

REVIEW ARTICLE

Open Access

Realizing the significance of noncoding functionality in clinical genomics

Brian S. Gloss^{1,2} and Marcel E. Dinger^{1,2,3}

Abstract

Clinical genomics promises unprecedented precision in understanding the genetic basis of disease. Understanding the impact of variation across the genome is required to realize this potential. Currently, clinical genomics analyses focus on protein-coding genes. However, the noncoding genome is substantially larger than the protein-coding counterpart, and contains structural, regulatory, and transcribed information that needs to be incorporated into genome annotations if the full extent of the opportunity to use genomic information in healthcare is to be realized. This article reviews the challenges and opportunities in unlocking the clinical significance of coding and noncoding genomic information and translating its utility in practice.

The evolution and revolution of clinical genomics

The rise of genomics

Resolving the genetic basis of disease seemed like a certainty once the human genome sequence was completed. This new comprehensive map of the body's operating system meant that the historically difficult task of mapping diseases with evidence of Mendelian inheritance patterns to a causative locus was now readily accessible to the scientific community.

This map also enabled the development of tools with unprecedented capacity to survey genomes at population scale. Microarray technology interrogation of common single-nucleotide polymorphisms (SNPs) revolutionized our understanding of genetic inheritance patterns through the hapmap project¹ (Fig. 1) and genome-wide association studies (GWAS) seemed set to unravel the genetic basis of monogenic and complex disease².

The increased power and precision of GWAS facilitated the mapping of monogenic traits but highlighted "missing heritability" where observed inherited traits could not be

explained by observed genetic variance. In the clinical genetics field, this was considered likely due to the inability of SNP-chip technology to adequately measure rare variants, structural defects, polygenic and/or complex inheritance patterns. Furthermore, epistatic interactions, where coinheritance of two or more variants could more adequately explain heritability, continues to be difficult to estimate due to computational limitations^{3,4}. At the same time, next-generation sequencing was poised to further revolutionize approaches to not only survey the genome, but also redefine the understanding of how the genome behaved and the diverse mechanisms through which genetic disease could manifest.

The rise of clinical genomics

Exome sequencing, where the coding regions of genes are enriched from DNA and sequenced, has allowed the direct measurement of variance at the genetic level. It has provided clinicians and researchers base-scale resolution of the coding genome, giving rise to a quantum leap in the resolving power of associating genetic variation directly to altered protein. The scalability and relatively low cost has resulted in large databases characterizing and annotating the variation in the coding genome such as ExAC⁵. But like many technological advances, exome sequencing has exposed its own limitations, namely technical artefacts of

Correspondence: Marcel E. Dinger (m.dinger@garvan.org.au)

¹Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia

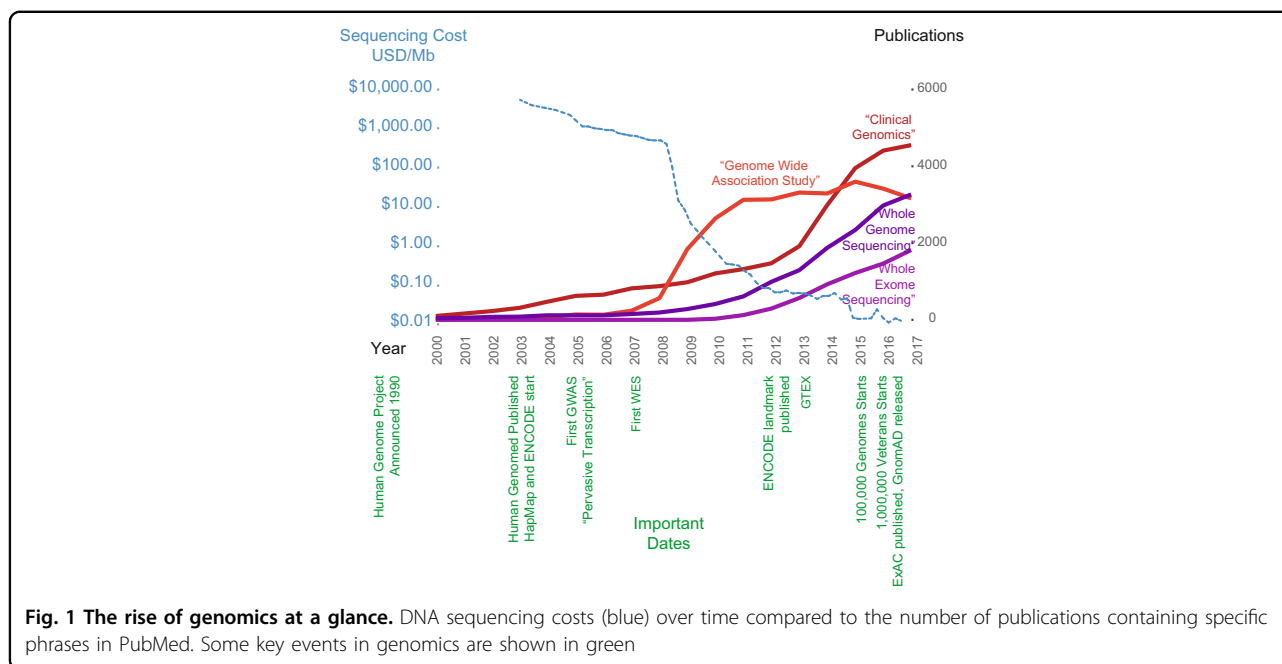
²St. Vincents Clinical School, UNSW Sydney, Sydney, NSW, Australia

Full list of author information is available at the end of the article

© The Author(s) 2018



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. If you remix, transform, or build upon this article or a part thereof, you must distribute your contributions under the same license as the original. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>.



the DNA capture used and overdependence on extant genome annotations. This latter limitation proved significant as the emergence of exome sequencing coincided with the observation of the pervasively transcribed genome⁶ and the rise of lncRNAs as important functional transcripts, which for the most part were overlooked by the approach.

Importantly, the understanding that the majority of informative variants identified by GWAS occurred within the noncoding genome, and a shift in how the genome was known to encode function through transcribed noncoding regulatory RNAs gave rise to many theories regarding the genetic basis of disease, particularly in providing an explanation for the source of missing heritability. Broadly, theories explaining missing heritability fell into two areas—(1) that variants in regulatory DNA sequences such as promoters, enhancers, and structural elements and regions encoding regulatory RNAs were responsible or (2) that large numbers of individual genetic features, potentially with complex interactions, contributed collectively to inherited traits.

Increasingly inexpensive whole-genome sequencing, particularly with PCR-free library preparations, has made it possible to overcome many of the technical artefacts of exome sequencing—thus yielding high-quality surveys of coding gene variants including SNP, copy-number, and structural variations, and insertion/deletion events. These technical advantages in analyzing coding regions alone enabled improved diagnostic yield of genome sequencing and has led to growing numbers of whole-genome clinical sequencing services worldwide.

Whole-genome sequencing consortia producing large databases of genomic variation, such as GnomAD⁵, 100,000 genomes⁷, and the Million Veterans Project⁸ (Fig. 1) are making publicly available their data for interrogation. As well as assisting in the distinction between rare pathogenic variants and those common in the population, this abundance of data provides measurements of the variation in the 98% of the genome that is non-protein-coding. Therefore, observed noncoding variants as well as protein-coding variants of unknown significance are increasing and there is potential now to advance the use of the noncoding genome to improve clinical diagnostic rates of genetic disease⁹.

The challenge

When the human genome project was completed, the implications of the complexity evident in the noncoding genome were staggering¹⁰. After more than a decade of research, considerable advances have been made in understanding how the genome instructs the development and function of organisms and it is increasingly pertinent that this knowledge is harnessed to maximize diagnoses in clinical genomics practice.

In essence, clinical genomics seeks to causatively associate a clinical feature (disease, drug response, risk) with one of the ~5 million variants (relative to a reference genome) present in every individual. This poses the challenge of effectively developing variant filtering algorithms that narrow the search space for variants to regions where pathogenicity can be most clearly determined, i.e., protein-coding regions related to well-described biological function.

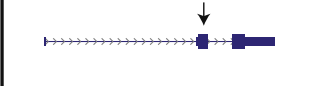
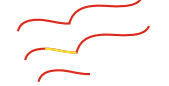
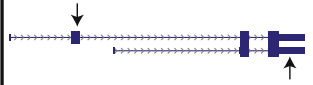

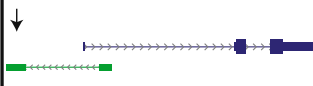

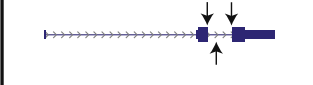
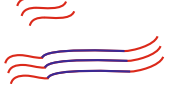
Variant Location	Transcript Map	Transcript Product	Transcript description	Potential Outcome
Coding (standard interpretation)			Synonymous/ Missense/ Nonsense	Homeostasis/ Altered Product/ Loss of function
Isoform specific/ Noncoding regulatory			Isoform loss/alteration Altered translation	Aberrant expression patterns
Promoter/Enhancer/ Looping/cis-regulatory lncRNA			Over/ Under expression	Aberrant expression patterns
Splice Donor/Acceptor Branchpoint			Skipped exon/ Retained intron	Altered product Nonsense Mediated Decay

Fig. 2 How coding and noncoding variation can impact gene function. Variants (arrows) at a hypothetical locus are shown along with potential functional impacts

Typical approaches for interpreting clinical genomes involve reducing a genome down to rare coding variants with the appropriate inheritance patterns in a gene list of interest. This approach typically yields a handful of variants for consideration. Various annotations of variant impact are then added including predicting the impact on protein structure (SIFT¹¹, Polyphen¹², and VEP¹³) and observation in disease (COSMIC¹⁴, ClinVar¹⁵, and HGMD¹⁶). The proliferation of these tools have led to aggregator services such as VarCards¹⁷ that allow multiple scores for a given variant to be interrogated in one place. A clinical molecular geneticist, molecular genetic pathologist, or other certified professional can interpret these data to assign a likely causal variant (Fig. 2). If a candidate is not apparent from these approaches even in cases where there is a strong genetic component, a diagnosis becomes difficult since biochemical testing of variants of unknown significance is not feasible in a typical pathology laboratory setting and may not be considered to be cost-effective. Furthermore, although expanding the search space to include more variants increases the number of candidates, there is typically insufficient evidence to associate any particular variant with the phenotype.

Efforts worldwide are attempting to expand the annotation of the genome beyond the pure coding and to better understand how variations in these regions can have biological impact to expand the understanding of genetic basis of disease¹⁸ and to thus fully realize the clinical utility of the whole genome.

Advances in functional annotation of the genome Resolving the annotation of gene-level variation

The interpretation of disease-associated variation at the level of the gene is undergoing a shift in understanding.

Protein-coding mutations have historically been considered deleterious where they lead to truncations (nonsense/deletions), amino acid alterations (missense/in-frame in/del), frame shifts (in/del) and splicing defects (splice-site donors/acceptors). However, these kinds of mutations have been shown to be relatively common, even in healthy genomes¹⁹. Furthermore, these variants can be difficult to interpret in a clinical setting if the mutation occurs in a region not previously reported, or in a gene whose function within the context of the disease in question has not been investigated²⁰. It is also becoming apparent that mutations that do not affect the encoded amino acid (synonymous) can affect gene products in the context of codon frequency and RNA structure^{21,22}. Furthermore, the concept of multiplicity, where gene expression can be impacted by combinations of genetic alterations²³ is only starting to be addressed. This implies that the even annotation of coding variants is far from complete.

It is also important to note that the coding proportion of a gene comprises a small percentage of the genetic information encoded by the locus and that alterations in the noncoding sequence can have impact on gene function (Fig. 2). Variation at gene promoters can impact the expression of the gene²⁴, e.g., the TERT promoter is frequently mutated, which leads to overexpression, and in turn, can be a pathogenic basis for causing or driving cancer development²⁵. Variation at imprinted loci can drive the deposition of epigenetic marks responsible for imprinting²⁶, which can lead to aberrant expression. Alterations in 5' and 3' untranslated regions of genes can impact transcript stability and translation primarily through RNA structural alterations^{27,28}. Introns can similarly contain important genetic information that can be influenced by mutation²⁹, e.g., disease-associated SNPs within branch points can be associated with altered

Table 1 Genome-wide tools for estimating impact of noncoding variation

Tool	Year	Method used to build model
CADD ⁴¹	2014	Support vector machine
GWAVA ⁴²	2014	Random forest algorithm
DeepSEA ⁷³	2015	Deep learning
FATHMM-MKL ⁷⁴	2015	Multiple Kernel learning + SVM
Eigen ⁴³	2016	Unsupervised partitioning
Basset ⁷⁵	2016	Deep convolutional neural network
LINSIGHT ⁷⁶	2017	Generalized linear model (LINSIGHT and fitCons)
Orion ⁴⁴	2017	Observed/expected variation

splicing patterns³⁰. Together, these investigations show that a significant proportion of the clinically relevant genetic information elucidated by whole-genome sequencing is not typically interpreted in diagnostic laboratories.

Resolving the transcriptional and regulatory landscape of the genome

Ever since the first observation of the pervasively transcribed genome more than a decade ago, there has been an explosion in the identification and functional characterization of long noncoding RNA (lncRNA)³¹ and other noncoding transcript types³². The encyclopedia of DNA elements consortium (ENCODE) raised considerable controversy in 2012 by using tissue-specific transcript profiling, supported by epigenetic profiling of the genome, to suggest that 82% of the human genome was functionally important³³. As the vast majority of transcribed species of the genome are noncoding, of which little is still known³¹, efforts are ongoing to describe the detail and regulation of noncoding RNA. lncRNAs are of particular interest to the field of clinical genomics as their exquisite tissue-specific expression and regulatory behavior³⁴ indicate that a role in disease will become apparent as more is understood about lncRNA biology.

As a result, several large-scale efforts have been undertaken to comprehensively annotate the noncoding transcriptional landscape, particularly through the FANTOM projects^{6,35,36}, ENCODE³³ and Roadmap Epigenomics³⁷. The large-scale GTEx project³⁸ has set out to further understand the genetic drivers of tissue-specific gene expression via expression quantitative trait loci (eQTL) analysis. Large-scale screens for noncoding RNA function have elucidated functional annotations for thousands of lncRNAs³⁹ and molecular tools tailored to the unique biology of lncRNA behavior are ongoing⁴⁰. These efforts have enhanced the understanding of gene transcription and hint at a complexity that requires

expanded resolution of functional annotation at the genetic level to inform interpretation in a clinical diagnostic setting.

Interpreting functionality at the whole-genome level

Traditional indicators of functionality (and thus of potential clinical utility), such as conservation, have thus been challenged by this expanding annotation of the genome. The volume of available data has fueled recent computational efforts to annotate functional parts of the genome without necessarily depending exclusively on the coding genome (Table 1). Early attempts used existing annotations to train computational models that could assess the potential function of a variant genome-wide (CADD⁴¹/GWAVA⁴²). Newer approaches have used genome-wide data itself to assign functional importance, either through association with DNA binding proteins (Eigen⁴³), or direct measures of resistance to variation (Orion⁴⁴), to provide comprehensive maps of coding and noncoding regions likely to be impacted by variation. These maps are expanding the pool of potentially clinically relevant variants and continue to evolve with growing interest and innovation.

Noncoding variation and disease

Structural alterations

The physical arrangement of the genome is also critical to homeostasis. Copy-number alterations are associated with many diseases, but can also have no pathogenic effect^{45,46}. The study of disease-associated genomic translocations has typically focused on the generation of gene fusions, which are particularly clinically relevant in cancer⁴⁷. However, studies of intergenic translocations can also perturb local gene expression, possibly by interrupting chromatin looping and by rearranging regulatory sequence^{48–50}. Moreover, chromatin looping⁵¹ and nucleosome occupancy⁵² are also susceptible to alteration by DNA mutation and structural rearrangement.

Localized DNA structures have been associated with genetic disease such as Huntington's disease mostly as recognition sites for genomic rearrangements⁵³. However, such quaternary structures recently gained traction as important mediators of biological information in themselves with left handed helices (z-DNA⁵⁴), G-quadruplexes^{55,56}, and DNA:DNA/DNA:RNA triplexes^{57,58} showing evidence of regulatory function. Indeed the interplay between the physical state of the DNA appears to be intimately associated with the process of gene expression⁵⁹ and transcription factor binding⁶⁰. Importantly, it was recently shown that disease-associated variations that disrupt G-quadruplex formation in RNA can affect post-transcriptional regulation of genes²⁷, suggesting that variants in structural features can directly impact cellular function.

Noncoding transcription at GWAS loci

The prevalence of intergenic, disease-associated SNPs from GWAS studies provoked diverse studies into how these variants were contributing to disease, revealing impacts on DNA conformation⁵¹, DNA-protein interactions⁶¹, and epigenetic marks⁶². Recent application of RNA-capture sequencing⁶³ to haplotype blocks associated with GWAS disease-associated SNPs revealed a multitude of transcripts of which less than half were in extant transcript databases⁶⁴. Combined with fine mapping of SNPs associated with breast cancer, this approach revealed enhancer alterations affecting novel transcript expression⁶⁵. These studies raise the possibility of direct and indirect impacts of disease-associated SNPs on tissue-specific transcription patterns and illustrate that both the resolution of disease-associated variants and genome annotation remain incomplete. The ongoing accumulation of whole-genome data worldwide will eventually resolve the exact disease associations and a greater understanding of the noncoding transcriptome will continue to provide context for elucidating the impact of these variants³⁴.

New classes of functional repeats in the human genome

In a similar vein, pseudogenes have classically been regarded as nonfunctional byproducts of

retrotransposition⁶⁶. With the observation of transcription and evidence of disease linkage⁶⁷, pseudogene biology is being revisited, however, consensus as to a generic biological role has not yet been reached^{68,69}. Indeed, the process of retrotransposition itself in shaping the genome is undergoing a renaissance through evidence of gene regulatory roles⁷⁰.

A place for noncoding annotations in clinical genomics

Rules of evidence

In 2015, the American College for Medical Genetics (ACMG) described a set of evidence lines that could be used to ascribe degrees of pathogenicity to a particular variant⁷¹. Importantly, these recommendations sought to distinguish deleterious impacts on a gene from contribution to disease. Predicting the impact of coding variation is a more mature process, especially in the case of missense and nonsense mutations. Tools like PolyPhen and VEP are commonly used to estimate genic pathogenicity, although the likely impact of the variant can be open to interpretation. Evidence for disease contribution is usually achieved by cross-referencing rare variants with lists of genes with known roles in the disease of interest, reports in the literature, and clinical databases such as COSMIC and ClinVar. The point at which there is

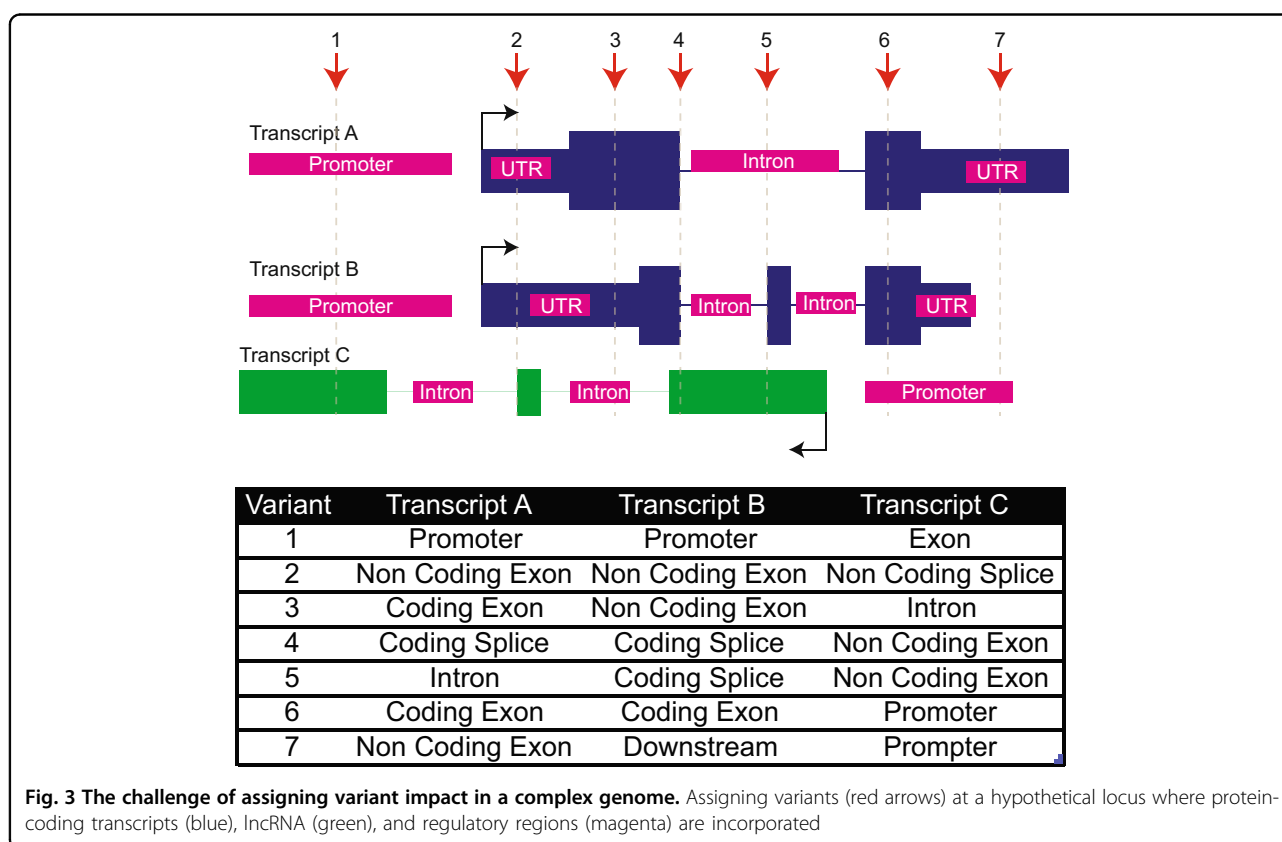


Fig. 3 The challenge of assigning variant impact in a complex genome. Assigning variants (red arrows) at a hypothetical locus where protein-coding transcripts (blue), lncRNA (green), and regulatory regions (magenta) are incorporated

sufficient evidence of a variant causing a disease is becoming refined²⁰. However, due to the complexities in the WGS data, interpretation, and phenotyping, associations can be subject to how the data are evaluated by genetic professionals and can still require in vitro testing. Including non-protein-coding into this framework would require extra complexity predominantly due to the lack of functional data to support impact of a particular variant with precision, given the ongoing genome annotations outlined above (Fig. 3). However, noncoding variants can clearly be clinically relevant and their inclusion into clinical genomics frameworks is necessary for realizing the full clinical utility of genomic information.

A framework for noncoding inclusion in clinical genomics

The clinical interpretation of variants typically begins strictly as an informatics exercise where variants are filtered and ranked according to likelihood of clinical trait association. One of the earliest steps is to omit variants that are noncoding, which in the light of the evidence outlined above may miss vital insights into the molecular basis of a disease. To address this limitation, existing frameworks that estimate noncoding impact such as the GTEX eQTLs and tools outlined in Table 1 should be integrated into existing variant interpretation frameworks such as GEMINI⁷². While less data is available for accurately calculating variant frequency in noncoding regions, growing whole-genome reference databases are now available for this purpose. These annotations can then be interpreted alongside existing lines of evidence within the context of disease.

The primary paradigm shift required by these additions to clinical genome interpretation workflows will be the expansion of the concept of what part of the genome constitutes a gene. Impacts on a specific gene function can theoretically occur anywhere within the genome. This represents a currently insurmountable computational obstacle for the same reason that epistasis remains an intractable issue in genomics. However, splicing and promoter variations are directly linked to genes and are currently well annotated. For this reason, we propose that variants occurring at splice sites and branch points as well as promoters annotated by ENCODE should be included in clinical genomics where they occur in disease relevant genes. We expect that a more inclusive approach to impacts on gene function will facilitate an improved picture of the clinical landscape, particularly in the case of disease with strong evidence of inheritance where no coding candidate can be found. For example, a promoter variant may be the second-hit in a recessive heterozygous locus leading to total loss of a gene product. Furthermore, as our knowledge of the biology of the genome grows, more interpretative power will become available in the context of clinical genomics. We contend that the potential to improve diagnostic rates using a multi-level

whole-genome annotation approach will outweigh the necessary increased time for manual variant review and ruling out of false positives.

The future

Understanding the genetic basis of disease has been an aim of science since heritable traits were first observed. Technological and conceptual progress have given rise to a picture of the genome that is as complex as one would expect from a four letter code that gives rise to living multicellular organisms. Research is currently at the point of attempting to describe and unravel this complexity as discussed above. We expect that tools and knowledge of the noncoding genome will continue to expand and that continued refinement of an integrated coding and non-coding genomic landscape through comprehensive genomic, transcriptomic, and epigenomic profiling will improve the prediction of variant outcomes. The computational issues of epistasis and polygenetic impacts will be improved as more data is generated and more powerful computational frameworks emerge, such as quantum computing to enable large combinatorial calculations that are currently unfeasible. These will go hand in hand with more widespread adoption of moderate throughput screens for rapid and direct measurements of the impact of candidate variants such as CRISPR-Cas9 tools in patient-derived iPSC cell lines. It will be important for clinical scientists involved in variant interpretation to remain mindful of the growing clinical significance of the whole genome and for developers of software and knowledgebases used to inform variant interpretation to consider non-protein-coding data sources and algorithms that act on noncoding genomic regions in their workflows.

Acknowledgements

We thank James Torpy for manuscript feedback, and Dr. Mark Pinese and Dr. Eric Lee for constructive criticism.

Author details

¹Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia. ²St. Vincents Clinical School, UNSW Sydney, Sydney, NSW, Australia. ³GenomeOne, Sydney, NSW, Australia

Conflict of interest

The authors declare that they have no conflict of interest.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 21 January 2018 Revised: 4 March 2018 Accepted: 9 March 2018.
Published online: 7 August 2018

References

1. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).

2. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
3. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
4. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
5. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
6. Carninci, P. et al. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
7. The 100,000 Genomes Project Protocolv4, Genomics England <https://doi.org/10.6084/m9.figshare.4530893.v4>, <https://www.genomicsengland.co.uk/100000-genomes-project-protocol/> (2017).
8. Gaziano, J. M. et al. Million veteran program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
9. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
10. Little, P. F. Structure and function of the human genome. *Genome Res.* **15**, 1759–1766 (2005).
11. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
12. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods.* **7**, 248–9 (2010).
13. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
14. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
15. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
16. Stenson, P. D. et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).
17. Li, J. et al. VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res.* **46**, D1039–D1048 (2018).
18. Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.* **25**, R157–R165 (2016).
19. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
20. MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
21. Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12**, 683–691 (2011).
22. Wan, Y. et al. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
23. Williams, R. B., Chan, E. K., Cowley, M. J. & Little, P. F. The influence of genetic variation on gene expression. *Genome Res.* **17**, 1707–1716 (2007).
24. Kwansieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl Acad. Sci. USA* **109**, 19498–19503 (2012).
25. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
26. Chuang, T. E., Tseng, Y. H., Chen, C. Y. & Wang, Y. D. Assessment of imprinting- and genetic variation-dependent monoallelic expression using reciprocal allele descendants between human family trios. *Sci. Rep.* **7**, 7038 (2017).
27. Zeraati, M. et al. Cancer-associated noncoding mutations affect RNA G-quadruplex-mediated regulation of gene expression. *Sci. Rep.* **7**, 708 (2017).
28. Pesole, G. et al. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **276**, 73–81 (2001).
29. Vaz-Drago, R., Custodio, N. & Carmo-Fonseca, M. Deep intronic mutations and human disease. *Hum. Genet.* **136**, 1093–1111 (2017).
30. Signal, B., Gloss, B. S., Dinger, M. E. & Mercer, T. R. Machine learning annotation of human branchpoints. *Bioinformatics* **34**, 920–927 (2018).
31. Quek, X. C. et al. lncRNAdbv2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* **43**, D168–D173 (2015).
32. Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nat. Rev. Genet.* **15**, 423–437 (2014).
33. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
34. Gloss, B. S. & Dinger, M. E. The specificity of long noncoding RNA expression. *Biochim. Biophys. Acta* **1859**, 16–22 (2016).
35. Katayama, S. et al. Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566 (2005).
36. Kawai, J. et al. Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).
37. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
38. GTEx Consortium et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
39. Liu, S. J., et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, 35–39 (2017).
40. Kashi, K., Henderson, L., Bonetti, A. & Carninci, P. Discovery and functional analysis of lncRNAs: methodologies to investigate an uncharacterized transcriptome. *Biochim. Biophys. Acta* **1859**, 3–15 (2016).
41. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
42. Ritchie, G. R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
43. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
44. Gussow, A. B. et al. Orion: detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLoS ONE* **12**, e0181604 (2017).
45. Lee, C. & Scherer, S. W. The clinical context of copy number variation in the human genome. *Expert Rev. Mol. Med.* **12**, e8 (2010).
46. Shaikh, T. H. et al. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res.* **19**, 1682–1690 (2009).
47. Rabbitts, T. H. Chromosomal translocations in human cancer. *Nature* **372**, 143–149 (1994).
48. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
49. Lupianez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
50. David, D. et al. Identification of OAF and PVRL1 as candidate genes for an ocular anomaly characterized by keratolenticular dysgenesis and ectopia lentis. *Exp. Eye Res.* **168**, 161–170 (2018).
51. Schierding, W., Cutfield, W. S. & O'Sullivan, J. M. The missing story behind genome wide association studies: single nucleotide polymorphisms in gene deserts have a story to tell. *Front. Genet.* **5**, 39 (2014).
52. Kaplan, N. et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).
53. Wells, R. D. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.* **32**, 271–278 (2007).
54. Rich, A. & Zhang, S. Timeline: Z-DNA: the long road to biological function. *Nat. Rev. Genet.* **4**, 566–572 (2003).
55. Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13**, 770–780 (2012).
56. Maizels, N. & Gray, L. T. The G4 genome. *PLoS Genet.* **9**, e1003468 (2013).
57. Bacolla, A., Wang, G. & Vasquez, K. M. New perspectives on DNA and RNA triplexes as effectors of biological activity. *PLoS Genet.* **11**, e1005696 (2015).
58. Jain, A., Wang, G. & Vasquez, K. M. DNA triple helices: biological consequences and therapeutic potential. *Biochimie* **90**, 1117–1130 (2008).
59. Levens, D., Baranello, L. & Kouzine, F. Controlling gene expression by DNA mechanics: emerging insights and challenges. *Biophys. Rev.* **8**, 23–32 (2016).
60. Zhou, T. et al. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl Acad. Sci. USA* **112**, 4654–4659 (2015).
61. Maurano, M. T. et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47**, 1393–1401 (2015).
62. Pai, A. A., Pritchard, J. K. & Gilad, Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet.* **11**, e1004857 (2015).
63. Mercer, T. R. et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30**, 99–104 (2012).
64. Bartonicek, N. et al. Intergenic disease-associated regions are abundant in novel transcripts. *Genome Biol.* **18**, 241 (2017).
65. Betts, J. A. et al. Long noncoding RNAs CUPID1 and CUPID2 mediate breast cancer risk at 11q13 by modulating the response to DNA damage. *Am. J. Hum. Genet.* **101**, 255–266 (2017).
66. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363–367 (2000).

67. Vinckenbosch, N., Dupanloup, I. & Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl Acad. Sci. USA* **103**, 3220–3225 (2006).
68. Grander, D. & Johnsson, P. Pseudogene-expressed RNAs: emerging roles in gene regulation and disease. *Curr. Top. Microbiol. Immunol.* **394**, 111–126 (2016).
69. Thomson, D. W. & Dinger, M. E. Endogenous microRNA sponges: evidence and controversy. *Nat. Rev. Genet.* **17**, 272–283 (2016).
70. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351**, aac7247 (2016).
71. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
72. Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.* **9**, e1003153 (2013).
73. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
74. Shihab, H. A. et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543 (2015).
75. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
76. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).