Minireview

Rare protection against type 1 diabetes

Robert M Plenge

Address: Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. Email: rplenge@partners.org

Abstract

A large population study using ultra-high-throughput DNA sequencing to re-sequence a genetic locus associated with type 1 diabetes reveals rare protective alleles.

Unlike inherited Mendelian diseases such as cystic fibrosis, the common complex diseases such as diabetes and heart disease have no single predisposing genetic factor. But, over the years, alleles at various genetic loci that either decrease or increase the risk of developing such diseases, in relation to their incidence in the general population, have been uncovered. Geneticists have made the somewhat artificial distinction between 'common' and 'rare' when describing alleles at polymorphic genetic loci. Classically, 'common' variants are those occurring at a frequency of more than 1% in any one continental population (for example, Europeans, Asians or Africans), whereas 'rare' variants are present at a frequency of less than 1%.

This distinction is used to frame the genetic approach to discovering and testing DNA variants and linking them to disease. For common variants, it is possible to screen a reference population to identify a catalog of variants (discovery phase), and then test these variants in collections of cases and matched controls (case-control studies) using high-throughput genotyping technologies (testing phase). For rare variants, both the discovery and testing phases can only be done in case-control collections themselves. In a recent paper in Science, Nejentsev et al. [1] describe the use of next-generation DNA sequencing of a large case-control population to search for, and find, rare protective alleles at a locus associated with susceptibility to type 1 diabetes, the gene IFIH1 (interferon induced with helicase C domain 1).

A variety of resources have been developed to test the role of common single nucleotide polymorphisms (SNPs) in common human diseases such as type 1 diabetes. Most notably, the International Haplotype Map (HapMap) Project has generated a catalog of more than 3 million variants in three continental populations (Europeans, Asians and Africans) [2,3]. On the basis of the number of individuals genotyped ($n = 90$ in each continental population), the HapMap Project was calibrated to study DNA variants of greater than 5% frequency. Phase III of HapMap, which is nearly complete, will push the allele-frequency spectrum to 1% (by genotyping more individuals), and extend the project to other continental populations (by genotyping individuals of different ancestries). High-throughput genotyping technologies [4] and statistical methods [5,6] have been developed to test common variants from the HapMap in large collections of patients and controls. A number of successes have emerged, including the identification of more than ten loci associated with type 1 diabetes [7].

Because of these evolving resources, a more accurate description of the allele-frequency spectrum is to classify DNA variants as those that are truly common (greater than 5% frequency), those that are truly rare (for example, private to a few families and of very recent mutational origin), and those that are low frequency (greater than 0.1% but less than 5%). Low-frequency variants are of recent mutational origin, but nonetheless segregate on a single ancestral haplotype and are amenable to testing with high-throughput genotyping technology - if they have been cataloged in a reference population. The number of low-frequency alleles available in reference populations such as HapMap depends upon the size of the reference population: the more subjects genotyped for known polymorphims, the larger the number of low-frequency variants that can be cataloged. The vast majority of common SNPs have been discovered and deposited in the dbSNP database, and the majority of these

have either been genotyped directly or adequately tagged by a SNP in HapMap. However, many low-frequency variants, and most rare variants, are not available in dbSNP and therefore cannot be tested in genome-wide association studies (GWAs). Thus, re-sequencing is required to discover and test low-frequency and rare variants.

Next-generation re-sequencing

Conventional Sanger sequencing technology still accounts for the overwhelming majority of DNA sequencing. However, cost constraints and throughput capacity limit its use in medical genetics. More recently, new sequencing systems based on massively parallel sequencing of short fragments by techniques such as pyrosequencing have been developed, which are poised to reduce DNA sequencing costs and raise capacity by several orders of magnitude. Often called next-generation sequencing, this technology is expected to mature rapidly over the next few years, prompting hope of the '$1,000 genome'. If successful, these sequencing advances will enable GWAs of common, rare and low-frequency variants.

To harness the full power of next-generation sequencing for testing the complete spectrum of alleles in human diseases, at least three challenges must be addressed. The first is to capture the target of interest; the second is to identify DNA variants from sequence data; and the third is to test the DNA variants for their role in disease. These three issues are now major bottlenecks in the widespread application of next-generation sequencing in medical genetics and I shall discuss the recent study of Nejentsev *et al.* [1] with those points in mind. Despite success in identifying two new protective alleles, their work illustrates well the limitations imposed by the three challenges.

Protective low-frequency variants in *IFIH1*

Nejentsev *et al.* [1] ask a simple question, which they answer quite convincingly by re-sequencing using the Roche/454 instrument: do low-frequency or private rare mutations predispose to common forms of type 1 diabetes? To address this, they sequenced the coding exons of ten genes with prior evidence of importance in type 1 diabetes; six of these genes harbor common SNPs that influence the risk of the disease (*PTPN22*, *PTPN2*, *IFIH1*, *SH2B3*, *CLEC16A*, and *IL2RA*). The authors capture 31 kb of target sequence using PCR-directed amplification of DNA pooled from either 10 cases or 10 controls and eventually identify 212 SNPs in this region (many important details on the sequencing and SNP calling are included in the Supplementary methods to [1]). This represents an average of 1 SNP per approximately 150 bp. From the pooled sequencing data, Nejentsev *et al.* [1] estimate that 33 SNPs have frequencies greater than 3%, and that 179 have frequencies less than 3%; of these 179, 156 were previously unknown.

In an association study using allele-frequency estimates, the authors then identify two low-frequency SNPs in the *IFIH1* gene that confer protection from type 1 diabetes. The SNP with the strongest association, rs35667974, was observed on an estimated 3 out of 960 case chromosomes, but 24 out of 960 control chromosomes (P = 0.00004); another SNP, rs35337543, was observed on 7 case chromosomes and 23 control chromosomes (P = 0.005). Both SNPs were genotyped in an additional 8,379 type 1 diabetes patients and 10,575 controls from Britain, and 3,165 families from Europe and the United States comprising one or more offspring with type 1 diabetes and their parents. In this extended genetic association study, rs35667974 was present at a frequency of around 1% among cases and 2% among controls (combined $P = 2.1 \times 10^{-16}$). This SNP resides in exon 14 and changes a conserved isoleucine at position 923 to valine. The other SNP, rs35337543, has a frequency of 1% versus 1.5% among cases and controls, respectively, and has a combined P-value of 1.4×10^{-4}. It resides within a conserved splice donor site at position +1 in intron 8 of *IFIH1*.

Remaining challenges

Several important details in the study of Nejentsev *et al.* [1] pertain to the three issues noted earlier: target capture; SNP calling from sequence data; and the statistical approach with which they test for association. To capture the coding exons of these ten genes, they used a PCR-based method, which is laborious and expensive. In the current study, 144 separate PCR fragments were required. It would be very difficult to scale this method to larger portions of the genome (for example, all coding exons). Other target-capture approaches also have limitations, including a lack of specificity of enrichment for the region of interest and the uniformity with which targets are captured [8,9].

The current study called SNPs and estimated allele frequency directly from the sequencing data. This was done from pools of ten individuals, without barcode identifiers. It is not clear how accurate this method will be. The study reports a strong correlation (r^2 = 0.99) between allele-frequency estimates from sequencing and genotyping data. However, these data are from only eight SNPs in twenty pooled DNA samples. Closer inspection of Supplementary Figure 1 of [1] shows that some of the pools have very different allele-frequency estimates, which raises questions about the accuracy of allele-frequency estimates from pooled sequencing data.

Finally, the association study design mimics that used for common SNPs: each individual SNP is tested for association in cases and controls separately, without regard to putative function and gene context. A more powerful approach, especially for truly private rare variants, might be to test whether any gene (or prespecified gene set) has an excess number of rare variants in either cases or controls [10].

This study shows that low-frequency variants can influence the risk of type 1 diabetes. Both rs35667974 and rs35337543 would be classified as low frequency rather than rare, private mutations, but even so re-sequencing was necessary to discover and test both of them. There are ongoing studies to expand SNP discovery and genotyping in reference populations (for example [11]). These resources should enable large-scale genetic association studies of putative functional SNPs such as rs35667974 and rs35337543. However, re-sequencing will be required to test the full spectrum of alleles, from truly common to truly rare. And, as demonstrated by Nejentsev *et al.* [1], next-generation re-sequencing represents a powerful tool.

References

1. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: **Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes.** *Science* 2009, **324:**387-389.
2. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, *et al.*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449:**851-861.
3. Consortium TIHM: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.
4. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES: **Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome.** *Science* 1998, **280:**1077-1082.
5. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559-575.
6. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38:**904-909.
7. Altshuler D, Daly MJ, Lander ES: **Genetic mapping in human disease.** *Science* 2008, **322:**881-888.
8. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J: **Multiplex amplification of large sets of human exons.** *Nat Methods* 2007, **4:**931-936.
9. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.** *Nat Biotechnol* 2009, **27:**182-189.
10. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR: **Power of deep, all-exon resequencing for discovery of human trait genes.** *Proc Natl Acad Sci USA* 2009, **106:**3871-3876.
11. **The 1000 Genomes Project** [http://www.1000genomes.org]