Power tools for human genetics

Leonid Kruglyak

Genome-wide association studies have the potential to identify systematically the contributions of common genetic variants to human disease, but the tools to carry out such studies have been incomplete. Assessments of the HapMap resource suggest that the tools are now in hand, but care is required in their use for study design, analysis and interpretation.

Humans show a great deal of heritable inter-individual variation in appearance, physiology, psychology, behavior and disease susceptibility. Geneticists and the general public alike want to understand the connection between this variation and variants in DNA sequence. In principle, this can be accomplished, at least for common variants, by comparing the frequency of every variant in individuals with and without a trait (or alternatively by comparing the trait distribution in individuals with and without a variant). In practice, although many of the ~7 million variants that exist with frequencies above 5% in the human population[1] have been identified[2], current genotyping technologies are limited to testing 100,000–500,000 variants in a collection of individuals in any given association study. Fortunately, the fact that nearby variants are correlated owing to a lack of historical recombination between them (such correlations are known as linkage disequilibrium) offers a shortcut alternative to genotyping every variant. As a result, the presence or absence of a variant at one site provides information about neighboring variants, allowing one variant to serve as a proxy for several others and reducing the total number of genotyping assays (**Fig. 1**). Several hundred thousand proxies can capture most common variants through strong correlations (the exact number depends on the study population and the criteria for proxy choice)[2,3]. But not just any proxy set of this size will do: variation in history among human populations and in recombination rates across the genome means that efficient proxy selection requires empirical measurement of linkage disequilibrium in multiple populations for a dense set of markers covering the genome[4,5]. This has now been accomplished by the International HapMap Consortium, as reported in *Nature*[2]. Two related studies in *Nature Genetics*, by de Bakker *et al.*[6] and Zeggini *et al.*[7], assess the utility of this resource for genome-wide association studies.

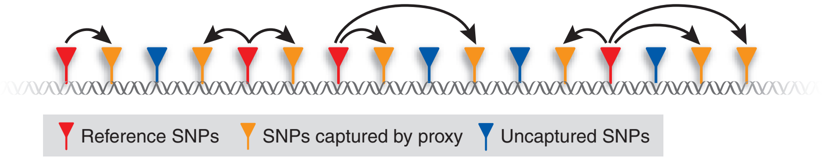

Figure 1 Schematic of a genomic region to be tested for association with a phenotype. The region contains 16 SNPs (triangles), any one of which could affect the phenotype. The four SNPs in red are genotyped directly (these are the reference or tag SNPs). The eight SNPs in orange are captured through correlations (linkage disequilibrium) with the SNPs in red (as denoted by arrows). The four SNPs in blue are neither genotyped nor correlated with genotyped SNPs, and so phenotypic association with any of these uncaptured SNPs would be missed.

Leonid Kruglyak is at the Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology, Carl Icahn Laboratory, Princeton University, Princeton, New Jersey 08544, USA.
e-mail: leonid@genomics.princeton.edu

What is the HapMap?

Most human sequence variation is in the form of single-nucleotide polymorphisms (SNPs), and these are the markers of choice for association studies. The International HapMap Consortium genotyped more than one million SNPs (roughly one per 3 kb) in 269 individuals of four ethnic origins[2]. A second phase of the project is slated to increase the density of genotyped SNPs to one per kilobase. de Bakker *et al.*[6] and Zeggini *et al.*[7] used combinations of empirical and simulated data to show that the current HapMap resource allows selection of reference panels with 250,000–500,000 SNPs that reliably capture a high fraction (70–80%) of all common variants (those with an allele frequency above 5%). The fraction captured is considerably lower in a panel of individuals of African origin, but this should be remedied when data from the second phase become available[2]. Thus, the HapMap project has largely succeeded in achieving its principal goal. Although the resource was not designed to capture variants with frequency below 5%, both groups considered its ability to do so and found that some rare variants are picked up (in the range of 20%). More of these variants can be recovered by exhaustive tests of association with multimarker haplotypes, but at the cost of lowering the power to detect association with common variants owing to the increased number of statistical tests. The relative contribution of rare versus common variants to phenotypic variation is a big question mark, but early successes[8] suggest that some common variants with important roles remain to be found.

How should a reference panel be assembled? The standard approach is to choose a minimal set of SNPs such that every common variant is in the panel or highly correlated with a SNP in the panel[9]. de Bakker *et al.* and Zeggini *et* *al.* show that panel size can be reduced by up to one-third if correlations between SNPs and multimarker haplotypes are considered. If this approach is pursued, a multimarker approach must also be used in association studies using the panel to avoid missing those variants correlated only with haplotypes (and not with any one SNP)[7]. de Bakker *et al.* propose a panel selection strategy in which SNPs are added to the panel in order of the number of SNPs they capture by proxy. In theory, this strategy can greatly reduce panel size while maintaining the power of an association study with a fixed number of individuals to detect a given effect. But this reduction can be achieved only by not including proxies for a substantial fraction of common variants, which means that power cannot be improved by increasing sample size. It is also worth noting that this strategy prioritizes the detection of variants strongly correlated with many others, making it difficult to identify the causal variant. In general, the average size of a group of strongly correlated SNPs is large (~10–20)[2]. Although it is precisely this observation that enables reduction of the number of assays by proxy selection, it has the flip side that distinguishing a causal variant from those merely strongly correlated with it will often not be possible by genetics alone[10].

HapMap revolutions?

The HapMap resource provides a rich data set for exploring demographic, selective and molecular forces that have shaped human sequence variation[2]. How will the resource be used by medical geneticists to find genetic determinants of disease, the main rationale for the project? In theory, researchers could use the principles laid out by de Bakker *et al.* and Zeggini *et al.*, along with considerations of cost, desired power and prior knowledge of the genetic architecture of the disease in question, to select study-specific sets of SNPs for genotyping in the relevant case and control samples. In practice, however, this approach will be out of reach for all but a small number of the best funded and technologically most sophisticated groups. Instead, most researchers will use off-the-shelf SNP panels provided by a few commercial suppliers. It is therefore essential that such panels be carefully assembled with the aim of maximizing statistical power for a given number of SNPs (the feasible number will increase as genotyping technologies continue to improve). Equally important, each panel must be accompanied by a precise, freely available description of the selection algorithms, the identities of the SNPs and their genotypes in the HapMap samples, the fraction of variants captured at different correlation levels and the expected power of the panel for a universal benchmark. One such benchmark, modified from that used by de Bakker *et al.*, is the power of the panel to detect, at a genome-wide $P <$ 0.05, an association due to a causal variant (chosen without regard to its inclusion in the panel) with an effect that would be detected with 95% power at nominal $P < 10^{-7}$ in 2,000 cases and 2,000 controls if tested directly (of course, the relevant parameters, including the frequency of the causal variant, should be varied to ensure robustness). Adherence to these standards will allow independent evaluation of the panels and facilitate development of software for association analysis appropriately tailored to each panel. Panel designers should also give priority to SNPs that change amino acids or otherwise alter gene function.

A new phase of human genetic research is upon us, one in which the contribution of common sequence variants to phenotypic variation will be tested systematically. This phase, relying on reference panels of common variants provided by the HapMap and similar resources[11], will in turn be supplanted by one of routine resequencing of entire genomes in large collections of phenotyped individuals[12,13]. Then we will finally be able to take an unbiased look at the influence of all types of DNA differences, common and rare, on heritable human traits.

1. Kruglyak, L. & Nickerson, D.A. *Nat. Genet.* **27**, 234–236 (2001).
2. The International HapMap Consortium. *Nature* **437**, 1299–1320 (2005).
3. Kruglyak, L. *Nat. Genet.* **22**, 139–144 (1999).
4. Kruglyak, L. *Proc. Natl. Acad. Sci. USA* **96**, 1170–1172 (1999).
5. Carlson, C.S. *et al. Nat. Genet.* **33**, 518–521 (2003).
6. de Bakker, P.I. *et al. Nat. Genet.* **37**, 1217–1223 (2005).
7. Zeggini, E. *et al. Nat. Genet.*, advance online publication 30 October 2005 (doi:10.1038/ng1670).
8. Klein, R.J. *et al. Science* **308**, 385–389 (2005).
9. Carlson, C.S. *et al. Am. J. Hum. Genet.* **74**, 106–120 (2004).
10. Rioux, J.D. *et al. Nat. Genet.* **29**, 223–228 (2001).
11. Hinds, D.A. *et al. Science* **307**, 1072–1079 (2005).
12. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
13. Shendure, J. *et al. Science* **309**, 1728–1732 (2005).

Gaining insight into PTPN22 and autoimmunity

Peter K Gregersen

The protein tyrosine phosphatase PTPN22 (also called LYP) is the leading example of a genetic variant that confers risk of developing diverse human autoimmune diseases, including type 1 diabetes, rheumatoid arthritis, autoimmune thyroid disease and systemic lupus. A new study now shows that the PTPN22 risk-associated variant, Trp620, results in a gain of PTPN22 phosphatase activity in T cells, opening up new avenues for exploring disease mechanisms.

Since the first report of an association between the PTPN22 Trp620 variant and type 1 diabetes[1], this intracellular tyrosine phosphatase has emerged as the strongest common genetic risk factor for human autoimmunity outside the major histocompatibility complex. Rheumatoid arthritis, Graves disease, Hashimoto thyroiditis, systemic lupus erythematosus and some forms of juvenile arthritis also show an association with this risk variant in populations of European ancestry[2–5]. These autoimmune phenotypes typically have a humoral component, with disease-specific autoantibiodies frequently appearing before the onset of clinical disease. It is not known whether PTPN22 Trp620 is a risk factor for autoantibody development *per se* or confers risk for progression to disease after the appearance of autoantibodies. Notably, autoimmune disorders in which autoantibodies are not a prominent feature, such as multiple sclerosis and Crohn disease, are not associated with the PTPN22 Trp620 variant[6,7].

Knockout of PEP, the mouse ortholog of PTPN22, leads to various immune abnormalities, including expansion of memory-effector T cell subsets and evidence of increased antibody production[8]. This overactivity of the immune system is accompanied by changes in the phosphorylation state of protein tyrosine kinases that regulate proximal events in T cell receptor (TCR) signaling, reflecting in part the

Peter K. Gregersen is at the Feinstein Institute for Medical Research, North Shore Long Island Jewish Health System, Manhasset, New York 11030, USA. email: peterg@nshs.edu