Understanding human genetic variation in the era of high-throughput sequencing

Julian C. Knight

The EMBO/EMBL symposium 'Human Variation: Cause and Consequence' highlighted advances in understanding the molecular basis of human genetic variation and its myriad implications for biology, human origins and disease. As high-throughput sequencing allows us to define genetic variation and its functional consequences at genome-wide resolution for a large number of people, important questions need to be asked about how to use new technologies to maximize the translational relevance of genetic research for society and the individual patient.

The amounts of human genetic data now being generated are unprecedented. Indeed, at the recent EMBO/EMBL symposium 'Human Variation: Cause and Consequence', held in Heidelberg between 20 and 23 June 2010, Richard Durbin (Wellcome Trust Sanger Institute), co-chair of the 1000 Genomes Project, described such data sets as astronomical. The growing use of massively parallel high-throughput sequencing (HTS) for applications ranging from rare variant discovery to functional genomic analyses raises significant issues in terms of the questions being asked and how we manage and use such data sets. Since 2008, sequence data for over ten trillion bases of DNA have been produced by the 1000 Genomes Project. As this work moves forward to include 2,500 DNA samples from 27 populations, the bioinformatic challenges raised will only increase. Our knowledge about the extent of sequence and structural genomic variation looks set to increase at an exponential rate. However, as this happens, there is a danger that public and scientific expectations about how such knowledge can be translated into advances in understanding biology, and in particular the genetic basis of disease, will increasingly be perceived as not being met.

Genome-wide association (GWA) studies using common genetic markers have led to enormous advances in our understanding of the genetic basis of common multifactorial traits, but can explain only a minority of expected genetic risk (Eichler *et al*, 2010). This is perhaps not surprising given that our knowledge of genetic variation,

notably of rarer variants and complex structural genomic variation, is incomplete. Moreover, it remains unclear how to define the complex interactions involving genetic and environmental factors that underpin specific traits; the role of epigenetics and, in particular, parent of origin effects; and the myriad ways in which genetic variation might have an impact on the regulation of gene expression or function of encoded proteins (Knight, 2009). The EMBO/EMBL symposium underlined the need to consider genetic variation in a functional genomic context in the hope of understanding the mechanisms by which mutations arise, are co-inherited and selected for in populations,

and might define specific molecular phenotypes or complex traits based on individual genetic and epigenetic differences together with environmental modulators. One of the many successes of the meeting was to bring together individuals with a range of expertise to bridge the gap that sometimes exists between Mendelian and common disease genetics, and between basic science research into mechanisms of mutagenesis and functional genomics. The resulting discussions were insightful and lively, appropriate for a field in which important decisions are being made about how we should move forward to understand the nature of genetic variation and use that information to improve human health.

Meiosis, mutations and cancer

Meiotic recombination is a key event in the generation of genomic diversity, with the crossovers clustering at particular genomic locations dependent on DNA sequence and epigenetic marks. Gil McVean (U. Oxford) described recent work which identified the zinc-finger protein PRMD9 as a key determinant of such crossover events (Myers *et al*, 2010). PRMD9 is expressed in germ cells during meiosis and has histone methyltransferase activity, consistent with histone marks seen at recombination hotspots. The zinc-finger domain is highly polymorphic and associated with variability in hotspot use, both in human populations and between species where the degree of observed divergence is extraordinary, consistent with differences in hotspot locations. The presence of a coding microsatellite in the zinc-finger array

is associated with a high intrinsic mutation rate and the exceptionally rapid evolution of this protein, with binding site specificity defined by positive selection.

Andrew Wilkie (U. Oxford) highlighted the concept of selfish mutations in spermatogenesis, in which mutational events in the same cell might lead to clonal expansion through selective growth advantage. These events can manifest as germline and somatic mutations with different phenotypes depending on the degree of activation of the mutations (Goriely *et al*, 2009). Wilkie's work offers unique insights into somatic and germline mutations, which normally involve separate cells at different times. The insights arise from Wilkie's observation that germline mutations underlying some congenital disorders are more common in older males. These 'paternal age effect mutations'—for example, the dominant gain-of-function mutations in fibroblast growth factor receptor 3 (*FGFR3*) and Harvey rat sarcoma viral oncogene (*HRAS*)—involve rare initial mutations that are moderately activating, become enriched as a result of clonal expansion and result in congenital disorders such as achondroplasia or Costello syndrome. Strongly activating mutations in spermatogonia involving the same genes might result in somatic mutation in the testes leading to spermatocytic seminoma when combined with secondary mutations, or fetal lethality when transmitted in the germline.

Some of the most dramatic insights from HTS have arisen from the study of the cancer genome. Mike Stratton (Wellcome Trust Sanger Institute) illustrated how sequencing data will advance our understanding of the evolution of the cancer genome, in particular of factors generating variation and how selective forces in the tissue micro-environment can act on the phenotypic diversity arising in cell populations as heritable somatic mutations are acquired by individual cells. The number of driver mutations required to result in cancer is thought to reflect the number of biological processes that need to be modified, with typically 10 driver mutations and up to 100,000 passenger somatic mutations found in the sequence of a cancer genome. So far, 385 protein-encoding genes are known to be somatically mutated and causally involved in oncogenesis. With sufficient coverage, the full catalogue of somatic mutations for an individual cancer genome can now be defined by HTS. The International Cancer

Genome Consortium (www.icgc.org/icgc) will define 25,000 cancer genomes worldwide for at least 50 cancers, combined with detailed transcriptomic and epigenomic profiling. Such extraordinary data sets will be of direct clinical utility and, as sequencing costs continue to fall, full or partial cancer genome sequencing should become part of routine clinical care for individual cancer patients.

Gene expression, common disease

The application of HTS to analyse the transcriptome, transcription factor binding, chromatin accessibility and histone modifications is allowing the functional genome to be mapped in remarkable detail. Emmanouil Dermitzakis (U. Geneva) highlighted the use of RNA-seq in expression quantitative trait (eQTL) mapping to identify a larger number of expression-associated variants at high resolution, as well as quantifying allele-specific gene expression and alternative splicing (Montgomery *et al*, 2010). He noted that gender-specific associations and evidence of tissue specificity are much more common than previously appreciated. This means that particular care is needed when integrating eQTL analyses with GWA studies of disease, to ensure that gene expression is analysed in relevant cell and tissue types.

...the role of rare variants in the genetic architecture of common diseases remains unresolved, but the advent of affordable large-scale whole-genome sequencing will increasingly allow this question to be addressed

Ewan Birney (EMBL-EBI, UK) described a detailed analysis of chromatin accessibility and CTCF (a zinc-finger protein) binding for lymphoblastoid cell lines showing that individual and allele-specific differences are common and heritable, with some of the variance due to *cis*-acting effects (McDaniell *et al*, 2010). The FAIRE assay (formaldehyde-assisted isolation of regulatory elements) for nucleosome depletion is technically easier than analysis of chromatin accessibility based on DNase I hypersensitivity and can be analysed by HTS (FAIRE-seq). The value of this approach was underlined by Jorge Ferrer (IDIBAPS, Barcelona) who used FAIRE to identify open chromatin specific to human pancreatic

islets (Gaulton *et al*, 2010). Strikingly, an intronic variant of *TCF7L2* associated with type 2 diabetes (T2D) was found in such a site to modulate accessibility and enhancer activity in an allele-specific manner. This study highlights the importance of analysing a disease-relevant tissue and demonstrates how overlaying functional genomic maps can inform the results of GWA studies and help resolve causal variants.

Mark McCarthy (U. Oxford) highlighted the progress that has been made in defining genetic susceptibility loci for T2D. The 38 confirmed loci show individually small effects and explain approximately only 10% of observed familial clustering (Voight *et al*, 2010). However, they provide important new insights into disease pathogenesis and when integrated with phenotypic and expression data, indicate an underlying mechanism and complexity, possibly involving imprinted loci and epigenetic modulators. Loci implicated in monogenic and multifactorial forms of diabetes support a role for rarer variants not currently resolved by GWA studies. A similar picture was presented by Mark Daly (Mass. General/Broad Institute) for Crohn disease, for which 71 distinct genomic loci have been defined but in which specific functional variants have not been resolved in most cases. Important novel insights into disease pathways such as autophagy have been established and an overlap is seen with loci associated with infectious disease susceptibility, linking specific viral infections with the development of Crohn disease. For both Crohn disease and T2D, the translation of studies into personalized medicine is not yet feasible but might be helpful diagnostically, for example in distinguishing between forms of inflammatory bowel disease.

Rare variants

So far, the role of rare variants in the genetic architecture of common diseases remains unresolved, but the advent of affordable large-scale whole-genome sequencing will increasingly allow this question to be addressed. Helen Hobbs (U. Texas Southwestern) noted that by focusing on rare variants there is a greater likelihood of finding alleles with a large effect size that provide a more direct link to function and translation. This was illustrated by work looking at extremes of distribution for phenotypes such as HDL cholesterol in the Dallas Heart Study. Subsequent sequencing of candidate genes for all participants

has highlighted the role of rare variants for a range of complex traits (Romeo *et al*, 2009).

Durbin gave an overview of pilot work from the 1000 Genomes Project. Sequencing to a depth of at least 20–30x is required to find almost all variants in a single sample, whereas a low coverage strategy across many samples is efficient to find shared variants. Of the 14.5 million single-nucleotide polymorphisms identified so far, 8 million are novel with most specific to distinct populations, and the majority found among Africans. Realignment and reassembly are crucial with inference of the correct reference sequence needed. The sequencing of family trios to high depth suggests a germline *de novo* substitution rate of 1×10^{-8} per generation with a 7–12-fold higher somatic rate. Plans were outlined to sequence 2,500 people from 27 populations by the end of 2011 at 4x, giving 95% power to find variants at 1% frequency. The public availability of data through the Amazon Web Services Cloud, announced while the symposium was being held, highlights the community's efforts to ensure the accessibility of the resources needed for analysis.

Application to the clinic

Translating insights from the study of human genetic variation into the clinic is a major goal of current research and is likely to have the most immediate impact through pharmacogenomics. This aims to maximize the treatment benefit for the individual patient while minimizing risk of harm. As our understanding of genetic risk and its relationship to other risk factors becomes clearer, the opportunities for personalized

medicine informed by improved screening with more accurate prognostic information will increase, while novel insights into disease pathogenesis will undoubtedly lead to new therapeutic opportunities.

...the pioneering genetic research carried out by [Leena] Peltonen also highlights how similar people are, and that genetics should be a force for good

The explosion of knowledge about the genetics of common disease achieved over the past decade was underlined by Kari Stefansson (deCODE genetics) in his keynote address describing the remarkable insights gained from study of the Icelandic population. His talk emphasized the importance of considering genetic variation in the context of the environment and natural selection, highlighting successes in identifying common variants associated with diverse medical conditions. Although in many instances the associated genetic risks were small, the insights into biology are potentially profound and reveal shared associations between diseases and specific pathways. In other instances, associations suggest direct clinical application, for example to define genetically corrected levels of prostate-specific antigen. The remarkable data repository underlying the genetic association studies conducted with the Icelandic population includes a detailed genealogy that, together with long-range phasing,

allows parent-of-origin effects to be defined. The risk associated with T2D for a sequence variant at chromosome 11p15 for example is dependent on whether the paternal or maternal allele was inherited, and this correlates with DNA methylation (Kong *et al*, 2009).

Finally, the insights gained from another genetically isolated population bordering the arctic circle, Finland, were highlighted by Stefansson and Durbin in a memorial session dedicated to the extraordinary life and work of Leena Peltonen (1952–2010), who was one of the original organizers of this EMBO/EMBL Symposium. It is perhaps appropriate to end with comments by Stefansson that despite the genetic variation that exists, the pioneering genetic research carried out by Peltonen also highlights how similar people are, and that genetics should be a force for good.

REFERENCES

- Eichler EE *et al* (2010) *Nat Rev Genet* **11**: 446–450
Gaulton KJ *et al* (2010) *Nat Genet* **42**: 255–259
Goriely A *et al* (2009) *Nat Genet* **41**: 1247–1252
Knight JC (2009) *Human Genetic Diversity. Functional Consequences for Health and Disease*. Oxford University Press
Kong A *et al* (2009) *Nature* **462**: 868–874
McDaniell R *et al* (2010) *Science* **328**: 235–239
Montgomery SB *et al* (2010) *Nature* **464**: 773–777
Myers S *et al* (2010) *Science* **327**: 876–879
Romeo S *et al* (2009) *J Clin Invest* **119**: 70–79
Voight BF *et al* (2010) *Nat Genet* **42**: 579–589

Julian C. Knight is at the Wellcome Trust Centre for Human Genetics at the University of Oxford, Oxford, UK.

E-mail: julian@well.ox.ac.uk

Published online 20 August 2010

EMBO reports (2010) **11**, 650–652.
doi:10.1038/embor.2010.126