Genome **Medicine**

RESEARCH HIGHLIGHT

From sequence to functional understanding: the difficult road ahead

Periklis Makrythanasis[1] and Stylianos E Antonarakis[1,2]*

Abstract

DNA sequencing has become cheap, rapid and accurate, allowing us to access thousands of genomes and reveal the extensive variation among individuals. The major problem that arises from this is distinguishing between neutral and pathogenic variants. A recent study by Davis *et al*., in which a functional screen of all the non-synonymous variants of a newly discovered gene was performed, highlights the value and necessity of characterizing the functional consequences of each genomic variant discovered. This is the main challenge for the advancement of genomic medicine in the years to come.

Human genomic variation: from sequence to function

The recent development of high-throughput sequencing (HTS) and its application for sequencing the exomes or genomes of thousands of people (including participants of the 1000 Genomes Project) has provided experimental evidence of the extensive variability of the human genome (both in single nucleotide polymorphisms (SNPs) and copy number variations (CNVs)). Within the coding fraction of the genome (the exome), each individual is estimated to have approximately 8,500 to 10,500 non-synonymous variants, 350 to 400 of which are predicted to cause loss-of-function alleles affecting 250 to 300 genes [1]. HTS data have also provided experimental evidence that the mutation rate of the human genome is 10^{-8} per nucleotide per generation, resulting in two to seven new variants in each individual exome [2].

The most difficult challenge for HTS projects aiming to discover pathogenic variants is the correct identification of the disease-causing mutations among thousands of additional variants that could be either contributing to unrecognized phenotypes or neutral [3]. At present, most HTS projects focus on the known functional elements of the genome. Protein-coding genes are at the heart of this analysis, along with non-coding transcripts and highly conserved non-coding sequences. The rules of heredity, gene expression data, evolutionary principles and protein structure-function relationships provide the current set of criteria for deciding between potential contributing and non-contributing variants relative to the phenotype in question. The phenotype is also an important consideration because identified variants may contribute to other phenotypes but not the one in question. Furthermore, the correlation between genome variation and phenotypic variation is relatively simple for monogenic/oligogenic phenotypes and highly penetrant variants, but is complicated for polygenic phenotypes and for medium or low-penetrance variants.

More precise examples of these criteria are: the presence of the variants and their allelic composition in affected and non-affected individuals according to the mode of inheritance imposed or hypothesized; the mapping position of the variants following linkage or association studies in families and populations; the predicted functional consequence of the variant (missense, nonsense, frameshift or splice-site); the evolutionary conservation of the affected codon; the expected disruption of the protein's structure; the frequency of the variant in the population without the phenotype in question; the potential disruption of a protein network; and the predicted 'recessive' or 'dominant' nature of variants in a gene of interest. There are computer prediction programs using some of these criteria for predicting the likely pathogenicity of non-synonymous variants [4].

Establishing the function of human genomic variants

However, the 'prior probability' for the pathogenicity of the majority of non-synonymous variants is not satisfactory, the gray zone of uncertainty is extensive, and most investigators ultimately require experimental evidence for the functionality of each variant. A recently

*Correspondence: Stylianos.Antonarakis@unige.ch
[1]Department of Genetic Medicine and Development, University of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland
Full list of author information is available at the end of the article

published paper by Davis *et al.* [5] provides an excellent example of such a functional screening study. The authors studied the *TTC21B* gene in 753 patients with ciliopathies and 398 controls in order to examine the spectrum and disease contribution of variants. The *TTC21B* gene encodes the IFT139 protein, which is involved in retrograde intraflagellar transport in cilia and negatively modulates Sonic Hedgehog signal transduction [6]. Forty non-synonymous variants of *TTC21B* were identified in patients, and all of these were studied in a functional assay using zebrafish embryos to establish pathogenicity. Briefly, the embryonic phenotype associated with reduced levels of the zebrafish *TTC21B* ortholog can be rescued using human *TTC21B* mRNA. Different mRNAs carrying HTS-identified non-synonymous variants of this gene either failed to rescue or partially or completely rescued the phenotype; these represent functional null, hypomorphic and benign alleles, respectively. The functional studies provided evidence for *TTC21B* causative variants in ciliopathies such as Jeune asphyxiating thoracic dystrophy (JATD) and nephronophthisis (NPHP); furthermore, other *TTC21B* variants function as modifier alleles in additional ciliopathies. The functional evidence for each allelic variant is pivotal in the understanding of the observed phenotype. A caveat, however, is that we cannot always predict the effect of a variant on the human phenotype from the experiments in model organisms. This is even more relevant in cases such as those studied by Davis *et al.* [5], in which a dysfunctional protein may result in different disorders.

For proteins for which there are functional assays, one could predict that databases will be developed with the functional results for all variants detected for specific proteins. Functional validation of non-synonymous variants could be performed using several laboratory models, using either whole organisms (such as yeast, *Drosophila*, fish or mice), or cells (such as cell-based models derived from humans or other organisms and *in-vitro*-differentiated cells). The advantage of such functional assays is that they provide not only the functional proof of the pathogenicity of a variant, but also novel insights into protein function and perhaps even the mechanism of disease. Unfortunately, there are no *TTC21B*-like functional assays for the majority of proteins, and most of the methods to test functionality are not amenable to large-scale screening approaches. Thus, considerable effort should be made to develop large-scale screening assays for all possible non-synonymous variants for all human proteins.

The challenges ahead

This is only the tip of the iceberg for the characterization of pathogenic variants. Assays need to be developed for the assessment of variants in all functional genomic elements outside the protein-coding genes. There is a sea of non-coding transcripts [7,8], hundreds of thousands of genomic regions with potential regulatory function [9], and hundreds of thousands of conserved non-coding regions with unknown but presumably important function [10,11]. This substantial fraction of the genome, for which we do not know the functional rules and constraints, could harbor variants for which functional assays need to be developed. This is obviously a major obstacle in the evaluation of the majority of the genomic variability. It is expected that the technology used in the ENCODE [9] and other projects will enhance our knowledge on the functional elements of our genomes. In addition, it is well known that the contribution of pathogenic variants to the phenotype is modified by the overall genomic variability of each individual, a notion in genetics known as 'penetrance'. Thus, an experimentally proven pathogenic allele may result in a phenotype in some individuals, but not in others.

We now have the ability to read almost entire individual genomes in a reasonable time-frame, and this is cause for celebration. On the other hand, the daunting task in front of us is the functional understanding of the extensive genomic variation (common and rare) that now populates the hard disks of supercomputers and biobanks. The next decade at the leading edge of genetic medicine will certainly be dedicated to this effort. And as the new graduate students and physicians in training now realize: sequencing is simple; functional understanding is still a dream.

Abbreviations

HTS, high-throughput sequencing.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PM and SEA wrote the manuscript.

Authors' information

PM and SEA are working on human genome variation and the detection of pathogenic mutations in monogenic disorders. Additional areas of interest and expertise are aneuploidies and non-coding conserved sequences.

Acknowledgments

PM is supported by a grant from the Bodossakis Foundation. SEA is funded by the Swiss National Science Foundation and the European Union ERC grant. The authors thank Richard Fish for editing and discussing the manuscript.

Author details

[1]Department of Genetic Medicine and Development, University of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland. [2]Service of Genetic Medicine, University Hospitals of Geneva, Rue Michel Servet 1, 1211 Geneva, Switzerland.

Published: 6 April 2011

References

1. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA: A map of human genome variation from population-scale sequencing. *Nature* 2010, **467**:1061-1073.
2. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B,

Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA: A *de novo* paradigm for mental retardation. *Nat Genet* 2010, **42:**1109-1112.

3. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009, **461:**272-276.
4. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: A method and server for predicting damaging missense mutations. *Nat Methods* 2010, **7:**248-249.
5. Davis EE, Zhang Q, Liu Q, Diplas BH, Davey LM, Hartley J, Stoetzel C, Szymanska K, Ramaswami G, Logan CV, Muzny DM, Young AC, Wheeler DA, Cruz P, Morgan M, Lewis LR, Cherukuri P, Maskeri B, Hansen NF, Mullikin JC, Blakesley RW, Bouffard GG; NISC Comparative Sequencing Program, Gyapay G, Rieger S, Tönshoff B, Kern I, Soliman NA, Neuhaus TJ, Swoboda KJ, *et al.*: TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum. *Nat Genet* 2011, **43:**189-196.
6. GeneCards [http://www.genecards.org]
7. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES: Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009, **458:**223-227.
8. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R: Long noncoding RNAs with enhancer-like function in human cells. *Cell* 2010, **143:**46-58.
9. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, *et al.*: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007, **447:**799-816.
10. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE: Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 2002, **420:**578-582.
11. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: Ultraconserved elements in the human genome. *Science* 2004, **304:**1321-1325.

doi:10.1186/gm235
Cite this article as: Makrythanasis P, Antonarakis SE: From sequence to functional understanding: the difficult road ahead. *Genome Medicine* 2011, **3**:21.