OXFORD UNIVERSITY PRESS | Briefings in Functional Genomics

Gene-specific metrics to facilitate identification of disease genes for molecular diagnosis in patient genomes: a systematic review

Journal:	*Briefings in Functional Genomics*
Manuscript ID	BFGP-18-0028.R1
Manuscript Type:	Review Paper
Date Submitted by the Author:	30-Aug-2018
Complete List of Authors:	Alyousfi, Dareen; University of Southampton, Human Genetics Collins, Andrew; University of Southampton, Human Genetics Baralle, Diana; University of Southampton, Human Genetics and Genomic Medicine Group
Keywords:	gene-specific metrics, disease genome, gene-level scores, gene essentiality, gene-specific filtering

SCHOLARONE™ Manuscripts

Gene-specific metrics to facilitate identification of disease genes for molecular diagnosis in patient genomes: a systematic review

Dareen Alyousfi[1], Diana Baralle[2], *Andrew Collins[1]

1. Genetic Epidemiology and Bioinformatics Research Group
University of Southampton, Southampton SO17 1BJ, UK

2[Human Development and Health, Faculty of Medicine, University of Southampton, SO16 6YD, UK and Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton, SO16 5YA.]

University of Southampton, ….etc.

Word count: 4673 words

*Corresponding author email: arc@soton.ac.uk

For Peer Review

1 Abstract

The evolution of next-generation sequencing (NGS) technologies has facilitated the detection of causal genetic variants in diseases previously undiagnosed at a molecular level. However, in genome sequencing studies, the identification of disease genes among a candidate gene list is often difficult because of the large number of apparently damaging (but usually neutral) variants. A number of *variant* prioritization tools have been developed to help detect disease-causal sites. However, the results may be misleading as many variants scored as damaging by these tools are often tolerated, and there are inconsistencies in prediction results among the different variant-level prediction tools. Recently, studies have indicated that understanding *gene* properties might improve detection of genes liable to have associated disease variation and that this information improves molecular diagnostics. The purpose of this systematic review is to evaluate how understanding gene-specific properties might improve filtering strategies in clinical sequence data to prioritise potential disease variants. Improved understanding of the "disease genome", which includes coding, non-coding and regulatory variation, might help resolve difficult cases. This review provides a comprehensive assessment of existing gene-level approaches, the relationships between measures of gene-pathogenicity and how use of these prediction tools can be developed for molecular diagnostics.

Key words: gene-specific metrics; disease genome; gene-level scores; gene essentiality; gene-specific filtering.

Author profiles

Dareen Alyousfi, Bachelor of Medicine and Bachelor of Surgery (MBBS), MSc Genomic Medicine, University of Southampton, UK and is undertaking a PhD in human genetics in development and application of methods for resolving molecular diagnoses from patient sequence data.

Diana Baralle is a Professor Genomic Medicine and Honorary Consultant in Clinical Genetics**,** Faculty of Medicine, University of Southampton and Programme Lead for the MSc in Genomic Medicine.

Andrew Collins is head of the Genetic Epidemiology and Bioinformatics Research Group at the University of Southampton and is involved in next generation sequencing-based studies in population genetics and research into the genetic basis of a number of diseases.

2 Introduction

The sequencing of whole genomes using next generation sequencing (NGS) yields vast data-sets which present significant analytical challenges for identification of disease-causal variants. It is known that a subset of human genes contain, or are associated with, rare and/or common variation which have a role in disease processes (the "disease genome"). However, recognition of causal variants amongst many thousands of mostly neutral variants is a huge challenge and a pressing problem. For example, Chong et al (2015) state that the genes underlying ~50% of all Mendelian phenotypes remain unknown and many more Mendelian conditions are still to be described.

For Peer Review

Alongside methods for predicting the potential pathogenicity of individual DNA variants a number of gene-specific metrics (scores) have been developed in recent years which may help facilitate recognition of disease causing variation. Understanding the properties of the disease genome and integrating existing gene-specific predictors may help in classifying genes based on their specific features to refine molecular diagnosis. Pathogenicity scores for individual DNA variants are often inconsistent in that different methods can provide conflicting evidence on potential pathogenicity. The degree of redundancy in the genome makes the task of picking out causal variation particularly challenging. We propose an integrated approach which evaluates evidence at both gene and variant levels. We recognize that variant prediction tools alone are currently not conclusive and that evidence at the gene-specific level has the potential to enhance the recognition of variant pathogenicity [1].

This systematic review considers the literature related to gene-specific scores and their applicability to improve filtering of genome sequence data. We set out to achieve a satisfactory answer to the following research question: "Can the use of gene-specific metrics facilitate the identification of disease genes in patient genomes?" Details of the methodology used in this systematic review are given in the Supplementary methods, Supplementary Figures 1 and 2 and Supplementary Table 1.

Findings: Key Models

From a set of 20 papers yielded by the systematic review methods were classified into three groups determined by the main focus of each method and the corresponding scores: (i) Essentiality and conservation (ii) Haploinsufficiency (iii) Selection.

4.1 Characteristics of essential and conserved genes.

Essential and conserved genes encode proteins which have core biological functions essential for an organism's viability. Genes vary in their degree of essentiality and a number of quantitative scores provide an approximation to essentiality. These include predictions of the extent to which a gene is tolerant or intolerant of loss of function (LoF) mutations and estimation of the expected rate of *de novo* mutations (Pengelly et al, 2017) [11]. Supplementary Table 2 outlines the key approaches in this category. The Residual Variation Intolerance Score (RVIS) (Petrovski et al.) ranks genes by probability of carrying more, or less, functional genetic variation than expected highlighting genes intolerant to common functional variation [12]. Genes with positive scores have more common functional variation, while negative scoring genes are less tolerant having reduced associated common functional variation. Genes containing variation involved in monogenic diseases have lower RVIS scores than other genes.

By examining the evolutionary conservation of protein sequences, Rackham et al. built the **Evo**lutionary in**Tol**erance score (EvoTol) to identify genes which are intolerant to mutation[13] [14]. Because only small areas of a gene may be intolerant, for example protein-coding domains, these sub-regions may be considered particularly essential [14].

For Peer Review

EvoTol allows identification of intolerant protein sub-domains alongside the identification of intolerant genes more generally.

The development of NGS makes possible the identification of newly arising (*de novo)* mutations (DNMs) and their potential roles in rare disease. Such mutations are not considered to play a significant role in the pathogenesis of complex diseases [15]. To accurately estimate the expected rate of *de novo* mutations in a given gene, careful assessment of gene mutability is required. Gene length and local sequence context are essential factors underlying mutation rate differences (11). Samocha et al. calculated per-gene probabilities of mutation which are correlated with observed counts of rare missense variants in the Exome Sequencing Project (ESP) data set. The Samocha et al. study extends a model which investigated *de novo* mutations in epileptic encephalopathy patients (Epi4K consortium) by considering depth of coverage (i.e., how many sequence reads were present on average per base) and the regional divergence in genes between humans and Macaques. Significant numbers of genes with missense variant deficits were observed, compared to expectation from predicted mutation rates, suggesting strong evolutionary constraint removing variants by negative selection [15] [16]. The Samocha et al. model utilizes exome sequence data to evaluate the DNM rate by gene set and on a single gene basis [15], this score is referred to as *de novo* excess (DNE). The metric is predictive of selective constraint in the human genome and they identified 1,003 constrained genes known to cause severe human disease[15]. It was found that constrained genes contain higher *de novo* LoF mutation rate than expected by chance[15].

The LoFtool measures the ratio of LoF mutations to synonymous mutations for every gene. The performance of LoFtool, compared to RVIS, DNE Z-score, and EvoTol, suggests enhanced performance for predicting *de novo* haploinsufficient disease-causing genes. The LoFtool represents values as intolerance percentiles: genes that are intolerant to LoF variation have low LoFtool percentiles [13]. The four measures of genic intolerance outlined so far were included by Bartha et al. who described them as essentiality scores [17].

In early 2016, using data from 1000 Genomes Project, Aggarwala et al. proposed the Substitution Intolerance Score (SIS) as a gene-level measurement of essentiality. The interpretation of this score is such that genes with high SIS scores are functionally constrained, while genes which score low are tolerant of functional changes in the protein which might arise through mutations in the DNA sequence [18].

Another scoring system by Gussow et al. evaluates intolerance in genic sub-regions proposing that more conserved regions within a gene are expected to contain more variants which are pathogenic [19]. Genes are divided into sub-regions and tiered by intolerance to functional variation. This 'subRVIS' score ranks regions using RVIS but with the addition of information on conservation. Regions intolerant to functional variation are scored low by the subRVIS scoring system. The method utilizes the GERP++ score to evaluate evolutionary constraint for bases in each sub-region [19].

For Peer Review

The Loss Intolerance probability (pLI) score quantifies the likelihood that a gene is intolerant to a mutation which produces LoF in the protein product [20]. The score is derived using the Exome Aggregation Consortium (ExAC) database which is an extensive catalogue of human genetic diversity. This catalogue identifies one variant every eight bases on average in the exome providing a powerful filter for analysis of candidate deleterious variants in severe Mendelian diseases [20]. Lek et al. proposed that genes with high pLI score (pLI >= 0.9) are most intolerant of LoF variation. Genes in this category are the most evolutionarily constrained. The least constrained genes (LoF tolerant) have low pLI scores (pLI< = 0.1) and typically contribute to the least constrained biological pathways, such as sensory perception, where high haplotype diversity is potentially advantageous [20].

It is challenging to assess the relationship between the DNM rate and genes involved in disease. In 2017, Jiang et al. utilized available DNM data to correct for the background mutation rate seen as one of the main limitations in the Samocha et al.[15] work. The problem arises because by sequencing more individuals, more DNMs are inevitably observed in the same gene by chance. Therefore, in a given disease, if a *de novo* mutation is related to pathogenesis, disease-genes might be expected to contain more DNMs than predicted from background rates. This work includes the development of a database which describes the background DNM rate (DNMR), acquired from population variation data [21].

4.2 Characteristics of Haploinsufficient genes

Haploinsufficiency (HI) occurs whenever there is a missing or damaged copy of a **gene** leaving a single copy insufficient to maintain normal function [22]. Haploinsufficiency is mostly caused by LoF mutations and results in dominant diseases. Recognition and prediction of genes which are haploinsufficient can facilitate the filtering of disease genome data wherever the phenotype is likely to have arisen through reduced levels of gene product.

In 2010, Haung et al. proposed a deletion-based HI score by identifying differences between HI and haplosufficient (HS) genes, aiming to better distinguish pathogenic from benign deletions which helps in variant prioritization [22]. The analysis develops a logarithm-of-odds (LOD) score to estimate the probability of a deletion causing a HI phenotype. A high LOD score suggests deletions are likely to be deleterious through HI and therefore potential candidates for causing dominant traits. The score assumes there are no statistical interactions between the genes [22]. Previously, and to try to assess the pathogenicity of a deletion, clinicians considered the length of a deletion or the number of genes deleted. The Haung et al. score provides a rational basis to classify pathogenic deletions by comparing deletions seen in patients with deletions in controls and calculating the fraction of controls with a deletion at least as deleterious as that seen in the patient [22].

To distinguish false-positive disease variants from the genuinely causal variants is crucial for accurate molecular diagnoses. MacArthur et al. developed the RECessive (REC) score for distinguishing genes involved in recessive diseases from genes which are LoF- variation

tolerant [23]. A "healthy" genome might contain 100 true LoF variants, the majority in a heterozygous state. Evidence suggests that the average human carries five recessive lethal alleles in single copy in their genome. Consequently, the majority of LoF variants are considered common variants. However, these variants might still have a phenotypic effect [23]. MacArthur et al. demonstrated differences in functional and evolutionary features between recessive disease and LoF-tolerant genes, allowing for the development of a predictive model to predict recessive disease variants [23].

Khurana et al. developed the "gene position in NETworks" (NET) indispensability score to investigate relationships between degree of network centrality of a gene and selection within biological networks [24]. They consider a range of biological networks (i.e., phosphorylation, signaling, protein-protein interaction, regulatory and genetic networks). Genes which are highly connected to many biological networks are the most functionally significant, therefore, mutations in those genes might have serious consequences[24]. However, genes connected to metabolic networks were found to have more duplicated copies through more paralogs with more LoF mutations[24] .This score was included as a predictor of haploinsufficient genes in the Hsu et al. study [2]

Ge et al. consider gene-specific pathogenicity using the ratio of non-synonymous to synonymous substitution rates (dN/dS) for X-chromosome genes [25]. Genes with unusually low ratios suggest intolerance to non-synonymous variation, suggesting these are susceptible to disease related variation. They found correlation between genomic regions depleted for missense variation with disease-causal variants [25].

Steinberg et al. proposed that study biases existing in many biological networks might affect the ability of previous HI prediction scores to recognize the genuinely haploinsufficient genes. For that reason they constructed a new, unbiased, HI score, the Genome-wide HaploInsufficiency Score (GHIS) which replaces biological networks with co-expression networks [26] [27]. They compared their model with the three pre-existing methods (i.e., HI [22], NET [24] and RVIS [12]) and demonstrated that GHIS provides a score for many genes not scored by other methods [26] with enhanced performance at classifying less well studied genes [26].

Scores have been developed to recognize Mendelian genes with different modes of inheritance. Hsu et al. considered Mendelian disease gene characteristics according to their mode of inheritance. Haploinsufficiency is an essential characteristic of Mendelian disease genes with an autosomal dominant (AD) mode of inheritance and sensitivity to *de novo* mutations was recognized for this group of genes [2]. In contrast disease genes with autosomal recessive (AR) modes of inheritance tend to have more non-synonymous variants and regulatory transcript isoforms [2]. However, the X-linked (XL) pattern of inheritance is associated with fewer non-synonymous and synonymous variants [2]. Based on these findings they create a new approach to prioritize Mendelian disease genes based on their mode of inheritance (AD, AR, and XL) termed Inheritance-mode Specific Pathogenicity Prioritization (ISPP) [2]. This score integrates pre-existing gene-specific prediction methods namely: HI (Huang et al., 2010) [23], REC (MacArthur et al., 2012) [24], RVIS (Petrovski et

For Peer Review

al., 2013) [13], NET (Khurana et al., 2013) [25], DNE (Samocha et al., 2014) [16] and GDI (Itan et al., 2015) [35] along with numerous genetic properties including global expression from RNA-Seq data, DNA replication time and the noncoding (intronic region) mutation rate [2].

Because the human genome contains an abundance of non-deleterious heterozygous variants, the identification of dominant mutations for monogenic disorders is challenging. Quinodoz et al. created DOMINO a method using machine learning to identify whether a given gene is liable to carry dominant changes [28].

Inevitably, well-studied genes are over-represented in most biological networks used to create scores that predict HI compared to less-studied genes, hence most biological networks are affected by study bias. Therefore the creation of unbiased HI score becomes essential[27]. Recently, Shibab et al. produced an integrated machine learning approach called (HIPred) merging functional annotations with genomic and evolutionary features to predict HI genes without study bias using data from NIH Roadmap Epigenomics [29] and the ENCODE [30] project. The performance of this approach is considered to exceed the pre-existing HI predictors [27]. Supplementary Table 3 outlines the key approaches in this category.

4.3 Characteristics of genes under selection.

Genetic variants may be subject to positive selection whereby, if they are advantageous, they may increase in frequency. Negative selection, in contrast, acts to remove deleterious alleles. Scores which quantify the intensity of negative selection acting on genes provide insights into which genes are more likely to have variation which may have damaging consequences. The pattern is complex because some essential genes are not known to have any associated disease variation and are perhaps subject to negative selection at particularly high intensity [31].

Bustamante et al. calculate the extent and directionality of Selection operating on a given gene, this score referred to here as “Sel”. They first compared fixed sequence differences, both synonymous and non-synonymous, between humans in the sample and Chimpanzees over 11.81 Mb region of aligned coding DNA. The ratio of non-synonymous to synonymous differences (divergence) was 23.76%. In contrast the ratio of non-synonymous to synonymous polymorphisms in the human subjects was 38.42%. This shows a significant excess of amino acid variation, relative to divergence, consistent with previous work stating that much amino acid variation in the human genome is slightly to moderately damaging [32].

Eilertson et al. create a model to identify genes under natural selection with a non-parametric approach (with no assumption of a specific population genetic model) which is robust to demography [33]. This approach, called Selection Inference using Poisson Random Effects (SnIPRE), utilizes polymorphism and divergence data from synonymous and non-synonymous sites within genes [33].

The Gene-level Integrated Metric of negative Selection (GIMS) was created by combining two meta-analyses into a single meta-analysis. The first meta-analysis combines comparative genomic metrics (GERP++) and functional genomic metrics (Poly-phen2), and the second meta-analysis combines mutation rates (as SNPs/kb) and allele frequencies (as % rare) from the 1000 Genomes Project. Meta-analysis was achieved by combining those metrics into GIMS scores for 20,079 genes [34]. Because the majority of genes are under purifying selection, the aim was to quantify the degree of negative selection applied to genes. Conservation and functional scores were initially combined as 'functional genomic metrics' integrated with mutation rates and fraction of rare variants as 'population genetic metrics'. The GIMS score combines these two metrics and provides a unified score per-gene. GIMS gives a probability distribution across the entire genome in quantiles. Genes under negative selection are scored low by GIMS [34].

The Gene Damage Index (GDI) is a gene-specific score which predicts the liability of a human protein-coding gene to contain disease-causing mutations considering the influences of selection and genetic drift. In GDI, Combined Annotation Dependent Depletion (CADD) scores are used as the variant-level damage prediction method because this method is efficient at distinguishing between benign and deleterious variants and is strongly dependent on evolutionary conservation [35]. Moreover, CADD scores can assess most types of variants while other methods, like Poly-Phen-2 and SIFT, can only predict missense variants. To construct the GDI score the cumulative predicted damage in exonic regions of the gene is calculated using the CADD score for each allele compared to the expected score for variants with similar allele frequencies. The homogenized Phred I-score is calculated for each metric to indicate the ranking of the targeted gene relative to all other genes. A low Phred score: indicates a human gene with a low GDI and high Phred score indicates a gene susceptible to contain damaging variation. Genes with high GDI tend to be under less intense purifying selective pressure. A low GDI score is associated with highly conserved genes (including genes enriched for ribosome, chemokine signaling proteasome and spliceosome functions) reflecting essentiality. Such genes tend to be under stronger purifying selection than the median selective pressure acting on human genes [35]. Supplementary Table 4 outlines the key approaches in this category.

3 Discussion

Considering approaches which score genes according to essentiality and conservation the DNE score offers some advantages. The main limitation of DNE is its validity only for interpretation of *de novo* mutations [2] but considers more variables related to mutation rate which goes beyond sequence context compared to other methods like RVIS and Sel. These additional variables include consideration of sequence depth of coverage and regional divergence in genes between humans and Macaques independently, which improve the predictive value of this model [15]. The DNE score has been compared to the RVIS and negative selection score Sel. The comparison showed that DNE and RVIS were equally effective emphasizing the benefits predicted from combining the two scores [15].

The strength of Samocha et al. model is enhanced by incorporation of the depth of coverage (i.e., how many sequence reads were present on average per base) and the regional divergence in genes between humans and Macaques independently. These strengths play a significant role in the improvement of their predictive model. The number of rare synonymous variants in the Exome Sequencing Project (ESP) is shown to be highly correlated with the probability of a synonymous mutation determined by their model.

EvoTol was compared to the RVIS and the DNE scores and shown to have increased performance at classifying intolerant genes compared to RVIS. EvoTol was shown to be highly sensitive and more powerful to characterize genes with high pathogenicity [14]. Although there was no significant correlation between RVIS and EvoTol, the application of the two scores simultaneously will likely be advantageous [14].

Considering approaches for scoring genes for potential roles in haploinsufficiency phenotypes the HIPred approach has been evaluated against five predictors (HI Score, NET, RVIS, EvoTol and GHIS, Supplementary Tables 2 and 3). HIPred was found to outperform all in predicting HI genes [27]. Using different perspectives across the 26 disease-associated gene lists, Hsu et al. estimates the power of several methods that predict gene pathogenicity showing a substantial positive correlation between HI and REC (correlation r= 0.77) while the six scores have a moderate relationship with each other (r= 0.46) [2]. Among these gene scores (DNE, GDI, HI, NET, RVIS, and REC) the best predictor of disease-predisposing genes was the REC score [2]. The performance of ISPP score was significantly superior for prioritizing AR and X-linked disease-associated genes [2]. The REC score is effective at predicting disease-associated genes generally but less successful in discriminating recessive and dominant disease genes [2].

DNE measures the rate of per-gene *de novo* mutation while RVIS ranks human genes based on the strength and consistency of the purifying selection acting against functional variation. Analysis has shown that GDI and RVIS capture unique sets of reciprocal information from population genetic data [35]. In essence, RVIS reflects selective pressure while DNE is based on *de novo* mutation rate estimates; both methods do not quantitatively estimate the mutational load for a gene in a healthy human population. For this reason, these methods are not optimal for filtering genes with high mutation rates and many residual false positives might be expected. GDI has proved to be the most efficient approach for filtering out false positive variants in genes known to contain damaging variation [35].

The Ge et al. X-linked scoring system is not limited by previous gene annotation and the dN/dS ratio can be calculated for any protein-coding gene. This score applies to all X-chromosome protein-coding genes and therefore can assess genes for multiple disease phenotypes [25]. Because the intra-human dN/dS ratio is not specific to the X-chromosome the analysis of more genomic data using dN/dS ratio is recommended for future studies to identify genes which may have disease variation [25].

For Peer Review

This work aims to bring together the growing evidence that gene properties, alongside variant scoring systems, can play an important role in filtering disease sequence data. As healthy individuals can have genetic variants that lead to disruption of protein-coding genes (with no clinical phenotype) [26][27][23][36], challenges remain to distinguish which loss of function variants are associated with disease phenotypes from those that do not cause any functional disturbance [26]. Data from the 1000 Genomes Project show that on average a healthy person might carry 250-300 LoF SNVs (1000 Genomes Project Consortium et al., 2010; The 1000 Genomes Project Consortium, 2012) [2].

The understanding of human genomes is advanced through the accumulation of sequence data in publically available databases. The ExAC resource provides a potent filter to aid recognition of pathogenic variants in severe Mendelian diseases. Using ExAC for filtering to remove false positive, but plausibly pathogenic, variants decreases the number of candidate protein-altering variants by 7-fold compared to the smaller Exome Sequencing Project database (ESP) which has fewer exome sequences [20].

Coupled with the previous evidence, another study suggests that the missense Z score which represents genes rather than variants adds more information than variant-specific Poly-phen2 and CADD classifications signifying that gene-level scores of constraints provide more details to variant-level scores in evaluating pathogenicity [20]. Furthermore, Haung et al. contend that variant level scores (e.g., SIFT and POLYPHEN) are limited by lacking the capability to determine , from cross-species alignments, if negative selection at a given site is acting in a recessive, additive or dominant mode [22].

The work proposed by Gussow et al., was based on dividing the genes into sub-regions to identify exactly where the pathogenic mutations are likely to present [19]. This study brought up an important question: is the whole gene the correct unit to judge patterns of intolerance? Future analyses may consider refinements to gene-specific scores which consider within-gene regional patterns of intolerance in more detail.

Another controversial issue is the difficulty in interpretation of benign LoF variants for which the nomenclature is still not unified. It is important to realize that there are overlaps in the interpretation of LoF variants in healthy people. In the literature, all the following categories are represent LoF variants in healthy individuals: true variants that do not seriously disrupt gene function, benign LoF variation in redundant genes, non-deleterious or less-deleterious variants that have an impact on risk of phenotype or disease [23].

Because each genic scoring approach considers only a specific property of genetic architecture, each individual score has limitations. For example, the (i) REC score does not consider dominant disease-predisposing genes (ii) Non-CNV (Non-Copy Number Variation) genetic variants were not included in HI prediction score. (iii) NET score lacks the systematic comparison of different known disease-associated genes (iv) RVIS score does not consider variations in allele frequencies across different populations (v)The DNE score has limited applicability for testing *de novo* mutations. (vi) The GDI score only considers mutation

For Peer Review

profiles [2]. Furthermore, a major limitation of the GHIS score is that the genetic background in individuals is not considered, which is an important issue since genetic variants do not act in isolation and disturbance of individual genes within a single biological pathway might affect the risk of a disease [26]. Accordingly, this analysis which provides a comprehensive review of each prediction scheme, may help establish new routes for prioritizing disease-causal variants.

Presented here are a range of well-studied gene-specific predictors with various independent genetic properties. Addressing the limitations of each score or perhaps exploiting the developed scores of pathogenicity and combining these scores in an integrated metrics might better predict disease-genes since there is currently no single method that is reliably predictive of gene pathogenicity.

Many advances were developed to assess whether a gene is tolerant or intolerant to common functional variation. Initially, scores were developed per gene then studies were published showing that dividing the gene into sub-regions might help in allocating the mutation accurately. At that time all scores that measure genic intolerance required disease knowledge, this limitation was addressed by developing a tool with no prior disease knowledge required, an essential step to better predict genic intolerance.

It is hoped that this review highlights existing work to identify and explain different gene-specific pathogenicity predictors, while pointing to the gaps in disease-gene prioritization and annotation issues to facilitate new scores and better prioritization of disease-causal genes.

Key points

1. A wide range of well-established models exist that prioritize genes based on their associated disease variation potential.
2. Integration of these strategies to represent individual genes could have a significant impact on our understanding of genic properties and the recognition of disease-related functional variation.
3. Evaluation and comparison of these individual scores and the development of integrated models to enhance NGS filtering strategies in disease genomes is a fertile area for future studies.

Funding

DA is funded by the Saudi Arabia Cultural Bureau, London, UK.
DB is funded by a NIHR Research Professorship.

References

[1] X. Huang, J. Lin, and D. Demner-Fushman, “Evaluation of PICO as a knowledge representation for clinical questions.,” *AMIA Annu. Symp. Proc.*, pp. 359–63, 2006.

For Peer Review

[2] I. T. Newsweekly, "Mendelian Disease ; Findings from University of Hong Kong Yields New Data on Mendelian Disease [(ISPP) for human protein coding genes]," vol. 32, no. 20, pp. 2016–2018, 2016.
[3] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *Engineering*, vol. 2, p. 1051, 2007.
[4] S. U. Khan, M. Niazi, and R. Ahmad, "Barriers in the selection of offshore software development outsourcing vendors: An exploratory study using a systematic literature review," *Inf. Softw. Technol.*, vol. 53, no. 7, pp. 693–706, 2011.
[5] C. Jalal, Samireh and Wohlin, "Systematic Literature Studies: Database Searches vs. Backward Snowballing Samireh." .
[6] D. Badampudi, "Experiences from using snowballing and database searches in systematic literature studies." .
[7] J. Gehanno, L. Rollin, and S. Darmoni, "Is the coverage of google scholar enough to be used alone for systematic reviews," no. December 2009, pp. 0–4, 2013.
[8] S. Becker, A. Bryman, and H. (Thomas H. Ferguson, *Understanding research for social policy and practice : themes, methods and approaches.* Policy, 2012.
[9] G. Craswell and M. Poore, *Writing for academic success.* SAGE, 2012.
[10] C. Thermes, "Ten years of next-generation sequencing technology," *Trends Genet.*, vol. 30, no. 9, pp. 418–426, 2014.
[11] R. J. Pengelly, A. Vergara-Lope, D. Alyousfi, M. R. Jabalameli, and A. Collins, "Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation," *Brief. Bioinform.*, no. June, pp. 1–7, 2017.
[12] S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein, "Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes," *PLoS Genet.*, vol. 9, no. 8, 2013.
[13] J. Fadista, N. Oskolkov, O. Hansson, and L. Groop, "LoFtool: A gene intolerance score based on loss-of-function variants in 60 706 individuals," *Bioinformatics*, vol. 33, no. 4, pp. 471–474, 2017.
[14] O. J. L. Rackham, H. A. Shihab, M. R. Johnson, and E. Petretto, "EvoTol : a protein-sequence based evolutionary intolerance framework for disease-gene prioritization," vol. 43, no. 5, 2014.
[15] K. E. Samocha *et al.*, "A framework for the interpretation of de novo mutation in human disease," vol. 46, no. 9, pp. 944–950, 2015.
[16] A. S. Allen *et al.*, "De novo mutations in epileptic encephalopathies," *Nature*, vol. 501, no. 7466, pp. 217–221, 2013.
[17] A. Bartha, István & di Iulio, Julia & Venter, J & Telenti, "Human gene essentiality. Nature Reviews Genetics." 2017.
[18] V. Aggarwala and B. F. Voight, "An expanded sequence context model broadly explains variability in polymorphism levels across the human genome," *Nat. Genet.*, vol. 48, no. 4, pp. 349–355, 2016.
[19] A. B. Gussow, S. Petrovski, Q. Wang, A. S. Allen, and D. B. Goldstein, "The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes," *Genome Biol.*, vol. 17, no. 1, pp. 1–11, 2016.
[20] Lek, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2017.
[21] Y. Jiang *et al.*, "MirDNMR: A gene-centered database of background de novo mutation rates in human," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D796–D803, 2017.
[22] N. Huang, I. Lee, E. M. Marcotte, and M. E. Hurles, "Characterising and predicting haploinsufficiency in the human genome," *PLoS Genet.*, vol. 6, no. 10, pp. 1–11, 2010.
[23] D. MacArthur, S. Balasubramanian, and A. Frankish, "A Systematic Survey of Loss-

For Peer Review

of-Function Variants in Human Protein-Coding Genes," *Science (80-.).*, vol. 335, no. 6070, pp. 1–14, 2012.

[24] E. Khurana, Y. Fu, J. Chen, and M. Gerstein, "Interpretation of Genomic Variants Using a Unified Biological Network Approach," *PLoS Comput. Biol.*, vol. 9, no. 3, 2013.

[25] X. Ge, P. Y. Kwok, and J. T. C. Shieh, "Prioritizing genes for X-linked diseases using population exome data," *Hum. Mol. Genet.*, vol. 24, no. 3, pp. 599–608, 2015.

[26] J. Steinberg, F. Honti, S. Meader, and C. Webber, "Haploinsufficiency predictions without study bias," *Nucleic Acids Res.*, vol. 43, no. 15, pp. 1–9, 2015.

[27] H. A. Shihab, M. F. Rogers, C. Campbell, and T. R. Gaunt, "HIPred: An integrative approach to predicting haploinsufficient genes," *Bioinformatics*, vol. 33, no. 12, pp. 1751–1757, 2017.

[28] M. Quinodoz *et al.*, "REPORT DOMINO : Using Machine Learning to Predict Genes Associated with Dominant Disorders," *Am. J. Hum. Genet.*, vol. 101, no. 4, pp. 623–629, 2017.

[29] Roadmap Epigenomics Consortium *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–329, 2015.

[30] Encode Consortium, N. Carolina, and C. Hill, "For Junk DNA," *Nature*, vol. 489, no. 7414, pp. 57–74, 2013.

[31] N. Spataro, J. A. Rodríguez, A. Navarro, and E. Bosch, "Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology," *Hum. Mol. Genet.*, vol. 26, no. 3, pp. 489–500, 2017.

[32] C. D. Bustamante *et al.*, "Natural selection on protein-coding genes in the human genome," *Nature*, vol. 437, no. 7062, pp. 1153–1157, 2005.

[33] K. E. Eilertson, J. G. Booth, and C. D. Bustamante, "SnIPRE: Selection Inference Using a Poisson Random Effects Model," *PLoS Comput. Biol.*, vol. 8, no. 12, 2012.

[34] M. G. Sampson, C. E. Gillies, W. Ju, M. Kretzler, and H. M. Kang, "Gene-level integrated metric of negative selection (GIMS) prioritizes candidate genes for nephrotic syndrome," *PLoS One*, vol. 8, no. 11, pp. 1–9, 2013.

[35] Y. Itan *et al.*, "The human gene damage index as a gene-level approach to prioritizing exome variants," *Proc. Natl. Acad. Sci.*, vol. 112, no. 44, pp. 13615–13620, 2015.

[36] P. C. Ng *et al.*, "Genetic variation in an individual human exome," *PLoS Genet.*, vol. 4, no. 8, 2008.

For Peer Review

Gene-specific metrics to facilitate identification of disease genes for molecular diagnosis in patient genomes: a systematic review

Dareen Alyousfi[1], Diana Baralle[2], *Andrew Collins[1]

1. Genetic Epidemiology and Bioinformatics Research Group
Human Development and Health,
Faculty of Medicine,
University of Southampton,
SO16 6YD, UK

2. Human Development and Health,
Faculty of Medicine,
University of Southampton, SO16 6YD, UK and Wessex Clinical Genetics Service,
Princess Anne Hospital,
Southampton, SO16 5YA.

Word count: 4586 words

*Corresponding author email: arc@soton.ac.uk

For Peer Review

1 Abstract

The evolution of next-generation sequencing (NGS) technologies has facilitated the detection of causal genetic variants in diseases previously undiagnosed at a molecular level. However, in genome sequencing studies, the identification of disease genes among a candidate gene list is often difficult because of the large number of apparently damaging (but usually neutral) variants. A number of *variant* prioritization tools have been developed to help detect disease-causal sites. However, the results may be misleading as many variants scored as damaging by these tools are often tolerated, and there are inconsistencies in prediction results among the different variant-level prediction tools. Recently, studies have indicated that understanding *gene* properties might improve detection of genes liable to have associated disease variation and that this information improves molecular diagnostics. The purpose of this systematic review is to evaluate how understanding gene-specific properties might improve filtering strategies in clinical sequence data to prioritize potential disease variants. Improved understanding of the "disease genome", which includes coding, non-coding and regulatory variation, might help resolve difficult cases. This review provides a comprehensive assessment of existing gene-level approaches, the relationships between measures of gene-pathogenicity and how use of these prediction tools can be developed for molecular diagnostics.

Key words: gene-specific metrics; disease genome; gene-level scores; gene essentiality; gene-specific filtering.

Author profiles

Dareen Alyousfi , is a Bachelor of Medicine and Bachelor of Surgery (MBBS) and has an MSc in Genomic Medicine from the University of Southampton, UK and is undertaking a PhD in human genetics studying development and application of methods for resolving molecular diagnoses from patient sequence data.

Diana Baralle is Professor Genomic Medicine and Honorary Consultant in Clinical Genetics**,** Faculty of Medicine, University of Southampton and Programme Lead for the MSc in Genomic Medicine.

Andrew Collins is head of the Genetic Epidemiology and Bioinformatics Research Group at the University of Southampton and is involved in next generation sequencing-based studies in population genetics and research into the genetic basis of a number of diseases.

2 Introduction

The sequencing of whole genomes using next generation sequencing (NGS) yields vast data-sets which present significant analytical challenges for identification of disease-causal variants. It is known that a subset of human genes contain, or are associated with, rare and/or common variation which have a role in disease processes (the "disease genome"). However, recognition of causal variants amongst many thousands of mostly neutral variants is a huge challenge and a pressing problem. For example, Chong et al. [1] state that the genes underlying ~50% of all Mendelian phenotypes remain unknown and many more Mendelian conditions are still to be described.

For Peer Review

Alongside methods for predicting the potential pathogenicity of individual DNA variants at least 20 gene-specific metrics (scores) have been developed in recent years which may help facilitate recognition of disease causing variation. An example of one of these methods is RVIS (residual variation intolerance score) which ranks genes by whether they have more or less common functional genetic variation relative to the genome wide expectation [2]. A candidate pathogenic variant found in a gene classed as intolerant of common functional variation might be worthy of follow-up as a potential causal variant.

Understanding the properties of the disease genome and integrating existing gene-specific predictors may help in classifying genes based on their specific features to refine molecular diagnosis. Pathogenicity scores for individual DNA variants are often inconsistent in that different methods can provide conflicting evidence on potential pathogenicity. The degree of redundancy in the genome makes the task of picking out causal variation particularly challenging. We recognize that variant prediction tools alone are currently not conclusive and that evidence at the gene-specific level has the potential to enhance the recognition of variant pathogenicity [3].

This systematic review considers the literature related to gene-specific scores and their applicability to improve filtering of genome sequence data. We set out to achieve a satisfactory answer to the following research question: "Can the use of gene-specific metrics facilitate the identification of disease genes in patient genomes?"

Gene-specific metrics are frequently based on properties of genic coding regions. The extent to which they provide information on the tendency of a gene to have associated disease causal variation outside the coding region is limited. Most of the tools analyzed in this review, with a few exceptions, are concerned with genomic coding variation.

Details of the methodology used in this systematic review are given in the Supplementary methods, Supplementary Figures 1 and 2 and Supplementary Table 1.

3 Findings: Key models

Each of the twenty gene-specific approaches identified by the systematic review were classified into one of three groups according to the main focus of each method. We consider below each of the three groups: (i) Essentiality and conservation (ii) Haploinsufficiency (iii) Selection. Supplementary Tables 2-4 give details of the main methods and scores allocated into each category.

3.1 Characteristics of essential and conserved genes.

Essential and conserved genes encode proteins which have core biological functions that are essential for an organism's viability. Genes vary in their degree of essentiality and a number

of different quantitative scores provide approximations to essentiality. These include predictions of the extent to which a gene is tolerant or intolerant of loss of function (LoF) mutations and estimation of the expected rate of *de novo* mutations [14]. Supplementary Table 2 outlines the key approaches in this category. The Residual Variation Intolerance Score (RVIS) ranks genes by probability of carrying more, or less, functional genetic variation than expected highlighting genes intolerant to common functional variation [2]. Genes with positive scores have more common functional variation, while negative scoring genes are less tolerant having reduced associated common functional variation. Genes containing variation involved in monogenic diseases have lower RVIS scores than other genes.

By examining the evolutionary conservation of protein sequences, Rackham et al. developed the Evolutionary inTolerance score (EvoTol) to identify genes which are intolerant to mutation[15] [16]. Because only small areas of a gene may be intolerant, for example protein-coding domains, these sub-regions might be particularly important domains of essentiality [16]. EvoTol allows identification of intolerant protein sub-domains alongside the identification of intolerant genes more generally.

The development of NGS makes possible the identification of newly arising (*de novo)* mutations (DNMs) and their potential roles in rare disease. Recognition of these variants is not without difficulty because of errors in alignment and poorly supported variant calls. Validation by re-sequencing and, in particular, sequencing of additional family members (often the parents of a patient) can help correctly resolve *de novo* variation which might be of disease significance. Such mutations are not considered to play a significant role in the pathogenesis of complex diseases [17]. To accurately estimate the expected rate of *de novo* mutations in a given gene, careful assessment of gene mutability is required. Gene length and local sequence context are essential factors underlying mutation rate differences [17]. Samocha et al. calculated per-gene probabilities of mutation which are correlated with observed counts of rare missense variants in the Exome Sequencing Project (ESP) data set. The Samocha et al. study extends a model which investigated *de novo* mutations in epileptic encephalopathy patients (Epi4K consortium) by considering depth of coverage (i.e., how many sequence reads were present on average per base) and the regional divergence in genes between humans and Macaques. Significant numbers of genes with missense variant deficits were observed, compared to expectation from predicted mutation rates, suggesting strong evolutionary constraint removing variants by negative selection [17] [18]. The Samocha et al. model utilizes exome sequence data to evaluate the DNM rate by gene set and on a single gene basis [17], this score is referred to as *de novo* excess (DNE). The metric is predictive of selective constraint in the human genome and identifies 1,003 constrained genes known to cause severe human disease[17]. It was found that constrained genes contain higher *de novo* LoF mutation rate than expected by chance[17].

The LoFtool measures the ratio of LoF mutations to synonymous mutations for every gene. The performance of the LoFtool, compared to RVIS, DNE Z-score, and EvoTol, suggests enhanced prediction of *de novo* haploinsufficient disease-causing genes. The LoFtool represents values as intolerance percentiles: genes that are intolerant to LoF variation have low LoFtool percentiles [15]. The four measures of genic intolerance outlined so far were included by Bartha et al. who described them as essentiality scores [19].

In early 2016, using data from 1000 Genomes Project, Aggarwala et al. proposed the Substitution Intolerance Score (SIS) as a gene-level measurement of essentiality. Genes with high SIS scores are functionally constrained, while genes which score low are tolerant of functional changes in the protein which might arise through mutations in the DNA sequence [20].

Another scoring system by Gussow et al. evaluates intolerance in genic sub-regions proposing that more conserved regions within a gene are expected to contain more variants which are pathogenic [21]. Genes are divided into sub-regions and tiered by intolerance to functional variation. This 'subRVIS' score ranks regions using RVIS but with the addition of information on conservation. Regions intolerant to functional variation are scored low by the subRVIS scoring system. The method utilizes the GERP++ [22] score to evaluate evolutionary constraint for bases in each sub-region [21].

The Loss Intolerance probability (pLI) score quantifies the likelihood that a gene is intolerant to a mutation which produces LoF in the protein product [23]. The score is derived using the Exome Aggregation Consortium (ExAC) database which is an extensive catalogue of human genetic diversity. This catalogue identifies one variant every eight bases on average in the exome providing a powerful filter for analysis of candidate deleterious variants in severe Mendelian diseases [23]. Lek et al. proposed that genes with high pLI score (pLI >= 0.9) are most intolerant of LoF variation. Genes in this category are the most evolutionarily constrained. The least constrained genes (LoF tolerant) have low pLI scores (pLI< = 0.1) and typically contribute to the least constrained biological pathways, such as sensory perception, where high haplotype diversity is potentially advantageous [23].

It is challenging to assess the relationship between the DNM rate and genes involved in disease. In 2017, Jiang et al. utilized available DNM data to correct for the background mutation rate seen as one of the main limitations of the Samocha et al.[17] model. The problem arises because by sequencing more individuals, more DNMs are inevitably observed in the same gene by chance. Therefore, in a given disease, if a *de novo* mutation is related to pathogenesis, disease-genes might be expected to contain more DNMs than predicted from background rates. This work includes the development of a database which describes the background DNM rate (DNMR), acquired from population variation data [24].

3.2 Characteristics of Haploinsufficient genes

Haploinsufficiency (HI) occurs whenever there is a missing or damaged copy of a gene leaving a single copy which is insufficient to maintain normal function [3]. Haploinsufficiency is mostly caused by LoF mutations and results in dominant diseases. Recognition and prediction of genes which are haploinsufficient can facilitate the filtering of disease genome data wherever the phenotype is likely to have arisen through reduced levels of gene product.

For Peer Review

In 2010, Haung et al. proposed a deletion-based HI score by identifying differences between HI and haplosufficient (HS) genes, aiming to better distinguish pathogenic from benign deletions which helps in variant prioritization [3]. The analysis develops a logarithm-of-odds (LOD) score to estimate the probability of a deletion causing a HI phenotype. A high LOD score suggests deletions are likely to be deleterious through HI and therefore potential candidates for causing dominant traits. The score assumes there are no statistical interactions between the genes [3]. Previously, and to try to assess the pathogenicity of a deletion, clinicians considered the length of a deletion or the number of genes deleted. The Haung et al. score provides a rational basis to classify pathogenic deletions by comparing deletions seen in patients with deletions in controls and calculating the fraction of controls with a deletion at least as deleterious as that seen in the patient [3].

Distinguishing false-positive disease variants from the genuinely causal variants is crucial for accurate molecular diagnoses. MacArthur et al. developed the RECessive (REC) score for distinguishing genes involved in recessive diseases from genes which are LoF- variation tolerant [25]. A "healthy" genome might contain 100 true LoF variants, the majority in a heterozygous state. Evidence suggests that the average human carries five recessive lethal alleles in single copy in their genome. Consequently, the majority of LoF variants are considered common variants. However, these variants might still have a phenotypic effect [25]. MacArthur et al. demonstrated differences in functional and evolutionary features between recessive disease and LoF-tolerant genes, allowing for the development of a predictive model to predict recessive disease variants [25].

Khurana et al. developed the "gene position in NETworks" (NET) indispensability score to investigate relationships between degree of network centrality of a gene and selection within biological networks [26]. They consider a range of biological networks relating to phosphorylation, signaling, protein-protein interaction and regulatory and genetic networks. Genes which are highly connected to many biological networks are the most functionally significant, therefore, mutations in those genes might have serious consequences[26]. However, genes connected to metabolic networks were found to have an excess of duplicated copies through more paralogs with LoF mutations[26] .This score was included as a predictor of haploinsufficient genes in the Hsu et al. study [5]

Ge et al. consider gene-specific pathogenicity using the ratio of non-synonymous to synonymous substitution rates (dN/dS) for X-chromosome genes [27]. Genes with unusually low ratios suggest intolerance to non-synonymous variation, indicating they may be susceptible to disease-related variation. The authors found correlation between genomic regions depleted for missense variation with disease-causal variants [27].

Steinberg et al. proposed that study biases existing in many biological networks might affect the ability of previous HI prediction scores to recognize the genuinely haploinsufficient genes. For that reason they constructed a new, unbiased, HI score, the Genome-wide HaploInsufficiency Score (GHIS) which replaces biological networks with co-expression

networks [28] [29]. They compared their model with the three pre-existing methods (i.e., HI [3], NET [26] and RVIS [2]) and demonstrated that GHIS provides a score for many genes not scored by other methods [28] with enhanced performance at classifying less well studied genes [28].

Scores have been developed to recognize Mendelian genes with different modes of inheritance. Hsu et al. considered Mendelian disease gene characteristics according to their mode of inheritance. Haploinsufficiency is an essential characteristic of Mendelian disease genes with an autosomal dominant (AD) mode of inheritance and sensitivity to *de novo* mutations was recognized for this group of genes [5]. In contrast disease genes with autosomal recessive (AR) modes of inheritance tend to have more non-synonymous variants and regulatory transcript isoforms [5]. However, the X-linked (XL) pattern of inheritance is associated with fewer non-synonymous and synonymous variants [5]. Based on these findings they create a new approach to prioritize Mendelian disease genes based on their mode of inheritance (AD, AR, and XL) termed Inheritance-mode Specific Pathogenicity Prioritization (ISPP) [5]. This score integrates pre-existing gene-specific prediction methods namely: HI [3], REC [25], RVIS [2], NET [26], DNE [17] and GDI [30] along with numerous genetic properties including global expression from RNA-Seq data, DNA replication time and the noncoding (intronic region) mutation rate [5].

Because the human genome contains an abundance of non-deleterious heterozygous variants, the identification of dominant mutations for monogenic disorders is challenging. Quinodoz et al. created DOMINO a method using machine learning to identify whether a given gene is liable to carry dominant changes [31].

Inevitably, well-studied genes are over-represented in most biological networks used to create scores that predict HI compared to less-studied genes, hence most biological networks are affected by study bias. Therefore the creation of unbiased HI score becomes particularly important [29]. Recently, Shihab et al. produced an integrated machine learning approach called (HIPred) merging functional annotations with genomic and evolutionary features to predict HI genes without study bias using data from NIH Roadmap Epigenomics [32] and the ENCODE [33] project. The performance of this approach is considered to exceed the pre-existing HI predictors [29]. Supplementary Table 3 outlines the key approaches in this category.

3.3 Characteristics of genes under selection.

Genetic variants may be subject to positive selection whereby, if they are advantageous, they may increase in frequency. Negative selection, in contrast, acts to remove deleterious alleles. Scores which quantify the intensity of negative selection acting on genes provide insights into which genes are more likely to have variation which may have damaging consequences. The pattern is complex because some essential genes are not known to have

For Peer Review

any associated disease variation and are perhaps subject to negative selection at particularly high intensity [34].

Bustamante et al. calculate the extent and directionality of Selection operating on a given gene, this score referred to here as "Sel". They first compared fixed sequence differences, both synonymous and non-synonymous, between humans in the sample and Chimpanzees over 11.81 Mb region of aligned coding DNA. The ratio of non-synonymous to synonymous differences (divergence) was 23.76%. In contrast the ratio of non-synonymous to synonymous polymorphisms in the human subjects was 38.42%. This shows a significant excess of amino acid variation, relative to divergence, consistent with previous work stating that much amino acid variation in the human genome is slightly to moderately damaging [35].

Eilertson et al. create a model to identify genes under natural selection with a non-parametric approach (with no assumption of a specific population genetic model) which is robust to demography [36]. This approach, called Selection Inference using Poisson Random Effects (SnIPRE), utilizes polymorphism and divergence data from synonymous and non-synonymous sites within genes.

The Gene-level Integrated Metric of negative Selection (GIMS) was created by combining two meta-analyses into a single meta-analysis. The first meta-analysis combines comparative genomic metrics (GERP++) [22] and functional genomic metrics (Poly-phen2) [37], and the second meta-analysis combines mutation rates (as SNPs/kb) and allele frequencies (as percentage rare) from the 1000 Genomes Project. Meta-analysis was achieved by combining those metrics into GIMS scores for 20,079 genes [38]. Because the majority of genes are under purifying selection, the aim was to quantify the degree of negative selection applied to genes. Conservation and functional scores were initially combined as 'functional genomic metrics' integrated with mutation rates and fraction of rare variants as 'population genetic metrics'. The GIMS score combines these two metrics and provides a unified score per-gene. GIMS gives a probability distribution across the entire genome in quantiles. Genes under negative selection are scored low by GIMS [38].

The Gene Damage Index (GDI) is a gene-specific score which predicts the liability of a human protein-coding gene to contain disease-causing mutations considering the influences of selection and genetic drift. In GDI, Combined Annotation Dependent Depletion (CADD) [39] scores are used as the variant-level damage prediction method because this method is efficient at distinguishing between benign and deleterious variants and is strongly dependent on evolutionary conservation [30]. Moreover, CADD scores can assess most types of variants while other methods, like Poly-Phen-2 [37] and SIFT [40], can only predict missense variants. To construct the GDI score the cumulative predicted damage in exonic regions of the gene is calculated using the CADD score for each allele compared to the expected score for variants with similar allele frequencies. The homogenized Phred I-score is calculated for each metric to indicate the ranking of the targeted gene relative to all other genes. A low Phred score: indicates a human gene with a low GDI and high Phred score indicates a gene susceptible to contain damaging variation. Genes with high GDI tend to be under less

For Peer Review

intense purifying selective pressure. A low GDI score is associated with highly conserved genes (including genes enriched for ribosome, chemokine signaling proteasome and spliceosome functions) reflecting essentiality. Such genes tend to be under stronger purifying selection than the median selective pressure acting on human genes [30]. Supplementary Table 4 outlines the key approaches in this category.

4 Discussion

Considering approaches which score genes according to essentiality and conservation the DNE score offers some advantages. The main limitation of DNE is its validity only for interpretation of *de novo* mutations [5] but it considers more variables related to mutation rate going beyond sequence context compared to other methods like RVIS and Sel. These additional variables include consideration of sequence depth of coverage and regional divergence in genes between humans and Macaques independently, which improve the predictive value of this model [17]. The DNE score has been compared to the RVIS and negative selection score Sel. The comparison showed that DNE and RVIS were equally effective emphasizing the benefits predicted from combining the two scores [17].

The strength of Samocha et al. model is enhanced by incorporation of the depth of coverage (i.e., how many sequence reads were present on average per base) and the regional divergence in genes between humans and Macaques independently. These strengths play a significant role in the improvement of their predictive model. The number of rare synonymous variants in the Exome Sequencing Project (ESP) which comprises a relatively small sample of 6700 exomes [41] is shown to be highly correlated with the probability of a synonymous mutation determined by their model. As rare variant allele frequencies are impacted by sample size evaluation in larger databases such as ExAC would be of interest [41].

EvoTol was compared to the RVIS and the DNE scores and shown to have increased performance at classifying intolerant genes compared to RVIS. EvoTol was shown to be highly sensitive and more powerful to characterize genes with high pathogenicity [16]. Although there was no significant correlation between RVIS and EvoTol, the application of the two scores simultaneously will likely be advantageous [16].

Considering approaches for scoring genes for potential roles in haploinsufficiency phenotypes the HIPred approach has been evaluated against five predictors (HI Score, NET, RVIS, EvoTol and GHIS, Supplementary Tables 2 and 3). HIPred was found to outperform all in predicting HI genes [29]. Using different perspectives across the 26 disease-associated gene lists, Hsu et al. estimates the power of several methods that predict gene pathogenicity showing a substantial positive correlation between HI and REC (correlation r= 0.77) while the six scores have a moderate relationship with each other (r= 0.46) [5]. Among these gene scores (DNE, GDI, HI, NET, RVIS, and REC) the best predictor of disease-predisposing genes was the REC score [5]. The performance of the ISPP score was significantly superior

for prioritizing AR and X-linked disease-associated genes [5]. The REC score is effective at predicting disease-associated genes generally but less successful in discriminating recessive and dominant disease genes [5].

DNE measures the rate of per-gene *de novo* mutation while RVIS ranks human genes based on the strength and consistency of the purifying selection acting against functional variation. Analysis has shown that GDI and RVIS capture unique sets of reciprocal information from population genetic data [30]. In essence, RVIS reflects selective pressure while DNE is based on *de novo* mutation rate estimates; both methods do not quantitatively estimate the mutational load for a gene in a healthy human population. For this reason, these methods are not optimal for filtering genes with high mutation rates and many residual false positives might be expected. GDI has proved to be the most efficient approach for filtering out false positive variants in genes known to contain damaging variation [30].

The Ge et al. X-linked scoring system is not limited by previous gene annotation and the dN/dS ratio can be calculated for any protein-coding gene. This score applies to all X-chromosome protein-coding genes and therefore can assess genes for multiple disease phenotypes [27]. Because the intra-human dN/dS ratio is not specific to the X-chromosome the analysis of more genomic data using dN/dS ratio is recommended for future studies to identify genes which may have disease variation [27].

The effort to improve the predictive ability of variant-level scores now includes combination of evidence from multiple pathogenicity scores and other data. An example is the "Mendelian Clinically Applicable Pathogenicity" (M-CAP) score [42] which uses machine learning classification based on existing pathogenicity scores and measures of evolutionary conservation. Such a combinatorial approach might usefully integrate evidence from both variant-level and gene-level metrics to improve predictive abilities overall [42].

This work aims to bring together the growing evidence that gene properties, alongside variant scoring systems, can play an important role in filtering disease sequence data. As healthy individuals can have genetic variants that lead to disruption of protein-coding genes (with no clinical phenotype) [25,28,29,43], challenges remain to distinguish which loss of function variants are associated with disease phenotypes from those that do not cause any functional disturbance [28]. Data from the 1000 Genomes Project show that on average a healthy person might carry 250-300 LoF SNVs (1000 Genomes Project Consortium et al., 2010; The 1000 Genomes Project Consortium, 2012) [5].

The ACMG guidelines consider *in silico* predictions of whether a variant is involved in disease, but without specifying which or how many variant interpretation algorithms to use. These data can be used only as 'supporting' evidence for variant interpretation. There are difficulties with respect to validation of these methods and there is a relatively high error rate with many pathogenic variants assessed as benign by some methods and many benign variants assessed as pathogenic [44]. The guidelines do not currently consider gene-specific

metrics which are the subject of this review but presumably could similarly constitute supporting evidence given alongside stronger independent evidence suggesting role or lack of role in disease. Ultimately, functional validation is optimal although is frequently not timely, practical or reimbursable [44,45].

The understanding of human genomes is advanced through the accumulation of sequence data in publically available databases. The ExAC resource provides a potent filter to aid recognition of pathogenic variants in severe Mendelian diseases. Using ExAC for filtering to remove false positive, but plausibly pathogenic, variants decreases the number of candidate protein-altering variants by 7-fold compared to the smaller Exome Sequencing Project database (ESP) which has fewer exome sequences [23].

Coupled with the previous evidence, another study suggests that the missense Z score which represents genes rather than variants adds more information than variant-specific Poly-phen2 and CADD classifications signifying that gene-level scores of constraints provide additional information for evaluating pathogenicity [23]. Furthermore, Huang et al. contend that variant level scores (e.g., SIFT [40] and poly-phen 2 [37]) are limited by lacking the capability to determine , from cross-species alignments, whether negative selection at a given site is acting in a recessive, additive or dominant mode [3].

The work proposed by Gussow et al. was based on dividing the genes into sub-regions to identify exactly where the pathogenic mutations are likely to present [21]. This study identified an important question: is the whole gene the correct unit by which to judge patterns of intolerance? Future analyses may consider refinements to gene-specific scores which consider within-gene regional patterns of intolerance in more detail.

Another controversial issue is the difficulty in interpretation of benign LoF variants for which the nomenclature is still not unified. It is important to realize that there are overlaps in the interpretation of LoF variants in healthy people. In the literature, all the following categories are represent LoF variants in healthy individuals: true variants that do not seriously disrupt gene function, benign LoF variation in redundant genes, non-deleterious or less-deleterious variants that have an impact on risk of phenotype or disease [25].

Because each genic scoring approach considers only a specific property of genetic architecture, each individual score has limitations. For example: (i) the REC score does not consider dominant disease-predisposing genes (ii) Non-CNV (Non-Copy Number Variation) genetic variants were not included in HI prediction score. (iii) the NET score lacks the systematic comparison of different known disease-associated genes (iv) the RVIS score does not consider variations in allele frequencies across different populations (v) the DNE score has limited applicability for testing *de novo* mutations. (vi) the GDI score only considers mutation profiles [5]. Furthermore, a major limitation of the GHIS score is that the genetic background in individuals is not considered, which is an important issue since genetic variants do not act in isolation and disturbance of individual genes within a single biological pathway might affect the risk of a disease [28]. Accordingly, this analysis which provides a

For Peer Review

comprehensive review of each prediction scheme, may help establish new routes for prioritizing disease-causal variants.

Many advances have been developed to assess whether a gene is tolerant or intolerant to common functional variation. Initially, scores were developed per gene then studies were published showing that dividing the gene into sub-regions might help in allocating the mutation accurately. At that time all scores that measure genic intolerance required disease knowledge, this limitation was addressed by developing a tool with no prior disease knowledge required, an essential step to better predict genic intolerance.

Reviewed here are a range of well-studied gene-specific predictors with various independent genetic properties. It is hoped that recognizing some of the limitations of each score and perhaps combining evidence from both variant-specific scores and gene-wise evidence might enable better prediction since there is currently no single method that is reliably predictive of gene pathogenicity. Therefore this hopefully might help to overcome one of the main challenges of 100,000 genome project which is variant annotation to prioritize important variants from harmless neutral variants. This review is intended to highlight existing work to identify and explain different gene-specific pathogenicity predictors, while pointing to the gaps in disease-gene prioritization and annotation issues to facilitate new scores and better prioritization of disease-causal genes.

Key points

1. A wide range of well-established models exist that prioritize genes based on their associated disease variation potential.
2. Integration of these strategies to represent individual genes could have a significant impact on our understanding of genic properties and the recognition of disease-related functional variation.
3. Evaluation and comparison of these individual scores and the development of integrated models to enhance NGS filtering strategies in disease genomes is a fertile area for future studies.

Funding

DA is funded by the Saudi Arabia Cultural Bureau, London, UK.
DB is funded through a NIHR Research Professorship.

References

[1] J. X. Chong *et al.*, "The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities," *Am. J. Hum. Genet.*, vol. 97, no. 2, pp. 199–215, 2015.
[2] S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein, "Genic

Intolerance to Functional Variation and the Interpretation of Personal Genomes," *PLoS Genet.*, vol. 9, no. 8, 2013.
[3] N. Huang, I. Lee, E. M. Marcotte, and M. E. Hurles, "Characterising and predicting haploinsufficiency in the human genome," *PLoS Genet.*, vol. 6, no. 10, pp. 1–11, 2010.
[4] X. Huang, J. Lin, and D. Demner-Fushman, "Evaluation of PICO as a knowledge representation for clinical questions.," *AMIA Annu. Symp. Proc.*, pp. 359–63, 2006.
[5] I. T. Newsweekly, "Mendelian Disease ; Findings from University of Hong Kong Yields New Data on Mendelian Disease [(ISPP) for human protein coding genes]," vol. 32, no. 20, pp. 2016–2018, 2016.
[6] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *Engineering*, vol. 2, p. 1051, 2007.
[7] S. U. Khan, M. Niazi, and R. Ahmad, "Barriers in the selection of offshore software development outsourcing vendors: An exploratory study using a systematic literature review," *Inf. Softw. Technol.*, vol. 53, no. 7, pp. 693–706, 2011.
[8] C. Jalal, Samireh and Wohlin, "Systematic Literature Studies: Database Searches vs. Backward Snowballing Samireh." .
[9] D. Badampudi, "Experiences from using snowballing and database searches in systematic literature studies." .
[10] J. Gehanno, L. Rollin, and S. Darmoni, "Is the coverage of google scholar enough to be used alone for systematic reviews," no. December 2009, pp. 0–4, 2013.
[11] S. Becker, A. Bryman, and H. (Thomas H. Ferguson, *Understanding research for social policy and practice : themes, methods and approaches.* Policy, 2012.
[12] G. Craswell and M. Poore, *Writing for academic success.* SAGE, 2012.
[13] C. Thermes, "Ten years of next-generation sequencing technology," *Trends Genet.*, vol. 30, no. 9, pp. 418–426, 2014.
[14] R. J. Pengelly, A. Vergara-Lope, D. Alyousfi, M. R. Jabalameli, and A. Collins, "Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation," *Brief. Bioinform.*, no. June, pp. 1–7, 2017.
[15] J. Fadista, N. Oskolkov, O. Hansson, and L. Groop, "LoFtool: A gene intolerance score based on loss-of-function variants in 60 706 individuals," *Bioinformatics*, vol. 33, no. 4, pp. 471–474, 2017.
[16] O. J. L. Rackham, H. A. Shihab, M. R. Johnson, and E. Petretto, "EvoTol : a protein-sequence based evolutionary intolerance framework for disease-gene prioritization," vol. 43, no. 5, 2014.
[17] K. E. Samocha *et al.*, "A framework for the interpretation of de novo mutation in human disease," vol. 46, no. 9, pp. 944–950, 2015.
[18] A. S. Allen *et al.*, "De novo mutations in epileptic encephalopathies," *Nature*, vol. 501, no. 7466, pp. 217–221, 2013.
[19] A. Bartha, István & di Iulio, Julia & Venter, J & Telenti, "Human gene essentiality. Nature Reviews Genetics." 2017.
[20] V. Aggarwala and B. F. Voight, "An expanded sequence context model broadly explains variability in polymorphism levels across the human genome," *Nat. Genet.*, vol. 48, no. 4, pp. 349–355, 2016.
[21] A. B. Gussow, S. Petrovski, Q. Wang, A. S. Allen, and D. B. Goldstein, "The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes," *Genome Biol.*, vol. 17, no. 1, pp. 1–11, 2016.
[22] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using GERP++," *PLoS Comput. Biol.*, vol. 6, no. 12, 2010.
[23] Lek, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol.

For Peer Review

536, no. 7616, pp. 285–291, 2017.
[24] Y. Jiang *et al.*, "MirDNMR: A gene-centered database of background de novo mutation rates in human," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D796–D803, 2017.
[25] D. MacArthur, S. Balasubramanian, and A. Frankish, "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes," *Science (80-.).*, vol. 335, no. 6070, pp. 1–14, 2012.
[26] E. Khurana, Y. Fu, J. Chen, and M. Gerstein, "Interpretation of Genomic Variants Using a Unified Biological Network Approach," *PLoS Comput. Biol.*, vol. 9, no. 3, 2013.
[27] X. Ge, P. Y. Kwok, and J. T. C. Shieh, "Prioritizing genes for X-linked diseases using population exome data," *Hum. Mol. Genet.*, vol. 24, no. 3, pp. 599–608, 2015.
[28] J. Steinberg, F. Honti, S. Meader, and C. Webber, "Haploinsufficiency predictions without study bias," *Nucleic Acids Res.*, vol. 43, no. 15, pp. 1–9, 2015.
[29] H. A. Shihab, M. F. Rogers, C. Campbell, and T. R. Gaunt, "HIPred: An integrative approach to predicting haploinsufficient genes," *Bioinformatics*, vol. 33, no. 12, pp. 1751–1757, 2017.
[30] Y. Itan *et al.*, "The human gene damage index as a gene-level approach to prioritizing exome variants," *Proc. Natl. Acad. Sci.*, vol. 112, no. 44, pp. 13615–13620, 2015.
[31] M. Quinodoz *et al.*, "REPORT DOMINO : Using Machine Learning to Predict Genes Associated with Dominant Disorders," *Am. J. Hum. Genet.*, vol. 101, no. 4, pp. 623–629, 2017.
[32] Roadmap Epigenomics Consortium *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–329, 2015.
[33] Encode Consortium, N. Carolina, and C. Hill, "For Junk DNA," *Nature*, vol. 489, no. 7414, pp. 57–74, 2013.
[34] N. Spataro, J. A. Rodríguez, A. Navarro, and E. Bosch, "Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology," *Hum. Mol. Genet.*, vol. 26, no. 3, pp. 489–500, 2017.
[35] C. D. Bustamante *et al.*, "Natural selection on protein-coding genes in the human genome," *Nature*, vol. 437, no. 7062, pp. 1153–1157, 2005.
[36] K. E. Eilertson, J. G. Booth, and C. D. Bustamante, "SnIPRE: Selection Inference Using a Poisson Random Effects Model," *PLoS Comput. Biol.*, vol. 8, no. 12, 2012.
[37] I. A. Adzhubei1, "A method and server for predicting damaging missense mutations," *October*, vol. 7, no. 4, pp. 248–249, 2010.
[38] M. G. Sampson, C. E. Gillies, W. Ju, M. Kretzler, and H. M. Kang, "Gene-level integrated metric of negative selection (GIMS) prioritizes candidate genes for nephrotic syndrome," *PLoS One*, vol. 8, no. 11, pp. 1–9, 2013.
[39] M. Kircher, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nat. g*, vol. 46, no. 3, pp. 310–315, 2014.
[40] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nat. Protoc.*, vol. 4, no. 7, pp. 1073–1082, 2009.
[41] P. L. Auer *et al.*, "Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project," *Am. J. Hum. Genet.*, vol. 99, no. 4, pp. 791–801, 2016.
[42] K. A. Jagadeesh *et al.*, "M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity," *Nat. Genet.*, vol. 48, no. 12, pp. 1581–1586, 2016.
[43] P. C. Ng *et al.*, "Genetic variation in an individual human exome," *PLoS Genet.*, vol. 4, no. 8, 2008.

For Peer Review

[44] L. J. H. Bean and M. R. Hegde, "Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines," *Genome Med.*, vol. 9, no. 1, pp. 9–11, 2017.
[45] K. D. Laboratories *et al.*, "HHS Public Access," *CA Cancer J Clin*, vol. 17, no. 5, pp. 405–424, 2015.

For Peer Review