel comprising a distinct list of genes, although this was relaxed later in recruitment to 100KGP to boost recruitment. Therefore, the CM cohort within 100KGP is likely a skewed population, depleted of 'typical' CM. Another reason for the low diagnostic rates may be the high number of proband-only cases (74.5%), possibly due to many CMs presenting in adulthood. Furthermore, for duos (17.8%), trios (6.1%), and more complex family structures (1.5%), filtering strategies risk missing diagnoses where relatives are inaccurately coded as affected or unaffected. For example, many CMs present later in life, can be asymptomatic and may not be detected unless screened for.

Despite syndromic cases having a higher diagnostic rate (19/67 [28%]) compared to other CM-subtypes, 12 cases were proband only, 5 were duos and only 2 were trios. One explanation for the higher diagnostic rate in the syndromic CM group could be their tendency to present earlier in childhood as part of syndromes that typically affect a wide range of tissues. Syndromic disease is also more likely to be monogenic, whereby many CMs are augmented by environmental factors, show variable expressivity, and incomplete penetrance.

For those in the small arrhythmia category, 4/9 (44.4%) received a molecular diagnosis, which represented the highest diagnostic rate amongst the subgroups of CMs. Three of the four diagnoses were in genes known to cause arrythmias, *SCN5A, SCN6A* and *KCNH2*. For CM participants recruited to the 'other' category, meaning they were recruited to a category unrelated to a CM, 5/36 (13.9%) of participants received a formal genetic diagnosis. Two of these were in genes associated with isolated CM. The remaining 3 diagnoses were in *SMAD3* (Loeys-Dietz syndrome type 3), *LDLR* (familial hypercholesterolaemia) and *PKD2* (polycystic kidney disease).

8.4.2 Diagnoses for participants without a CM phenotype

One-hundred and eighty-two patients without a CM listed in their HPO terms were identified harbouring a pathogenic variant in a gene on the CM super panel. Eighteen percent (33/182) had a pathogenic variant in *NF1,* which causes neurofibromatosis type 1 and where CM is not ubiquitous in those affected; therefore, this result is not entirely surprising. Fifteen percent (27/182) had a pathogenic variant in *PTPN11*, associated with Noonan syndrome, whereby CM is common but not completely penetrant. Six percent (10/182) had a variant in *ALMS1*, known to cause Alstrom syndrome, whereby 70% of patients in the literature are reported to have CMs. Interestingly, across all GEL participants with a diagnosis in *ALMS1* returned, only 1/11 (9.1%) had CM listed as an HPO term. Several patients had pathogenic variants in *MYO6* which can cause autosomal deafness with or without CM. One patient had a variant in *TTN* and their phenotype was proximal muscle weakness, consistent with the spectrum of disorders caused by *TTN* variants. Two patients had variants in *MYBPC3*, a well-characterised cardiomyopathy gene, and both participants had just one HPO term recorded. One had "palpitations" and the other had "arrhythmia". This highlights the challenges in interpretating data using HPO terms, which are inconsistently recorded amongst clinicians. Reassuringly, however, the HPO terms recorded did suggest a potential cardiac cause. One patient had a variant in *MYH7*, another isolated cardiomyopathy gene, however they presented with a complex syndrome and the variant was returned as a partial diagnosis and is likely an incidental finding.

8.4.3 Missed diagnoses in participants with a CM phenotype

Eighty-seven individuals with a CM phenotype were recruited as trios where the benefit of detecting *de novo* variants increases sensitivity for elucidating molecular diagnoses. Three *de novo* variants in the CM super panel were identified that were not returned by the 100KGP. After contacting the referring clinicians for these cases, two variants have since been confirmed as missed diagnoses (in *RYR2* and *NKX2-5*), and one variant in *TTN* is a putative diagnosis being followed up with functional work. *RYR2* and *NKX2-5* are known CM genes, therefore it may seem surprising that *de novo* variants in CM genes were missed by GEL in individuals that had CM as an HPO term. This highlights the challenges of 100KGP using pre-selected gene panels. For 100KGP, any variants outside of the gene panel(s) applied were ignored, even if *de novo*.[354] This approach has since been changed for the new Genomic Medicine Service, whereby all coding *de novo* variants, irrespective of gene panel, are assessed. This is arguably a better approach and is likely to increase the diagnostic rate for patients with CMs, whereby CMs are phenotypically heterogenous and caused by a broad range of genes.

After removing the *de novo* filter, additional 36 pLoF variants were identified in the CM super panel found in participants with a CM phenotype. Three of these have now been confirmed as diagnostic, following contact with the participant's referring clinician, and a further 3 have been curated as variants of uncertain significance by NHS accredited laboratories. The remainder may still represent missed diagnoses and have been returned to Genomics England through their Diagnostic Discovery pathway. When looking for missed diagnoses, only pLoF variants were assessed. This means that there may be further missed diagnoses in missense and splicing variants not captured by the applied filtering approach. Predicted LoF variants were selected as these are more likely to be classified as pathogenic/likely pathogenic by ACMG-AMP guidelines. Missense and splicing variants, even when found in established genes, are likely to represent variants of uncertain significance (VUSs) without prior functional evidence. Because such a large proportion of the 100KGP cohort with a CM phenotype were proband-only cases, it was expected that many missense and splicing variants would be VUSs and hard to validate as pathogenic without additional and expensive experiments.

8.4.4 Novel *de novo* candidate – *CYFIP1*

De novo pLoF variants in novel genes, i.e. absent from OMIM, were assessed in participants with a CM phenotype. A stop gained variant in *CYFIP1* was identified*,* which upon further interrogation of the literature is a gene linked to various neurological disorders. CYFIP1 is a binding partner of FMRP, the fragile X mental retardation protein. Expression levels of *CYFIP1* in neurons, both *in vivo* and *in vitro*, influence dendritic complexity, implicating the protein in the nervous system.[353] This is further supported in work by Bozdagi *et al*.[355] showing that haploinsufficiency of Cyfi1 produces a Fragile X-like phenotype in mice. Whilst *CYFIP1* is ubiquitously expressed across tissues, there is no evidence to date implicating *CYFIP1* in cardiomyopathies, rendering it a poor candidate. Rather, the protein is more likely to be implicated disorders of the nervous system.

8.4.5 Limitations

This work is not without limitations. The CM super panel included all 'cardiomyopathy' gene panels available from PanelApp. Whilst PanelApp is a peer-reviewed and open-source repository of gene panels and forms the basis for which panels are selected by clinicians when ordering clinical genome tests in the UK, the genes included on these panels are far from exhaustive and may even be inaccurate. Some known CM genes, such as *SCN5A*, are missing from PanelApp CM panels. Within the GEL PanelApp gene lists, there is inconsistency in the inclusion of genes that cause CM as part of a syndrome; for example, *DMD* (Duchenne muscular dystrophy) and *LZTR1* (Noonan) are included as green genes on the PanelApp list (for *DMD*: '*Cardiomyopathies – including childhood onset*', '*Dilated cardiomyopathy – adult and teen'*, and '*Dilated cardiomyopathy and conduction defects';* and for *LZTR1: 'Cardiomyopathies – including childhood onset' and 'Hypertrophic cardiomyopathy – teen and adult').* Conversely, some syndromic disorders associated with CM are not represented on the CM super panel, e.g. *SCGC* (Muscular dystrophy, limb-girdle, autosomal recessive 5; MIM: 253700).

One of the major limitations of the 100KGP panel-based approach is that the choice of panel determined which variants were assessed, and the panel selection was based on HPO terms provided by the referring clinician. This highlights the importance of accurate phenotyping, but equally highlights the limitations of using HPO terms which represent a cross-sectional clinical representation of a patient at an unspecified time point in the evolution of their disease. Similarly, affection status for relatives may be unreliable but is often used as a hard filter.

Medicine is often complex, and it is perfectly plausible for a participant to have two unrelated diagnoses. For example, a patient could present with a paediatric syndrome associated with cardiomyopathy but have a CM that is entirely pathogenetically distinct to their syndrome. If a gene panel only includes a list of syndromic genes, a CM caused by another gene (causing an isolated CM) could easily be missed. This issue is somewhat mitigated by the new Genomic Medicine Service which now includes a panel agnostic filtration step whereby *de novo* variants in any gene, in addition to Exomiser top 3 hits are routinely assessed. However, this 'panel-agnostic' approach mostly detects missed *de novo* diagnoses. For CMs, which often segregate through families, gene panel selection still remains critical in ensuring that inherited variants are identified. Application of the CM super panel identified 8 (independently verified) pathogenic variants meeting ACMG criteria that were absent on the primary report returned by 100KGP for patients with a CM-phenotype. A further 33 pathogenic variants were identified for patients without a CM phenotype on the CM super panel. Whilst this demonstrates diagnostic uplift, it also risks the detection of unwanted secondary or incidental findings. Indeed, one variant on the super panel was identified in *MYH7*, a gene causing isolated CM in a patient with no recorded CM (or CM-related) phenotype. This variant was returned by GEL.

8.5 Conclusion

This chapter utilises the value of a large-scale genomics project in describing a heterogenous group of participants presenting with CM phenotypes. On average, the molecular diagnostic rate was 12%, far below expected diagnostic rates for familial CMs. Twenty-five percent of diagnoses returned to participants with a CM phenotype were not on the initial gene panel applied by Genomics England. Nine percent were not on a permissive *in silico* gene panel encompassing CM genes from PanelApp. Three variants in known CM genes detected in this study have been independently validated as missed diagnoses. A further 36 pLoF variants in CM genes were identified and may represent missed diagnoses. Eighty-eight percent (1239/1412) of participants with a CM phenotype remain undiagnosed. CMs are heterogenous disorders; gene panels based on HPO terms recorded by clinicians are ineffective at diagnosing CM cases. It is likely that more diagnoses would be captured through use of a CM-super panel, inclusive of both syndromic and non-syndromic CM. Application of panel-agnostic strategies, such as the HiPPo protocol[356] as described in **Chapter 10**, may further uplift diagnosis rates for cases where HPO terms are particularly limited, and may even identify cases with subclinical cardiomyopathies.

Chapter 9 | 'GenePy' identifies biallelic diagnoses in the 100,000 Genomes Project

9.0 Contribution statement

This chapter presents the application of GenePy, a piece of software developed at the University of Southampton by Dr Enrico Mossotto and colleagues. This software has been updated since initial publication with input from myself, Dr Guo Cheng, and Dr Imogen Stafford. Using OMIM, I curated a list of 2682 recessive disease genes for the purposes of this results chapter. In the Genomics England Research Environment, I, with Drs Gary Leggatt and Guo Cheng, applied the GenePy software to 2862 recessive genes in 78,000 individuals from the 100,000 Genomes Project. The computational jobs were split equally to maximise use of the Genomics England High Performance Cluster. This work resulted in a GenePy matrix of 2862 genes and 78,000 individuals. After this, all work is entirely my own including development and implementation of all analytical scripts, and data interpretation. This work is available as a preprint and is in press at Genetics in Medicine: *SEABY, E. G., Leggatt, G., Cheng, G., Thomas, N. S., Ashton, J. J., Stafford, I., ... & Ennis, S. (2023). A gene pathogenicity tool 'GenePy' identifies missed biallelic diagnoses in the 100,000 Genomes Project. medRxiv, 2023-03* (**Appendix Paper 13**).[357]

9.1 Introduction

The 100,000 Genomes Project was a UK government funded research project led by Genomics England (GEL) to sequence 100,000 whole genomes for families predominantly presenting with rare disease.[132] The project utilised a phenotype to genotype approach, whereby genome sequencing data were filtered using a pre-selected PanelApp[118] gene panel or panels chosen by Genomics England based on the Human Phenotype Ontology (HPO)[167] terms recorded at recruitment.[132,354] The project was completed in 2020 and yielded an overall diagnostic rate of ~25% across all rare disease categories.[132,326] However, as ever-increasing numbers of researchers gained access to anonymised whole genome sequencing data from the 100,000 Genomes Project, additional diagnoses were made using methods that extended variant analysis beyond gene panels across more coding and non-coding regions, which have subsequently been returned to participants.[354] As of 2022, 26% of all diagnoses returned by the 100,000 Genomes Project were from diagnoses not on the pre-selected gene panel applied, with many being pathogenic *de novo* coding variants.[326,354] However, assessing other variants such as biallelic variants is more burdensome, particularly without the use of gene panels due to the sheer number of variants that require scrutiny. This is because many are inherited from unaffected relatives and are carried at non-trivial allele frequencies in population databases. Furthermore, biallelic variation is often hard to interpret especially for compound heterozygotes where one variant may be pathogenic, and another may be a copy number variant, non-coding variant, or other variant of uncertain significance. This is where gene panels show their greatest utility since they can help narrow down variants to clinically relevant genes.[118] However, this approach must be balanced against the potential of missing diagnoses outside of the original gene panel applied.

This project aimed to identify potential missed biallelic diagnoses in recessive disease genes independently of the gene panel applied using a whole genome pathogenicity metric called GenePy, pronounced "Jenni-pea (d͡ʒˈɛnɪpˌiː)". GenePy (https://github.com/UoS-HGIG/GenePy-1.3) is a gene pathogenicity prioritisation tool developed at the University of Southampton that transforms the interpretation of next generation sequencing data from the variant level to the gene or pathway level.[81] GenePy incorporates allele frequency, individual zygosity (where a heterozygote scores one point and a homozygote scores two points), and a user-defined deleterious metric (such as the Combined Annotation Dependent Depletion (CADD) score[358]) into a single variant score.

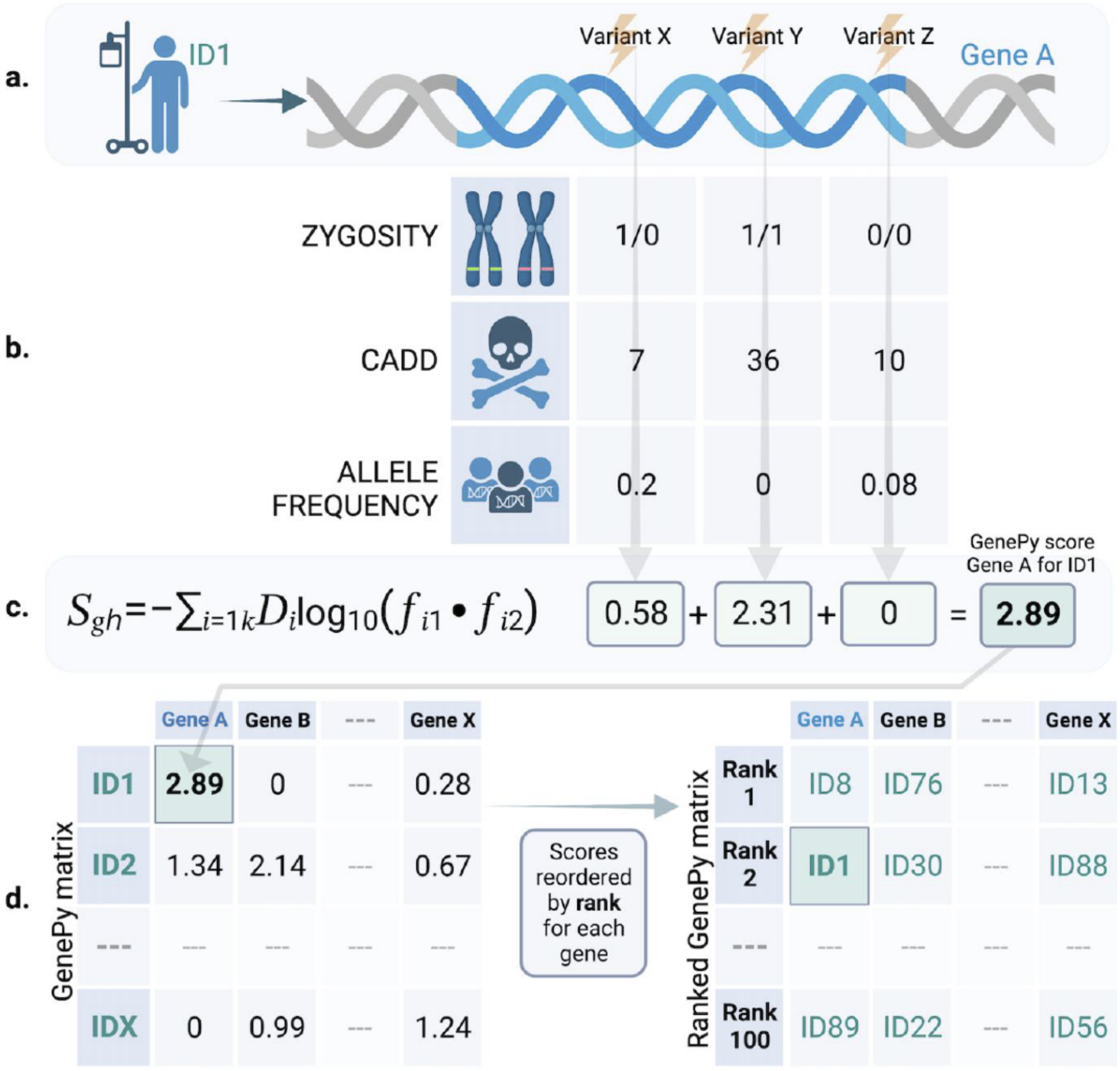
FIGURE 9.1 | GENEPY FORMULA

$$S_{gh} = -\sum_{i=1k} D_i \log_{10}(f_{i1} \bullet f_{i2})$$

[Where h = individual; g = gene; k = variants; i = locus; D_i = allele deleteriousness; f_i = allele frequency; f_{i1} = allele 1; f_{i1} = allele 2]

GenePy then aggregates variant scores across genes in an additive manner, generating a single score, per gene, per individual that is represented in a GenePy matrix table (**Figure 9.2**). However, for large genes and intronic regions there is a potential to accumulate noise from low scoring variants. To mitigate this, GenePy can be customised to filter variants with high *in silico* scores only e.g. CADD score above a particular threshold. Additionally, GenePy can be applied across any defined interval and variant scores do not have to be summed across genes, e.g. one may choose to sum variants across a particular biological pathway or genomic region.

FIGURE 9.2 | OVERVIEW OF GENEPY PATHOGENICITY SOFTWARE AND OUTPUT



	Variant X	Variant Y	Variant Z
ZYGOSITY	1/0	1/1	0/0
CADD	7	36	10
ALLELE FREQUENCY	0.2	0	0.08

$$S_{gh}=-\sum_{i=1k}D_i\log_{10}(f_{i1}\bullet f_{i2})$$

$$\boxed{0.58}+\boxed{2.31}+\boxed{0}=\boxed{2.89}$$

	Gene A	Gene B	---	Gene X
ID1	2.89	0	---	0.28
ID2	1.34	2.14	---	0.67
---	---	---	---	---
IDX	0	0.99	---	1.24

	Gene A	Gene B	---	Gene X
Rank 1	ID8	ID76	---	ID13
Rank 2	ID1	ID30	---	ID88
---	---	---	---	---
Rank 100	ID89	ID22	---	ID56

***a.** Patient's DNA undergoes sequencing and subsequent processing to produce a file listing all variants identified in their data. **b.** Each variant is individually annotated with biological information reflecting: zygosity i.e. the allele inherited from each parent; deleteriousness (D - CADD v1.6 was applied but any deleterious metric can be specified) and; frequency of the observed alleles (f) in gnomAD – one of the largest population database resources reporting the observed occurrence of alleles across very large population datasets. **c.** These data are input into the GenePy algorithm for each variant and then summed across all variants observed within that gene for that individual. This step is run in parallel for all genes across all patients within the cohort. **d.** The output is a matrix of all individuals by all genes. For certain applications, this matrix can be transposed such that for each gene, individuals are ordered by highest pathogenic variant loading.*

Upon generation of a GenePy matrix, GenePy scores can be compared across individuals in a cohort; GenePy scores are intuitive in that higher GenePy scores correlate with higher pathogenic variant burden such that individuals can be ranked for their score for any given gene, relative to all individuals with comparable input genomic data. GenePy scores are not easily compared between genes, without normalisation and adjustment for gene length. Even then, genes with alternative tolerance to dysfunctional variation are likely to exhibit very different GenePy score profiles. Instead, GenePy demonstrates the greatest utility when individual gene scores are compared across large numbers of individuals. Since GenePy is an additive score, individuals in large cohorts with the highest ranked GenePy scores will be enriched for biallelic disease. Given the potential for missed biallelic diagnoses in the 100,000 Genomes Project, GenePy was applied at scale in a panel-agnostic way to uplift diagnostic rates.

9.2 Methods

9.2.1 Access to 100,000 Genomes Project Data

Participants were recruited to the 100,000 Genomes Project with written consent. The full protocol is available here: https://doi.org/10.6084/m9.figshare.4530893.v7. Deidentified data from the project held are in the secure Genomics England Research Environment (RE). Access to the 100KGP dataset is restricted and only available as a registered GeCIP member in the Genomics England Research Environment. All data shared in this chapter were approved for export by Genomics England. The datasets and code supporting the current study, unavailable for export, are fully accessible within the Genomics England Research Environment in the shared directory: re_gecip/machine_learning/Ellie_Seaby/

Access to 100,000 Genomes Project data was obtained following governance training and through membership of the '*Quantitative Methods, Machine Learning, and Functional Genomics*' Genomics England Clinical Interpretation Partnership.

In 2022, 78,216 whole genomes (release V15) were accessed from affected and unaffected participants recruited to the 100,000 Genomes Project. Participants' affection status (i.e. whether they were coded as affected with disease or not) and any HPO terms associated with participants' records were extracted. Using the package RLabKey in R, the 'GMC Exit Questionnaire' SQL table was queried. Any likely pathogenic/pathogenic variants returned to participants by the project, present in the Exit Questionnaire, were extracted.

9.2.2 Curating a list of recessive disease genes

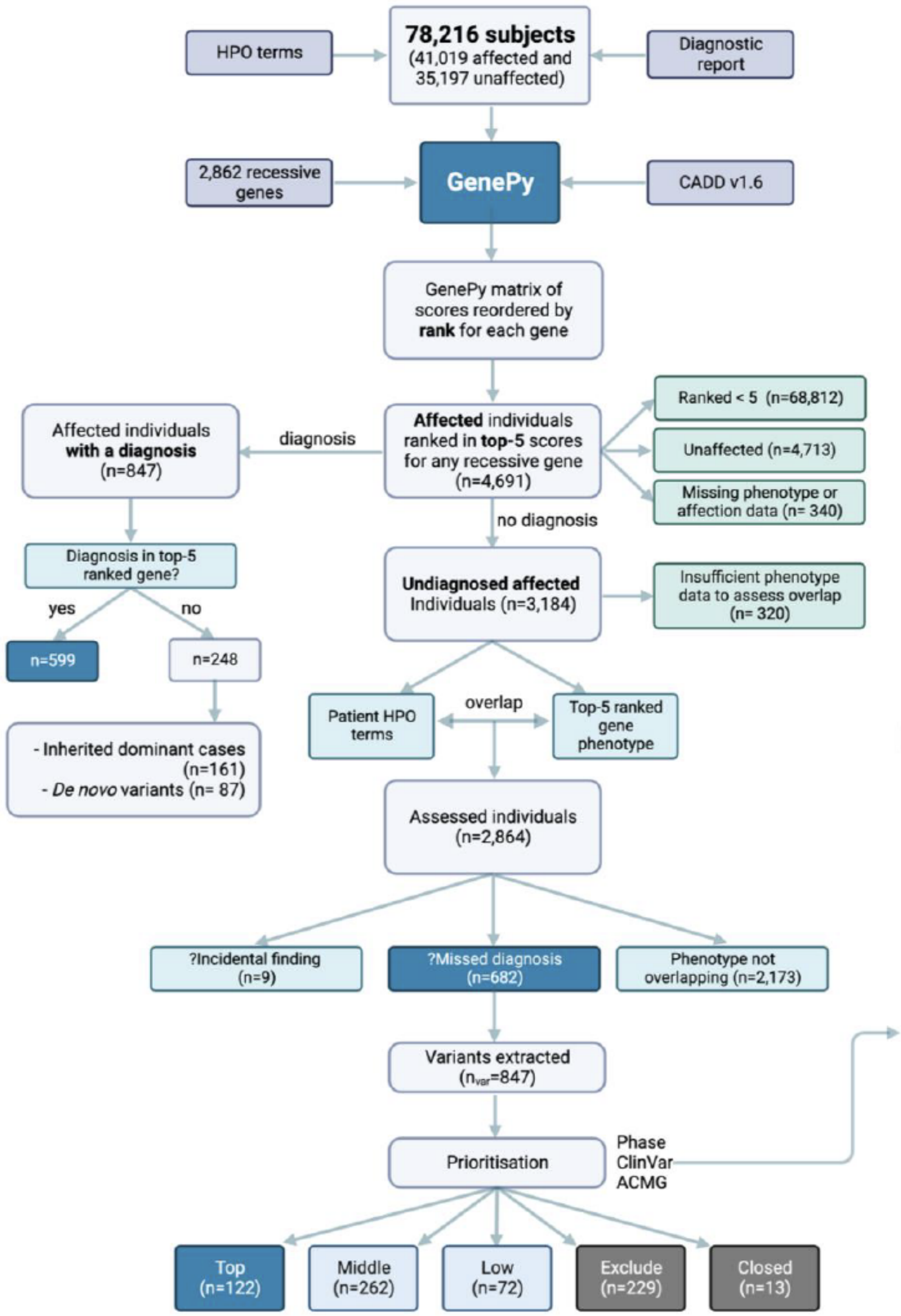
To target the method towards potential missed biallelic diagnoses, a list of 2862 recessive disease genes was curated using the Online Inheritance in Man (OMIM)[328] database (downloaded in May 2022) and cross checked with the Gene Curation Coalition (GenCC) database, whereby discrepancies in inheritance were examined more carefully.[322] A bed file of gene coordinates (in reference to GRCh38) was generated using the UCSC Genome Browser. The full gene list is available in **Supplementary Dataset SD11**.

9.2.3 Application of GenePy

Within the Genomics England RE, myself, Dr Gary Leggatt and Dr Guo Cheng, applied GenePy v.1.3 (https://github.com/UoS-HGIG/GenePy-1.3) software to 78,216 participants in the 100,000 Genomes Project using CADD[358] v1.6 as the deleterious metric and gnomAD v.2.1.1 and V3[133] databases as the allele frequency reference. CADD was specified as the input metric because it scores the greatest variety and number of variant types. Variants were selected with a minimum depth of 10, minimum genotype quality (GQ) of 20, and mean GQ > 35 using vcftools. A call-rate filter was applied, whereby each variant had to be genotyped in at least 70% of the cohort. For downstream analysis, only participant variants annotated as coding +/- 8 base pairs (on any transcript) and with a CADD score ≥15 were scored. GenePy scores were generated for 2862 recessive disease genes to create a matrix comprising GenePy scores for 2862 genes across 78,216 individuals. Of note, in addition to 'affected' participants, this cohort included many 'control' type individuals that represented unaffected parents of affected children and germline genomes of cancer patients.

All analyses from this point onwards were conducted by myself with input from my supervisory team. For each of the 2862 recessive genes, every Genomic England participant's GenePy score was ranked relative to one another e.g. the person with the highest GenePy score for *CFTR* would be ranked 1, and the person with the lowest GenePy score in *CFTR* would be ranked 78,216. After ranking, only individuals who ranked amongst the top-5 GenePy score for each gene were assessed. If two individuals had identical scores, all participants with a rank of 5 or less were included. Individuals were further removed if they were coded as unaffected or were affected individuals with insufficient phenotypic data in the form of recorded HPO terms. Affected participants were separated into those with a confirmed diagnosis returned by the 100,000 Genomes Project and those with a negative result. If the participant had a diagnosis returned, the established pathogenic variant was assessed as to whether it was in a top-5 ranked GenePy score (**Figure 9.3**).

FIGURE 9.3 | WORKFLOW OF GENEPY APPLIED TO 78,216 PARTICIPANTS IN THE 100,000 GENOMES PROJECT


HPO terms
78,216 subjects (41,019 affected and 35,197 unaffected)
Diagnostic report
2,862 recessive genes
GenePy
CADD v1.6
GenePy matrix of scores reordered by rank for each gene
Affected individuals with a diagnosis (n=847)
diagnosis
Affected individuals ranked in top-5 scores for any recessive gene (n=4,691)
Ranked < 5 (n=68,812)
Unaffected (n=4,713)
Missing phenotype or affection data (n= 340)
no diagnosis
Diagnosis in top-5 ranked gene?
yes
no
n=599
n=248
- Inherited dominant cases (n=161)
- De novo variants (n= 87)
Undiagnosed affected Individuals (n=3,184)
Insufficient phenotype data to assess overlap (n= 320)
Patient HPO terms
overlap
Top-5 ranked gene phenotype
Assessed individuals (n=2,864)
?Incidental finding (n=9)
?Missed diagnosis (n=682)
Phenotype not overlapping (n=2,173)
Variants extracted (n_{var}=847)
Prioritisation
Phase
ClinVar
ACMG
Top (n=122)
Middle (n=262)
Low (n=72)
Exclude (n=229)
Closed (n=13)

Priority	Rules for prioritisation
Top	Variants supported by phase and at least 1 variant (if compound het) is LP/P in ClinVar or by ACMG guidance.
Middle	Variants supported by phase but all VUS **OR** cases where there are >1 variants and no parental data for phasing and at least one variant is LP/P by ClinVar or ACMG guidance **OR** supported by phase and one variant is non-coding (with a CADD score >20) **OR** the variant is non-coding, homozygous and shared with an affected sibling.
Low	Variants supported by phase and at least one variant is extended splice site with a low SpliceAi score or at least one variant is non-coding (and CADD < 20) **OR** two het variants (both at least VUS) without parental samples to phase.
Exclude	Variants are *in cis* **OR** variants do not segregate across affected individuals **OR** no data available to phase and one variant is non-coding **OR** common in call-set **OR** no apparent second hit.
Closed	Participant variants are missing in the Research Environment.

GenePy scores were created for 2862 autosomal recessive genes in 78,216 participants, using CADD v.1.6 and gnomAD v.2.1.1. Participants scores were ranked across the cohort per gene, whereby those who ranked in the top 5 GenePy score for each gene were retained for downstream analysis. Unaffected individuals were removed. HPO terms from unaffected individuals without a diagnosis returned by the 100,000 Genomes Project were compared with the clinical features described for the autosomal recessive gene that the participant scored in the top 5 for. If the participant's HPO terms overlapped with the gene that the person ranked in the top 5 for, the individual participant variants were extracted and assessed phase, ClinVar status, and applied ACMG guidelines. The findings were prioritised according to the prioritisation rules, with 'Top' priority. being putative missed diagnoses, 'Middle' and 'Low' priority being of interest but lacking sufficient evidence, 'Exclude' being not diagnostic and 'Closed' being when the participants had been withdrawn from the Project.

For affected participants with a negative genome result, HPO terms from RLabkey were extracted and manually compared with the clinical features associated with the disease gene for which their GenePy score ranked in the top 5. For example, if the participant had the HPO terms ‘pancreatic insufficiency’, ‘failure to thrive’ and ‘recurrent chest infections’ and they ranked third for *CFTR*, their HPO terms would be compared with the clinical features of cystic fibrosis for phenotypic overlap. This process was completed manually using clinical acumen, phenotypic descriptions from the literature and HPO terms listed in OMIM. If the participant’s HPO terms were consistent with those for a gene that the same participant was ranked in the top 5 GenePy scores for (e.g. the participant had pancreatic insufficiency and recurrent chest infections and was ranked 3rd in *CFTR)*, this was considered a potential missed diagnosis. If the disease-gene phenotype was unrelated to the participant’s clinical phenotype but represented a gene in the American College of Human Genetics and Genomics (ACMG) 78[359] list or may represent an adult onset disease, this was considered a potential incidental finding. For these, the recruiting clinician was contacted to discuss the findings. If there was no correlation between the participant’s HPO terms and the clinical phenotype for the implicated disease gene, no phenotypic overlap was identified, and the gene for that participant was excluded from further consideration.

9.2.4 Assessing potential missed diagnoses

When the participant’s phenotype was overlapping with the disease gene for which the participant ranked in the top 5, all variants from the participant’s variant call file, with a CADD score ≥ 15, were extracted. These variants were then prioritised by likelihood of being a missed biallelic diagnosis, taking into consideration variant phase where possible, ClinVar[113] status, and variant curation by ACMG-AMP[123] guidelines (**Figure 9.3**). Variants prioritised as ‘Top’ priority were considered putative missed diagnoses and these commonly represented homozygous likely pathogenic/pathogenic variants or likely pathogenic/pathogenic compound heterozygous variants.

9.3 Results

GenePy was applied to 2862 recessive disease genes in 78,216 participants recruited to the 100,000 Genomes Project (**Figure 9.3**). For each gene, the top 5 ranked participants were arbitrarily selected by

GenePy score, which yielded a total of 9,404 unique participants, with some participants ranking top 5 for more than one recessive gene. Of the top ranked participants, 4713/9404 (50.1%) were unaffected and 4691/9404 (49.9%) were affected. Unaffected participants (rare disease or cancer germline) represented 45% of the entire cohort. Of the 4691 affected participants with a top-5 ranked GenePy score, 847/4691 (18.1%) already had a diagnosis returned by the 100,000 Genomes Project up to 2022. Of these, 599/847 (70.7%) had diagnoses in one of the top 5 ranked genes. Twenty-nine percent (248/847) of individuals had a diagnosis returned by GEL in an alternative gene and all these diagnoses were returned as complete diagnoses (i.e. they explained the entire phenotype). Of these, 87 individuals had a *de novo* pathogenic variant and 161 had a pathogenic variant in a dominant gene (either inherited from an affected individual or the participant was a from a singleton family).

In total, there were 3184 affected individuals who had a 'no diagnosis' genome report returned by the 100,000 Genomes Project who were ranked in the top 5 GenePy scores for the 2862 recessive disease genes. For these cases, the participant's reported HPO terms were compared with the clinical phenotype of the GenePy disease gene implicated in the participant. For 340 participants, there was missing phenotype data – typically this was an affected relative with no HPO terms. For 320 participants, there was insufficient HPO terms recorded to assess for phenotypic overlap between the participant's clinical phenotype and that of the implicated disease gene. These were either due to a very limited number of non-specific HPO terms or only one HPO term recorded. Therefore, these individuals were removed from downstream analysis. There were 2864 individuals who had sufficient HPO terms to assess phenotype overlap and for 682/2864 (23.8%) of these cases, the participant's HPO terms overlapped with the clinical presentation associated with the top 5 ranked GenePy disease gene. For 2173/2864 (75.9%) of cases, the phenotypes were non-overlapping and for 9/2864 (0.3%) of cases the phenotypes were not overlapping but the implicated gene was one of the ACMG 78 incidental finding genes.

For the 682 participants with a potential missed diagnosis, variants in their top 5-ranked gene (with a CADD score ≥ 15) were directly extracted from their variant call file. In total, 847 unique variants were extracted. Following prioritisation (**Figure 9.2**), 122 top priority, putative missed diagnoses were identified supported by phase, ClinVar[113] classifications and ACMG/AMP guidelines (**Supplementary Dataset SD12**).[123] A total of 262 individuals were assigned 'Middle' priority demonstrating supportive evidence for a potential missed diagnosis, whereby for many there was lack of phased data limiting ACMG diagnostic potential. Seventy-two individuals had some, but weak evidence for a potential missed diagnosis for example due to one variant being non-coding on the matched annotation from ECBI and EMBL-EBI (MANE)[330] transcript and were assigned 'Low' priority. There were 229 cases ruled as non-diagnostic, typically due to the variants being *in cis*, being non-coding on the MANE transcript, not segregating with affected and related individuals, and being common in the 100,000 Genomes call-set (**Table 9.1**). There were three cases whereby one variant was a pLoF and the second variant was non-coding on the MANE transcript (**Supplementary Dataset SD12**). Alternative transcripts were considered for these three cases, however the coding transcripts had poor overall expression in gnomAD. In 13 cases, no variants were extracted because the individual had withdrawn from the 100,000 Genomes Project.

TABLE 9.1 | FLAGS APPLIED TO DE-PRIORITISE VARIANTS

Variant priority (no. of variants)	**At least one non-coding variant**	**Common in call-set**	**Does not segregate**	***In cis***	**No second hit**
Top (122)	NA	NA	NA	NA	NA
Middle (262)	12	NA	NA	NA	NA
Low (72)	48	NA	NA	NA	NA
Exclude (229)	73	22	63	71	61

Variant pairs were deprioritised when at least one variant was non-coding on the MANE transcript, any variant was common in the 100,000 Genomes Project call-set (>5%), the variant(s) did not segregate between affected individuals from the same family, variants were in cis, or when only one heterozygous variant was identified.

9.4 Discussion

A gene pathogenicity score, GenePy, was applied to a cohort of 78,216 individuals recruited to the 100,000 Genomes Project. Utilising ranked individuals' GenePy scores for 2862 recessive disease genes, outliers with the highest GenePy scores per gene were identified. Individuals who ranked in the top 5 scores for each gene were selected, with an expectation that these individuals may harbour missed biallelic diagnoses.

Eight-hundred and forty-seven individuals with a top 5 ranked GenePy score had a diagnosis returned by the 100,000 Genomes Project. Seventy-one percent (599/847) of these individuals had a diagnosis in a top 5 ranked gene, demonstrating how GenePy was able to rapidly recover 71% of diagnoses, showing potential diagnostic utility for both known and novel disease genes. The remaining 248 cases had diagnoses in dominant genes, with 81 diagnoses being *de novo* and 161 being inherited from an affected individual or the individual represented a singleton.

In total, 2864 undiagnosed individuals were identified with top 5 ranked GenePy scores, of which 682/2864 (24%) had phenotypes overlapping with the clinical features of their top ranked recessive disease gene. Following prioritisation and removing 13 cases whereby participants had withdrawn from the 100KGP, 122/669 (18%) of the phenotype-matched cases had a putative missed diagnosis supported by phase, ClinVar classifications and ACMG-AMP guidelines. All these findings have since been returned to Genomics England through their Diagnostic Discovery Pathway. For 334/669 (50%) of individuals, variants of interest in a disease gene consistent with the participant's phenotype were identified with some supportive evidence for pathogenicity, but often phase could not be determined due to missing parental data. Additionally, for many of these cases, the variants contributing to the high GenePy scores were classified as VUSs and therefore require additional functional work-up. Whilst follow-up of these variants is outside the scope of this chapter, many of these variants, even those prioritised in the low category, may represent pathogenic variants. For example, non-coding variants were assigned to a lower priority grouping, despite them having a CADD score ≥ 15. It is hoped that many of these variants may be functionally investigated in the future as high-throughput methods to model VUSs advance.

Application of GenePy has identified putative missed diagnoses, which raises the question as to why these were not detected and returned by the 100KGP. For the 100KGP, referring clinicians recorded HPO terms, but the number recorded was very variable; some patients only had 1 or 2 non-specific terms recorded. The *in silico* gene panel selection used to analyse genomes in the 100KGP was made by GEL based on the HPO terms provided. This was a major limitation of the 100KGP; and indeed, 26% of all diagnoses made from the project were not on the original panel applied.[5] This showcases the limitations of panel-based strategies and highlights the need for panel-agnostic methods such as GenePy to recover missed diagnoses.

In total, GenePy identified potential missed diagnoses in 456/2864 (16%) of undiagnosed individuals who had a top-5 ranked GenePy score in a recessive disease gene. Forty-eight variants were previously identified as VUSs by GEL (**Supplementary Dataset SD12**). On average this resulted in the curation of 1.2 additional variants per participant. Therefore, the application of GenePy successfully uplifted diagnosis rates without adding large variant numbers requiring time-consuming manual curation for diagnostic laboratories to assess and classify.

GenePy[81] is an open-source transferrable piece of software that can be successfully applied at scale. GenePy matrices can be used as reference datasets for other cohorts applying the same GenePy methods i.e. when applying the same deleterious metric, population reference database and quality control thresholds. For example, GenePy may be applied to a cohort of 10 samples, whereby these 10 individuals' GenePy scores could be ranked against a larger GenePy matrix comprising 100,000 individuals. However, GenePy matrices for genome sequencing data should only be compared with other genome sequencing datasets, unless restricted to the same target regions of exome data.

9.4.1 Limitations and opportunities

The application of GenePy to the 100,000 Genomes Project is not without its limitations. For one, an entirely arbitrary cut-off of 5 was applied to ranked individuals. It is entirely possible that a more permissive value may capture a wider range of diagnoses; however, this must be balanced with the additional number of variants, per individual, that would require further scrutiny by clinical laboratories.

Phenotype overlap was assessed between the participants' HPO terms, and the clinical features described for the disease-gene in which the participants ranked in the top 5 GenePy scores. For 320 cases, the HPO terms were so limited (sometimes only one HPO term was recorded) that it was not possible to reliably assess overlap. This represents a real-world limitation of sequencing studies whereby there is often variability in how submitters record phenotype data and highlights the importance of accurate phenotyping. This phenotype comparison step was performed manually on 2864 cases. This large number of cases required 4 weeks of manual curation. Application of automated methods to compare participant HPO terms with disease gene phenotypes may, in the future, has the potential to increase efficiency for GenePy applied at scale. However, it is unlikely that clinical or diagnostic laboratories applying GenePy would be reviewing thousands of individuals at once, but rather on a case-by-case basis. Additionally, automated methods lack the clinical knowledge and experience of a clinician or clinical scientist that may be better able to intelligently compare groups of similar phenotypes.

In the application of GenePy, CADD v.1.6 was used to capture and model the greatest breadth of variation in an unbiased way, but it may be that incorporation of other metrics for different variant types (e.g. REVEL[47] for missense) may prove more sophisticated in an improved model. However, this is likely to require machine learning to apportion *in silico* weightings fairly for different variant types. A CADD cut off of ≥ 15 was applied to avoid individuals accruing high GenePy scores in genes of increasing length, where there was a higher chance of finding multiple ultrarare variants by pure chance that would score highly in GenePy. Whilst using a CADD score of ≥ 15 reduced significant noise and helped isolate pathogenic variants, this approach risks missing some pathogenic variants with lower CADD scores.

Fifty percent of individuals with a top 5 ranked GenePy score were unaffected. GenePy currently does not utilise phased data, meaning that some high scores may represent variants inherited *in cis*; indeed, this was observed in 71 cases (**Table 9.1**). However, there was a conscious effort not to limit GenePy to nuclear families with parental data since this does not represent real-world examples and would disadvantage non-parent/child families, where phase cannot be determined. In the future, this could perhaps be mitigated with long read sequencing data.

Whilst the application of GenePy herein focused on identification of potential missed recessive disease, there may also be opportunities to apply it in autosomal dominant diseases. When scrutinising the variants of individuals with potential missed diagnoses, 61 individuals were identified that ranked in the top 5

GenePy scores for a given gene, yet they only had one variant with a CADD score ≥ 15 in that gene. Most commonly these individuals harboured predicted loss-of-function variants which are upweighted in the GenePy statistic. Therefore, there may be utility of GenePy applied to haploinsufficient disease genes, but it is likely that a more stringent CADD cut off, such as ≥ 20, or limiting the GenePy statistic to the highest scoring variant is necessary to apportion lower GenePy scores to individuals who would otherwise accrue high scores from multiple rare, but benign variants with lower CADD scores.

GenePy also has potential to identify novel disease genes. If multiple top-ranking individuals across the same novel gene share similar clinical features, this may support the discovery of new disease genes. For novel haploinsufficient genes, unpublished data from our research group suggest that GenePy performs best when limited to high CADD scores e.g. CADD >20, whereas recessive genes may benefit from a more permissive CADD cut off.

9.5 Conclusion

The application of GenePy to ~78,000 individuals in the 100,000 Genomes Project has identified 122 putative missed biallelic diagnoses in known autosomal recessive disease genes that are being returned to participants through the Genomics England diagnostic discovery pathway. Selecting the top 5 ranked individuals for 2864 autosomal recessive genes yielded review of only 1.2 additional variants per individual, rendering GenePy a useful tool to identify biallelic variants of interest without significantly burdening diagnostic laboratories with additional variants to assess. A dilemma for many diagnostic laboratories is how to limit number of variants requiring assessment without missing diagnoses. Whilst strategies to prioritise dominant diseases are well established e.g. *de novo* analysis or Exomiser[134], there are limited tools for prioritising recessive conditions. GenePy is a useful panel-agnostic adjunct to exome and genome analysis pipelines to uplift diagnoses of recessive disease.

Chapter 10 | 'HiPPo' improves diagnostic efficiency in the Genomic Medicine Service

10.0 Contribution statement

This chapter describes and tests a new panel-agnostic strategy 'HiPPo'. I developed the concept of this method and designed and managed a project that tested it on research exomes, comparing the approach to results provided through the NHS Genomic Medicine Service. I was provided with secure NHS access to a local Southampton database of patients and families with rare genetic diseases who were interested in a research exome; this was provided by P. Costello (research nurse) and Dr D. Hunt (consultant geneticist). I approached 10 families, of which seven families agreed to join the study. I consented all seven families and obtained blood samples for those who did not already have DNA stored. N. Grahame performed DNA extraction from these whole blood samples. One family was recruited by Dr R. Gilbert. Stored DNA on families was provided by Dr S. Thomas at the Wessex Regional Genetics Laboratory. DNA was sent on dry ice to the USA, whereby whole exome sequencing was performed at the Broad Institute of MIT and Harvard and later made available for analysis. Dr S. Thomas provided patients' 'tiered' variant results that were returned to the Wessex Regional Genetics Laboratory from the NHS Genomic Medicine Service. All other work was entirely my own. This work is available as a preprint: *SEABY, E.G., Thomas, NS., Hunt, D., Baralle, D., Rehm, H.L., O'Donnell-Luria, A.L., & Ennis, S. (2023). A panel-agnostic strategy" HiPPo" improves diagnostic efficiency in the UK Genome Medicine Service. medRxiv, 2023-01* (**Appendix Paper 14**).[356]

10.1 Introduction

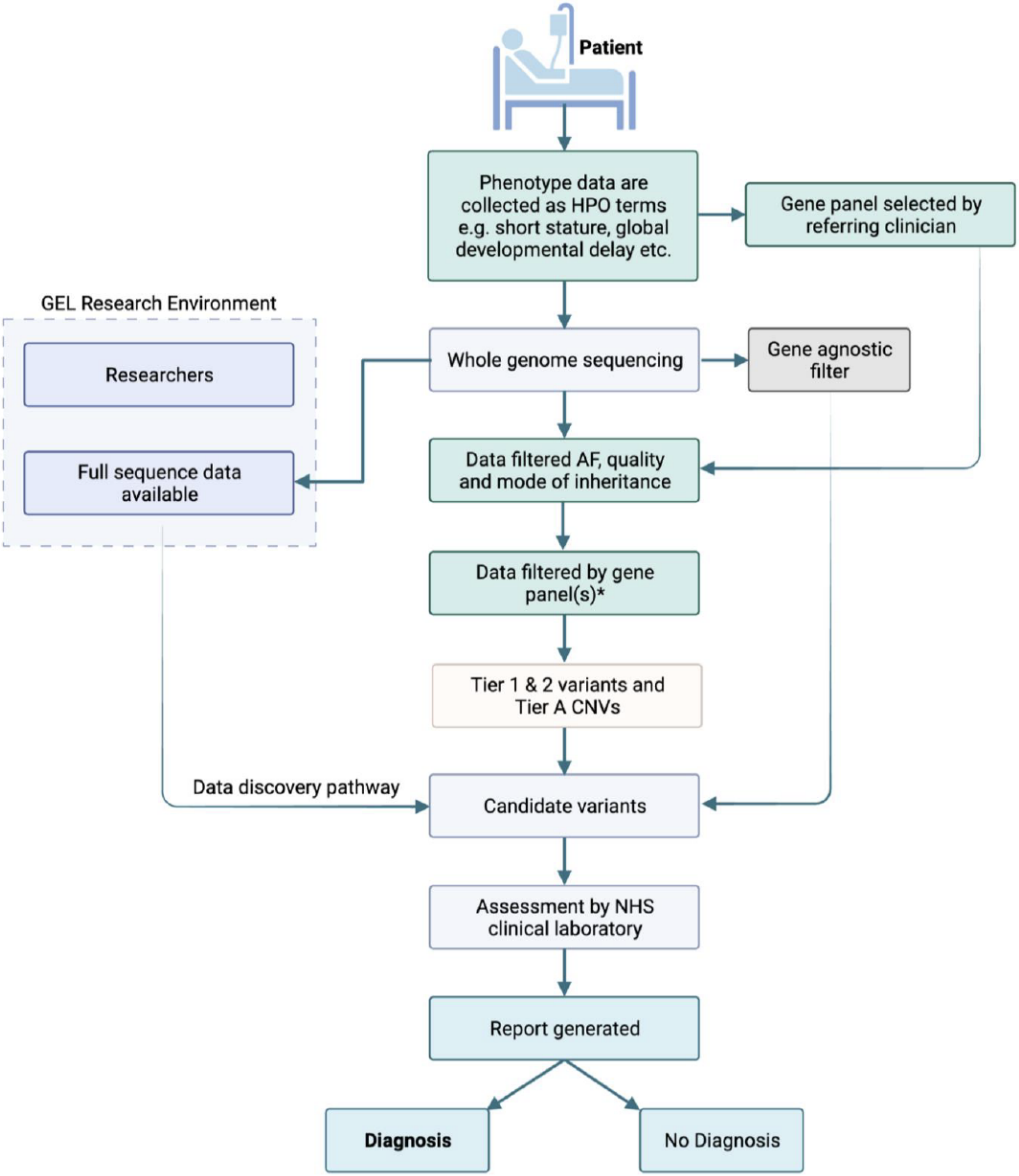
Genome sequencing is now available as a diagnostic test on the National Health Service (NHS) in the UK, offered through their Genomic Medicine Service (GMS). With the cost of genome sequencing becoming ever competitive, genome sequencing is beginning to supersede exome sequencing in some institutes, including in the NHS.[132] However, one of the challenges in diagnosing patients with rare disease is the expanded scope of analysis and need to correlate results with phenotype.[1] Genome sequencing produces

3-4 million variants per individual; therefore, strategies to reduce noise and focus on the most salient regions of DNA have been adopted, including use of virtual gene panels.[118,329] For the NHS, this is their primary analytical strategy, meaning that despite sequencing and storing an entire genome, only a fraction of the genome is actually analysed. Consequently, this risks missing pathogenic variants that would have been identified if more regions of the genome had been assessed.

All that said, there remains a trade-off between utilising the breadth of sequencing data available (such as for a genome) and the number of variants that require assessment by clinical laboratories. Filtering is necessary to reduce the number of variants identified to a manageable number that NHS laboratories can analyse, classify with respect to pathogenicity, and interpret with respect to causality of the patient's symptoms in a reasonable and acceptable timeframe.

The GMS, which primarily sequences trios, adopts a workflow similar to that used in the 100,000 Genomes Project, which predated the GMS.[3,132] First the data are filtered by inheritance pattern(s), data quality, and allele frequency. Following this, the remaining variants are filtered by a gene panel(s) selected by the clinician when the test is ordered, meaning that only coding regions are considered. Short variants and copy number variants (CNVs) overlapping the gene panel are returned for analysis ("Tiered variants"). The only variants mandated to be assessed outside of the gene panel(s) are 'gene agnostic variants' comprising *de novo* coding variants and Exomiser[134] top 3 ranked variants which are not filtered on quality (**Figure 10.1**).

FIGURE 10.1 | GENOME MEDICINE SERVICE WORKFLOW FOR GENOME SEQUENCING ON THE NHS


Patient
Phenotype data are collected as HPO terms e.g. short stature, global developmental delay etc.
Gene panel selected by referring clinician
GEL Research Environment
Researchers
Full sequence data available
Whole genome sequencing
Gene agnostic filter
Data filtered AF, quality and mode of inheritance
Data filtered by gene panel(s)*
Tier 1 & 2 variants and Tier A CNVs
Data discovery pathway
Candidate variants
Assessment by NHS clinical laboratory
Report generated
Diagnosis
No Diagnosis

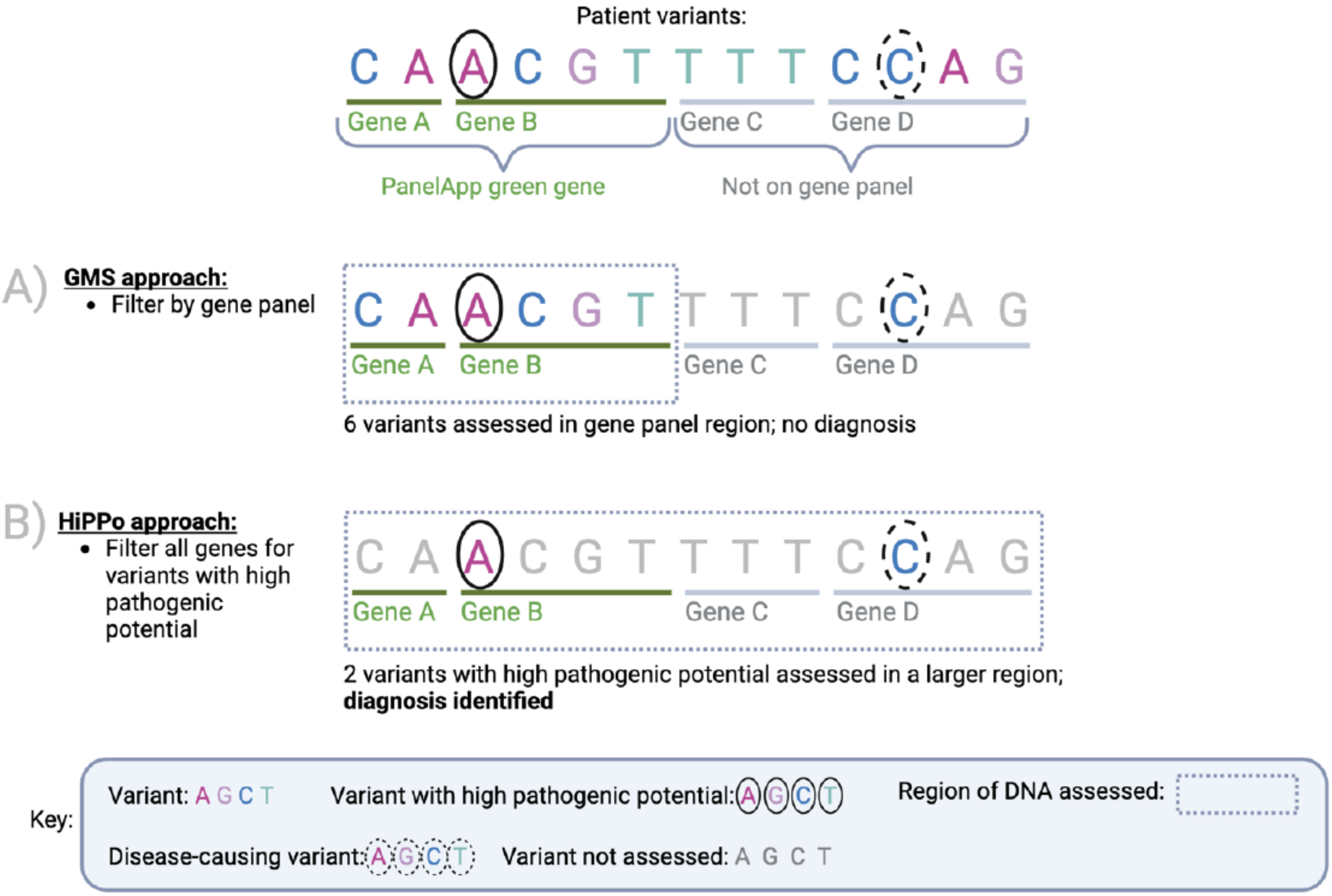
*Tier 1 variants are defined as predicted loss of function variants or de novo variants in a green gene on the gene panel(s) applied. Tier 2 variants are defined as coding variants +/- 8bp (excluding synonymous) on any transcript in the panel applied. The gene agnostic filter includes top 3 Exomiser rank variant with score of ≥0.95 and any de novo (coding) variant. Tier A is defined by a CNV (>10KB) overlapping a ClinGen curated pathogenic region relevant to a panel applied or a CNV overlapping with a green gene in the panel applied. Anonymised sequencing data are available for some patients in the Genomics England (GEL) Research Environment. *Gene panels are selected using PanelApp by the referring clinician.*

In contrast to genome sequencing, exome sequencing targets only coding regions of DNA. However, most variants filtered in the GMS strategy (Tier 1, Tier 2, and gene agnostic variants) would be captured by an exome alone. Given the known limitations of exome sequencing[360], genomes offer better coverage (even for coding regions) than exomes do and are far superior for identifying CNVs and other structural variants.[361] All that said, genome data are costly to store and process computationally and this should be considered alongside the benefits to having access to non-coding data, particularly if those data are mostly ignored.[13,362]

Panel based approaches that restrict analyses to clinically relevant genes clearly have merit, yet 26% of diagnoses made through the 100,000 Genomes Project were not on the original gene panel applied.[326] Therefore, complementary approaches that look beyond gene panels are warranted. However, this must be balanced with the potential of increasing the number of variants that require assessment by reporting laboratories. Currently the GMS assess every variant that is in a 'green' gene in the PanelApp[118] gene panel applied, regardless of *in silico* predictions. Metrics such as CADD[358], REVEL[47], and SpliceAI[64] can help reduce noise, facilitating the assessment of variants across a wider spectrum of genes without too much additional burden. This principle was exploited by adopting a panel agnostic approach that filters variants of **Hi**gh **P**athogenic **Po**tential (HiPPo) across the exome by utilising *in silico* prediction scores, allele frequency, and ClinVar[113] (**Figure 10.2**).

This study compares two different filtering approaches; one applied to exome sequencing data performed in a research setting and another approach applied to genome sequencing performed on the same patients through the GMS in a clinical setting. A gene-agnostic approach, HiPPo, is applied to exome data, whereby the resultant diagnostic yield is compared with the strategy applied by the GMS. The chapter aims to improve upon both the efficiency and diagnostic rates of current GMS standards, whilst trying to minimise the number of variants requiring assessment by clinical laboratories.

FIGURE 10.2 | A PROPOSED METHOD FOR IMPROVING DIAGNOSTIC YIELD AND EFFICIENCY


Patient variants:
C A A C G T T T T C C A G
Gene A
Gene B
Gene C
Gene D
PanelApp green gene
Not on gene panel
A)
GMS approach:
• Filter by gene panel
6 variants assessed in gene panel region; no diagnosis
B)
HiPPo approach:
• Filter all genes for variants with high pathogenic potential
2 variants with high pathogenic potential assessed in a larger region; diagnosis identified
Key:
Variant: A G C T
Variant with high pathogenic potential: A G C T
Region of DNA assessed:
Disease-causing variant: A G C T
Variant not assessed: A G C T

*Comparison of the current NHS approach versus the proposed HiPPo method for an example case (FAM 6). Patient variants are identified. In this example, there is a pathogenic variant (dashed circle) within the identified list of variants. To minimise the number of variants assessed, the NHS has adopted a method (**A**) that looks in small regions of the DNA (a panel of genes) and assesses the variants within that region. If the causal variant is in a region of the DNA not assessed, then the diagnosis is missed. The revised approach (**B**) captures a larger region of DNA (including all genes), but only looks at variants predicted to be damaging or submitted as P/LP to ClinVar (solid black circle). As a result, a larger area of DNA is assessed, whilst assessing fewer variants overall. This aims to result in a higher diagnostic rate per number of variants assessed, despite analysing a larger region of the genome than typically applied in a gene panel.*

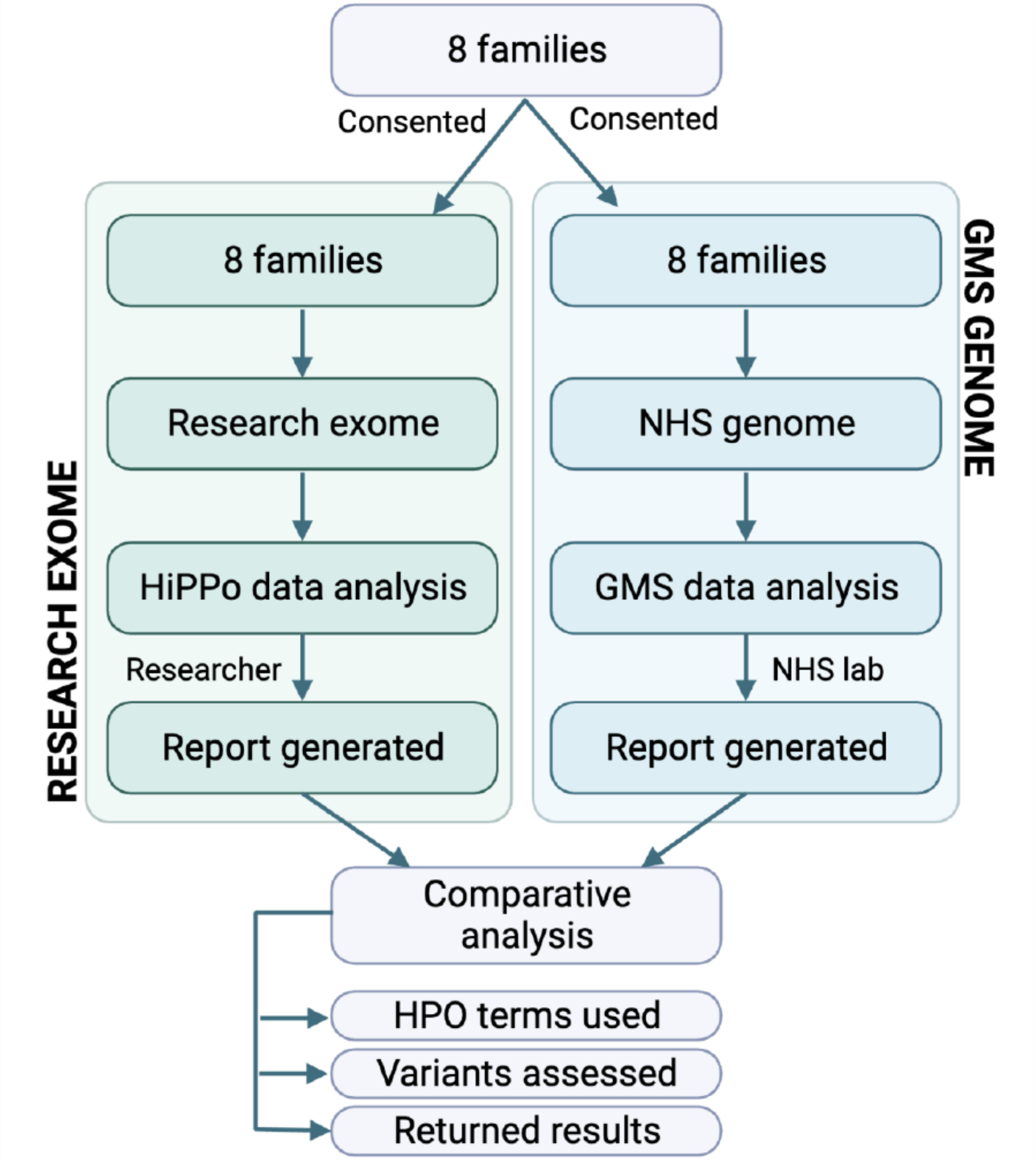
10.2 Methods

10.2.1 Recruitment and patient demographics

Clinical Geneticists at University Hospital Southampton were invited to identify and approach patients and families with suspected monogenic disease for recruitment to a research study '*Use of NGS technologies for resolving clinical phenotypes*' (IRAS: 212945; REC: 17/YH/0069). Identified individuals that showed interest were inputted into a spreadsheet stored behind the NHS firewall within University Hospital Southampton's 'CHARTS' software. These individuals gave permission to be contacted for formal recruitment. Recruited individuals were eligible for a research exome through the Centre for Mendelian Genomics[128] at the Broad Institute.

Twenty-five individuals from eight families were recruited to the research exome study. All participants were also recruited for genome sequencing on the NHS through the GMS, facilitating a parallel comparison study (**Figure 10.3**), providing an opportunity to evaluate these two sequencing and analysis strategies. All participants consented for their data to be shared.

FIGURE 10.3 | OVERVIEW OF PATIENT RECRUITMENT AND ANALYSIS



Eight families were recruited for a parallel WGS (on the NHS through the GMS) and a research exome. Different data analysis strategies were applied to the exome (HiPPo) vs genome sequencing data (adopting a panel-based strategy as outlined by the GMS). Variants reported were compared between analysis strategies including the time taken, HPO terms used, number of variants assessed and results reported.

For the research exome study, patient phenotypes were extracted from the clinical notes and recorded as Human Phenotype Ontology (HPO) terms in a manually encrypted database. The patients' clinicians also independently recorded HPO terms when requesting the GMS genome sequencing test. Both the clinician and I were blinded to each other's curated HPO terms. The family structures of the 8 families (7 trios and a quad), individual IDs, and phenotypes are described in **Table 10.1.**

10.2.2 Exome sequencing and pipeline

Quality control assessment of the DNA from the 25 samples revealed that individual (FAM_4_12), the mother in the family (FAM_4) comprising a quad of parents and monozygotic twins, had insufficient DNA quality. It was not possible to obtain a repeat sample in time for inclusion in the research exome portion of this study. However, this participant had genome sequencing through the GMS. In the GMS, quads are sequenced as two separate trios, therefore family FAM_4 was exome sequenced without maternal data (father, twin A, twin B) for the research portion of the study, but was genome sequenced through the GMS as two separate trios (mother, father, twin A) and (mother, father, twin B).

A total of 24 samples from 8 families met quality standards necessary for research exome sequencing at the Broad Institute (**Supplementary Table S10**). Libraries from DNA samples were created with an Illumina exome capture (37 Mb target) and sequenced on a NovaSeq 6000 machine using the NovaSeq XP workflow to cover >85% of targets at >20x, comparable to ~55x mean coverage. The samples underwent QC as previously described and were processed through the GATK best practices pipeline.[363] The samples were joint called with >15,000 other samples and added to seqr[364] (https://seqr.broadinstitute.org), an exome/genome analysis software hosted on the cloud platform Terra (https://app.terra.bio).

TABLE 10.1 | SAMPLES AND PHENOTYPES OF PATIENTS RECRUITED FOR A PARALLEL RESEARCH EXOME AND NHS GENOME

Samples					Clinical data				
Fam_ID	**Pro_ID**	**Pat_ID**	**Mat_ID**	**Sib_ID**	**Age**	**Sex**	**WES phenotype - HPO terms identified from clinical notes**	**WGS phenotype - HPO terms identified by clinician**	**WGS - Gene Panel applied**
FAM_1	1	2	3		0-5	M	Gastroesophageal reflux, **Myopia**, Delayed eruption of primary teeth, Triangular face, Prominent forehead, **Cow milk allergy**, **Egg allergy**, **Nut food product allergy**, **Sacral dimple**, **Clinodactyly of the 5th finger**, **Short 5th toe**, **2-3 toe syndactyly**, Mild global developmental delay, Delayed speech and language development, Oligohydramnios	Global developmental delay; Delayed speech and language development; Triangular face; Prominent forehead; Feeding difficulties; Delayed gross motor development; Oligohydramnios; Delayed eruption of primary teeth.	Intellectual disability (**R.29.4**), Congenital malformation and dysmorphism syndromes (**R27.3**), Skeletal dysplasia (**R104.3**), Likely inborn error of metabolism (**R98.2**)
FAM_1	2				41-45	M	unaffected	unaffected	
FAM_1	3				41-45	F	unaffected	unaffected	
FAM_2	4	5	6		6-10	M	**Simple ear**, **Astigmatism**, Obesity, **Patchy hypo- and hyperpigmentation**, **2-3 toe syndactyly**, **Short finger**, **Specific learning disability**, Global developmental delay, Intellectual disability, **Delayed speech and language development**	Chronic otitis media, Obesity, Severe intellectual disability, Autistic behaviour, Global developmental delay	Intellectual disability (**R.29.4**), severe early onset obesity (**R149.1**)
FAM_2	5				31-35	M	unaffected	unaffected	
FAM_2	6				31-35	F	unaffected	unaffected	
FAM_3	7	8	9		6-10	F	**Low-set ears**, **Hypermetropia**, Abnormality of the palmar creases, **Broad distal phalanges of all fingers**, **Shallow orbits**, **Cranial asymmetry**, Plagiocephaly, **Mild global developmental delay**, Intellectual disability	Thin upper lip vermillion, **Long philtrum**, **Downslanted palpebral fissures**, Deep palmar crease, Intellectual disability, Plagiocephaly	Intellectual disability (**R29.4**)
FAM_3	8				61-65	M	unaffected	unaffected	
FAM_3	9				46-50	F	unaffected	unaffected	
FAM_4	10	11	12	13	46-50	F	**Delayed ability to walk**, **Delayed speech and language development**, Spastic paraparesis, Global developmental delay	Global developmental delay, **Intellectual disability**, and Spastic paraparesis	Intellectual disability (**R29.4**)
FAM_4	11				76-80	M	unaffected	unaffected	
FAM_4	12				76-80	F	unaffected	unaffected	
FAM_4	13				46-50	F	**Delayed ability to walk**, **Delayed speech and language development**, Seizure, Spastic paraparesis, **Global developmental delay**	Developmental delay, **Intellectual disability**, Spastic paraparesis, and Seizure	Intellectual disability (**R29.4**)
FAM_5	14	15	16		0-5	F	Prominent forehead, Low hanging columella, **Prominent fingertip pads**, **Preauricular pit**, **Hypopigmented macule**, Frontal bossing, **Flat occiput**, **Joint hypermobility**, **Confluent hyperintensity of cerebral white matter on MRI**, Mild global developmental delay, **Polydipsia**	Prominent forehead, Moderate global developmental delay, **Relative macrocephaly**, **Anxiety**, Low hanging columella	Intellectual disability (**R29.4**)
FAM_5	15				26-30	M	unaffected	unaffected	
FAM_5	16				21-25	F	unaffected	unaffected	
FAM_6	17	18	19		0-5	F	Epicanthic folds, **Joint hypermobility**, Global developmental delay, **Intellectual disability**, Increased nuchal translucency	Global developmental delay, Increased prenatal nuchal translucency, **Short toenails**, Epicanthic folds	Intellectual disability (**R29.4**)
FAM_6	18				31-35	M	unaffected	unaffected	
FAM_6	19				31-35	F	unaffected	unaffected	
FAM_7	20	21	22		6-10	M	**Hypertelorism**, Bilateral polymicrogyria, Global developmental delay, Delayed speech and language development, **Delayed fine motor development**, **Delayed gross motor development**, Focal seizures, Generalised seizures, Intellectual disability	Focal seizures, Generalised seizures, **Infantile encephalopathy**, Polymicrogyria, Delayed speech and language development, Intellectual disability severe, Global developmental delay	Early onset or syndromic epilepsy (**R59.3**), Cerebral malformation (**R87.3**)
FAM_7	21				36-40	M	unaffected	unaffected	
FAM_7	22				36-40	F	unaffected	unaffected	
FAM_8	23	24	25		0-5*	F	Microphthalmia, Cataract, Retinal dystrophy, Congenital nephrotic syndrome, Microcephaly	**Intrauterine growth restriction**, Microcephaly, Congenital nephrotic syndrome, **Renal failure**, Bilateral congenital cataract, **Cerebellopontine hypoplasia**, Retinal dysfunction, **Thrombocytopaenia**, **Giant platelets**, **Howell-Jolly bodies**	Congenital malformation and dysmorphic syndromes (**R27**), Structural eye disease (**R36**), Unexplained paediatric onset end-stage renal disease (**R257**), Cerebellar anomalies (**R84**), Severe microcephaly (**R88**), Proteinuric renal disease (**R195**)
FAM_8	24				36-40	F	unaffected	unaffected	
FAM_8	25				41-45	M	unaffected	unaffected	

*Discrepancies between phenotypes underlined in bold. Ages given in age ranges. Fam_ID = Family ID, Mat_ID = Maternal ID, Pat_ID = Paternal ID, Pro_ID = Proband ID, Sib_ID = sibling ID. *Patient deceased*

10.2.3 Genome Medicine Service pipeline

Twenty-five patients in 8 families were consented for GMS clinical genome sequencing; however, as one family (FAM_4) comprised a quad, the parents were sequenced with each child as two independent trios. Sequencing was performed on an Illumina NovaSeq 6000 machine, with ≥95% of the autosomal genome covered at ≥15x calculated from reads with mapping quality >10 and >85x10^9 bases with Q≥30, after removing duplicate reads and overlapping bases after adaptor and quality trimming. Cross-sample contamination was checked using VerifyBamID and samples with >3% contamination failed QC. Sequencing alignment was performed using the DRAGEN aligner, with ALT-aware mapping and variant calling to improve specificity. Detection of small variants (single nucleotide variants (SNVs) and indels) and CNVs were performed using the DRAGEN small variant caller and DRAGEN CNV respectively. Short tandem repeat expansions were detected using ExpansionHunter (v2.5.6) as part of the DRAGEN software. The DRAGEN software v3.2.22 was used for alignment and variant calling and structural variants were detected using Manta (v1.5).

10.2.4 Data analysis

Different filtering analysis strategies were applied to the research exome and the GMS genome data (**Table 10.2**). The research exome adopted the HiPPo strategy, and the GMS adopted a panel-based approach.

Table 10.2 | Filtering criteria for the research exome and NHS genome

	Research exome HiPPo strategy		NHS genome panel-based strategy	
	Dominant	**Recessive**	**Dominant**	**Recessive**
Inheritance	*De novo*/dominant search	Recessive search	*De novo*/dominant search	Recessive search
AF (gnomAD exomes, gnomAD genomes, TOPMED*, ExAC, 1000g)	<0.001	<0.05	<0.001	<0.01
Cohort^ AF	<0.01	<0.01	No filter applied	No filter applied
Variant type	All coding +/- 20bp excluding synonymous, on any transcript	All coding +/- 20bp, excluding synonymous, on any transcript	All coding +/- 8bp on any transcript, excluding synonymous	All coding +/- 8bp on any transcript, excluding synonymous
SpliceAI (for splicing variants)	>0.2	>0.2	No filter applied	No filter applied
CADD (all variants)	>15	>15	No filter applied	No filter applied
ClinVar	Remove benign/likely benign	Remove benign/likely benign	No filter applied	No filter applied
Genes	All genes and later restricted to GenCC definitive and strong genes	All genes and later restricted to GenCC definitive and strong genes	Green in PanelApp Panel(s)	Green in PanelApp Panel(s)
Allele balance	>0.2	>0.2	N/A	N/A
Genotype Quality	>40	>40	>30	>30
QC	all variants	all variants	pass	pass
Other	Pathogenic variants in ClinVar retained even if in unaffected parents	N/A	In any gene: Exomiser top 3 rank variant (coding) with score of ≥0.95 or any *de novo* (coding)	In any gene: Exomiser top 3 rank variant (coding) with score of ≥0.95 or any *de novo* (coding)
SV/CNV	Not assessed	Not assessed	CNV (>10KB) overlaps a ClinGen curated pathogenic region relevant to a panel applied or the CNV overlaps with a green gene in the panel applied.	CNV (>10KB) overlaps a ClinGen curated pathogenic region relevant to a panel applied or the CNV overlaps with a green gene in the panel applied.

*Comparison of filtering criteria between the research exome and NHS genome. AF – maximum allele frequency, QC – quality control, N/A – not applicable. *TOPMED allele frequency was only applied to the research exome. ^Cohort AF is the frequency of any given variant as a frequency of the total number of individuals in that cohort (~6000 individuals for the research study).*

10.2.4.1 Research exome analysis

For the research exome, each family was analysed as a unit to utilise segregation data. The same *de novo*/dominant and recessive filtering strategies were applied to all families, applying a gene-agnostic filtering strategy by selecting variants with the highest pathogenic potential (HiPPo) using allele frequency, *in silico* prediction scores, and ClinVar (**Table 10.2**). The HiPPo strategy was later restricted to GenCC[322] genes with a definitive or strong disease association.

10.2.4.2 Reporting on exome variants

Variants remaining following HiPPo filtering were reviewed in seqr[364] using a wealth of in-built annotations. Variants that did not meet any of the below exclusion criteria were considered 'reportable' and returned to the referring clinician following application of ACMG-AMP guidelines[123]. Because the exome data was obtained in a research setting, novel discoveries which would not meet diagnostic criteria in a clinical setting could be considered. Any novel discoveries were discussed with the referring clinician before submission to the Matchmaker Exchange (MME).[129,157,365]

Exclusion criteria:

1. The variant is heterozygous in a known autosomal recessive disease gene and no second hit (coding or non-coding) is identified
2. The variant is found in a disease gene and is not associated with the phenotype presented by the patient, as assessed using OMIM[328], GenCC[322] and the medical literature <u>and</u> the variant is not likely pathogenic/pathogenic in ClinVar[113]
3. The variant is in a known disease gene but that gene is poorly expressed as indicated in GTEx[82] in the tissue relevant to the patient's phenotype or in an exon of the gene with poor expression as determined by the per base expression metric, pext[87]
4. The variant is in a novel gene (currently not associated with disease) and
    a. the gene is poorly expressed in the relevant disease tissue as indicated in GTEx[82] OR

   b. the gene is explicitly **not** involved in the relevant biological pathway as evidenced in Monarch[174]

5. A predicted loss-of-function (LoF) variant as called by Variant Effect Predictor[252] that is deemed 'not LoF' or 'likely not LoF' after application of LoF manual curation guidance as described in **Chapter 3**.[133]
6. The variant appears artefactual upon visualisation of the read data in Integrative Genomics Viewer (IGV).[253]

10.2.4.3 Taking novel exome candidates forward

Where the referring clinician agreed, candidate variants in novel genes were submitted to MME, sharing anonymised genotype and phenotype data. Any potential matches were discussed in detail with the patient's clinician and explicit consent was obtained (at recruitment) from the participants prior to joining case series.

10.2.4.4 GMS clinical genome analysis pipeline

Variants called through the GMS pipeline were prefiltered on mode of inheritance, quality, and allele frequency. These variants were then restricted to 'green' genes on the pre-selected PanelApp[118] gene panels for review (**Table 10.2**). A complementary gene agnostic filter was also applied to the data, which included all *de novo* variants and Exomiser[134] top 3 ranked coding variants (of any quality). Variants passing filtering were returned to the Wessex Regional Genetics Laboratory for reporting.

10.2.4.5 Reporting of genome variants

GMS variant classification was carried out according to the ACMG-AMP guidelines with ACGS[366] modifications. This included an assessment of the gene-phenotype match based on the HPO[167] terms supplied. Variants in genes with no known disease association (determined using OMIM[328], HGMDPro, ClinGen[367] and PanelApp[118]) were discounted and not assessed. Classified variants were reported according to standard ACMG/AMP guidelines: i.e pathogenic and likely pathogenic variants were always reported, variants of uncertain significance (VUSs) were only reported if there was significant evidence for pathogenicity and/or with the prior agreement of the clinician following a multidisciplinary team discussion (typically via email).

10.2.5 Comparison of two filtering approaches

The diagnostic yield and the number of variants requiring assessment after variant filtering were compared for both the research exome HiPPo approach (which included novel discoveries) and the GMS clinical genome panel-based approach, which was restricted to reporting variants that met diagnostic standards only. Specifically, the number of variants passing HiPPo filtering criteria in the research exome study were counted and compared with the number of Tier 1 and 2 variants reported in the same patients' GMS genome results, in addition to the 'gene agnostic' variants (*de novo* and Exomiser[134] Top 3 hits) as provided in an anonymised spreadsheet by the Wessex Regional Genetics Laboratory. CNVs were omitted since these were not assessed in the exome data and no diagnoses were made from structural variants in the GMS clinical genome data. Variants reported from the research exome were compared with the variants interpreted and reported by the NHS on the patient's GMS genome report. For the GMS, the reporting threshold is high with novel genes and nearly all VUSs not reported, however as part of the GMS, patients have their deidentified data deposited into a genomics library for researchers to access. Therefore, to test the efficiency of the methods applied, the diagnostic rate per number of variants assessed was calculated across the cohort.

10.3 Results

Twenty-five individuals from 8 families were consented for a GMS clinical genome on the NHS. Twenty-four individuals from the same 8 families underwent exome sequencing at the Broad Institute Centre for Mendelian Genomics. There were a total of 24 individuals in 8 families who completed parallel research exome and GMS clinical genome sequencing.

10.3.1 GMS clinical genome analysis strategy

In the 8 families who underwent GMS clinical genome sequencing, a total of 77 single nucleotide and indel variants were returned for analysis as 'Tiered variants' including the gene agnostic variants (Exomiser and coding *de novos*). A further 108 CNVs passed filtering. Five variants in total from 4 patients were included on the final reports issued by the NHS: two diagnoses, one variant of uncertain significance, and compound heterozygous variants (pathogenic and VUS); all five reported variants were also identified by HiPPo in the research exome (**Table 10.3**).

10.3.2 Research exome HiPPo strategy

HiPPo identified a total of 109 variants (**Supplementary Dataset SD13**) from 8 families (8 trios) passing filtering criteria as outlined in **Table 10.2**. However, one family, FAM_4, comprising a mother, father and monozygotic twins was sequenced as a trio (father, twin A and twin B) in the research exome study as there was insufficient maternal DNA. For the genome performed through the GMS, there was available maternal DNA and thus the twin daughters were sequenced as separate trios, with the parents sequenced twice in accordance with GMS policy. This meant more variants were identified in the research exome than the GMS genome (68 vs 11 respectively) for this family, since no maternal DNA was available for segregation analysis in the exome. Of the 109 variants identified by HiPPo across the 8 families, 38 variants were in genes reported as having definitive or strong evidence for disease association as classified by GenCC.

In addition to the 2 pathogenic variants identified by the GMS and deemed causal, HiPPo identified a further pathogenic variant in a known disease gene (*ABCC8*), representing a partial diagnosis that was filtered out by the GMS strategy due to not being on the chosen gene panel. HiPPo also identified compound heterozygous variants in a known disease gene, *INTS1* in participant FAM_2_4 which is known to cause an autosomal recessive neurodevelopmental disorder with cataracts, poor growth, and dysmorphic facies (MIM: 618571). These variants were discounted by the GMS as weak VUSs with limited evidence but remain under review by the clinical team.

HiPPo detected a further 6 VUSs in 5 novel (currently not associated with disease) genes, in addition to the same compound heterozygous variants in *SDCCAG8* and the VUS in *HMGB1* reported by the GMS (**Table 10.3**). In total, the research exome identified 109 variants using HiPPo, of which 38/109 (34.9%) were in

GenCC disease genes. After application of exclusion criteria to all HiPPo variants, independent of GenCC disease status, a total of 14 variants from the research exome were curated against ACMG/AMP criteria and returned as shown in **Table 10.4**.

TABLE 10.3 | COMPARISON OF VARIANTS REPORTED BY THE RESEARCH EXOME SEQUENCING STUDY VS THE GMS GENOME SEQUENCING

Samples		Research exome				GMS genome					
FamID	**ProID**	**Reported variants**	**Status**	**No. HiPPo variants**	**No. HiPPo variants in GenCC genes**	**Reported variants**	**No. of variants passing filtering**	***De Novo***	**Exomiser**	**Additional HiPPo Variants**	**GMS interpretation of HiPPo variants**
FAM_1	1	**VUS:** *HMGB1*: 13:30462666:CT:C; c.342del; p.Gly115GlufsTer37 (frameshift, *de novo*).	Potential new disease gene, submitted to MME. Variant also detected by NHS.	8	3	**VUS**: *HMGB1*: 13:30462666:CT:C; c.342del; p.Gly115GlufsTer37	9	*REST* *HMGB1*	*1.HMGB1* *2.KDM4* *3.ROBO1*	None	N/A
FAM_2	4	**VUS:** *INTS1*: 7:1480876:G:C; c.3908C>G; p.Thr1303Ser (missense). **VUS:** *INTS1*: 7:1497193:C:G; c.1547G>C; p.Cys516Ser (missense). (Variants *in trans)*	Phenotype partially fitting with disease gene – undergoing clinical review.	4	4	No variants reported	5	*None*	*1.KDM5A* *2.RPS3A* *3.COL16A1*	*INTS1* - VUS x 2	*INTS1* (Tier 2) discounted as weak evidence
FAM_3	7	**Pathogenic:** *PPP1CB*: 2:28776944:C:G; c.146C>G; p.Pro49Arg (missense, *de novo*).	Confirmed diagnosis (also detected by NHS).	4	2	**Pathogenic**: *PPP1CB*: 2:28776944:C:G; c.146C>G; p.Pro49Arg	7	*MYO7B* *PPP1CB* *EXOC7*	*1.PPP1CB* *2.SELENBP1* *3.EFCAB11*	None	N/A
FAM_4	10	**VUS:** *ADGRB2*: 1:31731030:G:A; c.4150C>T; p.Arg1384Ter (*de novo*, nonsense).	Potential new disease gene. Confirmed *de novo* by Sanger sequencing and in identical twin (FAM_4_13). Functional work underway.	68	23	No variants reported	14	*ADGRB2* *CRNN* *PCDHB7* *NFYB* *PIEZO1*	*1.FBXO46* *2.CEP290* *3.NFYB*	*ADGRB2* - VUS x 2	*De novo* (*ADGRB2*) variant discounted as in novel gene
FAM_4	13	**VUS**: *ADGRB2*: 1:31731030:G:A; c.4150C>T; p.Arg1384Ter (*de novo*, stop gained).	Same variant as in present in identical twin (FAM_4_10)			No variants reported					
FAM_5	14	**Pathogenic:** *ABCC8*: 11:17413408:G:A; c.2464C>T; p.Gln822Ter (nonsense, inherited from parent)	Clinically agreed as partial diagnosis.	8	1	No variants reported	5	*GOLGA8T*	*1.PTPRF* *2.NPHP4* *3.PRKDC*	*ABCC8* - **Pathogenic**	*ABCC8* not analysed as untiered and gene absent from R29 panel
FAM_6	17	**Pathogenic:** *CHAMP1*: 13:114325034:C:T; c.1192C>T; p.Arg398Ter (*de novo*, nonsense).	Confirmed diagnosis (also detected by NHS).	2	1	**Pathogenic**: *CHAMP1*: 13:114325034:C:T; c.1192C>T; p.Arg398Ter	6	*KRTAP5-5*	*1.CHAMP1* *2.MDK* *3.CRAC2RA*	None	N/A
FAM_7	20	**VUS:** *FOXB2*: 9:77020700:A:G; c.1046A>G; p.Lys349Arg (missense, *de novo*). **VUS:** PKD1L3: 16:71951734:T:G; c.3020A>C; p.Glu1007Ala (missense).	Both FOXB2 and PFK1L3 are potential novel disease genes and have been submitted to MME.	10	1	No variants reported	7	*FOXB2* *RP1L1*	*1.IGFN1* *2.ZXDA* *3.CADNA1F*	*FOXB2* - VUS *PKD1L3* - VUS x 2	*FOXB2 de novo* variant - Discounted *PKD1L3* - Not analysed - Tier

		VUS: *PKD1L3*: 16:71951734:T:G; c.3020A>C; p.Glu1007Ala (missense). *PFK1L3* variants are *in trans.*									3; Exomiser rank 33
FAM_8	23	**VUS:** *ZNF91*: 19:23361341:G:C; c.1638C>G; p.Tyr546Ter (*de novo*, nonsense). **VUS:** *SDCCAG8*: 14:92449109:A>C, c.1552A>G, p.Arg518Gly (missense). **Pathogenic:** *SDCCAG8*: 1:243341070:TG>T, c.1255del, p.Glu419ArgfsTer43 (frameshift).	*ZNF91* is a novel disease gene. A group is working on this gene and we have joined their case series. The *SDCCAG8* variants are *in trans* but are not felt to fit with the clinical phenotype.	5	3	**VUS:** *SDCCAG8*: 14:92449109:A>C, c.1552A>G, p.Arg518Gly (missense). **Pathogenic:** *SDCCAG8*: 1:243341070:TG>T, c.1255del, p.Glu419ArgfsTer43 (frameshift).	24	*ZNF91* *ZNF91*	1. *RIN3* 2. *ERAP2* 3. *ZNF91*	None	*ZNF91* variant discounted as no established disease association

FamID – Family ID; MME – matchmaker exchange; NA – not applicable; ProID – Proband ID; VUS – variant of uncertain significance

Table 10.4 | Details of 14 variants reported by the research exome study meeting prioritisation criteria

Variant	Gene	Consequence	gnomAD	cadd	revel	hgvsc	hgvsp	ClinVar	ACMG	FamID	ProbandID	P_AC	Sample_2	S2_AC	Sample_3	S3_AC	Returned by GMS?
13:30462666:CT:C	*HMGB1*	frameshift	0	38		ENST00000341423.9:c.342del	p.Gly115GlufsTer37		VUS	FAM_1	1	1	2	0	3	0	Yes
7:1480876:G:C	*INTS1*	missense	5.56^{-4}	23.5	0.243	ENST00000404767.7:c.3908C>G	p.Thr1303Ser		VUS	FAM_2	4	1	5	1	6	0	No
7:1497193:C:G	*INTS1*	missense	7.76^{-5}	24	0.315	ENST00000404767.7:c.1547G>C	p.Cys516Ser		VUS	FAM_2	4	1	5	0	6	1	No
2:28776944:C:G	***PPP1CB***	**missense**	**0**	**26.7**	**0.438**	**ENST00000395366.2:c.146C>G**	**p.Pro49Arg**	**P**	**P**	**FAM_3**	**7**	**1**	**8**	**0**	**9**	**0**	**Yes**
1:31731030:G:A	*ADGRB2*	stop_gained	0	38		ENST00000373655.6:c.4150C>T	p.Arg1384Ter		VUS	FAM_4	13	1	11	0	13	1	No
1:31731030:G:A	*ADGRB2*	stop_gained	0	38		ENST00000373655.6:c.4150C>T	p.Arg1384Ter		VUS	FAM_4	10	1	11	0	13	1	No
13:114325034:C:T	***CHAMP1***	**stop_gained**	**0**	**35**		**ENST00000643483.1:c.1192C>T**	**p.Arg398Ter**	**P**	**P**	**FAM_6**	**17**	**1**	**18**	**0**	**19**	**0**	**Yes**
11:17413408:G:A	***ABCC8***	**stop_gained**	**0**	**43**		**ENST00000302539.9:c.2464C>T**	**p.Gln822Ter**		**LP**	**FAM_5**	**14**	**1**	**15**	**0**	**19**	**1**	**No**
16:71951734:T:G	*PKD1L3*	missense	5.09^{-4}	23.1		ENST00000620267.1:c.3020A>C	p.Glu1007Ala		VUS	FAM_7	20	1	21	1	22	0	No
16:71973386:C:T	*PKD1L3*	missense	1.02^{-4}	22		ENST00000620267.1:c.1891G>A	p.Ala631Thr		VUS	FAM_7	20	1	21	0	22	1	No
9:77020700:A:G	FOXB2	missense	0	25.3	0.534	ENST00000376708.1:c.1046A>G	p.Lys349Arg		VUS	FAM_7	20	1	21	0	22	0	No
19:23361341:G:C	*ZNF91*	stop_gained	0	32		ENST00000300619.11:c.1638C>G	p.Tyr546Ter		VUS	FAM_8	23	1	24	0	25	0	No
1:243341070:TG:T	*SDCCAG8*	frameshift	0	26		ENST00000366541.7:c.1255del	p.Glu419ArgfsTer43		P	FAM_8	23	1	24	1	25	0	Yes
1:243378799:A:G	*SDCCAG8*	missense	9.55^{-5}	22.2	0.195	ENST00000366541.7:c.1552A>G	p.Arg518Gly	VUS	VUS	FAM_8	23	1	24	0	25	1	Yes

Families separated by colour. Bold indicates confirmed diagnoses. FamID – Family ID; hgvsc – HGVS coding consequence; hgvsp – HGVS protein consequence; LP – likely pathogenic; P – pathogenic; P_AC – proband allele count; S2_AC – sample2 allele count; S3_AC – sample3 allele count; VUS – variant of uncertain significance. Sample_2 and sample_3 refers to parental DNA.

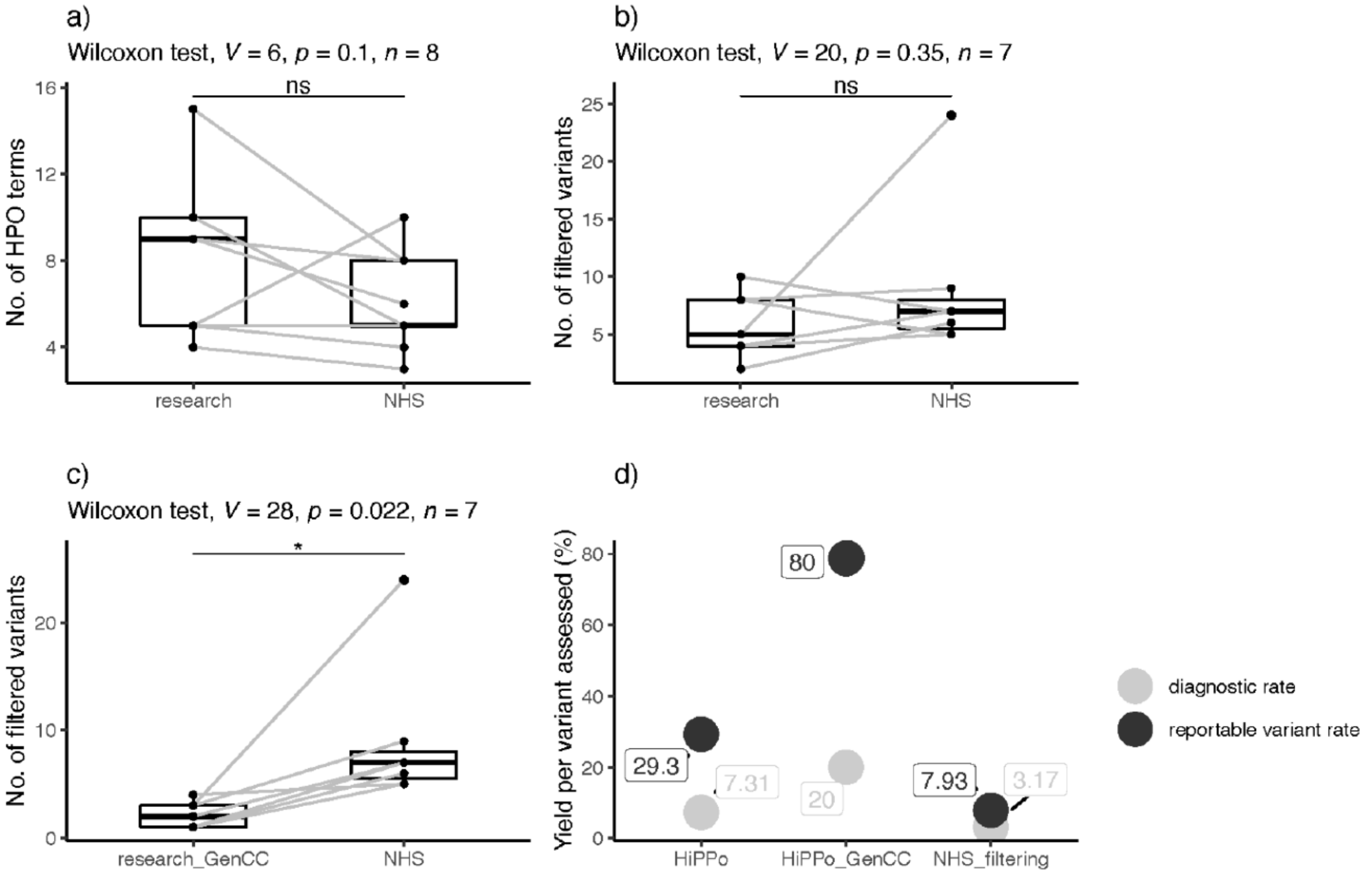
10.3.3 Comparison between exome study and GMS clinical genome results

On average, more HPO terms were recorded in the research exome study compared to the GMS genome (**Table 10.1** and **Figure 10.4**) although this was not statistically significant (p-value = 0.1, Wilcox signed rank test).

When comparing the 8 families who underwent parallel research exome and GMS clinical genome sequencing, one family (FAM_4) was removed from analysis as the mother was not sequenced in the research study but was sequenced by the GMS. There was no statistical difference between the number of variants (excluding CNVs) assessed by the GMS panel-based strategy and the HiPPo method (p-value = 0.35, Wilcoxon signed rank test), although HiPPo identified more reportable variants (**Table 10.4**), of which 6 variants in 5 unique genes have been taken forward as candidates to MME. Three of these variants were identified but discounted by the GMS as disease gene discovery is outside of the remit of clinical reporting. However, when restricting the HiPPo analysis to GenCC strong and definitive genes, there was a statistical difference between groups (p-value = 0.022, Wilcoxon signed rank test), with the research study assessing fewer variants overall (**Figure 10.4**) yet still identifying an additional pathogenic variant in *ABCC8* that did not pass filtering thresholds by the GMS.

The efficiency of the relative analytical methods varied between the groups. The diagnostic rate per number of variants assessed was higher for the HiPPo approach applied to the research exome (3/41 [7.31%]) compared with the panel-based GMS strategy (2/63 [3.17%]), although the result was not statistically significant (p value = 0.39, Fisher's exact test). When limiting HiPPo analysis to GenCC disease genes, the diagnostic rate per variant assessed improved further to 3/15 (20%) (**Figure 10.4**), but again the results did not reach statistical significance (p value = 0.06). The reportable variant rate per number of variants assessed was higher for the HiPPo approach when limited to GenCC disease genes (12/15 [80.0%]) compared with the panel-based GMS strategy (5/63 [7.93%]).

FIGURE 10.4 | COMPARISON OF RESULTS BETWEEN THE RESEARCH EXOME AND CLINICAL GENOME (NHS) SEQUENCING



(a) Number of HPO terms recorded between the exome and genome studies. (b) Number of variants assessed by the NHS reporting laboratory following GMS genome sequencing versus number of variants passing HiPPo filtering (in any gene) in the exome study. (c) Number of variants assessed by the NHS reporting laboratory following GMS genome sequencing versus the number of filtered HiPPo variants in GenCC disease genes assessed by the exome study. (d) Plot showing the diagnostic rate per variant assessed and the reported variant rate per variant assessed for the HiPPo research approach, HiPPo restricted to GenCC disease genes approach, and the GMS panel-based filtering strategy.

10.4 Discussion

Genome sequencing is available as a clinical test on the NHS through the GMS. Following sequencing, data are filtered by a pre-selected gene panel chosen by the referring clinician, in addition to CNVs overlapping the panel applied, *de novo* variants, and Exomiser[134] top 3 ranked variants. This predominantly 'panel-based' approach attempts to minimise noise and efficiently identify pathogenic variants in disease-relevant genes.

However, panel-based strategies are not without limitations. PanelApp[118] is open-source but gene reviews and updates of the approved gene content relies on volunteer efforts and comes with a significant lag time. Panels represent a snapshot in time and their application is contingent on clinicians selecting the optimal gene panel(s) with variable levels of genetics training. This is particularly problematic for clinicians in non-genetics specialties lacking adequate familiarity with gene panel selection. If the "wrong" panel is chosen, pathogenic variants can easily be missed. With only 20% of rare disease patients receiving a diagnosis through the 100,000 Genomes Project[132] (the precursor to the UK's GMS), there is clear need to investigate variants beyond a limited gene list but without significantly increasing the number of variants for review.

This study compares the GMS' data analysis filtering strategy using genome sequencing to a gene-agnostic HiPPo approach targeting variants with high pathogenic potential as applied to exome sequencing in a research setting. Twenty-four individuals from 8 families underwent parallel clinical genome and research exome sequencing, providing an opportunity to compare different filtering approaches. With many factors influencing differences between the research and NHS studies, such as different technical pipelines, capture, and timescales, the fairest comparison of efficiency of the two approaches was the number of variants that required review following filtering and the corresponding diagnostic rates. On average the research exome study reviewed fewer variants than the GMS yet identified more diagnostic variants, although this was not statistically significant (p-value = 0.35). The number of reportable variants per variant assessed was higher for HiPPo (29.3%) versus the GMS (7.9%), however the threshold for what constituted a reportable variant differed between the research exome and the GMS genome strategies. The research exome reported variants that would not be reportable in the current NHS setting, notably variants in novel disease genes and variants of uncertain significance, although it is worth noting that some international diagnostic labs do report variants in novel genes. However, when restricting the exome HiPPo filtering approach to GenCC disease genes (genes strongly associated with disease that would be reportable in the NHS setting), statistically fewer variants required assessment when compared with the GMS' panel-based approach (Wilcoxon signed rank p-value = 0.022). Despite this, more pathogenic variants were identified; including a pathogenic variant in *ABCC8* representing a partial diagnosis which was missed by the GMS as it was not on the selected gene panel. For the 8 families undergoing parallel exome and genome sequencing, the GenCC disease gene HiPPo analysis strategy identified 15 variants that required further assessment, compared with 41 variants for the GMS approach. Overall, the diagnostic rate per number of variants assessed between the GenCC disease gene HiPPo analysis and the GMS' panel-based approach was 3/15 [20%] vs 2/63 [3%] respectively. Although the sample size is modest, there is a strong argument that genotype-to-phenotype methods, focused on variants with high pathogenic potential in known disease genes could prove more effective and less resource-intensive than panel-based approaches, despite covering a wider range of the genome. Indeed, in the GMS very few Tier 2 variants are reported, meaning that Tier 1 + HiPPo may prove an efficient alternative strategy and could also be used to prioritise the interpretation of gene agnostic variants and/or determine which should be reported and/or taken to multidisciplinary team meetings. There is also a further argument that genome sequencing is not being optimally utilised by the NHS due to resource limitations and that exome sequencing may prove similarly effective; however, this comparison is beyond the scope of this limited study, whereby no pathogenic CNVs were identified, and a time-cost-analysis could not be fairly undertaken.

The number of HPO terms did not vary significantly between those selected for the research study versus those submitted by clinicians working in the NHS (p-value = 0.1) (**Table 10.1**). A recent study by Kingsmore *et al.*[368] showed that more HPO terms may not increase diagnostic yield, but that a more focused list of key terms may support analysis more effectively.

In total, HiPPo identified 3 diagnoses (compared with 2 diagnoses by the GMS) and a further 10 unique variants of interest in 7 unique genes, of which 3 genes were discounted by the GMS (**Table 10.4**) as they did not meet the threshold for clinical reporting. In FAM_2_4, HiPPo identified compound heterozygous variants in *INTS1* (7:1480876G:C and 7:1497193:C:G), a disease gene associated with an autosomal recessive disorder (MIM: 618571) presenting with cataracts, poor growth, developmental delay, and dysmorphic facies. Whilst FAM_2_4 shares some features with the *INTS1* related syndrome, he does not have cataracts and is large (with his weight tracking along the 99th percentile) opposed to being small. These variants are being reviewed by his clinical team.

In FAM_8_23, both HiPPo and the GMS identified compound heterozygous variants (one pathogenic and one VUS) in *SDCCAG8* (1:243341070:TG:T and 1:243378799:A:G), a disease gene associated with an autosomal recessive retinal-renal ciliopathy (MIM: 615993 and MIM: 613615). These variants have been discussed at length with the clinical team and are not felt to explain the nephrotic phenotype. On renal biopsy, the patient had immature glomerular development diffuse foot process effacement on electron microscopy which is inconsistent with a retinal-renal ciliopathy. Furthermore, there were additional inconsistent features including microcephaly, cerebellopontine hypoplasia and functional asplenia. In the same individual, a *de novo* variant in *ZNF91* was identified. Through MME, a collaboration has been established with a group performing functional studies on this gene, whereby they also have a patient with microcephaly and nephrotic syndrome.

In total, 6 variants in 5 novel genes were submitted to MME from the exome study, which is beyond the remit of the NHS diagnostic capacity. No matches were established for *PKD1L3* and *FOXB2.* In addition to *ZNF91* (as described above in FAM_9_26), a matches were made with collaborators working on *HMGB1* (*de novo* variant found in FAM_1_1) and *ADGRB2* (*de novo* variant found in monozygotic twin sisters FAM_4_13 and FAM_4_10). In 2021, a paper was published on *HMGB1* predicted loss-of-function variants in 6 patients.[369] Common features included developmental delay, language delay, microcephaly, obesity and dysmorphic features, some of which overlap with FAM_1_1. This variant has been returned to the patient's clinician, whereby they are directly collaborating with the authors of the 2021 paper. Whilst the *HMGB1* variant was also reported as a VUS through the GMS, there is no time provision for clinicians to consider and follow up any unreported novel candidates. Furthermore, most *de novo* candidates in novel genes are disregarded by the GMS and so are seldom investigated further. That said, anonymised patient data are eventually deposited in the Genomics England Research Environment, meaning that novel variants may be identified and later investigated through research.

In 2017, a paper was published in Human Mutation describing a missense variant in *ADGRB2* in a patient presenting with developmental delay and progressive spastic paraparesis; features shared with identical twins FAM_4_13 and FAM_4_10 harbouring a *de novo* pLoF in the same gene.[370] The authors showed that their specific variant demonstrated gain of function. The authors of the original paper have been contacted and are now modelling FAM_4's variant *in vitro* and *in vivo.*

10.4.1 Limitations

This study is small, representing 25 individuals from 8 families, of which 24 participants received parallel exome and genome sequencing which compared two different filtering strategies. Inevitably a larger study is needed to test the value of gene-agnostic approaches utilising pathogenicity scores compared with gene panel approaches. This is not easily feasible within the NHS, as it is not possible to access an individual patient's sequencing data through the GMS to test alternative strategies that would enable me to access HPO terms, variants returned to the diagnostic labs and the diagnostic clinical reports. Therefore, the only way to compare methods was in a study that independently sequenced the same patients. Whilst this study compares an exome with a genome, it should be noted that the QC was excellent, and that the panel-based strategy of the GMS essentially limits data analysis to exonic regions.

Data analysed in a research setting is not comparable with data analysed for diagnostic purposes as the threshold for variant follow-up and investigation differs in a clinical setting, with inconsistency in reporting on novel discoveries. It is important not to conflate a potential novel discovery identified in the research setting with a 'missed' diagnosis. That said, the research exome picked up a pathogenic variant missed by the GMS due to being outside the applied gene panels.

This study was further biased by predominantly neurodevelopmental phenotypes, due to the nature of intellectual disability (ID) being one of the most common reasons for GMS referral. The biggest bias with ID is enrichment for *de novo* variants[371], although both the GMS and HiPPo filtering methods assessed for these, therefore it is not expected that the performance would drastically change between approaches.

No pathogenic variants were identified by GMS clinical genome sequencing that were not captured by the research exome, although a larger sample size is needed to test the diagnostic uplift gained from structural variants detected using genome sequencing versus potential missed diagnoses from using panel-based approaches.

10.5 Conclusion

This study compared a gene agnostic filtering strategy called HiPPo as applied to research exome data with a gene panel-based analysis strategy applied to genome sequencing data. Despite HiPPo being pan-exomic, a similar number of variants were assessed per patient to the panel-based strategy of the GMS and more variants of interest were identified; this includes a pathogenic variant in *ACDCC8* and *de novo* variants in 3 novel genes, whereby case series and functional experiments are underway, which highlights the added potential that research studies can offer. When restricting HiPPo to GenCC disease genes, statistically fewer variants required assessment to identify the same diagnoses as identified by the GMS (20% vs 3% respectively), representing a greater diagnostic yield per variant assessed. This preliminary work suggests that panel-based approaches are limited and that they could be improved by incorporating specific variant prioritisation metrics. Further testing is required to integrate these complementary approaches to optimise the analytical strategy for genome sequencing within the NHS.

Chapter 11 | Summary of key findings

11.0 Introduction

This chapter provides a summary of the key findings of this doctoral thesis. The following chapter (**Chapter 12**) addresses future directions.

11.1 Loss of function constraint

It is well established that some genes are more biologically essential than others; in other words, some genes are redundant, and humans can tolerate biallelic knockout, whereas other genes are embryonic lethal with just one affected allele i.e. the gene causes lethality though haploinsufficiency. Whilst some genes lie at the extremes of constraint for- and against- loss-of-function, the majority reside along a spectrum of *some* intolerance to inactivation.[26] Unsurprisingly, autosomal recessive disease genes occupy the middle of this spectrum.[26,305] Being able to differentiate between essential and redundant genes is critical in elucidating which genes may be implicated in disease, and which genes may be used as safe drug targets.

The LOEUF method (**methods 2.6.1.2**), developed during time spent at the Broad Institute of MIT and Harvard, models exome and genome data from 141,000 individuals in the population reference database, gnomAD.[26] LOEUF is a constraint score, that in simple terms, compares the expected number of pLoF variants in a given gene versus the number of pLoF variants observed.[133] Genes with the lowest LOEUF scores are associated with haploinsufficient disease genes and genes with the highest LOEUF scores are highly mutable, e.g. olfactory receptor genes.[305] Therefore, it is reasonable to assume that pLoF variants identified in low scoring LOEUF genes may cause disease, and that pLoF variants in high scoring LOEUF genes would have no phenotypic consequence. However, variant calling software applied to sequencing data cannot reliably discriminate a true LoF variant from an erroneous LoF variant. Whilst automated methods, such as LOFTEE (see **methods 2.6.1.3**), exist to identify some erroneous pLoF calls, the current best method is curation.[284] This involves manually assessing pLoF variants for errors including annotation, transcript, and rescue errors.[327]

Chapter 3 describes a set of manual curation rules that were developed to systematically classify the likelihood that a predicted loss-of-function variant is truly loss-of-function. Data were accessed from 141,000 adults in gnomAD, a dataset depleted for monogenic paediatric disease.[25] Predicted LoF variants in 61 highly penetrant genes known to cause monogenic paediatric-onset disease were identified.[214] The theoretical expectation was that no individuals in gnomAD would harbour pLoF variants in this gene-set. However, many variants were identified, but unsurprisingly, the majority (76%) were not true LoF variants following manual curation. The most common reason to exclude a variant was a transcript error, whereby the pLoF variant was on a poorly expressed transcript, likely of little biological relevance.[214] These data highlight how pLoF variants must be interpreted cautiously in a clinical setting. This has led to additional work, in collaboration with the Broad Institute, that advances interpretation of pLoF variants and provides recommendations on adjustments to ACMG-AMP guidelines, now available as a publication in the *American Journal of Human Genetics* (**Appendix Paper 6**).[284]

11.2 Application of LOEUF to identify novel disease genes

Being able to differentiate genes tolerant to LoF and genes intolerant to LoF has significant clinical implications. Genes most constrained for LoF, as quantified by a very low (<0.2) LOEUF score are highly enriched for haploinsufficient disease genes ($p < 0.00001$). Yet, ~65% of these genes are yet to have a disease-gene relationship.[2] Indeed, the function of most genes remains unknown, suggesting that there are many disease genes yet to be discovered. This is perhaps further supported by suboptimal diagnosis rates for putative monogenic diseases; currently, approximately 70% of rare disease patients do not receive a molecular diagnosis. And whilst the non-coding genome may resolve a proportion of these cases, evidence suggests there is still much to be gleaned from the exome.

In **Chapter 4**, a list of genes with a LOEUF score <0.2 were identified that, at the time, were yet to have a disease-gene relationship. With access to deidentified genotype and phenotype data from participants sequenced through the 100,000 Genomes Project, 20,050 rare disease families were queried for individuals with *de novo* pLoF variants in LOEUF constrained genes. Specifically, cases were prioritised whereby more than one unrelated individual had a pLoF variant in the same gene and shared HPO terms. This approach identified 18 putative novel disease genes; whereby multiple unrelated individuals had overlapping phenotypes. This showcases the value of accessing rich genomic datasets of rare disease patients, whereby there is increased power to identify multiple individuals with extremely rare variants in the same genes. Since identifying these 18 putative novel disease genes, 12/18 (67%) have been independently published as novel disease genes. This demonstrates that gene constraint metrics can predict novel disease genes, even when the function of the gene is unknown. The 6 remaining genes are being investigated further. Two of these genes, *DDX17* and *HDLBP*, are close to publication and are described in **Chapters 5** and **6,** and briefly below. Four genes, *RIF1, CASZ1, ANKRD12* and *CLASP1,* have experiments planned in *Xenopus*.

11.2.1 *DDX17*

DDX17 is an RNA helicase shown to be involved in critical processes during the early phases of neuronal differentiation. **Chapter 5** describes the collation of a global case series of 11 patients with neurodevelopmental phenotypes harbouring *de novo* monoallelic variants in *DDX17*. All 11 patients have intellectual disability, delayed speech and language, and motor delay.

Working with colleagues in France, *in utero* cortical electroporation were performed in the brain of developing mice, assessing axon complexity and outgrowth of electroporated neurons, comparing wild-type and Ddx17 knockdown. *Ex vivo* cortical electroporation were then conducted on neuronal progenitors to quantitively assess axonal development at a single cell resolution. Homozygous and heterozygous *ddx17* crispant knockouts in *Xenopus tropicalis* were generated whereby morphology was assessed, behavioural assays performed, and neuron outgrowth examined. Further transcriptomic analysis was undertaken of neuroblastoma SH-SY5Y cells, looking for differentially expressed genes in DDX17-KD cells compared to controls.

Knockdown of Ddx17 in electroporated mouse neurons *in vivo* showed delayed neuronal migration as well as decreased cortical axon complexity. Mouse primary cortical neurons revealed reduced axon outgrowth upon knockdown of *Ddx17 in vitro*. The axon outgrowth phenotype was replicated in crispant *ddx17* tadpoles, including in a heterozygous model. Crispant tadpoles had clear functional neural defects and showed an impaired neurobehavioral phenotype. Transcriptomic analysis identified a statistically significant number of differentially expressed genes involved in neurodevelopmental processes in DDX17-KD cells compared to control cells.

Evidence from mouse, frog, and transcriptomic analysis strongly support the observation in humans that monoallelic variants in *DDX17* cause a neurodevelopmental disorder, Seaby-Ennis Syndrome.

11.2.2 *HDLBP*

HDLBP is a gene shown to contribute to biological processes including protein segregation and translation. **Chapter 6** describes 7 patients with *de novo* monoallelic variants in *HDLBP*, all of whom have a neurodevelopmental phenotype.

Working with the European Xenopus Resource Centre, homozygous and heterozygous *hdlbp* crispant knockouts were generated. These crispants were assessed for gross morphological abnormalities and subject to neurobehavioural assays. Markus Landthaler's group in Germany have assessed the RNA-binding of HDLBP and two patient-specific mutants to an RNA TFRC2 probe.

Crispant *hdlbp Xenopus* (F_0 homozygotes) injected at the one cell stage died early in development. Highly mosaic surviving F_0 crispants showed gross morphological differences and impairment in their neurobehavioural function compared to wild type. RNA-binding of two HDLBP mutants (patient-specific variants) to a TFRC2 RNA probe was reduced when compared to the wild type protein.

Evidence from *Xenopus* suggest that *HDLBP* is a critical gene in development and that severely mutated crispants in the homozygous state are embryonic lethal. Mosaic homozygous crispants displayed a neurodevelopmental phenotype, in keeping with the phenotype observed in human patients. RNA binding studies support evidence that patient-specific missense variants in *HDLBP* reduce the function of the protein. F_1 studies of heterozygous *Xenopus* crispants are underway and it is hoped these will add to the evidence that de novo monoallelic variants in *HDLBP* cause a neurodevelopmental phenotype.

11.3 Application of LOEUF to improve diagnostic rates

The 100,000 Genomes Project diagnosed a quarter of affected participants, but 26% of diagnoses were in genes not on the applied *in silico* gene panel(s).[326] Gene panels rely on clearly characterised phenotypes and risk missing diagnoses outside of the panel(s) applied. **Chapter 7** describes the development and testing of a panel-agnostic filtering strategy called DeNovoLOEUF that exploits gene constraint to identify potential diagnoses. Low LOEUF scores (<0.2) are highly enriched for haploinsufficient disease genes. Firstly, known disease genes with a LOEUF score <0.2 were selected. Participants from the 100,000 Genomes Project (100KGP) were computationally queried for *de novo* pLoF variants in these LOEUF-constrained disease genes. DeNovoLOEUF rapidly identified 332 putative diagnostic variants. Over a two-year period, 324/332 (98%) of these variants were independently confirmed as diagnostic or partially diagnostic by the 100KGP. A further 39 diagnoses were identified that were missed by the 100KGP's standard analysis. All these diagnoses have been shared with Genomics England and have been returned to patients. **Chapter 7** demonstrates a highly scalable, panel-agnostic, open-access tool (available at https://github.com/lecb/DeNovoLOEUF) that has a 98% positive predictive value when applied to large dataset of patients with rare diseases. Globally, as more patients are offered exome and genome sequencing, it is anticipated that DeNovoLOEUF can be used to rapidly identify diagnoses and facilitate iterative reanalyses when new disease genes are discovered.

11.4 Diagnosing cardiomyopathies in the 100,000 Genomes Project

Many of the disorders represented in the 100,000 Genomes Project are neurodevelopmental, whereby *de novo* variation is the expected mode of inheritance.[132,371] Cardiomyopathies represent a heterogenous group of disorders and are commonly inherited; this is because unlike many neurodevelopmental disorders, they do not affect fecundity. **Chapter 8** looked at cardiomyopathies in the 100KGP. Of all 100KGP participants, 1.6% had a cardiomyopathy phenotype, and despite the project favouring trio analysis, 75% of participants with cardiomyopathies were sequenced as singletons. The diagnostic rate for cardiomyopathy patients was low at 12% and significantly below expected rates for cardiomyopathies and below rates reported by the 100KGP for other phenotypes.[132,337,340,351] This highlights the limitations of singleton sequencing, in addition to the added complexity of interpreting segregation data which may be inaccurate for cardiac phenotypes that can present sub-clinically.

In total, 25% of all diagnoses returned by 100KGP for cardiomyopathy participants were not on the original gene panel selected by GEL. However, 16% would have been identified if a cardiomyopathy super panel had been applied. Application of a cardiomyopathy super panel to undiagnosed individuals with cardiomyopathies revealed 3 confirmed missed diagnoses and a potential 36 additional diagnoses awaiting review by NHS diagnostic laboratories.

While many rare disease studies focus on neurodevelopmental disorders, there are still many rare, monogenic diseases that affect specific organ systems such as the heart. Work in **Chapter 8** highlights how these can potentially be challenging to diagnose molecularly, and that gene panel selection is an essential consideration. With moves towards long read sequencing, it is hoped that some of the limitations of proband-only sequencing can be mitigated.

11.5 Missed biallelic diagnoses in the 100,000 Genomes Project

The DeNovoLOEUF method identified 39 missed diagnoses using a very simple filtering strategy, but all the variants were *de novo* and of high impact. These variants are typically rare, most often novel, and quick to assess. On the other hand, assessing biallelic variants is more challenging. This is, in part, due to the sheer number of variants requiring scrutiny, and that many are inherited and thus have higher allele frequencies than pathogenic *de novo* variants. **Chapter 9** describes the application of a whole gene pathogenicity metric, GenePy, to identify potential missed biallelic diagnoses in the 100KGP. GenePy is a dimensionality reduction algorithm that rescales complex variant-level data into more intuitive gene-level data. GenePy applies a statistical formula to every variant in an individual, incorporating allele frequency, variant zygosity, and a user-defined deleterious metric. GenePy then sums all variant scores across a gene, generating an aggregate GenePy score per gene, per participant. Higher GenePy scores in any given gene indicate increased burden of pathogenicity.

GenePy scores were calculated for 2862 known recessive disease genes in 78,216 (affected and unaffected) individuals in 100KGP. It was predicted that individuals with the highest ranked GenePy scores would be enriched for biallelic disease. For each gene, every person's GenePy scores were ranked relative to everyone else's in the cohort. Affected individuals without a diagnosis whose scores ranked amongst the top-5 for each gene were scrutinised. If someone who ranked in the top-5 scores had a phenotype overlapping the gene of interest, rare variants were extracted for detailed review. In 122/669 (18%) of the phenotype-matched cases, a putative diagnosis was identified in a top-ranking gene supported by variant phase, ClinVar status, and ACMG-AMP classification. These diagnoses were missed by 100KGP and have been returned through the diagnostic discovery pathway. A further 334/669 (50%) of cases have a possible missed diagnosis but require functional validation to prove pathogenicity. Applying GenePy at scale only added 1.2 additional variants (per individual) for assessment, suggesting that GenePy is a rapid and efficient tool in identifying biallelic diagnoses.

11.6 A panel-agnostic strategy 'HiPPo' improves diagnostic efficiency

Chapters 4, 7, 8 and 9 showcase new methods to improve diagnosis rates for rare diseases in the 100,000 Genomes Project, with a particular focus on panel-agnostic approaches. The new NHS Genomic Medicine Service (GMS) improves upon some of the shortcomings of the 100,000 Genomes Project, including a gene-agnostic filtering strep that assesses all *de novo* variants regardless of gene panel, and inclusion of Exomiser top 3 hits. This means that diagnoses, such as the 39 identified in **Chapter 7**, should no longer be missed. However, the GMS analytical strategy still predominantly uses *in silico* gene panels. Whilst this reduces variants requiring assessment by reporting laboratories significantly and avoids overloading the

NHS with incidental findings, pathogenic variants outside applied panels may still be missed if inherited, and variants in novel genes are largely ignored.

Chapter 10 describes a study comparing the analysis of a research exome versus a GMS clinical genome for the same set of patients. For the research exome, a panel-agnostic approach was applied filtering for variants in any gene with **Hi**gh **P**athogenic **Po**tential (HiPPo) using ClinVar, allele frequency, and *in silico* prediction tools. Resultant HiPPo variants were then further restricted to Gene Curation Coalition (GenCC) disease genes only. These results were compared with the GMS panel-based approach. Twenty-four participants from 8 families underwent parallel research exome and GMS genome sequencing. Exome HiPPo analysis identified a similar number of variants as the GMS panel-based approach despite covering many more genes. GMS genome analysis returned 2 pathogenic variants and 1 *de novo* variant. Exome HiPPo analyses returned the same three variants plus an additional pathogenic variant and 3 further *de novo* variants in novel genes. Case series are underway for the novel discoveries. When HiPPo was restricted to GenCC disease genes, statistically fewer variants required assessment to identify more pathogenic variants than reported by the GMS. This gave a diagnostic rate per variant assessed of 20% for HiPPo versus 3% for the GMS. This work demonstrates that panel-agnostic strategies do not necessarily mean assessment of more variants if *in silico* metrics are incorporated into filtering steps. With UK plans to sequence 5 million genomes, strategies are needed to optimise genome analysis beyond gene panels whilst minimising the burden of variants requiring clinical assessment. This can help ease pressures on stressed NHS systems and improve diagnostic efficiency.

Chapter 12 | Future challenges

12.0 Introduction

Undoubtedly, the utility of contemporary sequencing technologies has catapulted clinical medicine into a genomics era. The sheer pace at which these technologies has emerged and the investment into their clinical application highlights their potential. The UK has pledged to create the most advanced genomic healthcare system with a Genomic Medicine Service that plans to sequence 5 million individuals in the coming years. Application of genomics technologies has the potential to transform healthcare by identifying diagnoses early, matching patients to the most effective treatments, and improving mechanistic understanding of disease that can feed into development of new therapies. However, the ability to rapidly sequence and process data far outpaces the ability to analyse it. The explosion of generated sequencing data is currently stressing healthcare systems, particularly the NHS, impeding diagnostic reporting, and limiting personalised therapies. Waiting times for the interpretation and feedback of genomic results is being negatively impacted by insufficient manpower and vast quantities of variants of uncertain significance. There is a current backlog of ~29,000 patients in England pending reporting of their data by the GMS (Genomic Medicine Service Alliance Forum; July 2023).

This chapter discusses some of the future challenges, considerations, and opportunities for genomics moving forward.

12.1 Moving beyond panel based strategies

To focus genomic analysis and significantly reduce the burden of variation requiring assessment by clinical laboratories, the 100,000 Genomes Project and the subsequent GMS have adopted *in silico* panel-based strategies. Whilst these certainly reduce the number of variants requiring scrutiny by diagnostic laboratories including incidental findings, they risk missing diagnoses in genes not included on the pre-selected gene panel that may still be clinically relevant. In the 100KGP, 26% of confirmed diagnoses were not on the gene panel applied and were later identified by researchers.[326] One of the major challenges with gene panels is that they are outdated at the point of use and require selection of the 'correct' gene panel. In the UK, clinicians must select the gene panel to be applied when requesting genome sequencing in the NHS. However, with the pace that genomics has infiltrated all aspects of clinical medicine, this has meant that little to no training has been provided to guide clinicians on gene panel selection, nor accurate recording of phenotype data.

Work presented in this thesis clearly demonstrates that panel-agnostic approaches have merit. However, a major concern of widening analysis beyond panels is the risk that many more variants will require assessment. Despite these concerns, the work presented in **Chapter 10** shows that panel-agnostic strategies do not have to mean excessive numbers of additional variants requiring review. When pathogenicity tools are integrated into variant analyses, many variants more can be excluded; this means that panel-agnostic strategies can assess more of the genome, yet still minimise the number of variants requiring assessment. The HiPPo protocol clearly demonstrates that variant analysis can be expanded to all disease genes (as defined with strong or definitive evidence in GenCC) and still return fewer variants than the GMS panel-based strategy. Another concern of panel-agnostic approaches is the identification of incidental findings. Whilst it may be desirable to identify actionable incidental findings, the NHS currently does not have capacity to deal with these, as was demonstrated in the 100KGP.

Whilst the HiPPo protocol shows promise, the sample size used was modest, limiting the validity of any conclusions. This highlights one of the major challenges of working with sensitive genomic data in clinical and research settings. For example, UK-based clinicians that refer patients to the GMS for genome sequencing are not able to view their own patient's genomic data; in the NHS they only receive the final report. Even if the clinician has access to deidentified data within a secure research environment, that may include some of their own patients, they are not permitted to attempt to identify or analyse their own patient's data. Therefore, to conduct a comparison of the HiPPo method versus the GMS strategy, there was no choice but to independently consent and re-sequence (in a research setting) the same patients that underwent genome sequencing through the GMS. To obtain access to study participants' GMS data, research consent was obtained to access their medical records, their GMS diagnostic reports, and the variant-level data returned to the Wessex Regional Genomics Laboratory. There is an argument that resequencing patients who have already undergone prior sequencing is not a good use of research funding. However, with the lack of fluid information governance structures facilitating secure data flow that benefits patients without compromising their privacy, such measures are sometimes necessary to access patient-specific variant level data. This is becoming an increasing problem when clinicians receive negative genome reports on their patients and then wish to view (or ask a colleague to view) their patient's raw data, perhaps to assess non-coding regions or newly discovered genes.

12.2 Balancing data sharing with data protection

The willingness for patients, participants, researchers, and clinicians to share data has facilitated a wealth of genomics repositories and tissue biobanks (**Table 12.1**). The need to protect patients' genotype and phenotype data is clearly of utmost importance; however, the degree to which data governance restricts data access for certain groups of people varies between countries, and this is demonstrated by varying levels of data access across numerous data repositories.

When working with research data from Genomics England, it is not possible to export any HPO terms or combination of HPO terms (regardless of genotype data) that is present in fewer than 5 individuals due to the potential that this may identify a patient. This means that if a researcher identifies a potential novel disease gene, they are unable to submit the gene to the Matchmaker Exchange, nor share any genotype or phenotype data. The only way for these data to be shared is through contact with the patient's clinician. In GEL, the researcher is blind to the clinician looking after the patient and must fill in a clinician contact request form, which is not an easy process, taking considerable time. Completion of the form prompts GEL to initiate contact with the clinician, however experience suggests that only 20% of clinical contact requests in GEL are successful. In contrast, the GREGoR Consortium in the US adopts a less restrictive approach to data sharing, whereby participating subjects are specifically consented at recruitment for sharing of their genotype and phenotype data for the purposes of novel gene discovery and diagnostic uplift. Currently, there is no agreed global consensus on: a) what constitutes patient identifiable data; and b) how and when data should be shared. Ultimately this is where patient and public involvement is critical to inform future policies and consent processes. Perhaps a solution may be a graded level of consent, whereby participants can opt into different levels of data sharing, depending on their personal preferences.

TABLE 12.1 | EXAMPLES OF PATIENT REGISTRIES AND TISSUE BIOBANKS

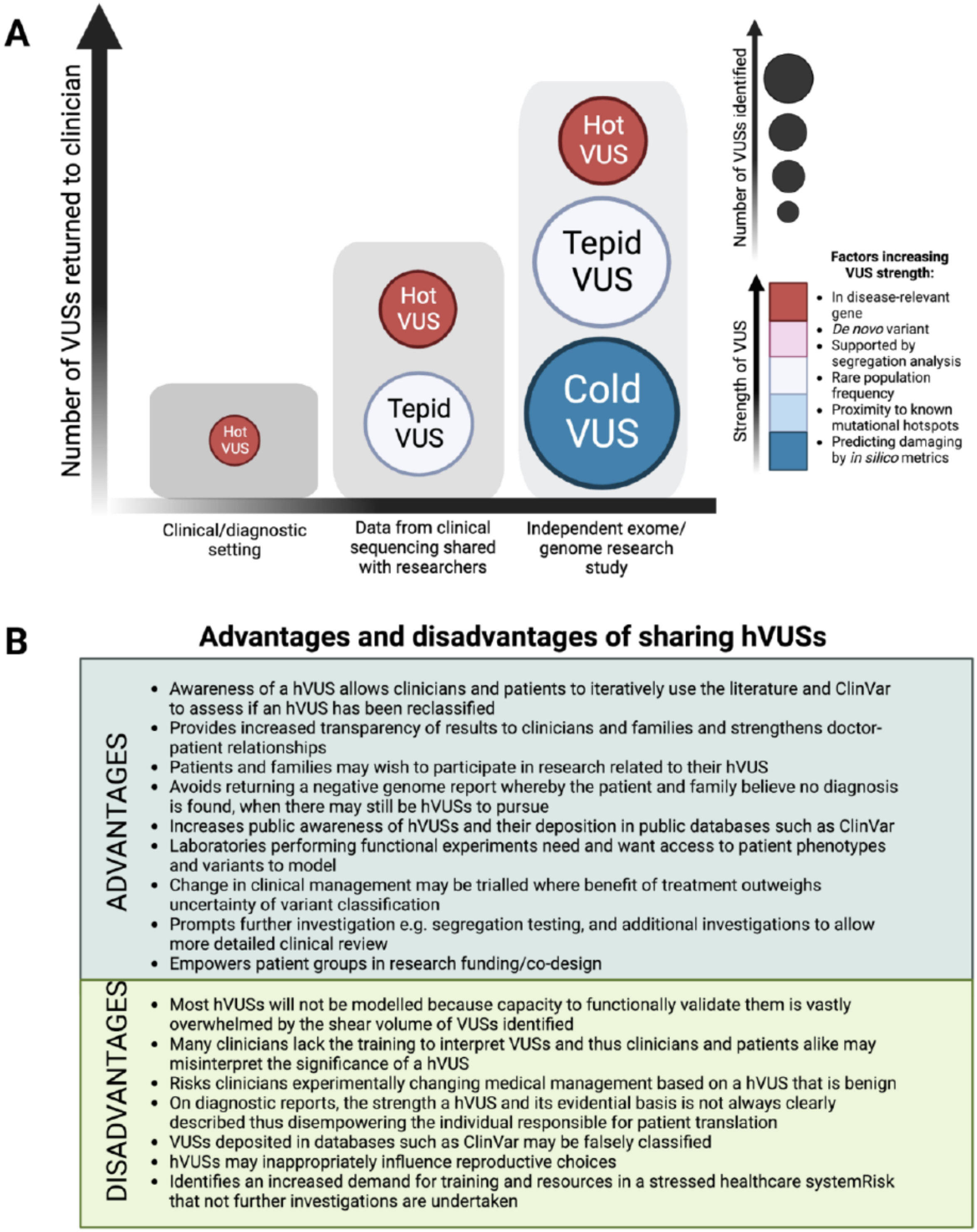
Registries	**Data type**	**Availability**	**Data access**	**Additional data access**
gnomAD	Population genotype data	Open access	https://gnomad.broadinstitute.org	Individual level data can be applied for through dbGaP
GTEx portal	RNA tissue expression data	Open access	https://gtexportal.org/home	Individual level data can be applied for through dbGaP
GREGoR Consortium	Genotype and health data	Requires consortium membership	https://gregorconsortium.org/data	Some data can be applied for without consortium membership through dbGaP
Genomics England	Genotype and health data	Requires data access approval	https://www.genomicsengland.co.uk/	Patient clinicians can be contacted using a form within the secure Genomics England Research Environment
DECIPHER	Individual level variants from patient cohorts with neurodevelopmental disorders. High level phenotype data available.	Open access	https://www.deciphergenomics.org	Patient clinicians can be contacted through an online form on the website
UKBioBank	Biobank data	Requires payment and data access approval	https://www.ukbiobank.ac.uk	N/A
All of Us	Biobank data	Open access to summary data only	https://allofus.nih.gov	Researchers can apply for access to individual level data
FinnGen	Biobank data	Open access to summary data only	https://www.finngen.fi	Individual level data is available to consortium partners
European Genome-Phenome Archive	List of studies containing genetic and phenotype data	Each study requires approval to access	https://ega-archive.org/	N/A
ClinVar	Genotype data	Open access	https://www.ncbi.nlm.nih.gov/clinvar	N/A
HGMD	Genotype data	Open access to limited dataset following registration	https://www.hgmd.cf.ac.uk/	HGMD professional is available commercially

12.3 VUSs in disease genes impede clinical translation

Increasing numbers of healthcare professionals and researchers are exposed to genomic data of expanding complexity. There is now an expectation on these individuals to interpret genomic diagnostic reports, often with limited or no formal training. For variants to reach a clinical threshold for diagnostic reporting by ACMG-AMP guidelines, they must be already established as disease-causing, predicted to truncate the protein product, or have robust evidence from established *in vivo* or *in vitro* studies (see **methods 2.10**).[123]

It may be assumed that any rare variant in a clinically relevant disease gene would be diagnostic. However, most humans have rare variants, private to them, in known disease genes; therefore, even if a variant is absent from population datasets and predicted deleterious by computational metrics, many of these variants remain as VUSs even when the phenotype perfectly fits the gene. Whilst all VUSs are ostensibly equal, it can be helpful to think of VUSs along a spectrum of those more likely to be pathogenic e.g. a 'hot' VUS (hVUS), versus a 'cold' VUS, more likely to be benign (**Figure 12.1**).[372]

FIGURE 12.1 | PATHOGENIC STRENGTH OF VUSS AND ADVANTAGES/DISADVANTAGES OF THEIR RETURN


A
Number of VUSs returned to clinician
Hot VUS
Clinical/diagnostic setting
Hot VUS
Tepid VUS
Data from clinical sequencing shared with researchers
Hot VUS
Tepid VUS
Cold VUS
Independent exome/ genome research study
Number of VUSs identified
Strength of VUS
Factors increasing VUS strength:
In disease-relevant gene
De novo variant
Supported by segregation analysis
Rare population frequency
Proximity to known mutational hotspots
Predicting damaging by *in silico* metrics
B
ADVANTAGES
DISADVANTAGES

Advantages and disadvantages of sharing hVUSs

ADVANTAGES

- Awareness of a hVUS allows clinicians and patients to iteratively use the literature and ClinVar to assess if an hVUS has been reclassified
- Provides increased transparency of results to clinicians and families and strengthens doctor-patient relationships
- Patients and families may wish to participate in research related to their hVUS
- Avoids returning a negative genome report whereby the patient and family believe no diagnosis is found, when there may still be hVUSs to pursue
- Increases public awareness of hVUSs and their deposition in public databases such as ClinVar
- Laboratories performing functional experiments need and want access to patient phenotypes and variants to model
- Change in clinical management may be trialled where benefit of treatment outweighs uncertainty of variant classification
- Prompts further investigation e.g. segregation testing, and additional investigations to allow more detailed clinical review
- Empowers patient groups in research funding/co-design

DISADVANTAGES

- Most hVUSs will not be modelled because capacity to functionally validate them is vastly overwhelmed by the shear volume of VUSs identified
- Many clinicians lack the training to interpret VUSs and thus clinicians and patients alike may misinterpret the significance of a hVUS
- Risks clinicians experimentally changing medical management based on a hVUS that is benign
- On diagnostic reports, the strength a hVUS and its evidential basis is not always clearly described thus disempowering the individual responsible for patient translation
- VUSs deposited in databases such as ClinVar may be falsely classified
- hVUSs may inappropriately influence reproductive choices
- Identifies an increased demand for training and resources in a stressed healthcare systemRisk that not further investigations are undertaken

***A.** Schematic illustrating the variable number of VUSs returned in a UK clinical setting versus two different research settings including: approved researchers gaining access to UK clinical sequencing data whereby established channels exist to enable return of VUSs to clinicians and variants are often screened; and independent exome or genome research studies whereby no formal policies exist to return VUSs. Far fewer VUSs are returned in a clinical setting compared with a research setting. Fewer VUSs are identified, even for hot VUSs, this is because the application of in silico gene panels applied to clinical sequencing in the NHS Genome Medicine Service significantly reduces the number of variants assessed. The number of VUSs identified and returned typically increases in a research setting but sometimes at the expense of the strength of the VUSs identified, i.e. more VUSs of limited strength are both identified and returned, particularly when VUSs are identified in novel genes. **B.** Advantages and disadvantages of returning hVUSs.*

Recently, as part of a doctoral side-project, a missense VUS (GATM; c.965G>C; p.Arg322Pro) was identified in an affected mother-daughter duo with idiopathic Fanconi syndrome and renal failure. The VUS was in *GATM*, a disease gene fully concordant with the patients' phenotypes. Despite the variant being novel (i.e. absent from population databases), having a CADD score of 32, and being in a mutational hotspot where all other pathogenic missense variants reside, the variant did not meet 'likely pathogenic' or 'pathogenic' classification by ACMG-AMP guidelines. Upon collaboration with the chemistry department at the University of Southampton, molecular dynamics (MD) simulations were undertaken which demonstrated a dynamic signature that differentiated the Arg322Pro variant, and two previously identified *GATM* pathogenic variants (confirmed by *in vivo* studies) from wildtype. Whilst the MD simulations provided supportive evidence that the Arg322Pro variant (and the proven other pathogenic variants) altered the protein function, (and these results have since been published, **Appendix Paper 15**[373]), the variant remains a VUS without biochemical or *in vivo* modelling.

The return of VUSs depends on whether the sequencing test was performed in a clinical or research setting (**Figure 12.1A**) and varies by national/international professional policies e.g. currently across 19 North American laboratories, all VUSs from multi-gene panel tests are returned.[374] Generally in the UK, VUSs are more commonly returned in a research setting as evidenced by the additional variants returned through the 100KGP.[132] However, as many patient undergoing clinical genome sequencing are consented for research, clinicians will increasingly encounter VUSs, yet many lack the training to interpret them.[375]

In the UK, most VUSs are not returned by clinical diagnostic laboratories. Whilst this avoids returning ambiguous results, it can be helpful for patients, families, and clinicians to be made aware of hVUSs. Sometimes hVUSs are reported, but their definition is subjective. Where possible, hVUSs should be discussed in multidisciplinary teams, ideally involving a triumvirate of a clinical geneticist, mainstream specialist, and clinical scientist. This can help identify missing evidence, e.g. phenotype data or segregation data which can help upweight a hVUS to likely pathogenic or pathogenic.

There is no robust consensus on whether hVUSs should be shared with patients.[376] Qualitative studies show that patients commonly misunderstand VUSs and respond both positively and negatively to the knowledge of them.[377,378] Ideally, families should be counselled and consented for the return of hVUSs prior to ordering genetics tests. Once consented, hVUSs would only be returned by fully trained individuals.[378] This is not common practice currently, whereby clinicians must use clinical judgement, on a family-by-family basis, about whether to discuss any VUSs they receive with families.

Whilst there are obvious risks to returning hVUSs, there are clearly advantages (**Figure 12.1B**). Patient sequencing data are rarely re-analysed,[379] meaning if a hVUS is functionally validated as pathogenic after the patient's clinical report is returned, that diagnosis would be missed. Awareness of hVUSs allows clinicians and patients to follow up variants (such as in ClinVar[113]) in the event they are ever reclassified, but no support is provided to help with this. However, clinicians should counsel patients that their specific hVUS will likely never be modelled, and if modelled, the variant may be down-classified to benign.[380] That said, functional research labs need and want access to human phenotype and genotype data for *in vivo* and *in vitro* modelling, thus knowledge of a hVUS can prompt patient involvement with research.

With an exponential increase in VUSs identified, there is clear need for better integration of guidelines and training on the interpretation and return of VUSs. Guidance is clear however that medical management should not change based on a VUS, although some clinicians may wish to try empirical therapies if they consider the risks low.

12.4 Functional validation of novel disease genes

Substantial progress in next generation sequencing technologies has led to the curation of long lists of novel putative disease-causing genes for a multitude of human genetic diseases. However, most of these genes are of unknown or poorly understood molecular, cellular, developmental, and homeostatic functions and therefore cannot be proven as disease-genes. Consequently, these novel genes are ignored in clinical settings, yet they require discovery to expand the number of disease genes tested for.

To translate a variant in a novel gene all the way through to a diagnosis is an extremely protracted and expensive process. Discovery and validation of novel disease genes usually necessitates the identification of multiple unrelated and affected families with similar phenotypes. This must be further supported by robust *in vivo* or *in vitro* functional evidence, typically involving a knockout animal model, whereby morphological features are compared with the patient phenotype and additional functional assays are undertaken. Some model organisms, such as *Xenopus* and zebrafish, can be grown rapidly, but these models best recapitulate a homozygous disease. Yet, many disease genes are haploinsufficient and a homozygous knockout may prove lethal. Heterozygous models are possible but require an F_1 generation and successful matings. This is not straightforward nor quick and adds complexity and increased risk of failure. Biochemical assays offer an alternative to animal models, but require basic understanding of the gene's function, which is not always apparent.

Chapter 4 describes the identification of many putative novel disease genes, including *RIF1, CASZ1, ANKRD12* and *CLASP1,* where international patient case series are underway. All these genes are predicted to cause disease through haploinsufficiency. Similar to the functional *Xenopus* work presented in **Chapters 5** and **6**, heterozygous knockout models in *Xenopus* are planned. However, experiments cannot start until additional resources become available. Further to *RIF1, CASZ1, ANKRD12* and *CLASP1*, a further 99 possible novel disease genes were identified through work described in **Chapter 4**.[306] These were deprioritised for functional modelling due to there being evidence in only a single individual/pedigree but these ultra-rare VUSs also require functional validation through animal modelling or biochemical assays. This is where there is demand for cross-disciplinary collaborations between research groups, research funders, clinicians, and commercial entities to reduce the time in translating discoveries from bench to bedside. Interest in this area is growing, and in 2022, ModelMatcher[381] was released which facilitates collaborations between scientists and other stakeholders of rare and undiagnosed disease research. The platform aims to address the unmet requirements for a global network devoted to pursuing functional studies relevant to rare disease. Furthermore, the US' GREGoR consortium has been set up with an infrastructure capable of feeding VUSs, in both known and novel genes, identified through GREGoR sequencing sites, to affiliated functional laboratories. Candidate variants and genes can be pitched to the most appropriate functional laboratory to answer the specific research question. Functional testing on offer may include massively parallel reporter assays; targeted CRISPR for rare variant interpretation; high-throughput CRISPR reporter gene screens; iPSC functional genomics; and CRISPRi mouse models.

12.5 Long read sequencing technologies

Short read sequencing (SRSeq) has been the major next generation sequencing technology applied to genomic studies over the last 15 years. SRSeq produces reads up to 600 bases and is cost-effective, accurate, and supported by many analytical tools and pipelines. However, SRSeq has its limitations, particularly with regards to genome mapping in repetitive regions; coverage bias around GC-rich regions; detection of large or complex structural variants; and in determining variant phase.

Long read sequencing (LRSeq) technologies, routinely generating reads in excess of 10 kbs, are beginning to overcome the limitations in short read sequencing. Compared to SRSeq, LRSeq boasts improved *de novo* assembly; telomere-to-telomere chromosome assembly, improved mapping certainty in poorly characterised regions of the genome; methylation data; detection of large and complex structural variants; accurate measurement of sequencing tandem repeat expansions; and in determining variant phase in the absence of parental data.[382-384]

Technologies currently dominating the LRSeq space include PacBio and Oxford Nanopore Technologies, however both are significantly more expensive that SRSeq per run, although prices are becoming more competitive.[383] Illumina also provide a LRSeq-like technology and Bionano offer optical mapping.[382] Whilst SRSeq is far from obsolete, the use of LRSeq is increasingly rapidly. In the rare disease space, its utility is particularly suited to patients who may harbour complex structural variants; may have tandem-repeat disease; or where parental/segregation data are unavailable, facilitating phasing of alleles for non-trio families.[385] The ability to phase data in the absence of parental samples increases equity of access to sequencing for non-nuclear families; this has been a critical limitation of many genomics studies to date, including the 100,000 Genomes Project.[294]

Whilst LRSeq shows promise, there are still considerations and limitations. LRSeq traditionally has a lower accuracy per read when compared with SRSeq[382,384], however PacBio's HiFi technology boasts accuracy comparable with SRSeq.[386] Further, the output from some LRSeq technologies requires more data storage than SRSeq, although this is not the case for PacBio's Revio machine which contains a supercomputer within it, capable of providing processed files, although the 2023 list price for a Revio machine is ~800,000 USD.[387] A further limitation is that the availability of tools and pipelines specific for LRSeq data are far fewer than for SRSeq. That said, with the ability of LRSeq to resolve many unanswered questions in genomics, it seems reasonable to predict that LRSeq may replace SRSeq in the future.

12.6 Integrating Artificial Intelligence (AI)

There is demand for a paradigm shift in the analysis of genomic data to ease the strain on healthcare systems and improve the efficiency of data interpretation and reporting. Despite a plan to sequence 5 million genomes through the GMS, there is lack of high-throughput pipelines for data interpretation and prioritisation. This is leading to severe delays in patients receiving results from genome sequencing tests. There is obvious demand for scalable and rapid digital solutions to improve the efficiency and interpretation of big data. Artificial intelligence can be used to extract knowledge from large complex datasets, input these data into complex algorithms and produce reproducible, interpretable, automatic, and translational results. The application of AI to genomics has potential to transform healthcare by expediting diagnoses; offering personalised therapies based on improved mechanistic understanding of disease; and improving cost-efficiency for stressed healthcare systems.

12.6.1 Reanalysis of genomic data

In the NHS and in many other healthcare systems, genome and exome data are seldom reanalysed. Therefore, if a new disease gene or pathogenic variant is discovered after the patient's report is returned, the diagnosis is missed. Reanalysis of data is time-consuming and resource intensive. Having automated methods to iteratively reanalyse data at frequent intervals would free up human resource and facilitate diagnostic uplift when novel discoveries are published. A screening tool, coded in hail, as described in

Chapter 7.4.1, has recently been developed that can automate the extraction of variants from annotated genomic data. This tool incorporates ClinVar and GenCC data, which are continually updated, meaning it can be iteratively applied. This tool is being tested on exome data made available through the GREGoR consortium to assess the sensitivity and specificity of the approach. There are plans to use supervised learning to refine the performance of this software. These results will be shared with Microsoft, who I am collaborating with in the development of an AI-reanalysis tool for genomic data at the Broad Institute.

12.6.2 Automated methods for phenotyping

Healthcare services are becoming increasingly digitised. Within the NHS, the importance of clinical coding and use of standardised terminology (e.g. ICD10, SNOMED-CT, OPCS-4 codes etc) to accurately record clinical data is ever apparent, particularly when hospitals make significant losses from suboptimal coding.[388,389] For genomics specifically, HPO terms (belonging to a coded hierarchical structure) are the means by which phenotype data are recorded.[167] However, a major weakness of HPO terms is that they only represent a snapshot in time and do not provide the full clinical patient narrative, which is far more nuanced. That said, in many cases, HPO terms are more than sufficient to interpret variants without the need for verbose clinical descriptions. Further, given the coded nature of HPO terms, and indeed other electronic coding structures such as ICD-10 etc, they serve as useful and measurable data in machine learning applications. Additionally, such data can be automatically and periodically extracted, recapitulating essential data from a patient's clinical history but in a manner that is amenable for integration into AI-software, and for use in research and public health statistics to improve overall healthcare.

12.6.3 Using AI to improve the interpretation of VUSs

As sequencing costs have fallen, more individuals than ever are being sequenced. This explosion of human sequencing data is causing a logarithmic expansion of identified novel variants that cannot be functionally interpreted (**Figure 11.2**). These variants of uncertain significance are typically not reported in a clinical setting. Yet, they represent valuable data that, when coupled with additional evidence, can provide a diagnosis for a patient; improve mechanistic understanding of disease; and lead to better personalised medicine.

FIGURE 12.2 | THE GROWING PROBLEM OF VUSS

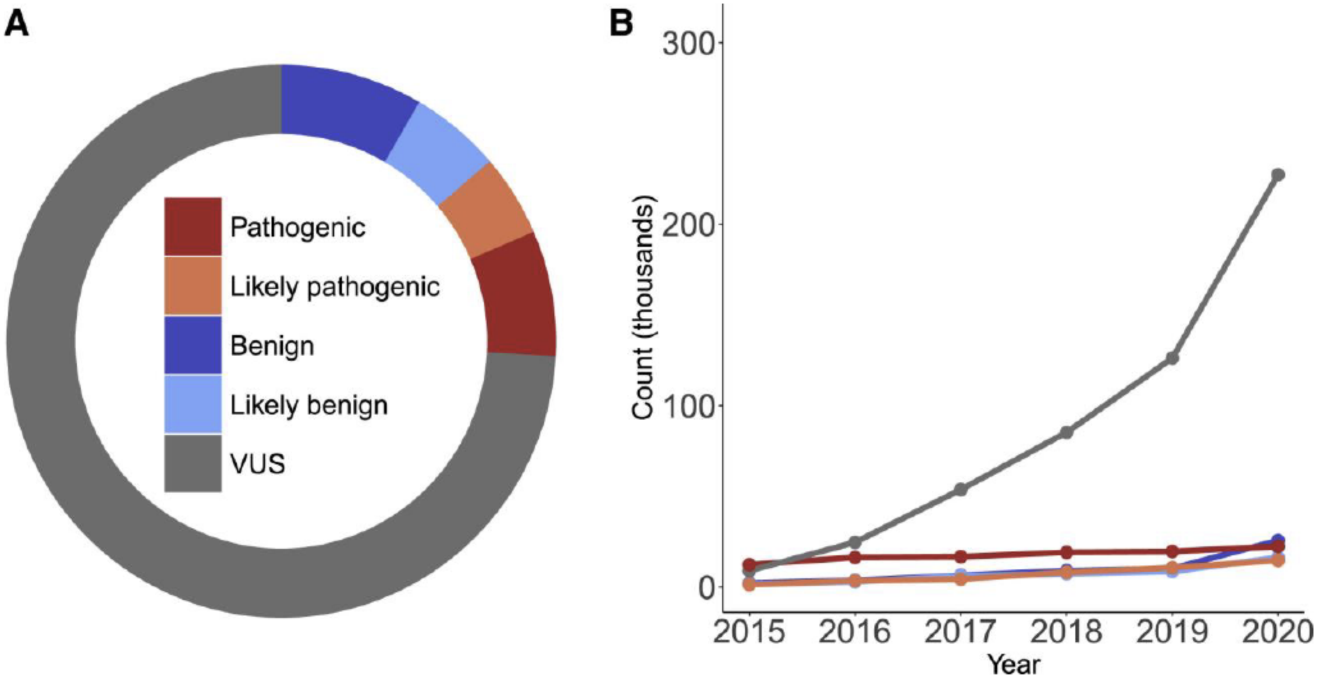


Image taken from Fayer et al.[390] *(A) Missense variants (downloaded from ClinVar in 2020) coloured by ClinVar Classification (benign = 25,707; likely benign = 16,377; VUSs = 227,365; likely pathogenic = 14,716; pathogenic = 22,489; conflicting = 20,026. (B) Missense variants in ClinVar from 2015 to 2020, coloured by clinical significance.*

Currently, the only way to up- or down-grade a VUS to ‘pathogenic’ or ‘likely pathogenic’, is through *in vivo* or *in vitro* experiments. Whilst there have been *in vitro* efforts to test the impact of all variants in a gene[246], this is not practical for all 20,000 human genes. *In silico* metrics can help predict pathogenicity of variants but these metrics are not sufficient to reclassify a VUS as benign or pathogenic for clinical reporting. Therefore, the interpretation of VUSs is proving a major bottleneck, that is impeding clinical translation to the detriment of patients.

Whilst, to date, *in silico* methods cannot prove pathogenicity, they can still be invaluable in prioritising which variants are best to functionally model. Moreover, with improved sensitivity and specificity of *in silico* models, it is possible that generated *in silico* data may be weighted differently in future ACMG-AMP classification guidelines. This is where AI shows huge potential; it can offer scalable, rapid solutions to the growing problem of VUSs, whereby supporting evidence of pathogenicity can be obtained at a pace not achievable through wet lab experiments. AlphaFold[391] is an excellent example of an AI tool capable of improving interpretation of genomic data; this open-source software can predict the 3D structure of a protein from its amino acid sequence with high accuracy.[392] The increasing use of AI in data science is thanks to availability of large-scale publicly available data libraries, whereby data can be used to train and improve machine-learning models. For example, in developing AI solutions to improve the interpretation of VUSs, training data could be obtained from pathogenic and benign variants in ClinVar and used alongside existing models of protein structure (AlphaFold). Integrating these data into a supervised learning could facilitate development and iterative improvement of a machine learning model for improved interpretation of VUSs.

12.7 Newborn screening

Currently in the NHS, nine conditions are screened for at birth from a heel prick test, predominantly using biochemical assays. With the advancements in next generation sequencing, there has been increased interest in the idea of performing whole genome sequencing at birth and using these data to inform healthcare throughput that individual’s life.[393] Genomics England, in partnership with the NHS, have designed a research study, now referred to as the Generation Project, which aims to deliver whole genome sequencing to 100,000 newborn babies and assess the potential value, costs, risks and benefits of such a service.[394] This is an exciting time for genomics and it is easily conceivable that in the near distant future, genomic data will be stored on a person's healthcare record and be used to: screen for rare diseases; identify risk alleles for complex diseases, guiding preventative medicine; and inform drug responses and optimal therapies.

12.8 Conclusion

In summary, the field of genomics is moving at an unprecedented pace. Within healthcare, there is a seismic shift to move away from reactive medicine to preventative medicine; and genomics has the potential to help drive this paradigm shift. As more genomic data are collected and stored, automated methods are needed to better manage and interpret these data in order to inform personalised healthcare and improve clinical outcomes. Methods will constantly need refining to keep up with the pace and scale of data being generated, but with increased digitisation of data and open data sharing, artificial intelligence can be utilised to develop digital solutions capable of handling big data on an unprecedented scale. Whilst the field of genomics has exploded over the last decade, this is only the beginning. Projects such as the Generation Study are needed to assess the feasibility, risks and benefits of integrating genomic data within routine healthcare records over a patient's lifetime. These results will be essential in informing future healthcare policies both nationally and internationally.

References

1. Seaby EG, Ennis S. Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Briefings in Functional Genomics*. 2020;doi:10.1093/bfgp/elaa009

2. Seaby EG, Rehm HL, O'Donnell-Luria A. Strategies to Uplift Novel Mendelian Gene Discovery for Improved Clinical Outcomes. Review. *Frontiers in Genetics*. 2021-June-17 2021;12(935)doi:10.3389/fgene.2021.674295

3. Turnbull C, Scott RH, Thomas E, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *Bmj*. 2018;361:k1687.

4. The Hidden Costs of Rare Diseases: A Feasibility Study. 2016. https://www.geneticalliance.org.uk/media/2502/hidden-costs-full-report_21916-v2-1.pdf

5. Rode J. Rare diseases: understanding this public health priority. *Paris: EURORDIS*. 2005;

6. Yoon PW, Olney RS, Khoury MJ, Sappenfield WM, Chavez GF, Taylor D. Contribution of birth defects and genetic diseases to pediatric hospitalizations: a population-based study. *Archives of pediatrics & adolescent medicine*. 1997;151(11):1096-1103.

7. Dodge JA, Chigladze T, Donadieu J, et al. The importance of rare diseases: from the gene to society. BMJ Publishing Group Ltd; 2011.

8. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. *Nature Reviews Genetics*. 2018;19(5):253.

9. Adams DR, Eng CM. Next-generation sequencing to diagnose suspected genetic disorders. *New England Journal of Medicine*. 2018;379(14):1353-1362.

10. Retterer K, Juusola J, Cho MT, et al. Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine*. 2016;18(7):696-704.

11. Clark M, Stark Z, Farnaes L, Tan T, White S, Dimmock D, Kingsmore S. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. npj Genomic Medicine 3, 1–10 (2018). *URL http://dx doi org/101038/s41525-018-0053-8*. 2018;

12. Srivastava S, Love-Nichols JA, Dies KA, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med*. 2019;21(11):2413-2421. doi:10.1038/s41436-019-0554-6

13. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. *New England Journal of Medicine*. 2014;370(25):2418-2425.

14. Siva N. 1000 Genomes project. Nature Publishing Group; 2008.

15. Sankar PL, Parker LS. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genetics in Medicine*. 2017;19(7):743-750.

16. Posey JE, O'Donnell-Luria AH, Chong JX, et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genetics in Medicine*. 2019;21(4):798-812.

17. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*. 2016/11/28 2016;17(1):241. doi:10.1186/s13059-016-1110-1

18. Taylor JL, Debost J-CP, Morton SU, et al. Paternal-age-related de novo mutations and risk for five disorders. *Nature Communications*. 2019;10(1):3043.

19. Rahbari R, Wuster A, Lindsay SJ, et al. Timing, rates and spectra of human germline mutation. *Nature Genetics*. 2016/02/01 2016;48(2):126-133. doi:10.1038/ng.3469

20. Arnadottir GA, Jensson BO, Marelsson SE, et al. Compound heterozygous mutations in UBA5 causing early-onset epileptic encephalopathy in two sisters. *BMC Medical Genetics*. 2017/10/02 2017;18(1):103. doi:10.1186/s12881-017-0466-8

21. Zernant J, Lee W, Collison FT, et al. Frequent hypomorphic alleles account for a significant fraction of ABCA4 disease and distinguish it from age-related macular degeneration. *Journal of Medical Genetics*. 2017;54(6):404-412. doi:10.1136/jmedgenet-2017-104540

22. Karczewski KJ, Weisburd B, Thomas B, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research*. 2017;45(D1):D840-D845.

23. Gudmundsson S, Carlston CM, O'Donnell-Luria A. Interpreting variants in genes affected by clonal hematopoiesis in population data. *Human Genetics*. 2023/02/04 2023;doi:10.1007/s00439-023-02526-4

24. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021;590(7845):290-299.

25. Gudmundsson S, Singer-Berk M, Watts N, et al. Variant interpretation using population databases: lessons from gnomAD. *arXiv preprint arXiv:210711458*. 2021;

26. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020/05/01 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

27. Whiffin N, Minikel E, Walsh R, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine*. 2017/10/01 2017;19(10):1151-1158. doi:10.1038/gim.2017.26

28. MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012;335(6070):823-828.

29. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. *Nature Genetics*. 1/2019 2019;51(1):88-95. doi:10.1038/s41588-018-0294-6

30. Samocha KE, Kosmicki JA, Karczewski KJ, et al. Regional missense constraint improves variant deleteriousness prediction. *BioRxiv*. 2017:148353.

31. Hurst LD. The sound of silence. *Nature*. 2011/03/01 2011;471(7340):582-583. doi:10.1038/471582a

32. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. Exposing synonymous mutations. *Trends in Genetics*. 2014;30(7):308-321.

33. Bao S, Moakley DF, Zhang C. The splicing code goes deep. *Cell*. 2019;176(3):414-416.

34. Wai HA, Lord J, Lyon M, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genetics in Medicine*. 2020:1-10.

35. Soemedi R, Cygan KJ, Rhine CL, et al. Pathogenic variants that alter protein code often disrupt splicing. *Nature genetics*. 2017;49(6):848.

36. Adamson SI, Zhan L, Graveley BR. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome biology*. 2018;19(1):71.

37. Lin M, Whitmire S, Chen J, Farrel A, Shi X, Guo J-t. Effects of short indels on protein structure and function in human genomes. *Scientific Reports*. 2017/08/24 2017;7(1):9313. doi:10.1038/s41598-017-09287-x

38. Shen H, Li J, Zhang J, et al. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. *PloS one*. 2013;8(4):e59494.

39. Jiang Y, Turinsky AL, Brudno M. The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection. *Nucleic acids research*. 2015;43(15):7217-7228.

40. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*. 2016/12/01 2016;48(12):1581-1586. doi:10.1038/ng.3703

41. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*. 2015;31(5):761-763.

42. Li J, Shi L, Zhang K, et al. VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic acids research*. 2018;46(D1):D1039-D1048.

43. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*. 2012;40(W1):W452-W457.

44. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nature methods*. 2010;7(4):248-249.

45. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature methods*. 2014;11(4):361-362.

46. Grantham R. Amino acid difference formula to help explain protein evolution. *Science*. 1974;185(4154):862-864.

47. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*. 2016;99(4):877-885.

48. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *The American Journal of Human Genetics*. 2018;103(4):474-483.

49. Wang M, Zhao X-M, Takemoto K, Xu H, Li Y, Akutsu T, Song J. FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PloS one*. 2012;7(8)

50. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*. 2015;31(10):1536-1543.

51. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nature methods*. 2014;11(3):294.

52. Gelfman S, Wang Q, McSweeney KM, et al. Annotating pathogenic non-coding variants in genic regions. *Nature Communications*. 2017/08/09 2017;8(1):236. doi:10.1038/s41467-017-00141-2

53. Ratan A, Olson TL, Loughran TP, Miller W. Identification of indels in next-generation sequencing data. journal article. *BMC Bioinformatics*. February 13 2015;16(1):42. doi:10.1186/s12859-015-0483-6

54. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics*. 2014;30(19):2813-2815. doi:10.1093/bioinformatics/btu376

55. Koboldt DC, Chen K, Wylie T, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. 2009;25(17):2283-2285.

56. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. *Genome research*. 2011;21(6):961-973.

57. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*. 2011;43(5):491.

58. Folkman L, Yang Y, Li Z, et al. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. *Bioinformatics*. 2015;31(10):1599-1606. doi:10.1093/bioinformatics/btu862

59. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*. 2018/11/01 2018;36(10):983-987. doi:10.1038/nbt.4235

60. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic acids research*. 2001;29(5):1185-1190.

61. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology*. 2004;11(2-3):377-394.

62. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research*. 2009;37(9):e67-e67.

63. Mort M, Sterne-Weiler T, Li B, et al. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome biology*. 2014;15(1):R19.

64. Jaganathan K, Panagiotopoulou SK, McRae JF, et al. Predicting splicing from primary sequence with deep learning. *Cell*. 2019;176(3):535-548. e24.

65. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*. 2010;20(1):110-121.

66. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*. 2005;15(7):901-913.

67. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*. 2005;15(8):1034-1050.

68. Tang H, Thomas PD. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*. 2016;32(14):2230-2232. doi:10.1093/bioinformatics/btw222

69. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*. 2015;12(10):931-934.

70. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific reports*. 2015;5:10576.

71. Lehmann K-V, Chen T. Exploring functional variant discovery in non-coding regions with SInBaD. *Nucleic Acids Research*. 2012;41(1):e7-e7. doi:10.1093/nar/gks800

72. Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research*. 2015;43(W1):W306-W313. doi:10.1093/nar/gkv359

73. Pires DEV, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research*. 2014;42(W1):W314-W319. doi:10.1093/nar/gku411

74. Liu M, Watson LT, Zhang L. HMMvar-func: a new method for predicting the functional outcome of genetic variants. *BMC bioinformatics*. 2015;16(1):351.

75. Ryan M, Diekhans M, Lien S, Liu Y, Karchin R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics*. 2009;25(11):1431-1432.

76. Giollo M, Martin AJ, Walsh I, Ferrari C, Tosatto SC. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC genomics*. 2014;15(4):S7.

77. López-Ferrando V, Gazzo A, de la Cruz X, Orozco M, Gelpí JL. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research*. 2017;45(W1):W222-W228. doi:10.1093/nar/gkx313

78. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*. 2014;46(3):310.

79. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics*. 2016;48(2):214.

80. Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*. 2018;50(8):1161-1170.

81. Mossotto E, Ashton JJ, O'Gorman L, Pengelly RJ, Beattie RM, MacArthur BD, Ennis S. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics*. 2019/05/16 2019;20(1):254. doi:10.1186/s12859-019-2877-3

82. Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. *Nature genetics*. 2013;45(6):580.

83. Ware JS, Li J, Mazaika E, et al. Shared genetic predisposition in peripartum and dilated cardiomyopathies. *New England Journal of Medicine*. 2016;374(3):233-241.

84. Roberts AM, Ware JS, Herman DS, et al. Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Science translational medicine*. 2015;7(270):270ra6-270ra6.

85. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS recommendations for the description of sequence variants: 2016 update. *Human mutation*. 2016;37(6):564-569.

86. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*. 2022;604(7905):310-315.

87. Cummings BB, Karczewski KJ, Kosmicki JA, et al. Transcript expression-aware annotation improves rare variant discovery and interpretation. *bioRxiv*. 2019:554444. doi:10.1101/554444

88. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human genetics*. 2013;132:1077-1130.

89. Kingdom R, Wright CF. Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. Review. *Frontiers in Genetics*. 2022-July-25 2022;13doi:10.3389/fgene.2022.920390

90. Chen R, Shi L, Hakenberg J, et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature biotechnology*. 2016;34(5):531.

91. Shawky RM. Reduced penetrance in human inherited disease. *Egyptian Journal of Medical Human Genetics*. 2014;15(2):103-111.

92. Ropers HH, Wienker T. Penetrance of pathogenic mutations in haploinsufficient genes for intellectual disability and related disorders. *European journal of medical genetics*. 2015;58(12):715-718.

93. Bresin E, Rurali E, Caprioli J, et al. Combined complement gene mutations in atypical hemolytic uremic syndrome influence clinical phenotype. *Journal of the American Society of Nephrology*. 2013;24(3):475-486.

94. Veitia RA, Caburet S, Birchler JA. Mechanisms of Mendelian dominance. *Clinical Genetics*. 2018;93(3):419-428. doi:10.1111/cge.13107

95. Gourraud J-B, Barc J, Thollet A, et al. The Brugada syndrome: a rare arrhythmia disorder with complex inheritance. *Frontiers in cardiovascular medicine*. 2016;3:9.

96. Marsh AP, Heron D, Edwards TJ, et al. Mutations in DCC cause isolated agenesis of the corpus callosum with incomplete penetrance. *Nature genetics*. 2017;49(4):511-514.

97. Fahsold R, Hoffmeyer S, Mischung C, et al. Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain. *The American Journal of Human Genetics*. 2000;66(3):790-818.

98. Oetjens M, Kelly M, Sturm A, Martin C, Ledbetter D. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nature communications*. 2019;10(1):4897.

99. Igarashi P, Somlo S. Genetics and pathogenesis of polycystic kidney disease. *Journal of the American Society of Nephrology*. 2002;13(9):2384-2398.

100. van der Kolk DM, de Bock GH, Leegte BK, et al. Penetrance of breast cancer, ovarian cancer and contralateral breast cancer in BRCA1 and BRCA2 families: high cancer incidence at older age. *Breast cancer research and treatment*. 2010;124:643-651.

101. Correa H. Li–Fraumeni syndrome. *Journal of pediatric genetics*. 2016:084-088.

102. Biller LH, Syngal S, Yurgelun MB. Recent advances in Lynch syndrome. *Familial cancer*. 2019;18:211-219.

103. Alessi DR, Sammler E. LRRK2 kinase in Parkinson's disease. *Science*. 2018;360(6384):36-37.

104. Wright CF, West B, Tuke M, et al. Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *The American Journal of Human Genetics*. 2019;104(2):275-286.

105. Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. *Nature Reviews Genetics*. 2018/07/01 2018;19(7):453-467. doi:10.1038/s41576-018-0007-0

106. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Human molecular genetics*. 2015;24(R1):R102-R110.

107. Spielmann M, Mundlos S. Looking beyond the genes: the role of non-coding variants in human disease. *Human molecular genetics*. 2016;25(R2):R157-R165.

108. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature genetics*. 2017;49(4):618-624.

109. Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. *Nature*. 2020/05/01 2020;581(7809):444-451. doi:10.1038/s41586-020-2287-8

110. Merker JD, Wenger AM, Sneddon T, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine*. 2018/01/01 2018;20(1):159-163. doi:10.1038/gim.2017.86

111. Marks P, Garcia S, Barrio AM, et al. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome research*. 2019;29(4):635-645.

112. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*. 2005;33(suppl_1):D514-D517.

113. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*. 2014;42(D1):D980-D985.

114. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*. 2009;84(4):524-533.

115. Manrai AK, Funke BH, Rehm HL, et al. Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine*. 2016;375(7):655-665. doi:10.1056/NEJMsa1507092

116. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291.

117. Shah N, Hou Y-CC, Yu H-C, Sainger R, Caskey CT, Venter JC, Telenti A. Identification of misclassified ClinVar variants via disease population prevalence. *The American Journal of Human Genetics*. 2018;102(4):609-619.

118. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature genetics*. 2019;51(11):1560-1565.

119. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genetics in Medicine*. 2017;19(2):209-214.

120. Nambot S, Thevenon J, Kuentz P, et al. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genetics in Medicine*. 2018;20(6):645-654.

121. Ewans LJ, Schofield D, Shrestha R, et al. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. *Genetics in Medicine*. 2018/12/01 2018;20(12):1564-1574. doi:10.1038/gim.2018.39

122. Posey JE, O'Donnell-Luria AH, Chong JX, et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genet Med*. 2019 04 2019;21(4):798-812. doi:10.1038/s41436-018-0408-7

123. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424. doi:10.1038/gim.2015.30

124. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, Biesecker LG. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine*. 2018;20(9):1054-1060.

125. Amendola LM, Jarvik GP, Leo MC, et al. Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the Clinical Sequencing Exploratory Research Consortium. *The American Journal of Human Genetics*. 2016;98(6):1067-1076.

126. Dewey FE, Grove ME, Pan C, et al. Clinical interpretation and implications of whole-genome sequencing. *Jama*. 2014;311(10):1035-1045.

127. Firth HV, Wright CF, study D. The deciphering developmental disorders (DDD) study. *Developmental Medicine & Child Neurology*. 2011;53(8):702-703.

128. Bamshad MJ, Shendure JA, Valle D, et al. The Centers for Mendelian Genomics: A new large-scale initiative to identify the genes underlying rare Mendelian conditions. *American Journal of Medical Genetics Part A*. 2012;158A(7):1523-1525. doi:https://doi.org/10.1002/ajmg.a.35470

129. Philippakis AA, Azzariti DR, Beltran S, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Human mutation*. 2015;36(10):915-921.

130. Snape K, Wedderburn S, Barwell J. The new genomic medicine service and implications for patients. *Clinical Medicine*. 2019;19(4):273.

131. Ainsworth C. Three ways genomics is already helping NHS patients—and three ways it will soon. *bmj*. 2022;379

132. 100 GPPI. 100,000 genomes pilot on rare-disease diagnosis in health care—preliminary report. *New England Journal of Medicine*. 2021;385(20):1868-1880.

133. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020/05/01 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

134. Smedley D, Jacobsen JO, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature protocols*. 2015;10(12):2004-2015.

135. Thevenon J, Duffourd Y, Masurel-Paulet A, et al. Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. *Clinical genetics*. 2016;89(6):700-707.

136. Partners ICH. *A preliminary assessment of the potential impact of rare diseases on the NHS*. 2018. Accessed July 2020. https://imperialcollegehealthpartners.com/wp-content/uploads/2019/05/ICHP-RD-Report-Nov-2018-APPROVED-002.pdf

137. Stark Z, Schofield D, Alam K, et al. Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. *Genetics in Medicine*. 2017;19(8):867-874.

138. Soden SE, Saunders CJ, Willig LK, et al. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Science translational medicine*. 2014;6(265):265ra168-265ra168.

139. Muir E. The rare reality-an insight into the patient and family experience of rare disease. *Rare Disease UK*. 2016;

140. Mirmiran A, Schmitt C, Lefebvre T, et al. Erythroid-progenitor-targeted gene therapy using bifunctional TFR1 ligand-peptides in human erythropoietic protoporphyria. *The American Journal of Human Genetics*. 2019;104(2):341-347.

141. Legendre CM, Licht C, Muus P, et al. Terminal complement inhibitor eculizumab in atypical hemolytic–uremic syndrome. *New England Journal of Medicine*. 2013;368(23):2169-2181.

142. Ribeil J-A, Hacein-Bey-Abina S, Payen E, et al. Gene therapy in a patient with sickle cell disease. *New England Journal of Medicine*. 2017;376(9):848-855.

143. Maguire AM, High KA, Auricchio A, et al. Age-dependent effects of RPE65 gene therapy for Leber's congenital amaurosis: a phase 1 dose-escalation trial. *The Lancet*. 2009;374(9701):1597-1605.

144. Pierce EA, Bennett J. The status of RPE65 gene therapy trials: safety and efficacy. *Cold Spring Harbor perspectives in medicine*. 2015;5(9):a017285.

145. Beck DB, Petracovici A, He C, et al. Delineation of a Human Mendelian Disorder of the DNA Demethylation Machinery: TET3 Deficiency. *Am J Hum Genet*. 2020 02 06 2020;106(2):234-245. doi:10.1016/j.ajhg.2019.12.007

146. Mendell JR, Al-Zaidy S, Shell R, et al. Single-dose gene-replacement therapy for spinal muscular atrophy. *New England Journal of Medicine*. 2017;377(18):1713-1722.

147. Habib A-RR, Kajbafzadeh M, Desai S, Yang CL, Skolnik K, Quon BS. A systematic review of the clinical efficacy and safety of CFTR modulators in cystic fibrosis. *Scientific reports*. 2019;9(1):1-9.

148. Wainwright CE, Elborn JS, Ramsey BW, et al. Lumacaftor–ivacaftor in patients with cystic fibrosis homozygous for Phe508del CFTR. *New England Journal of Medicine*. 2015;373(3):220-231.

149. Ly CV, Miller TM. Emerging antisense oligonucleotide and viral therapies for ALS. *Current opinion in neurology*. 2018;31(5):648.

150. Smith RA, Miller TM, Yamanaka K, et al. Antisense oligonucleotide therapy for neurodegenerative disease. *The Journal of clinical investigation*. 2006;116(8):2290-2296.

151. Watkins D, Schwartzentruber JA, Ganesh J, et al. Novel inborn error of folate metabolism: identification by exome capture and sequencing of mutations in the MTHFD1 gene in a single proband. *Journal of medical genetics*. 2011;48(9):590-592.

152. Burda P, Kuster A, Hjalmarson O, et al. Characterization and review of MTHFD1 deficiency: four new patients, cellular delineation and response to folic and folinic acid treatment. *Journal of inherited metabolic disease*. 2015;38(5):863-872.

153. Ramakrishnan KA, Pengelly RJ, Gao Y, et al. Precision molecular diagnosis defines specific therapy in combined immunodeficiency with megaloblastic anemia secondary to MTHFD1 deficiency. *The Journal of Allergy and Clinical Immunology: In Practice*. 2016;4(6):1160-1166. e10.

154. Chong JX, Buckingham KJ, Jhangiani SN, et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *The American Journal of Human Genetics*. 2015;97(2):199-215.

155. Bamshad MJ, Nickerson DA, Chong JX. Mendelian Gene Discovery: Fast and Furious with No End in Sight. *The American Journal of Human Genetics*. 2019/09/05/ 2019;105(3):448-455. doi:https://doi.org/10.1016/j.ajhg.2019.07.011

156. Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics*. 2010/09/01 2010;42(9):790-793. doi:10.1038/ng.646

157. Azzariti DR, Hamosh A. Genomic Data Sharing for Novel Mendelian Disease Gene Discovery: The Matchmaker Exchange. *Annual Review of Genomics and Human Genetics*. 2020;21(1):null. doi:10.1146/annurev-genom-083118-014915

158. Directors ABo. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. *Genetics in Medicine*. 2017;19(7):721-722.

159. Phillips M, Molnár-Gábor F, Korbel JO, et al. Genomics: data sharing needs an international code of conduct. Nature Publishing Group; 2020.

160. O'Donnell-Luria AH, Pais LS, Faundes V, et al. Heterozygous Variants in KMT2E Cause a Spectrum of Neurodevelopmental Disorders and Epilepsy. *The American Journal of Human Genetics*. 2019/06/06/ 2019;104(6):1210-1222. doi:https://doi.org/10.1016/j.ajhg.2019.03.021

161. Kaye J, Curren L, Anderson N, et al. From patients to partners: participant-centric initiatives in biomedical research. *Nature Reviews Genetics*. 2012;13(5):371-376.

162. Lambertson KF, Damiani SA, Might M, Shelton R, Terry SF. Participant-Driven Matchmaking in the Genomic Era. *Human Mutation*. 2015;36(10):965-973. doi:10.1002/humu.22852

163. Macnamara EF, D'Souza P, Undiagnosed Diseases N, Tifft CJ. The undiagnosed diseases program: Approach to diagnosis. *Translational Science of Rare Diseases*. 2019;4:179-188. doi:10.3233/TRD-190045

164. Might M, Might CC. What happens when N= 1 and you want plus 1? *Prenatal diagnosis*. 2017;37(1):70-72.

165. Might M, Wilsey M. The shifting model in clinical diagnostics: how next-generation sequencing and families are altering the way rare diseases are discovered, studied, and treated. *Genetics in Medicine*. 2014;16(10):736-737.

166. Chong JX, Yu J-H, Lorentzen P, et al. Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. *Genetics in Medicine*. 2016/08/01 2016;18(8):788-795. doi:10.1038/gim.2015.161

167. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*. 2008;83(5):610-615.

168. Bauer S, Köhler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics*. 2012;28(19):2502-2508.

169. Schulz MH, Köhler S, Bauer S, Robinson PN. Exact score distribution computation for ontological similarity searches. *BMC bioinformatics*. 2011;12(1):441.

170. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*. 2009;85(4):457-464.

171. Sifrim A, Popovic D, Tranchevent L-C, et al. eXtasy: variant prioritization by genomic data fusion. *Nature methods*. 2013;10(11):1083-1084.

172. Singleton MV, Guthery SL, Voelkerding KV, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *The American Journal of Human Genetics*. 2014;94(4):599-610.

173. Javed A, Agrawal S, Ng PC. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nature methods*. 2014;11(9):935-937.

174. Mungall CJ, McMurry JA, Köhler S, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*. 2017;45(D1):D712-D722.

175. Alliance of Genome Resources Portal: unified model organism research platform. *Nucleic acids research*. 2020;48(D1):D650-D658.

176. Wang J, Al-Ouran R, Hu Y, et al. MARRVEL: integration of human and model organism genetic resources to facilitate functional annotation of the human genome. *The American Journal of Human Genetics*. 2017;100(6):843-853.

177. Smedley D, Schubach M, Jacobsen JO, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *The American Journal of Human Genetics*. 2016;99(3):595-606.

178. Bone WP, Washington NL, Buske OJ, et al. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine*. 2016;18(6):608-617.

179. Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Group tMGD. Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Research*. 2017;46(D1):D836-D842. doi:10.1093/nar/gkx1006

180. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. Mouse genome database (MGD) 2019. *Nucleic acids research*. 2019;47(D1):D801-D806.

181. Austin CP, Battey JF, Bradley A, et al. The knockout mouse project. *Nature genetics*. 2004;36(9):921.

182. Meehan TF, Conte N, West DB, et al. Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nature genetics*. 2017;49(8):1231-1238.

183. Muñoz-Fuentes V, Cacheiro P, Meehan TF, et al. The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conservation Genetics*. 2018;19(4):995-1005.

184. Cacheiro P, Haendel MA, Smedley D, et al. New models for human disease from the International Mouse Phenotyping Consortium. *Mammalian Genome*. 2019/06/01 2019;30(5):143-150. doi:10.1007/s00335-019-09804-5

185. Bowl MR, Simon MM, Ingham NJ, et al. A large scale hearing loss screen reveals an extensive unexplored genetic landscape for auditory dysfunction. *Nature Communications*. 2017/10/12 2017;8(1):886. doi:10.1038/s41467-017-00595-4

186. Rozman J, Rathkolb B, Oestereicher MA, et al. Identification of genetic elements in metabolism by high-throughput mouse phenotyping. *Nature communications*. 2018;9(1):1-16.

187. Moore BA, Leonard BC, Sebbag L, et al. Identification of genes required for eye development by high-throughput screening of mouse knockouts. *Communications biology*. 2018;1(1):1-12.

188. Gruber C, Bogunovic D. Incomplete penetrance in primary immunodeficiency: a skeleton in the closet. *Hum Genet*. 2020;139(6):745-757.

189. Castel SE, Cervera A, Mohammadi P, et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet*. Sep 2018;50(9):1327-1334. doi:10.1038/s41588-018-0192-y

190. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biology*. 2017/05/05 2017;18(1):83. doi:10.1186/s13059-017-1215-1

191. Stranger BE, Brigham LE, Hasz R, et al. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease The eGTEx Project. *Nature genetics*. 2017;49(12):1664.

192. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318-1330.

193. Kremer LS, Bader DM, Mertes C, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature communications*. 2017;8(1):1-11.

194. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science translational medicine*. 2017;9(386)

195. Brechtmann F, Mertes C, Matusevičiūtė A, et al. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am J Hum Genet*. Dec 6 2018;103(6):907-917. doi:10.1016/j.ajhg.2018.10.025

196. Green CJ, Gazzara MR, Barash Y. MAJIQ-SPEL: web-tool to interrogate classical and complex splicing variations from RNA-Seq data. *Bioinformatics*. Jan 15 2018;34(2):300-302. doi:10.1093/bioinformatics/btx565

197. Mertes C, Scheller IF, Yépez VA, et al. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat Commun*. Jan 22 2021;12(1):529. doi:10.1038/s41467-020-20573-7

198. Deelen P, van Dam S, Herkert JC, et al. Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nature communications*. 2019;10(1):1-13.

199. Aicher JK, Jewell P, Vaquero-Garcia J, Barash Y, Bhoj EJ. Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genet Med*. Jul 2020;22(7):1181-1190. doi:10.1038/s41436-020-0780-y

200. Rowlands CF, Taylor A, Rice G, et al. MRSD: a novel quantitative approach for assessing suitability of RNA-seq in the clinical investigation of mis-splicing in Mendelian disease. *medRxiv*. 2021:2021.03.19.21253973. doi:10.1101/2021.03.19.21253973

201. Satterlee JS, Chadwick LH, Tyson FL, et al. The NIH common fund/roadmap epigenomics program: Successes of a comprehensive consortium. *Science advances*. 2019;5(7):eaaw6507.

202. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH roadmap epigenomics mapping consortium. *Nature biotechnology*. 2010;28(10):1045-1048.

203. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012/09/01 2012;489(7414):57-74. doi:10.1038/nature11247

204. Aref-Eshghi E, Bend EG, Colaiacovo S, et al. Diagnostic Utility of Genome-wide DNA Methylation Testing in Genetically Unsolved Individuals with Suspected Hereditary Conditions. *American journal of human genetics*. 2019;104(4):685-700. doi:10.1016/j.ajhg.2019.03.008

205. Turinsky AL, Choufani S, Lu K, et al. EpigenCentral: Portal for DNA methylation data analysis and classification in rare diseases. *Human Mutation*. 2020;41(10):1722-1733. doi:https://doi.org/10.1002/humu.24076

206. LaCroix AJ, Stabley D, Sahraoui R, et al. GGC Repeat Expansion and Exon 1 Methylation of XYLT1 Is a Common Pathogenic Variant in Baratela-Scott Syndrome. *Am J Hum Genet*. 2019 Jan 03 2019;104(1):35-44. doi:10.1016/j.ajhg.2018.11.005

207. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209.

208. Unlu G, Qi X, Gamazon ER, et al. Phenome-based approach identifies RIC1-linked Mendelian syndrome through zebrafish models, biobank associations and clinical studies. *Nat Med*. Jan 2020;26(1):98-109. doi:10.1038/s41591-019-0705-y

209. Almontashiri NAM, Zha L, Young K, Law T, Kellogg MD, Bodamer OA, Peake RWA. Clinical Validation of Targeted and Untargeted Metabolomics Testing for Genetic Disorders: A 3 Year Comparative Study. *Sci Rep*. Jun 10 2020;10(1):9382. doi:10.1038/s41598-020-66401-2

210. Graham E, Lee J, Price M, et al. Integration of genomics and metabolomics for prioritization of rare disease variants: a 2018 literature review. *J Inherit Metab Dis*. May 2018;41(3):435-445. doi:10.1007/s10545-018-0139-6

211. Claussnitzer M, Dankel SN, Kim K-H, et al. FTO obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*. 2015;373(10):895-907.

212. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*. 2014/09/01 2014;46(9):944-950. doi:10.1038/ng.3050

213. Kosmicki JA, Samocha KE, Howrigan DP, et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nature genetics*. 2017;49(4):504.

214. Cummings BB, Karczewski KJ, Kosmicki JA, et al. Transcript expression-aware annotation improves rare variant interpretation. *Nature*. 2020/05/01 2020;581(7809):452-458. doi:10.1038/s41586-020-2329-2

215. Pérez-Palma E, Gramm M, Nürnberg P, May P, Lal D. Simple ClinVar: an interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database. *Nucleic Acids Research*. 2019;47(W1):W99-W105. doi:10.1093/nar/gkz411

216. Hayeck TJ, Stong N, Wolock CJ, Copeland B, Kamalakaran S, Goldstein DB, Allen AS. Improved Pathogenic Variant Localization via a Hierarchical Model of Sub-regional Intolerance. *Am J Hum Genet*. Feb 7 2019;104(2):299-309. doi:10.1016/j.ajhg.2018.12.020

217. Pérez-Palma E, May P, Iqbal S, et al. Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res*. Jan 2020;30(1):62-71. doi:10.1101/gr.252601.119

218. Abramovs N, Brass A, Tassabehji M. GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. *Nature Genetics*. 2020/01/01 2020;52(1):35-39. doi:10.1038/s41588-019-0560-2

219. Turner TN, Douville C, Kim D, Stenson PD, Cooper DN, Chakravarti A, Karchin R. Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. *Human molecular genetics*. 2015;24(21):5995-6002.

220. Orenstein N, Goldberg-Stern H, Straussberg R, et al. A de novo GABRA2 missense mutation in severe early-onset epileptic encephalopathy with a choreiform movement disorder. *European Journal of Paediatric Neurology*. 2018;22(3):516-524.

221. Posey JE, O'Donnell-Luria AH, Chong JX, et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genetics in Medicine*. 2019/04/01 2019;21(4):798-812. doi:10.1038/s41436-018-0408-7

222. The 100,000 Genomes Project. Parliamentary Office of Science and Technology; 2015.

223. Kerem B-s, Rommens JM, Buchanan JA, et al. Identification of the cystic fibrosis gene: genetic analysis. *Science*. 1989;245(4922):1073-1080.

224. Ramsey BW, Davies J, McElvaney NG, et al. A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *New England Journal of Medicine*. 2011;365(18):1663-1672.

225. Farrell PM, Rock MJ, Baker MW. The Impact of the CFTR Gene Discovery on Cystic Fibrosis Diagnosis, Counseling, and Preventive Therapy. *Genes*. 2020;11(4):401.

226. Timms KM, Wagner S, Samuels ME, et al. A mutation in PCSK9 causing autosomal-dominant hypercholesterolemia in a Utah pedigree. *Hum Genet*. 2004/03/01 2004;114(4):349-353. doi:10.1007/s00439-003-1071-9

227. Kereiakes DJ, Robinson JG, Cannon CP, Lorenzato C, Pordy R, Chaudhari U, Colhoun HM. Efficacy and safety of the proprotein convertase subtilisin/kexin type 9 inhibitor alirocumab among high cardiovascular risk patients on maximally tolerated statin therapy: the ODYSSEY COMBO I study. *American heart journal*. 2015;169(6):906-915. e13.

228. Cannon CP, Cariou B, Blom D, et al. Efficacy and safety of alirocumab in high cardiovascular risk patients with inadequately controlled hypercholesterolaemia on maximally tolerated doses of statins: the ODYSSEY COMBO II randomized controlled trial. *European heart journal*. 2015;36(19):1186-1194.

229. Roth EM, Taskinen M-R, Ginsberg HN, et al. Monotherapy with the PCSK9 inhibitor alirocumab versus ezetimibe in patients with hypercholesterolemia: results of a 24 week, double-blind, randomized Phase 3 trial. *International journal of cardiology*. 2014;176(1):55-61.

230. Blom DJ, Hala T, Bolognese M, et al. A 52-week placebo-controlled trial of evolocumab in hyperlipidemia. *N Engl J Med*. 2014;370:1809-1819.

231. Tarailo-Graovac M, Shyr C, Ross CJ, et al. Exome sequencing and the management of neurometabolic disorders. *New England Journal of Medicine*. 2016;374(23):2246-2255.

232. Anderson M, Elliott EJ, Zurynski YA. Australian families living with rare disease: experiences of diagnosis, health services use and needs for psychosocial support. *Orphanet journal of rare diseases*. 2013;8(1):22.

233. Zurynski Y, Deverell M, Dalkeith T, et al. Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. *Orphanet Journal of Rare Diseases*. 2017;12(1):68.

234. Lingen M, Albers L, Borchers M, et al. Obtaining a genetic diagnosis in a child with disability: impact on parental quality of life. *Clinical genetics*. 2016;89(2):258-266.

235. Sexton A, Sahhar M, Thorburn D, Metcalfe S. Impact of a genetic diagnosis of a mitochondrial disorder 5–17 years after the death of an affected child. *Journal of genetic counseling*. 2008;17(3):261-273.

236. Wojcik MH, Stewart JE, Waisbren SE, Litt JS. Developmental Support for Infants With Genetic Disorders. *Pediatrics*. 2020;145(5)

237. Tan TY, Dillon OJ, Stark Z, et al. Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. *JAMA pediatrics*. 2017;171(9):855-862.

238. Strande NT, Riggs ER, Buchanan AH, et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource. *The American Journal of Human Genetics*. 2017;100(6):895-906.

239. Bean LJH, Funke B, Carlston CM, et al. Diagnostic gene sequencing panels: from design to report-a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. Mar 2020;22(3):453-461. doi:10.1038/s41436-019-0666-z

240. Rehder C, Bean LJH, Bick D, et al. Next-generation sequencing for constitutional variants in the clinical laboratory, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. Apr 29 2021;doi:10.1038/s41436-021-01139-4

241. Mnookin S. One of a Kind. The New Yorker2014.

242. Birney E, Vamathevan J, Goodhand P. Genomics in healthcare: GA4GH looks to 2022. *bioRxiv*. 2017:203554. doi:10.1101/203554

243. Lupiáñez DG, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. May 21 2015;161(5):1012-1025. doi:10.1016/j.cell.2015.04.004

244. Small KW, DeLuca AP, Whitmore SS, et al. North Carolina Macular Dystrophy Is Caused by Dysregulation of the Retinal Transcription Factor PRDM13. *Ophthalmology*. Jan 2016;123(1):9-18. doi:10.1016/j.ophtha.2015.10.006

245. Green DJ, Lenassi E, Manning CS, et al. North Carolina macular dystrophy: phenotypic variability and computational analysis of disease-implicated non-coding variants. *medRxiv*. 2021:2021.03.05.21252975. doi:10.1101/2021.03.05.21252975

246. Matreyek KA, Starita LM, Stephany JJ, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature genetics*. 2018;50(6):874-882.

247. Jepsen MM, Fowler DM, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Classifying disease-associated variants using measures of protein activity and stability. *Protein Homeostasis Diseases*. Elsevier; 2020:91-107.

248. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science*. 2003;300(5617):286-290.

249. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*. 2017;109(2):83-90.

250. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat*. 2015 Oct 2015;36(10):928-30. doi:10.1002/humu.22844

251. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842.

252. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome biology*. 2016;17(1):1-14.

253. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nature Biotechnology*. 2011/01/01 2011;29(1):24-26. doi:10.1038/nbt.1754

254. Faber J, Nieuwkoop PD. *Normal table of Xenopus Laevis (Daudin): a systematical & chronological survey of the development from the fertilized egg till the end of metamorphosis*. Garland Science; 2020.

255. Moreno-Mateos MA, Vejnar CE, Beaudoin J-D, Fernandez JP, Mis EK, Khokha MK, Giraldez AJ. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature methods*. 2015;12(10):982-988.

256. Shen MW, Arbab M, Hsu JY, et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*. 2018;563(7733):646-651.

257. Nakayama T, Blitz IL, Fish MB, Odeleye AO, Manohar S, Cho KW, Grainger RM. Cas9-based genome editing in Xenopus tropicalis. *Methods in enzymology*. Elsevier; 2014:355-375.

258. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3—new capabilities and interfaces. *Nucleic acids research*. 2012;40(15):e115-e115.

259. Sambrook J, Fritsch EF, Maniatis T. *Molecular cloning: a laboratory manual*. Cold spring harbor laboratory press; 1989.

260. Guille M. *Molecular Methods in Developmental Biology*. Springer; 1999.

261. Ismail V, Zachariassen LG, Godwin A, et al. Identification and functional evaluation of GRIA1 missense and truncation variants in individuals with ID: An emerging neurodevelopmental syndrome. *Am J Hum Genet*. Jul 7 2022;109(7):1217-1241. doi:10.1016/j.ajhg.2022.05.009

262. Cleal M, Fontana BD, Ranson DC, McBride SD, Swinny JD, Redhead ES, Parker MO. The Free-movement pattern Y-maze: A cross-species measure of working memory and executive function. *Behav Res Methods*. Apr 2021;53(2):536-557. doi:10.3758/s13428-020-01452-x

263. Hand R, Polleux F. Neurogenin2 regulates the initial axon guidance of cortical pyramidal neurons projecting medially to the corpus callosum. *Neural development*. 2011;6(1):1-16.

264. Courchet J, Lewis TL, Lee S, Courchet V, Liou D-Y, Aizawa S, Polleux F. Terminal axon branching is regulated by the LKB1-NUAK1 kinase pathway via presynaptic mitochondrial capture. *Cell*. 2013;153(7):1510-1525.

265. Polleux F, Ghosh A. The Slice Overlay Assay: A Versatile Tool to Study the Influence of Extracellular Signals on Neuronal Development. *Science's STKE*. 2002;2002(136):pl9-pl9. doi:doi:10.1126/stke.2002.136.pl9

266. Meyer-Dilhet G, Courchet J. In utero cortical electroporation of plasmids in the mouse embryo. *STAR protocols*. 2020;1(1):100027.

267. Hand R, Bortone D, Mattar P, et al. Phosphorylation of Neurogenin2 Specifies the Migration Properties and the Dendritic Morphology of Pyramidal Neurons in the Neocortex. *Neuron*. 2005/10/06/ 2005;48(1):45-62. doi:https://doi.org/10.1016/j.neuron.2005.08.032

268. Dardenne E, Polay Espinoza M, Fattet L, et al. RNA helicases DDX5 and DDX17 dynamically orchestrate transcription, miRNA, and splicing programs in cell differentiation. *Cell Rep*. Jun 26 2014;7(6):1900-13. doi:10.1016/j.celrep.2014.05.010

269. Terrone S, Valat J, Fontrodona N, et al. RNA helicase-dependent gene looping impacts messenger RNA processing. *Nucleic Acids Res*. Sep 9 2022;50(16):9226-9246. doi:10.1093/nar/gkac717

270. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. Sep 1 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560

271. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. Aug 2019;37(8):907-915. doi:10.1038/s41587-019-0201-4

272. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. Feb 16 2021;10(2)doi:10.1093/gigascience/giab008

273. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. Jan 15 2015;31(2):166-9. doi:10.1093/bioinformatics/btu638

274. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

275. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*. Apr 15 2020;36(8):2628-2629. doi:10.1093/bioinformatics/btz931

276. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*. 05/2015 2015;17(5):405-424. doi:10.1038/gim.2015.30

277. Strande NT, Brnich SE, Roman TS, Berg JS. Navigating the nuances of clinical sequence variant interpretation in Mendelian disease. *Genetics in Medicine*. 2018/09/01/ 2018;20(9):918-926. doi:https://doi.org/10.1038/s41436-018-0100-y

278. Zearfoss NR, Ryder SP. End-labeling oligonucleotides with chemical tags after synthesis. *Recombinant and in vitro RNA synthesis: methods and protocols*. 2012:181-193.

279. Mao C, Flavin KG, Wang S, Dodson R, Ross J, Shapiro DJ. Analysis of RNA–protein interactions by a microplate-based fluorescence anisotropy assay. *Analytical biochemistry*. 2006;350(2):222-232.

280. Gudmundsson S, Karczewski KJ, Francioli LC, et al. Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2021:1-2.

281. Riggs ER, Nelson T, Merz A, et al. Copy number variant discrepancy resolution using the ClinGen dosage sensitivity map results in updated clinical interpretations in ClinVar. *Human mutation*. 2018;39(11):1650-1659.

282. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*. 2011;27(13):i275-i282.

283. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome research*. 2002;12(6):996-1006.

284. Singer-Berk M, Gudmundsson S, Baxter S, et al. Advanced variant classification framework reduces the false positive rate of predicted loss of function (pLoF) variants in population sequencing data. *medRxiv*. 2023:2023.03.08.23286955. doi:10.1101/2023.03.08.23286955

285. Seaby EG, Bunyan DJ, Ennis S, Gilbert RD. Sporadic, isolated Fanconi syndrome due to a mutation of EHHADH: a case report. *Journal of Clinical Nephrology and Renal Care*. 2017;3(2)

286. Brunet T, Berutti R, Dill V, et al. Clonal hematopoiesis as a pitfall in germline variant interpretation in the context of Mendelian disorders. *Human Molecular Genetics*. 2022;31(14):2386-2395. doi:10.1093/hmg/ddac034

287. Pich O, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. Discovering the drivers of clonal hematopoiesis. *Nature Communications*. 2022;13(1):4267.

288. Janssen BD, van den Boogaard MJH, Lichtenbelt K, et al. De novo putative loss-of-function variants in TAF4 are associated with a neuro-developmental disorder. *Human Mutation*. 2022;43(12):1844-1851.

289. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet*. 2015 Aug 06 2015;97(2):199-215. doi:10.1016/j.ajhg.2015.06.009

290. Kaplanis J, Samocha KE, Wiel L, et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*. 2020/10/01 2020;586(7831):757-762. doi:10.1038/s41586-020-2832-5

291. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.

292. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Human mutation*. 2018;39(11):1517-1524. doi:10.1002/humu.23626

293. Liu P, Meng L, Normand EA, et al. Reanalysis of clinical exome sequencing data. *New England Journal of Medicine*. 2019;380(25):2478-2480.

294. Seaby EG, Baralle D, Rehm HL, O'Donnell-Luria A, Ennis S. Response to Ramos et al. *Genetics in Medicine*. 2022;24(12):2593-2594.

295. Wilfert AB, Sulovari A, Turner TN, Coe BP, Eichler EE. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Medicine*. 2017/11/27 2017;9(1):101. doi:10.1186/s13073-017-0498-x

296. Zhang X, Wakeling M, Ware J, Whiffin N. Annotating high-impact 5′ untranslated region variants with the UTRannotator. *Bioinformatics*. 2021;37(8):1171-1173.

297. Schatz MC, Philippakis AA, Afgan E, et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). *bioRxiv*. 2021;

298. Bourgeois CF, Mortreux F, Auboeuf D. The multiple functions of RNA helicases as drivers and regulators of gene expression. *Nat Rev Mol Cell Biol*. Jul 2016;17(7):426-38. doi:10.1038/nrm.2016.50

299. Bohnsack KE, Yi S, Venus S, Jankowsky E, Bohnsack MT. Cellular functions of eukaryotic RNA helicases and their links to human diseases. *Nat Rev Mol Cell Biol*. Jul 20 2023;doi:10.1038/s41580-023-00628-5

300. Xing Z, Ma WK, Tran EJ. The DDX5/Dbp2 subfamily of DEAD-box RNA helicases. *Wiley Interdiscip Rev RNA*. Mar 2019;10(2):e1519. doi:10.1002/wrna.1519

301. Giraud G, Terrone S, Bourgeois CF. Functions of DEAD box RNA helicases DDX5 and DDX17 in chromatin organization and transcriptional regulation. *BMB Rep*. Dec 2018;51(12):613-622. doi:10.5483/BMBRep.2018.51.12.234

302. Fuller-Pace FV. The DEAD box proteins DDX5 (p68) and DDX17 (p72): multi-tasking transcriptional regulators. *Biochim Biophys Acta*. Aug 2013;1829(8):756-63. doi:10.1016/j.bbagrm.2013.03.004

303. Lambert M-P, Terrone S, Giraud G, et al. The RNA helicase DDX17 controls the transcriptional activity of REST and the expression of proneural microRNAs in neuronal differentiation. *Nucleic acids research*. 2018;46(15):7686-7700.

304. Suthapot P, Xiao T, Felsenfeld G, Hongeng S, Wongtrakoongate P. The RNA helicases DDX5 and DDX17 facilitate neural differentiation of human pluripotent stem cells NTERA2. *Life Sciences*. 2022;291:120298.

305. Seaby EG, Thomas NS, Webb A, et al. Targeting de novo loss-of-function variants in constrained disease genes improves diagnostic rates in the 100,000 Genomes Project. *Human Genetics*. 2023;142(3):351-362.

306. Seaby EG, Smedley D, Tavares ALT, et al. A gene-to-patient approach uplifts novel disease gene discovery and identifies 18 putative novel disease genes. *Genetics in Medicine*. 2022;24(8):1697-1707.

307. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Human mutation*. 2015;36(10):928-930.

308. Nakayama T, Fish MB, Fisher M, Oomen-Hajagos J, Thomsen GH, Grainger RM. Simple and efficient CRISPR/Cas9-mediated targeted mutagenesis in Xenopus tropicalis. *Genesis*. Dec 2013;51(12):835-43. doi:10.1002/dvg.22720

309. Panthi S, Chapman PA, Szyszka P, Beck CW. Characterisation and automated quantification of induced seizure-related behaviours in Xenopus laevis tadpoles. *J Neurochem*. May 2 2023;doi:10.1111/jnc.15836

310. Hewapathirane DS, Haas K. The Albino Xenopus laevis Tadpole as a Novel Model of Developmental Seizures. *Animal Models of Epilepsy: Methods and Innovations*. 2009:45-57.

311. Fontana BD, Alnassar N, Parker MO. The zebrafish (Danio rerio) anxiety test battery: comparison of behavioral responses in the novel tank diving and light-dark tasks following exposure to anxiogenic and anxiolytic compounds. *Psychopharmacology (Berl)*. Jan 2022;239(1):287-296. doi:10.1007/s00213-021-05990-w

312. Ngo TD, Partin AC, Nam Y. RNA Specificity and Autoregulation of DDX17, a Modulator of MicroRNA Biogenesis. *Cell Rep*. Dec 17 2019;29(12):4024-4035.e5. doi:10.1016/j.celrep.2019.11.059

313. Song QX, Liu NN, Liu ZX, Zhang YZ, Rety S, Hou XM, Xi XG. Nonstructural N- and C-tails of Dbp2 confer the protein full helicase activities. *J Biol Chem*. May 2023;299(5):104592. doi:10.1016/j.jbc.2023.104592

314. Baek ST, Kerjan G, Bielas SL, Lee JE, Fenstermaker AG, Novarino G, Gleeson JG. Off-target effect of doublecortin family shRNA on neuronal migration associated with endogenous microRNA dysregulation. *Neuron*. 2014;82(6):1255-1262.

315. Volvert M-L, Prevot P-P, Close P, et al. MicroRNA targeting of CoREST controls polarization of migrating cortical neurons. *Cell reports*. 2014;7(4):1168-1183.

316. De Robertis EM, Gurdon JB. A Brief History of Xenopus in Biology. *Cold Spring Harb Protoc*. Dec 1 2021;2021(12)doi:10.1101/pdb.top107615

317. Kostiuk V, Khokha MK. Xenopus as a platform for discovery of genes relevant to human disease. *Curr Top Dev Biol*. 2021;145:277-312. doi:10.1016/bs.ctdb.2021.03.005

318. Hellsten U, Harland RM, Gilchrist MJ, et al. The genome of the Western clawed frog Xenopus tropicalis. *Science*. Apr 30 2010;328(5978):633-6. doi:10.1126/science.1183670

319. Blum M, Ott T. Xenopus: An Undervalued Model Organism to Study and Model Human Genetic Disease. *Cells Tissues Organs*. 2018;205(5-6):303-313. doi:10.1159/000490898

320. Macken WL, Godwin A, Wheway G, et al. Biallelic variants in COPB1 cause a novel, severe intellectual disability syndrome with cataracts and variable microcephaly. *Genome Med*. Feb 25 2021;13(1):34. doi:10.1186/s13073-021-00850-w

321. Rouf MA, Wen L, Mahendra Y, et al. The recent advances and future perspectives of genetic compensation studies in the zebrafish model. *Genes & Diseases*. 2023/03/01/ 2023;10(2):468-479. doi:https://doi.org/10.1016/j.gendis.2021.12.003

322. DiStefano MT, Goehringer S, Babb L, et al. The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources. *Genet Med*. May 4 2022;doi:10.1016/j.gim.2022.04.017

323. Bowes JB, Snyder KA, Segerdell E, et al. Xenbase: a Xenopus biology and genomics resource. *Nucleic acids research*. 2007;36(suppl_1):D761-D767.

324. Zinnall U, Milek M, Minia I, et al. HDLBP binds ER-targeted mRNAs by multivalent interactions to promote protein synthesis of transmembrane and secreted proteins. *Nature Communications*. 2022/05/18 2022;13(1):2727. doi:10.1038/s41467-022-30322-7

325. Danilchik MV, Brown EE, Riegert K. Intrinsic chiral properties of the Xenopus egg cortex: an early indicator of left-right asymmetry? 2006;

326. Rehm HL. Time to make rare disease diagnosis accessible to all. *Nature Medicine*. 2022/02/01 2022;28(2):241-242. doi:10.1038/s41591-021-01657-3

327. Gudmundsson S, Karczewski KJ, Francioli LC, et al. Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2021/09/01 2021;597(7874):E3-E4. doi:10.1038/s41586-021-03758-y

328. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids research*. 2015;43(D1):D789-D798.

329. Seaby EG, Smedley D, Taylor Tavares AL, et al. A gene-to-patient approach uplifts novel disease gene discovery and identifies 18 putative novel disease genes. *Genetics in Medicine*. doi:10.1016/j.gim.2022.04.019

330. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*. 2022:1-6.

331. Bush LW, Beck AE, Biesecker LG, et al. Professional responsibilities regarding the provision, publication, and dissemination of patient phenotypes in the context of clinical genetic and genomic testing: points to consider—a statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*. 2018;20(2):169-171.

332. Miller DT, Lee K, Chung WK, et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*. 2021/08/01 2021;23(8):1381-1390. doi:10.1038/s41436-021-01172-3

333. Austin-Tse CA, Jobanputra V, Perry DL, et al. Best practices for the interpretation and reporting of clinical whole genome sequencing. *NPJ Genomic Medicine*. 2022;7(1):1-13.

334. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nature genetics*. 2014;46(9):944-950.

335. McNally EM, Mestroni L. Dilated cardiomyopathy: genetic determinants and mechanisms. *Circulation research*. 2017;121(7):731-748.

336. Marian AJ, Braunwald E. Hypertrophic cardiomyopathy: genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circulation research*. 2017;121(7):749-770.

337. Sabater-Molina M, Pérez-Sánchez I, Hernández Del Rincón J, Gimeno J. Genetics of hypertrophic cardiomyopathy: A review of current state. *Clinical genetics*. 2018;93(1):3-14.

338. Akhtar M, Elliott P. The genetics of hypertrophic cardiomyopathy. *Glob Cardiol Sci Pract*. Aug 12 2018;2018(3):36. doi:10.21542/gcsp.2018.36

339. Hershberger RE, Jordan E. Dilated cardiomyopathy overview. 2022;

340. Pugh TJ, Kelly MA, Gowrisankar S, et al. The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. *Genetics in Medicine*. 2014;16(8):601-608.

341. Schafer S, De Marvao A, Adami E, et al. Titin-truncating variants affect heart function in disease cohorts and the general population. *Nature genetics*. 2017;49(1):46-53.

342. Santiago CF, Huttner IG, Fatkin D. Mechanisms of TTN tv-related dilated cardiomyopathy: insights from zebrafish models. *Journal of Cardiovascular Development and Disease*. 2021;8(2):10.

343. Basso C, Corrado D, Marcus FI, Nava A, Thiene G. Arrhythmogenic right ventricular cardiomyopathy. *The Lancet*. 2009;373(9671):1289-1300.

344. Marcus FI, Edson S, Towbin JA. Genetics of arrhythmogenic right ventricular cardiomyopathy: a practical guide for physicians. *Journal of the American College of Cardiology*. 2013;61(19):1945-1948.

345. SEN-CHOWDHRY S, Syrris P, McKenna WJ. Genetics of right ventricular cardiomyopathy. *Journal of cardiovascular electrophysiology*. 2005;16(8):927-935.

346. Vimalanathan AK, Ehler E, Gehmlich K. Genetics of and pathogenic mechanisms in arrhythmogenic right ventricular cardiomyopathy. *Biophysical Reviews*. 2018;10(4):973-982.

347. Towbin JA, Lorts A, Jefferies JL. Left ventricular non-compaction cardiomyopathy. *The Lancet*. 2015;386(9995):813-825.

348. Finsterer J, Stoellberger C, Towbin JA. Left ventricular noncompaction cardiomyopathy: cardiac, neuromuscular, and genetic factors. *Nature Reviews Cardiology*. 2017;14(4):224-237.

349. Rojanasopondist P, Nesheiwat L, Piombo S, Porter Jr GA, Ren M, Phoon CK. Genetic basis of left ventricular noncompaction. *Circulation: Genomic and Precision Medicine*. 2022;15(3):e003517.

350. Dong X, Fan P, Tian T, et al. Recent advancements in the molecular genetics of left ventricular noncompaction cardiomyopathy. *Clinica Chimica Acta*. 2017;465:40-44.

351. Lorca R, Martín M, Pascual I, et al. Characterization of left ventricular non-compaction cardiomyopathy. *Journal of Clinical Medicine*. 2020;9(8):2524.

352. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020 05 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

353. Pathania M, Davenport E, Muir J, Sheehan D, López-Doménech G, Kittler J. The autism and schizophrenia associated gene CYFIP1 is critical for the maintenance of dendritic complexity and the stabilization of mature spines. *Translational psychiatry*. 2014;4(3):e374-e374.

354. Seaby EG, Thomas NS, Webb A, et al. Targeting de novo loss-of-function variants in constrained disease genes improves diagnostic rates in the 100,000 Genomes Project. *Hum Genet*. 2022:1-12.

355. Bozdagi O, Sakurai T, Dorr N, Pilorge M, Takahashi N, Buxbaum JD. Haploinsufficiency of Cyfip1 produces fragile X-like phenotypes in mice. 2012;

356. Seaby EG, Thomas S, Hunt D, Baralle D, Rehm HL, O'Donnell-Luria AL, Ennis S. A panel-agnostic strategy" HiPPo" improves diagnostic efficiency in the UK Genome Medicine Service. *medRxiv*. 2023:2023.01. 31.23285025.

357. Seaby EG, Leggatt G, Cheng G, et al. A gene pathogenicity tool 'GenePy'identifies missed biallelic diagnoses in the 100,000 Genomes Project. *medRxiv*. 2023:2023.03. 21.23287545.

358. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014 Mar 2014;46(3):310-5. doi:10.1038/ng.2892

359. Gambin T, Jhangiani SN, Below JE, et al. Secondary findings and carrier test frequencies in a large multiethnic sample. *Genome Med*. 2015 2015;7(1):54. doi:10.1186/s13073-015-0171-1

360. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application. *Briefings in functional genomics*. 2016;15(5):374-384.

361. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Human mutation*. 2015;36(8):815-822.

362. Petersen B-S, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and-exome sequencing. *BMC genetics*. 2017;18(1):1-13.

363. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*. 2010;20(9):1297-1303.

364. Pais LS, Snow H, Weisburd B, et al. seqr: A web-based analysis and collaboration tool for rare disease genomics. *Human Mutation*. 2022;

365. Buske OJ, Schiettecatte F, Hutton B, et al. The Matchmaker Exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles. *Hum Mutat*. 2015 Oct 2015;36(10):922-7. doi:10.1002/humu.22850

366. Ellard S, Baple EL, Owens M, Eccles DM, Abbs S, Deans ZC, McMullan D. ACGS best practice guidelines for variant classification 2019. *ACGS Guidelines*. 2019;

367. Rehm HL, Berg JS, Brooks LD, et al. ClinGen—the clinical genome resource. *New England Journal of Medicine*. 2015;372(23):2235-2242.

368. Peterson BD, Hernandez EJ, Hobbs C, et al. Automated Prioritization of Sick Newborns for Rapid Whole Genome Sequencing Using Clinical Natural Language Processing and Machine Learning. *medRxiv*. 2022;

369. Uguen K, Krysiak K, Audebert-Bellanger S, et al. Heterozygous HMGB1 loss-of-function variants are associated with developmental delay and microcephaly. *Clinical genetics*. 2021;100(4):386-395.

370. Purcell RH, Toro C, Gahl WA, Hall RA. A disease-associated mutation in the adhesion GPCR BAI2 (ADGRB2) increases receptor signaling activity. *Human mutation*. 2017;38(12):1751-1760.

371. Kaplanis J, Samocha KE, Wiel L, et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*. 2020;586(7831):757-762.

372. Joynt AC, Axford MM, Chad L, Costain G. Understanding genetic variants of uncertain significance. *Paediatrics & Child Health*. 2022;27(1):10-11.

373. Seaby EG, Turner S, Bunyan DJ, Seyed-Rezai F, Essex J, Gilbert RD, Ennis S. A novel variant in GATM causes idiopathic renal Fanconi syndrome and predicts progression to end-stage kidney disease. *Clinical Genetics*. 2023;103(2):214-218.

374. Rehm HL, Alaimo JT, Aradhya S, et al. The landscape of reported VUS in multi-gene panel and genomic testing: Time for a change. *Genetics in Medicine*. 2023:100947.

375. Macklin SK, Jackson JL, Atwal PS, Hines SL. Physician interpretation of variants of uncertain significance. *Familial cancer*. 2019;18:121-126.

376. Han PK, Umstead KL, Bernhardt BA, et al. A taxonomy of medical uncertainties in clinical genome sequencing. *Genetics in Medicine*. 2017;19(8):918-925.

377. Makhnoon S, Garrett LT, Burke W, Bowen DJ, Shirts BH. Experiences of patients seeking to participate in variant of uncertain significance reclassification research. *Journal of community genetics*. 2019;10:189-196.

378. Clift K, Macklin S, Halverson C, McCormick JB, Abu Dabrh AM, Hines S. Patients' views on variants of uncertain significance across indications. *Journal of community genetics*. 2020;11:139-145.

379. Mensah NE, Sabir AH, Bond A, Roworth W, Irving M, Davies AC, Ahn JW. Automated reanalysis application to assist in detecting novel gene–disease associations after genome sequencing. *Genetics in Medicine*. 2022;24(4):811-820.

380. Zouk H, Yu W, Oza A, et al. Reanalysis of eMERGE phase III sequence variants in 10,500 participants and infrastructure to support the automated return of knowledge updates. *Genetics in Medicine*. 2022;24(2):454-462.

381. Harnish JM, Li L, Rogic S, et al. ModelMatcher: A scientist-centric online platform to facilitate collaborations between stakeholders of rare and undiagnosed disease research. *Human Mutation*. 2022;43(6):743-759.

382. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome biology*. 2020;21(1):1-16.

383. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. *Frontiers in genetics*. 2019;10:426.

384. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*. 2020;21(10):597-614.

385. Mitsuhashi S, Matsumoto N. Long-read sequencing for rare human genetic diseases. *Journal of Human Genetics*. 2020;65(1):11-19.

386. Mahmoud M, Huang Y, Garimella K, et al. Utility of long-read sequencing for All of Us. *bioRxiv*. 2023:2023.01. 23.525236.

387. Kovaka S, Ou S, Jenike KM, Schatz MC. Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nature Methods*. 2023;20(1):12-16.

388. Dong H, Falis M, Whiteley W, et al. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*. 2022;5(1):159.

389. O'Dowd A. Coding errors in NHS cause up to£ 1bn worth of inaccurate payments. *BMJ: British Medical Journal (Online)*. 2010;341

390. Fayer S, Horton C, Dines JN, et al. Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. *The American Journal of Human Genetics*. 2021;108(12):2248-2258.

391. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021/08/01 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

392. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*. 2021;50(D1):D439-D444. doi:10.1093/nar/gkab1061

393. Bick D, Ahmed A, Deen D, et al. Newborn screening by genomic sequencing: opportunities and challenges. *International Journal of Neonatal Screening*. 2022;8(3):40.

394. Spiekerkoetter U, Bick D, Scott R, Hopkins H, Krones T, Gross ES, Bonham JR. Genomic newborn screening: are we entering a new era of screening? *Journal of Inherited Metabolic Disease*. 2023;