

EDITORIAL

Open Access



Human disease genomics: from variants to biology

Mark I. McCarthy^{1,2,3*} and Daniel G. MacArthur^{4,5*}

Abstract

We summarize the remarkable progress that has been made in the identification and functional characterization of DNA sequence variants associated with disease.

Editorial

The central objectives of human genetic research are to identify the sequence variation that plays a causal role in the development of disease, and then to use this information to generate insights into the biology of health and disease that can support clinical translation. These objectives have long been realized for a subset of diseases attributable to rare, high-penetrance alleles but, until recently, the challenges inherent in extending this success to the common, multifactorial diseases that account for most human illness have often seemed insurmountable. Nevertheless, over the past decade, the advent of genome-scale approaches for testing variant association to disease, and their application to increasingly large sample sets, has transformed our ability to identify alleles underlying rare and common diseases alike. At the same time, the arrival of a growing battery of sequence-based genomic assays has accelerated the high-throughput characterization of variant and transcript function, and these approaches are increasingly able to highlight the mechanisms through which disease-risk alleles operate. This is the best of times for human disease genomics, and this special issue of *Genome Biology* charts the remarkable progress that the community has made.

The first theme that emerges from the manuscripts published in this issue is the tremendous progress now

being made in understanding the genetic basis of rare, typically monogenic, diseases, thanks in large part to rapid advances in the development and uptake of high-throughput DNA-sequencing methods. The advent of cost-effective exome sequencing in particular has ushered in a second golden age of Mendelian disease gene discovery, in large part due to the ability to discover causal variants that were inaccessible with previous technologies. This includes a wave of discovery of newly occurring (de novo) mutations that were largely invisible in the era of family linkage studies [1], as well as the detection of complex structural rearrangements that were difficult or impossible to characterize with previous array-based methods [2]. In this issue, authors report the highly successful application of large-scale sequencing approaches to the diagnosis of ciliopathies [3] and disorders of sexual development [4], and to the discovery, through linkage and exome sequencing, of a novel gene for a complex neuropsychiatric disorder [5].

One of the crucial challenges in the genomic era of rare disease diagnosis is determining exactly which of the many potentially functional variants found in any patient's genome actually contributes to their disease. This process is complicated by the fact that existing databases of reported disease-causing mutations are heavily contaminated by false-positive reports of pathogenicity, in large part a hangover from an era of discovery in the absence of large databases of variant frequency in the general population. In this issue, Abouelhoda and colleagues [6] demonstrate that exome-sequencing data from Saudi Arabian samples can help reclassify many such variants, due to the largely unexplored nature of the variation in the Middle Eastern region and the presence of autozygosity. Nevertheless, much work remains to be done to clean up variant databases and to empower accurate variant classification as we move into an era of increasingly pervasive sequencing in both disease patients and healthy individuals.

The second theme that emerges is the increasing recognition that monogenic and complex disease are not

* Correspondence: mark.mccarthy@dr1.ox.ac.uk; danmac@broadinstitute.org

¹Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK

⁴Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

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



discrete entities, but rather lie along a general spectrum of human disease. For a growing number of alleles implicated in monogenic disease, wider access to sequence data has revealed the extent to which ascertainment through densely affected pedigrees has led to a general overestimation of penetrance. Genotypes previously thought to be tightly coupled to significant early-onset pathology have often turned out to be compatible with normal health. In some cases, the variants are genuinely innocuous, and there has been false attribution of their relationship to disease. In others, the variants concerned have variable penetrance, presumably reflecting the contribution of other genetic, environmental, or stochastic factors that modulate their phenotypic impact. From the perspective of common disease research, there are ever more examples of genes that harbor an extended series of disease-associated alleles, ranging widely in allele frequency and effect size. In this issue, for instance, Jansen and colleagues [7] describe how whole-exome sequencing in unrelated subjects with non-familial Parkinson's disease has revealed multiple genes that contain homozygous (or compound heterozygous) loss-of-function alleles, several of which displayed disease-relevant phenotypes when manipulated in appropriate cellular models.

The historical distinction between 'Mendelian' and 'polygenic' alleles is a throwback to 19th century arguments between the Mendelian and biometrician schools of evolutionary thought, a debate sustained into the modern era by the restricted discovery range of the earliest genome-wide techniques—linkage analysis in multiplex pedigrees on the one hand, common variant genome-wide association analysis on the other. With sequence-based discovery now possible across a variety of sample types—from pedigrees and trios to unselected populations—the continuum of risk allele effects and frequencies has been exposed, and can be exploited for mechanistic and translational benefit [8]. In this issue, Delahaye-Duriez and colleagues [9] demonstrate how sets of genes implicated in both rare and common forms of epilepsy converge on a pathway of proteins involved in synaptic function, and Yu and colleagues [10] show that a set of genes in which sequence variants are associated with serum amino acid levels overlap strikingly with known Mendelian metabolic disease loci. A detailed understanding of both rare, early-onset disease and common, later-onset disease will require a more holistic view of genetic risk. As outlined in two other contributions, the phenotypic consequences of sequence variation can not only be observed at the level of the individual, but also play out over many generations through their impact on fecundity, as revealed through evidence of selection [11, 12].

The identification of risk-alleles is a largely sterile exercise until it is possible to use them to drive

improvements in the understanding of disease pathogenesis. The third theme represented in the manuscripts published in this special issue is the accelerating pace with which the integration of genome-scale annotation is delivering biological meaning for the discovered alleles. For many common diseases, it is becoming increasingly clear that most of the genetic variance is attributable to common variants. The variants responsible overwhelmingly map in non-coding sequence, and presumably act through the transcriptional regulation of downstream effector transcripts, often in tissue-specific ways.

The advent of high-throughput genome-scale technologies for mapping sites of regulatory impact (through assays of chromatin accessibility and modification, methylation status, transcription factor binding, and the like) and their implementation across a wide range of tissues (pioneered by the ENCODE consortium and the Epigenome Road Map) has provided the Rosetta Stone that links risk variant localization to functional impact. The integration of genome-wide genetic association data with genomic-scale annotation data has often resulted in a virtuous cycle of mutual advantage, defining the tissues, and sometimes the specific cell types, that are central to the pathogenesis of the disease of interest, supporting more accurate fine-mapping of risk alleles and characterizing key regulatory circuitry. Papers in this issue provide several examples, including studies in schizophrenia [13] and in inflammatory conditions [14, 15].

In addition, by linking regulatory elements to their downstream targets, for example through detection of *cis*-expression signals [16] and physical DNA–DNA interactions [17], researchers have been able to make more confident assignments regarding the effector transcripts through which the various risk-alleles exert their phenotypic impact. Additional insights have flowed from the ability to compare the genome-wide association profiles of apparently diverse phenotypes, highlighting unsuspected overlaps, for example between the innate immune system and autism spectrum disorder [18]. Key to these advances has been the development of novel tools for the statistical analysis and visualization of these complex data sets; two such tools are described in this issue [19, 20].

The next few years will bring increasingly massive genomic data sets from patients and controls, as well as more sophisticated mechanisms for the analysis and integration of a wide variety of genomic data types. We can expect to see not only continued growth in the number of disease genes identified, but also a deepening of our understanding of the fundamental genetic architecture of human disease states, and a transformation in our ability to move from associated genes, to pathways,

to biology and clinical translation. The articles in this special issue illustrate our community's exciting progress along each of these avenues.

Authors' contributions

Both authors wrote and approved the final manuscript.

Competing interests

MIM serves on advisory panels for Pfizer and NovoNordisk; has received honoraria from Pfizer, NovoNordisk, and Eli Lilly; and has received research funding from Pfizer, Eli Lilly, NovoNordisk, Sanofi-Aventis, Boehringer Ingelheim, Astra Zeneca, Janssen, Takeda, Roche, Merck, Abbvie, and Servier as part of precompetitive research initiatives supported by the Innovative Medicines Initiative and the Accelerating Medicines Partnership. DGM is a founder with equity in Goldfinch Biopharma, and has received research funding from Biogen, Merck, Pfizer, and Eisai.

Author details

¹Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK. ²Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. ³Oxford NIHR Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK. ⁴Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. ⁵Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA.

Published online: 30 January 2017

References

- Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* 2016;17:241.
- Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR et al. Defining the spectrum of large complex structural variation and extreme chromothripsis in the Morbid Human Germline. *Genome Biol.* 10.1186/s13059-017-1158-6
- Shaheen R, Szymanska K, Basu B, Patel N, Ewida N, Faqeh E, et al. Characterizing the morbid genome of ciliopathies. *Genome Biol.* 2016;17:242.
- Eggers S, Sadedin S, van den Bergen JA, Robevska G, Ohnesorg T, Hewitt J, et al. Disorders of sex development: insights from targeted gene sequencing of a large international patient cohort. *Genome Biol.* 2016;17:243.
- Shamseldin HE, Masuho I, Alenizi A, Alyamani S, Patil DN, Ibrahim N, et al. GNB5 mutation causes a novel neuropsychiatric disorder featuring attention deficit hyperactivity disorder, severely impaired language development and normal cognition. *Genome Biol.* 2016;17:195.
- Abouelhoda M, Faquih T, El-Kalioby M, Alkuraya FS. Revisiting the morbid genome of Mendelian disorders. *Genome Biol.* 2016;17:235.
- Jansen IE, Ye H, Heetveld S, Lechler M, Michels H, Seinstra RI et al. Discovery and functional prioritization of Parkinson's disease candidate genes from large-scale whole exome sequencing. *Genome Biol.* 10.1186/s13059-017-1147-9
- Katsanis N. The continuum of causality in human genetic disorders. *Genome Biol.* 2016;17:233.
- Delahaye-Duriez A, Srivastava P, Shkura K, Langley SR, Laaniste L, Moreno-Moral A, et al. Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. *Genome Biol.* 2016;17:245.
- Yu B, de Vries PS, Metcalf GA, Wang Z, Feofanova EV, Liu X, et al. Whole genome sequence analysis of serum amino acid levels. *Genome Biol.* 2016;17:237.
- Quintana-Murci L. Understanding rare and common diseases in the context of human evolution. *Genome Biol.* 2016;17:225.
- Sams AJ, Dumaine A, Nédélec Y, Yotova V, Alfieri C, Tanner JE, et al. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol.* 2016;17:246.
- Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol.* 2016;17:176.
- Meddens CA, Harakalova M, van den Dungen NAM, Asl HF, Hijma HJ, Cuppen EPJG, et al. Systematic analysis of chromatin interactions at disease associated loci links novel candidate genes to inflammatory bowel disease. *Genome Biol.* 2016;17:247.
- Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol.* 2016;17:255.
- Joehanes R, Zhang X, Huan T, Yao C, Ying S-X, Nguyen QT et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* 10.1186/s13059-016-1142-6
- McGovern A, Schoenfelder S, Martin P, Massey J, Duffus K, Plant D, et al. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.* 2016;17:212.
- Nazeen S, Palmer NP, Berger B, Kohane IS. Integrative analysis of genetic datasets reveals a shared innate immune component in autism spectrum disorder and its co-morbidities. *Genome Biol.* 2016;17:228.
- Liang S, et al. iRegNet3D: 3D integrated regulatory network for the genomic analysis of coding and non-coding disease mutations. *Genome Biol.* 2017;18:10.
- Chen L, Jin P, Qin ZS. DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biol.* 2016;17:252.