Meeting report

Mining the HapMap to dissect complex traits

Sung K Kim* and Justin Borevitz†

Addresses: *Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA. †Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA.

Correspondence: Sung K Kim. Email: sungkkim@usc.edu

Published: 28 March 2006

Genome ***Biology*** 2006, **7:**310 (doi:10.1186/gb-2006-7-3-310)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/3/310

A report on the Keystone Symposium 'Genome Sequence Variation and the Inherited Basis of Common Disease and Complex Traits', Big Sky, USA, 8-13 January 2006.

The Keystone Symposium 'Genome Sequence Variation and the Inherited Basis of Common Disease and Complex Traits' held in Montana in January showcased recent advances in the understanding of complex traits in both model systems and humans, thus highlighting the utility of model systems in dissecting the difficult etiology of complex trait phenotypes. Major topics included patterns of genome sequence variation and the underlying evolutionary mechanisms that have shaped this pattern, as well as computational approaches to linking genotype to phenotype so as to reveal the genetic architecture of complex traits.

Complex traits are the result of intricate interactions of genes and environmental factors, and are exceptionally difficult to analyze in humans. Model systems such as mouse and *Drosophila* have proved invaluable as archetypes by demonstrating that complex traits are influenced by a large number of individual quantitative trait loci (QTL), are affected by gender and environment, and are characterized by pervasive pleiotropy - that is, they usually affect more than one aspect of morphology or physiology. Furthermore, these model systems serve to facilitate excellent candidate gene selection for higher organisms. For example, Lawrence Marsh (University of California, Irvine, USA) discussed studies in *Drosophila* that could lead to potential treatments for Huntington's disease. Marsh demonstrates that inhibition of the post-translational modification process SUMOylation of Huntingtin can reduce the pathology of neurodegeneration on *Drosophila*.

The human haplotype-mapping project HapMap, which is aimed at detecting and characterizing millions of single-nucleotide polymorphisms (SNPs) in human genomes, promises to facilitate the linking of phenotypes to genotypes and to identify genetic determinants of human disease. At the same time, the human HapMap and accompanying whole-genome association-mapping methods can be applied and tested in model organisms. Association mapping can be viewed as the localization of a polymorphic marker that is associated with a genomic region and disease. David Altshuler (Massachusetts General Hospital, Boston, USA) suggested that large-scale systematic genome-wide studies using HapMap data may be required to identify all genetic variants that influence complex traits. Pennelope Bonnen (Rockefeller University, New York, USA) evaluated the potential of a genome-wide scan study of metabolic syndrome, a complex trait characterized by obesity, glucose intolerance, hypertension and blood lipid abnormalities, in an isolated population on the island of Kosrae in Melanesia. Bonnen reported that 99% of all Kosraen haplotypes were found in the HapMap samples. This finding validates the general use of the HapMap data for genome studies even for isolated populations. In addition, Augustine Kong (deCODE genetics, Reykjavik, Iceland) reported that in a study of type 2 diabetes in an Icelandic population, five HapMap SNPs were correlated with a candidate region associated with the disease. This study demonstrated the feasibility of mapping medium-risk common variants in a genome-wide scan using HapMap data.

Detecting deletions

The HapMap provides a great opportunity to study the types of genome variation in humans. As current genotyping technology does not target deletions, SNP base calls are falsely called homozygous when they may be hemizygous because of a deletion at that location on the other chromosome. Work addressing this issue was presented, dealing with the distribution of structural variation, primarily deletion polymorphisms or copy-number variants, within the HapMap data. Jonathan Pritchard (University of Chicago, USA) and Steve

McCarroll (Massachusetts General Hospital) have both used unfiltered HapMap calls to identify unusual runs of sequences and used them to indicate deletion polymorphisms. Pritchard's method focused on sequences apparently incompatible with Mendelian inheritance whereas McCarroll's method also applied deviations from Hardy-Weinberg equilibrium (the expected frequencies of homozygotes and heterozygotes for a locus in a population). Kelly Fraser (Perlegen Sciences, Mountain View, USA) presented data showing that SNPs are found in normal patterns of linkage disequilibrium with deletions. Some deletions were found by all of these three studies, but there was surprisingly little overlap. Combined methods, including a more detailed analysis of the raw genotype signals, should reveal additional natural deletion polymorphisms.

Array-based comparative genome hybridization (CGH), in which one genome is hybridized against another in the form of a microarray, can detect copy-number variation. Charles Lee (Brigham and Women's Hospital, Boston, USA) reported evidence of genomic imbalances in the cell lines of the HapMap found while studying copy-number variation using CGH. A detailed analysis using fluorescent *in situ* hybridization showed that 14 of 100 tested samples displayed abnormal karyotypes. This should serve as a warning that variation in phenotypes such as gene expression from these cell lines may not reflect natural variation found in the populations from which the lines came.

Determinants of variation

Recombination plays a vital role in the local patterns of haplotype block structure. Peter Donnelly (University of Oxford, UK) revealed the highly uneven distribution of recombination rates across the human chromosomes at a fine scale, which suggests that most crossovers occur in a small fraction of the genome. Donnelly also described evidence for a DNA motif that is enriched at recombination hotspots.

Evidence of natural selection may indicate the presence of functional variation relevant to complex phenotypes. Mutations that confer fitness benefits are expected to be driven to fixation under selection, in so-called selective sweeps. Common assumptions found among standard selective-sweep models include random mating, constant population size, and selection on new co-dominant mutations. Molly Pzreworski (University of Chicago, USA) described how relaxation of any of these assumptions on the model showed that selective sweeps are highly stochastic, often dependent on demography and the mode of selection, and that the ability to detect a sweep is much more difficult when selection has acted on standing variations rather than new mutations. Benjamin Voight (University of Chicago) presented a statistical measure for detecting positive selection within the HapMap data. His method measures the differences between ancestral and derived alleles by comparing the decay of haplotype homozygosity. In this connection, Jared Drake (Children's Hospital, Boston, USA) described evidence that conserved noncoding sequences are selectively constrained and are not mutation 'cold spots', by comparing allele frequency distributions to the remaining noncoding sequence within the HapMap data.

Linking genotype to phenotype

Association-mapping methods that use linkage disequilibrium have been shown to identify rare genetic variants with large effects on clinical phenotypes. Nevertheless, failures to replicate association studies have plagued many complex trait analyses. Using four cardiovascular disease cohorts from four different cities, Andrew Clark (Cornell University, Ithaca, USA) has tested the hypothesis that the samples are consistent with four random draws from the same population. Despite the population heterogeneity of the samples, he found consistent association of cardiovascular disease variants within the cities sampled.

It is often the case that fine genetic mapping requires the indirect association of a marker that does not contribute to the phenotype but is in linkage disequilibrium with the 'causal variant' for the phenotype (the allele or mutation that is responsible for the phenotype). By treating linkage-disequilibrium mapping as a missing-data problem, the state of an unknown causal polymorphism can be inferred from the underlying patterns of linkage disequilibrium of associated markers. In this regard, Heather Cordell (University of Cambridge, UK) presented a novel method that applies a multiple imputation algorithm, in which one assigns a posterior probability to the imputed data and updates the posterior probability at each iteration according to the current state of the data. Similarly, Jonathan Marchini (University of Oxford, UK) discussed another implementation of data imputation for disease mapping using a Bayesian logistic regression framework. Interestingly Marchini's method can take advantage of the full HapMap data when only 500,000 SNPs are typed. Methods that require data imputation may not be so effective if the genetic determinants of complex traits have allelic heterogeneity, where different alleles from the same gene cause similar phenotypes.

Traditionally, genetic variants that affect coding regions and alter protein structure, and presumably function, were thought to explain most of the variation in phenotypes. Recent evidence, however, suggests that variation in noncoding sequences is just as important. In this regard, the quantification of allele-specific transcripts is a powerful tool for determining the functional role of *cis*-acting genetic variation in the differential expression of genes. Julian Forton (University of Oxford) described allele-specific expression assays using the HapMap cell lines, while Rachel Brem (Fred Hutchinson Cancer Research Center, Seattle, USA) described

similar assays in *Saccharomyces cerevisiae*, both studies resulting in evidence for several *cis*-regulatory polymorphisms. By calculating the allele-specific expression ratio and testing for significant deviation from the expected value of 1, Forton studied the interleukin (IL)-13 locus within the 5q31 region in order to localize *cis*-acting polymorphisms. Brem found 44 of the 77 surveyed genes to have significant allele-specific expression distortion, suggesting *cis*-regulation. Brem explained the important difference between *cis*-regulatory variation that differentially affects the expression of different alleles at a gene, and what she calls "local regulatory variation", which simply maps to the affected gene. This local regulatory variation may be due either to *cis* variation or to variation in components acting in *trans* in an autoregulatory loop.

Modeling human complex traits

One limiting factor in human studies of complex traits is the lack of confirmatory genetic tests. Model organisms provide a solution by allowing direct hypothesis testing for a disease model. Mouse strains in which just one chromosome has been replaced with the chromosome from a different inbred strain (chromosome-substituted or consomic mice) are a powerful tool for testing the effects of interaction between specific loci. Joseph Nadeau (Case Western Reserve University, Cleveland, USA) described the use of consomic mice for mapping genes involved in obesity. Previous studies using recombinant inbred mouse lines identified few QTLs, each with marginal effects, each explaining on average 5-6% of the variation in this trait. With consomic mice, a large number of QTLs, each explaining on average 51% of the phenotype variation, were identified. Nadeau reasons that the genomic buffering, or the collective ability of the genome to compensate for mutations, is lost in consomic mice, as replacement of genes found in sensitive steps in a metabolic pathway can drastically alter the functioning of the pathway.

Association studies would benefit tremendously if the causal variant were part of the marker dataset used for mapping. Single-feature polymorphisms (SFP) can help to achieve this. These are polymorphisms that are defined by the use of high-density oligonucleotide tiling arrays for genotyping. For each feature (typically a DNA sequence of 25 bp) on an array, differences between a query sequence and a reference sequence are revealed through differences in hybridization intensities. Thus SFPs can be used for the joint discovery and typing of SNPs, microsatellites and larger deletions. SFPs provide the opportunity to resolve common and rare haplotypes, potentially including the causative polymorphism, which can then be used directly for whole-genome linkage-disequilibrium mapping. One of us (J.B.) outlined the use of SFP mapping to look at the genetic basis of adaptation variation in *Arabidopsis*, while Chao-Qiang Lai (Tufts University, Medford, USA) suggested QTL mapping using SFPs on arrays to find genes involved with *Drosophila* lifespan, by determining hybridization signal differences between young and old flies.

The meeting successfully bridged the gap between those who study complex traits in humans and those who use model organisms, and emphasized the joint progress in elucidating complex traits. As genotypic and phenotypic databases get larger, however, it is evident that statistical methods need to be developed that comprehensively incorporate the complex network of gene-gene, gene-environment, and *cis*- or *trans*-regulatory factors that influence a complex phenotype. We eagerly anticipate the maturation of such novel ideas and their application to current and emerging datasets for the next symposium.

Acknowledgements

We thank Christopher Toomajian and Keyan Zhao for helpful comments and suggestions.