Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation

Reuben J. Pengelly, Alejandra Vergara Lope, Dareen Alyousfi, M Reza Jabalameli, Andrew Collins

Genetic Epidemiology and Genomic Informatics,

Faculty of Medicine,

University of Southampton,

Duthie Building (808),

Tremona Road, Southampton, SO16 6YD, UK.

Tel: 44(0)2381206939

Email: arc@soton.ac.uk

Abstract

Despite the identification of many genetic variants contributing to human disease (the 'disease genome') establishing reliable molecular diagnoses remain challenging in many cases. The ability to sequence the genomes of patients has been transformative but difficulty in interpretation of voluminous genetic variation often confounds recognition of underlying causal variants. There are numerous predictors of pathogenicity for individual DNA variants but their utility is reduced because many plausibly pathogenic variants are probably neutral. The rapidly increasing quantity and quality of information on the properties of genes suggests that gene-specific information might be useful for prediction of causal variation when used alongside variant-specific predictors of pathogenicity. The key to understanding the role of genes in disease relates in part to gene essentiality which has recently been approximated, for example, by quantifying the degree of intolerance of individual genes to loss-of-function variation. Increasing understanding of the interplay between genetic recombination, selection and mutation and their relationship to gene essentiality suggests that gene-specific information may be useful for the interpretation of sequenced genomes. Considered alongside additional distinctive properties of the disease genome, such as the timing of the evolutionary emergence of genes and the roles of their products in protein networks, the case for using gene-specific measures to guide filtering of sequenced genomes seems strong.

Keywords

Disease genome

Gene essentiality

Gene-specific filtering

Next generation sequencing

Author profiles

Reuben J. Pengelly is a Research Fellow in the Genetic Epidemiology and Bioinformatics research group at the University of Southampton and is involved in method development and analysing next generation sequencing data sets encompassing diverse clinical phenotypes.

Alejandra Vergara Lope is a PhD student in the Genetic Epidemiology and Bioinformatics research group at the University of Southampton and is involved in developing classifiers for genes with associated disease variation.

Dareen Alyousfi, MBBS, is an MSc student in the Genetic Epidemiology and Bioinformatics research group at the University of Southampton and is involved in research towards improving the filtering of disease sequence data.

M Reza Jabalameli is a PhD student in the Genetic Epidemiology and Bioinformatics research group at the University of Southampton and is involved in next generation sequencing data analyses in human disease and population genomics.

Andrew Collins is head of the Genetic Epidemiology and Bioinformatics Research Group at the University of Southampton and is involved in next generation sequencing studies of a number of diseases.

Introduction

Next generation sequencing (NGS) of exomes (the protein coding regions of the sequence) or whole genomes from clinical patient samples typically yields tens of thousands of coding DNA variants. The volume and complexity of these data presents many challenges for identification of underlying disease causal mutations. The diagnostic rate for rare diseases by whole exome sequencing is generally in the range 25-50%, varying by phenotype [1, 2]. Chong et al [3] indicate that the genes underlying ~50% (3,152) of all known Mendelian phenotypes are still unknown, and many more Mendelian conditions have yet to be recognized. Establishment of gene-disease relationships is complicated by pleiotropy where genetic loci harbour multiple variants associated with multiple and sometimes distinct traits. Therefore many gene–disease relationships remain poorly understood, even for single gene disorders and, particularly, for common, complex diseases where multiple causal gene variants of small effect are difficult to recognise.

Part of the difficulty with the interpretation of genome sequences arises through the large number of plausibly damaging variants which are actually tolerated [4]. A ‘healthy’ human genome is estimated to contain ~100 loss-of-function variants [5]. Efforts to predict disease causal variation amongst sequenced genomes usually focus on the properties of individual DNA variants. To predict variant pathogenicity a number of metrics have been developed, based on, for example, conservation scores, changes in amino acid sequence, or predicted effect on protein function (for example, SIFT, PolyPhen and GERP) [6, 7, 8] . Both SIFT and PolyPhen use sequence homology of related proteins to predict whether an amino acid change might damage protein function. The conservation of the specific base though evolution is considered through multiple sequence alignment across species. The SIFT algorithm (Sorting Intolerant From Tolerant) uses only homology for prediction whereas PolyPhen also considers whether an amino acid change occurs in an important functional or structural site in the

protein. GERP (Genomic Evolutionary Rate Profiling) considers evolutionary constraint at specific positions in the sequence using a maximum likelihood approach to compute evolutionary rates. There are also evolutionary and functional prediction tools such as CADD (Combined Annotation Dependent Depletion) [9], which integrates predictive scores from multiple annotations into one metric.

However, many variants scored as apparently damaging by these methods are likely to be tolerated and firmly establishing a molecular diagnosis from a sequenced genome can be challenging. Therefore, the use of gene-specific measures alongside these variant-specific annotations has been suggested as a way to improve the ability to identify causal mutations [10].

Understanding the nature of the 'disease genome', which we define as the set of genes which contain coding variation and/or have associated non-coding regulatory variation contributing to disease, is important for developing strategies which best exploit gene-specific information. The complex mechanisms underlying the creation and persistence of the disease genome and pathogenic variation therein are not clearly understood but depend substantially on interactions between genetic recombination, selection and mutation. The pattern of linkage disequilibrium (LD) is an outcome of these processes and may have a close relationship to the disease genome [11, 12]. Increased understanding of how these underlying processes define patterns of disease variation will go some way towards resolving the causes of disease in individual genomes [13]. The interplay between these processes and outcomes, and how they underlie disease variation in the genome, is fascinating and the main focus of this review.

Selection and the disease genome

The role of selection in shaping genomes is defined in three ways: 1. 'Hard' selective sweeps where new, advantageous, mutations are driven to fixation by positive selection, 2. 'Soft' selective sweeps in which there is more gradual fixation of weakly beneficial variation by positive selection and 3. By negative (purifying) selection in which there is elimination of deleterious mutations. The relative impact of negative versus positive selection in shaping the human genome is uncertain [14], however, most mutations affecting phenotypes must be deleterious [15]. Lohmueller et al [16] stress the relative importance of negative over positive selection acting on the genome. Although selective sweeps tend to locally reduce genetic variation they are not considered to be a dominant factor explaining patterns of variability across the genome. Variants contributing to disease most likely arise by random mutation and, at least for highly penetrant monogenic variants, are maintained at low frequencies by purifying selection [17]. However, variants involved in complex traits (common disorders in which a disease allele contributes only a small fraction of disease risk) must be subject to only extremely weak negative selection [18].

The efficiency of selection may be reduced under certain conditions through the mechanism of Hill-Robertson interference (HRI) [19, 20, 21]. Considering variants subject to positive selection there may be a situation where an advantageous mutation arises and starts to spread through the population. However, before this mutation achieves fixation a second advantageous mutation at a nearby locus emerges in an individual who lacks the first mutation. The two advantageous alleles are effectively in competition. Recombination enables the creation of haplotypes carrying both advantageous alleles, with increased fitness assuming it is more advantageous to carry both alleles. However, in weakly recombining genomic regions this haplotype is much less likely to be generated. Therefore the efficacy of selection acting on linked sites simultaneously can be reduced in the presence of limited recombination.

The impact of HRI in weakly recombining genome regions can also be seen for variants subject to purifying selection. Hussin et al [13] considered the impact of HRI on the distribution of damaging variants. If there are many sites (for example, damaging nonsynonymous variants) in a small genomic region which has a low recombination rate, HRI may allow potentially deleterious variation to achieve high frequencies [15]. The impact is greater with an increasing number of sites subject to purifying selection. Meiotic recombination acts to break down this interference allowing these sites to segregate independently and form new haplotypes leading to reduction in the accumulation of damaging alleles [22].

Recombination and the disease genome

During meiosis the creation of DNA double-stranded breaks is followed by repair through homologous recombination. This process enables allele/haplotype shuffling with significant evolutionary advantage through the breakdown of associations between alleles at linked loci (in LD), which arise by genetic drift [22]. The close alignment between the recombination structure and patterns of LD enabled the recognition of the exquisite and remarkable mechanism which promotes narrow, intense, regions of recombination (hotspots). This process involves the binding of histone methyltransferase PR domain containing 9 (PRDM9*)*. This mechanism results in histone methylation before creation of a double-stranded break and is associated with biased gene conversion or 'hotspot drive' [14]. Selective bias in favour of the non-recombinogenic allele eventually drives the extinction of the recombination hotspot [23]. However, the highly evolving zinc finger domain of *PRDM9* changes the motif it recognises with subsequent generation of new hotspots [24]. Recombination may also influence the evolution of the genome through GC-biased gene conversion in which there is biased introduction of G and C nucleotides during mismatch repair following recombination [25] and also there may be biased transmission of the shorter or longer allele of an insertion-deletion polymorphisms (indels) during meiosis [14]. However, Webster and Hurst [14] suggest there is no

evidence that these indirect effects of variation in recombination rate across the genome impacts the efficiency of selection.

Meiotic recombination has a significant role in determining the abundance and location of disease associated variation in the genome [13]. Where recombination is absent (and there is no mutation back to the original allele) a process termed Muller's ratchet [22, 26] has an important impact. In the absence of recombination, deleterious variants arising by mutation cannot be eliminated because the original haplotypes which lack the mutation cannot be re-generated. Suppressed recombination and the build-up of deleterious variation may explain why most Y chromosome genes are inactive [20]. Given the highly variable recombination rates across chromosome regions the pattern of recombination provides insights into the processes underlying the distribution of disease variation across the genome. Hussin et al. [13] contrasted levels of potentially damaging variation in highly recombining parts of the genome with weakly recombining regions. They provide clear evidence that purifying selection removes damaging variation more efficiently in highly recombining regions.

The possibility that recombination is itself mutagenic has been considered. It is known that recombination underlies sequence structural changes due to non-allelic homologous recombination [14] but there is limited evidence it can introduce point mutations [27, 28]. Schaibley et al [27] found wide variation in mutation rates related to local GC content, but not to the recombination rate. Overall the available data suggest that the recombination rate has limited effect on the frequency of mutation.

Mutation and the disease genome

Genes with high mutation rates might appear to be disease candidates simply because multiple patient genomes are likely to contain mutations in these genes. Variability in mutation rate provides a particular challenge to interpretation of sequenced genomes [28]. Mutations arise through copying errors during replication, spontaneous DNA changes and DNA instability [29]. It is known that mutation rates vary widely on different scales from single nucleotides through to whole chromosomes [30]. There are powerful context effects in which the mutation rate is influenced by adjacent nucleotides causing mutation rate variability of more than 650 fold [31]. For example, CpG dinucleotides constitute less than 2% of the genome but account for ~19% of the *de novo* mutations [29, 32] and are the most mutable sites in the genome [33]. During replication DNA mispairing is frequent with G-T and A-C mispairing the most common. This creates a twofold rate of transitions compared to transversions, when the opposite would be expected if all changes were equally likely [29]. There is also evidence for more cryptic context-independent variation with some sites appearing hyper-mutable [34]. The sequence context of each gene is a powerful predictor of mutation rates. Aggarwala and Voight [35] introduce "substitution intolerance scores" for genes demonstrating that a heptanucleotide context accounts for more than 81% of variability in substitution probabilities. They identify mutation-promoting motifs at ApT dinucleotides, CAAT and TACG sequences. Based on this

7-mer sequence the substitution intolerance score quantifies the difference between expected and observed functional variants in a gene given the sequence context.

Linkage disequilibrium and the disease genome

The pattern of linkage disequilibrium (LD) is broadly conserved among different populations [36] and known to be highly determined by recombination, but is also impacted by selection and mutation. Recombination and mutation tend to increase the diversity of haplotypes therefore act to reduce LD locally, in contrast selection tends to increase LD, although its effects are complex [37]. Remarkable alignment between the structure of the linkage map in centimorgans (which quantifies meiotic recombination over a few generations) and the 'historical' pattern of recombination in LD maps (reflecting accumulated recombination over many generations) has been demonstrated [36]. The X chromosome shows an excess of LD reflecting either reduced recombination or, more significantly, increased selective pressure on the haploid X [38, 39]. Lek et al [40] note that genes on the X chromosome are significantly more constrained (having fewer rare variants per gene than expected under a selection neutral model) compared to genes on the autosomes.

Gibson et al [11] and Collins [12] have shown, by constructing LD maps of individual genes from exome data, that there is enrichment of disease variation amongst genes with 'average' levels of LD. This pattern is distinct from genes with strong LD which are enriched for essential functions (e.g. phosphorylation, cell division, cellular transport and metabolic processes) and genes with weak LD which are enriched for functions related to sensory perception and some immune functions.

Gene essentiality and the disease genome

Essential genes are critical for cell viability. The degree of gene essentiality is likely to have a direct bearing on the tolerance a gene has for damaging/disease variation. Quantifying gene essentiality is challenging and the essentiality of individual genes has traditionally been evaluated from mouse knock-out experiments for the orthologous genes. Dickerson et al [41] questioned whether knock-outs, which remove the protein-coding region of the gene, are a valid representation since less severe changes (such as point mutations) are more typical with likely less damaging effects. More recently a range of techniques, such as large scale short hairpin RNA (shRNA) screens of diverse cell lines, ChIP-seq and computational predictions, through integration of gene expression, molecular alterations and pathways, have been developed [42]. CRISPR-Cas9 genome editing has also emerged as a technique to allow large-scale studies into genome-wide essentiality [43, 44]. The latter approach has enabled refined determination of some of the distinct features of essential genes suggesting that protein interaction networks, integrated with gene expression or histone marks, are predictive of gene essentiality.

The substitution intolerance score [35], in which higher scores indicate functionally constrained genes, is a measure correlated with essentiality. As expected, genes that were classed as likely to be essential or ubiquitously expressed scored highly for intolerance of functional variation. Genes related to keratin pathways or with olfactory functions were highly tolerant of functional changes whilst OMIM disease genes had more intermediate tolerance (Table 1).
Similarly, the loss intolerance probability (pLI) score, described from the ExAC data set of 60,706 exomes [40], has been used as an approximation to gene essentiality. pLI defines the probability of a gene being intolerant to variation causing loss of gene function. Lek et al. [40] identified 3,230 genes as intolerant (pLI > 0.9) and 10,374 as tolerant (pLI <0.1). Dominant disease genes were found to be enriched for loss-of-function (LoF) intolerant genes whereas recessive disease genes were found to include a smaller proportion of LoF intolerant genes. Genes found to be intolerant of LoF variation had almost complete absence of protein truncating variants suggesting strong purifying selection. The gene-specific pLI metric is positively correlated with degree of interconnectivity in protein-protein networks and the most constrained pathways include core biological processes (spliceosome, ribosome, proteasome components) whereas olfactory receptors are the least constrained.

A number of studies have evaluated the relationship between gene essentiality and human disease. Tu et al [46] recognised that essential genes are distinct from other 'non-disease' genes. They compared ubiquitously expressed human genes (housekeeping genes), as a group likely to contain many essential genes, with disease genes and other non-disease genes. Ubiquitously expressed genes are presumed essential for fundamental cellular physiology but essential genes with more tissue-specific functions will not be included in this set. Essential genes might be regarded as the most severe 'disease' genes in that disruption of function is likely to be developmentally lethal. Housekeeping genes were found to have shorter coding sequence lengths than disease genes consistent with earlier evidence [49] of shorter introns, untranslated regions and coding sequences, suggesting selection for more compact sequences (Table 1). Interestingly there is some evidence that disease genes are longer on average than other genes [12, 46] (Table 1).

Spataro et al [17] analysed gene properties based on roles in protein networks, rates of protein evolution and tests of neutrality. They identified three gene groups with distinct degrees of essentiality:
1. Genes which are neither essential nor associated with disease (non-disease non-essential genes, NDNE) which have the least functional relevance and are under the weakest levels of purifying selection .

2. Human disease (HD) genes, from a curated version of OMIM (hOMIM) [48], which are functionally relevant but less than essential non-disease genes. These genes are under stronger and longer lasting purifying selection than NDNE genes.

3. Essential non-disease (END), based on orthologues of mouse essential genes from knock-out experiments. These genes have no association with human disease because functionally relevant mutations are likely to have lethal consequences such as a miscarriage or early death.

We compared two alternative representations of essentiality by evaluating pLI scores [40] in each of the three Spataro et al [17] gene groups (Table 2). Although there is a significant trend towards higher pLI (greater intolerance of functional variation) from NDNE genes, HD genes through to END genes (correlation p=0.17, $p<0.0001$), in line with assumptions about essentiality, there is wide overlap between the three groups. This suggests only limited consistency between the three-group classification and pLI scores as measures of gene essentiality. Inconsistency in classification could arise in a number of ways. For example, the classification of END genes, based on mouse knock-outs has been criticised [41] and pLI essentiality scores consider only functional variation in coding regions of the genome whereas disease variation is known to extend to non-coding regulatory regions, particularly for complex diseases. However, integrative analysis of alternative measures of essentiality may form a basis for the development of models which enhance recognition of disease variation.

Mendelian or complex trait genes?

It is known that most variants associated with complex traits are regulatory in function and their target genes are difficult to ascertain requiring challenging functional investigation [50]. Therefore, the understanding of variation underlying complex phenotypes is far less complete than for monogenic disorders. Spataro et al [17] find that genes with variants for Mendelian disorders, which are also associated with variation underlying complex traits ("Complex-Mendelian" genes), have higher functional relevance in protein networks and higher expression levels than genes associated only with complex traits. In this sense they might be seen as intermediate between Mendelian-only and complex trait-only genes.

Synthesis

We propose a scheme representing the opposing and interacting processes which define the disease and non-disease genomes (Figure 1). We assume an underlying increasing measure of gene essentiality in which the most essential genes are those which are required for survival and reproduction such that functional disruption is lethal [17, 46]. The intensity of recombination and selection varies across the spectrum of gene essentiality. The nature of the relationships between these processes is not known but Figure 1 indicates trends supported by published studies.

Genes with low essentiality tend to have high recombination rates (for example, quantified as centimorgans per kilobase) and are weakly impacted by selection. They have high haplotype diversity and correspondingly weak linkage disequilibrium. Genes at this end of the essentiality scale may be

more tolerant of mutation and include genes involved in sensory perception, such as genes encoding olfactory receptors [12, 40, 51, 52]. The high recombination rate may enable re-generation of less damaging haplotypes but residual variation is presumably tolerated and unlikely to contribute to disease.

Genes with high essentiality, however, tend to have low recombination rates but the impact of selection is intense because, with increasing essentiality, any damaging variation is associated with lethality. As a result they have limited haplotype diversity and strong linkage disequilibrium. Previous studies [40, 51] have found genes involved in DNA and RNA metabolism, response to DNA damage and the cell cycle may fall into this category. The most essential genes might be regarded as the most severe 'disease' genes.

Genes which contain, or are impacted by, disease variation are suggested to occupy an intermediate place in this scheme. There is evidence that disease genes show intermediate levels of linkage disequilibrium [12]. Genes may be exposed to recombination and selection of reduced intensity which enables retention of some damaging variation associated with disease. The impact of Hill Robertson interference in reducing the efficiency of selection and Muller's ratchet in enabling accumulation of damaging variation may be significant for this class of genes. Arguably genes impacted by variation involved in common diseases might be discriminated from genes involved in more severe monogenic disease through the monogenic forms being closer to the essential gene end of the spectrum.

Discussion

The dramatic growth in the number of human genomes sequenced (now likely to be in the hundreds of thousands [40]) is underpinning a developing understanding of genes with disease-related variation in their coding or regulatory regions. Increased knowledge of the processes which generate this variation and allow it to persist is likely to improve the efficiency with which patient genomes can be screened to identify the molecular basis of disease. The interplay of selection, recombination and mutation underlies the pattern of disease variation and understanding these processes may enhance resolution of more cases with monogenic disease. The extent to which these processes can be informative for complex disease is less clear given the extremely small effect size of the variants involved which have mostly been identified in very large genome-wide association studies [50]. However, analyses of gene properties may be enhanced through consideration of additional gene characteristics (Table 1). Gene age was highlighted by Cai et al [45] where age is defined through models of evolutionary emergence times [53]. Younger genes, for example, are more likely to have primate or human-specific functions contrasting with older genes which have more ancient phylogenetic origins. They found that Mendelian disease genes tend to be a more ancient group compared with non-disease genes whilst complex disease genes tended to have intermediate ages.

Coding sequence length is reduced in genes with greater essentiality [49]. These genes are subject to intense selection but have reduced recombination rates (Figure 1). Where the outcome of selection is not lethality the efficiency of selection may be impacted by HRI and damaging variation might accumulate by Muller's ratchet. Conceivably the smaller coding sequence length in these genes reduces the target size and therefore the probability of a deleterious mutation occurring in the sequence, offsetting the impact of these processes. Sequence context analysis, for example through substitution probabilities [35] may provide insights into differential mutation rates across genes and their interaction with other contributing mechanisms.

Further distinctions include degree of connectivity of the protein product [41], position in the protein network [45] and cellular localisation [41] (Table 1) although these may be more informative of the essential gene: non-essential gene categorisation.

As might be expected broad gene categories are not independent. Genes which contain lethal null alleles can have non-lethal disease alleles [41] complicating efforts to categorise genes. The use of gene-specific measures to filter sequenced genomes to identify causal variation can only be successful when used alongside variant-specific analyses with conclusions supported by functional tests. However, there is good evidence that integrated models using emerging approximations for essentiality and gene-specific data on recombination, mutation and selection, may contribute to improved molecular diagnostics in the analysis of patient sequence data.

Key points

1. The identification of causal disease variation from patient genome sequences is challenging and confounded by plausibly damaging variation which is actually neutral.
2. Methods which predict whether a variant is damaging or not might be misleading and recent studies have suggested that information about the properties of genes might improve molecular diagnoses.
3. There is evidence that genes which have associated disease variation have intermediate essentiality between the extremes of genes of low essentiality (which are tolerant of functional variation) and genes of high essentiality (in which functional variation may be lethal).
4. Modelling gene essentiality and its relationship to variable recombination and mutation rates, along with variation in intensity of selection, may provide a basis for developing models which improve gene-specific predictors of disease variation.

References

1. Smith ED, Radtke K, Rossi M et al. Classification of Genes: Standardized Clinical Validity Assessment of Gene–Disease Associations Aids Diagnostic Exome Analysis and Reclassifications. Human mutation. 2017 May 1;38(5):600-8.

2. Cummings BB, Marshall JL, Tukiainen T et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science translational medicine. 2017 Apr 19;9(386):eaal5209.

3.Chong JX, Buckingham KJ, Jhangiani SN et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* 2015, **97**:199–215.

4. Itan Y, Casanova JL. Can the impact of human genetic variations be predicted?. Proceedings of the National Academy of Sciences. 2015 Sep 15;112(37):11426-7.

5. MacArthur DG, Balasubramanian S, Frankish A et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17;335(6070):823-8.

6. Adzhubei IA, Schmidt S, Peshkin L et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

7. Cooper GM, Stone EA, Asimenos G et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15:901–13.

8. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.

9. Kircher M, Witten DM, Jain P et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.

10. Petrovski S, Wang Q, Heinzen EL et al. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013 Aug 22;9(8):e1003709.

11. Gibson J, Tapper W, Ennis S, Collins A. Exome-based linkage disequilibrium maps of individual genes: functional clustering and relationship to disease. Human genetics. 2013 Feb 1;132(2):233-43.

12. Collins A. The genomic and functional characteristics of disease genes. Briefings in bioinformatics. 2015 Jan 1;16(1):16-23.

13. Hussin JG, Hodgkinson A, Idaghdour Y et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nature genetics. 2015 Apr 1;47(4):400-4.

14. Webster MT, Hurst LD. Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends in Genetics. 2012 Mar 31;28(3):101-9.

15. Charlesworth B. The effects of deleterious mutations on evolution at linked sites. Genetics. 2012 Jan 1;190(1):5-22.

16. Lohmueller KE, Albrechtsen A, Li Y et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 2011 Oct 13;7(10):e1002326.

17. Spataro N, Rodriguez JA, Navarro A, Bosch E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. Human Molecular Genetics. 2017 Jan 4:ddw405.

18. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular biology and evolution. 2009 Sep 1;26(9):2097-108.

19. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetical research. 1966 Dec 1;8(03):269-94.

20. Charlesworth B, Charlesworth D. The degeneration of Y chromosomes. Philosophical Transactions of the Royal Society B: Biological Sciences. 2000 Nov 29;355(1403):1563.

21. Comeron JM, Williford A, Kliman RM. The Hill–Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity. 2008 Jan 1;100(1):19-31.

22. Felsenstein J. The evolutionary advantage of recombination. Genetics 1974, 78, 737-756.

23. Jeffreys AJ, Neumann R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nature genetics. 2002 Jul 1;31(3):267-71.

24. Oliver PL, Goodstadt L, Bayes JJ et al. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet. 2009 Dec 4;5(12):e1000753.

25. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annual review of genomics and human genetics. 2009 Sep 22;10:285-311.

26. Muller HJ. The relation of recombination to mutational advance. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1964 May 31;1(1):2-9.

27. Schaibley VM, Zawistowski M, Wegmann D et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome research. 2013 Dec 1;23(12):1974-84.

28. Fuentes Fajardo KV, Adams D, Mason CE et al. Detecting false-positive signals in exome sequencing. Human mutation. 2012 Apr 1;33(4):609-13.

29. Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annual review of genomics and human genetics. 2014 Aug 31;15:47-70.

30. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics. 2011 Nov 1;12(11):756-66.

31. Carlson J, Scott LJ, Locke AE et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. bioRxiv. 2017 Jan 1:108290.

32. Fryxell KJ, Moon WJ. CpG mutation rates in the human genome are highly dependent on local GC content. Molecular Biology and Evolution. 2005 Mar 1;22(3):650-8.

33. Cooper DN, Youssoufian H. The CpG dinucleotide and human genetic disease. Human genetics. 1988 Feb 1;78(2):151-5.

34. Hodgkinson A, Ladoukakis E, Eyre-Walker A. Cryptic variation in the human mutation rate. PLoS Biol. 2009 Feb 3;7(2):e1000027.

35. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nature genetics. 2016 Feb 15.

36. Lonjou C, Zhang W, Collins A et al. Linkage disequilibrium in human populations. Proceedings of the National Academy of Sciences. 2003 May 13;100(10):6069-74.

37. Jacobs GS, Sluckin TJ, Kivisild T. Refining the use of linkage disequilibrium as a robust signature of selective sweeps. Genetics. 2016 Aug 1;203(4):1807-25.

38. Vicoso B, Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. Nature Reviews Genetics. 2006 Aug 1;7(8):645-53.

39. Wang ET, Kodama G, Baldi P, Moyzis RK. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceedings of the National Academy of Sciences of the United States of America. 2006 Jan 3;103(1):135-40.

40. Lek M, Karczewski KJ, Minikel EV et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285-91.

41. Dickerson JE, Zhu A, Robertson DL, Hentges KE. Defining the role of essential genes in human disease. PloS one. 2011 Nov 11;6(11):e27368.

42. Jiang P, Wang H, Li W, Zang C et al. Network analysis of gene essentiality in functional genomics experiments. Genome biology. 2015 Nov 6;16(1):239.

43. Shalem O, Sanjana NE, Hartenian E et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014 Jan 3;343(6166):84-7.

44. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014 Jan 3;343(6166):80-4.

45.Cai JJ, Borenstein E, Chen R, Petrov DA. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome biology and evolution. 2009;1:131-44.

46. Tu Z, Wang L, Xu M et al. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC genomics. 2006 Feb 21;7(1):31.

47.Goh KI, Cusick ME, Valle D et al. The human disease network. Proceedings of the National Academy of Sciences. 2007 May 22;104(21):8685-90.

48.Blekhman R, Man O, Herrmann L et al. Natural selection on genes that underlie human disease susceptibility. Current biology. 2008 Jun 24;18(12):883-9.

49.Eisenberg E, Levanon EY. Human housekeeping genes are compact. TRENDS in Genetics. 2003 Jul 31;19(7):362-5.

50.Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics. 2015 Apr 1;16(4):197-212.

51. Smith AV, Thomas DJ, Munro HM, Abecasis GR. Sequence features in regions of weak and strong linkage disequilibrium. Genome research. 2005 Nov 1;15(11):1519-34.

52.Pierron D, Cortés NG, Letellier T, Grossman LI. Current relaxation of selection on the human genome: tolerance of deleterious mutations on olfactory receptors. Molecular phylogenetics and evolution. 2013 Feb 28;66(2):558-64.

53. Domazet-Loso T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol. 2008;25:2699–2707.

Table 1. Some comparative functional and sequence characteristics among gene classes

Characteristic	Non-disease, non-essential	Complex disease genes	Monogenic disease genes	Essential "non-disease"*	References
Gene age	+	++	+++	++++	[45]
Cellular localisation of encoded protein	plasma membrane/ extracellular	plasma membrane/ extracellular	plasma membrane/ extracellular	nuclear localisation	[41]
Gene expression, position in protein network	Not ubiquitously expressed	Not ubiquitously expressed, peripheral functions in protein networks	Not ubiquitously expressed, peripheral functions in protein networks	Ubiquitous expression, protein network hub	[45, 46, 47]
Degree of connectivity in protein-protein interaction networks	+	++	++	+++	[41]
Intensity of purifying selection	+	+/++?	+++ (more for dominant)	++++	[17, 48]
Coding sequence length	++	+++?	+++?	+	[12, 46, 49]
Substitution intolerance score	+	?	++	+++	[35]
Gene intolerance of rare variation	+	++?	++	+++	[40]

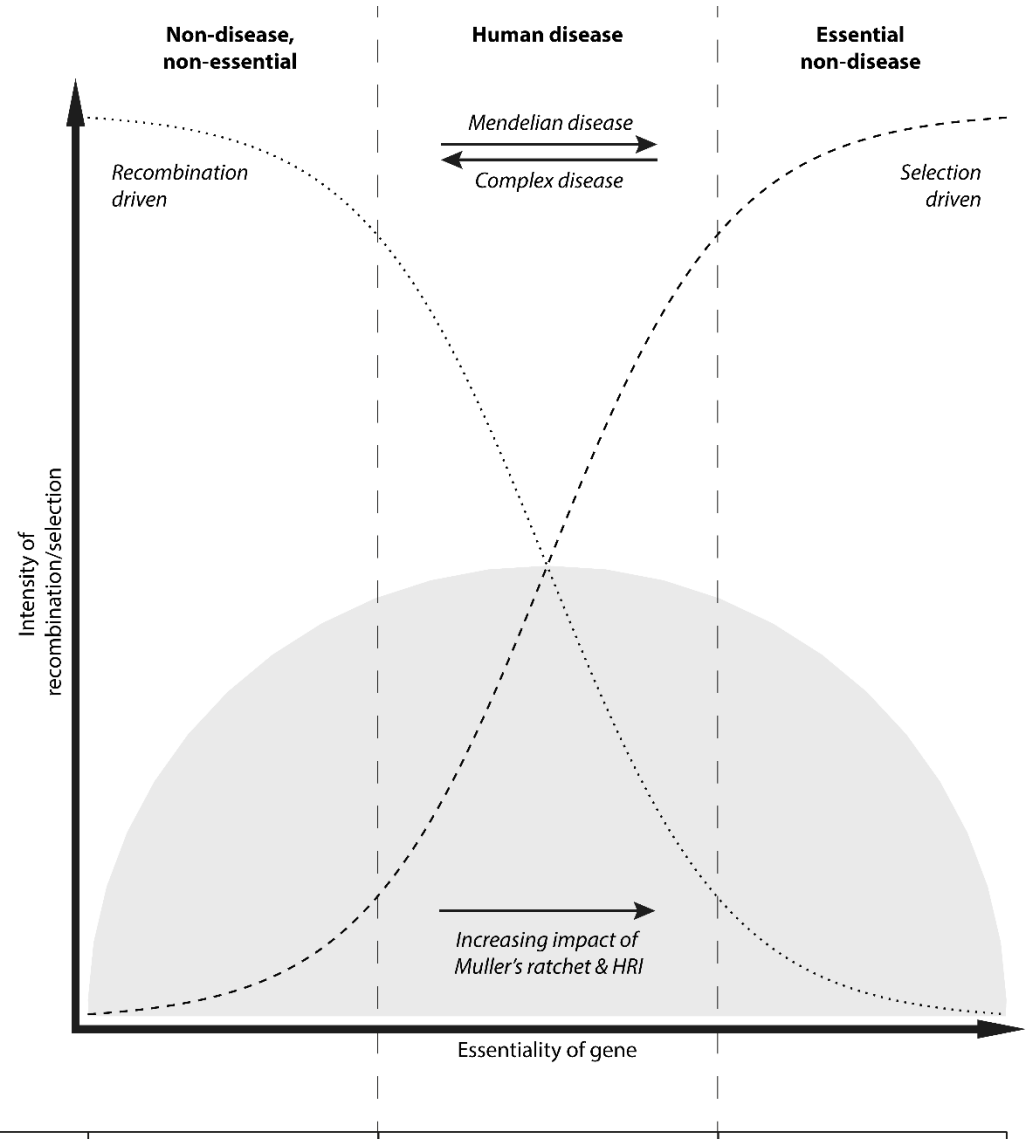
+, ++, +++ = relative magnitude of specific gene property

*Any damaging mutations likely to be lethal

Table 2. Gene essentiality pLI scores [40] within gene essentiality groups [17]

Gene class [17]	Number of genes [17]	Number of genes with pLI score	1st quartile pLI score	Median pLI score	3rd quartile pLI score	Mean pLI score
NDNE	13,135	12,062	0.000	0.010	0.475	0.251
HD	3,275	3,165	0.000	0.041	0.820	0.339
END	1,572	1,509	0.022	0.704	0.991	0.554

NDNE= Non-disease, non-essential genes; HD= Human disease genes; END= Essential, non-disease genes.



Examples of gene function	Immune response, sensory perception	Disease genes	Cell cycle, DNA repair
Strength of LD	Weak	Moderate	Strong

Figure 1. Outline of hypothetical relationships between gene essentiality, recombination (dotted line) and selection (dashed line). Deleterious variation (shaded area) is presumed to be depleted through recombination for "non-disease, non-essential" gene groups and intense selection for "essential non-disease" gene groups. Relatively weaker recombination and selection intensities may allow persistence of damaging variation for genes with associated disease variation.