The nature, pattern and function of human sequence variation

Evan E Eichler* and Kelly A Frazer[†]

Addresses: *Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [†]Perlegen Sciences, 2021 Stierlin Court, Mountain View, CA 94043, USA.

Correspondence: Evan Eichler. E-mail: eee@po.cwru.edu. Kelly Frazer. E-mail: Kelly_Frazer@perlegen.com

---

A report on the 2004 Keystone Symposium 'Human Genome Sequence Variation and the Inherited Basis of Common Disease', Breckenridge, USA, 8-13 January 2004.

The Keystone Symposia on Human Genome Sequence Variation and the Inherited Basis of Common Disease and Quantitative Genetics in Model Organisms were held concurrently at the Beaver Run Resort this year. The dual nature of the meeting created a unique atmosphere in which participants were encouraged to cross-attend sessions. Indeed, several sessions were held jointly between the two meetings, creating the opportunity for interaction among a larger body of scientists than typically attend a Keystone meeting. This unusual format encouraged cross-fertilization of ideas among population, quantitative and human geneticists as well as epidemiologists and genome scientists. This report is dedicated primarily to the proceedings within the Human Sequence Variation sessions.

Polymorphisms and haplotype mapping

Two presentations confronted the prevailing hypothesis that common diseases are likely to be due solely to common genetic variants, such as those that can be identified by mapping frequent polymorphisms. Aravinda Chakravarti (Johns Hopkins University, Baltimore, USA) challenged the community by raising, once again, the unsettling specter that rare variants might in fact underlie a significant fraction of common disease. He presented data from his long-standing work on the complex genetics of Hirschsprung Disease. The data strongly suggest that both common and rare single-nucleotide polymorphisms (SNPs) should be considered if the true nature of a complex genetic disease is to be understood. Similarly, David Altshuler (Massachusetts General Hospital, Boston, USA) described a genetic analysis of diabetes. He showed preliminary results suggesting that mitochondrial DNA is involved in type II diabetes and hypothesized that primary alterations in mitochondrial oxidative phosphorylation pathways contribute to type II diabetes. Both common and rare variants are equally important in this case.

Several presentations focused on the way that the structure of the human genome can be understood as being made up of haplotype blocks, and on how this may be useful as well as its biological origin and significance. Mark Daly (Massachusetts Institute of Technology, Cambridge, USA) provided insight into the progress and potential early fruits of the human HapMap project, which aims to produce a complete human haplotype map. A central question is how well the current (200,000 genotyped SNPs) and projected (500,000 SNPs by April 2004) haplotype maps recapitulate existing genetic variation within the human population. Extrapolating from existing data, Daly concluded that a density of one SNP marker every 5 kilobases should be sufficient to tag around 80% of human variation. Andy Clark (Cornell University, Ithaca, USA) stressed the need to impute missing genotypes on the basis of flanking SNP information. He presented a powerful Bayesian approach, which used the linkage disequilibrium between SNP pairs from a region of high marker density to infer the genotypes of missing data. He suggested that with a sufficient density of markers, information about the haplotype phase (the paternal or maternal origin) is unnecessary. David Goldstein (University College London, UK) similarly emphasized the need to identify hidden SNPs and further suggested that when the minor allele frequency drops below 7% there would be insufficient power to predict a hidden SNP within a second sample (based on a sample size of 64 individuals). His simulation studies predicted a loss in performance when the SNP density drops below one SNP per 6 kilobases.

One of the highlights of the meeting was the analysis by Peter Donnelly (University of Oxford, UK) of the landscape of fine-scale recombination. He presented a composite likelihood approach with which to estimate rates of recombination between pairs of SNPs. Using polymorphism data to infer properties of fine-scale recombination in a well-studied 10 megabase region of chromosome 20, he showed that the level of recombination varied as much as three orders of magnitude. His results suggest that 80% of all recombination occurs in around 25% of the sequence. While no clear sequence properties of 'hotspots' and 'coldspots' of recombination emerged, in general coldspots were found to be larger than hotspots, and hotspots tend to locate outside genes. There was a general agreement among several speakers that the only convincing correlation with the boundaries of some, but not all, haplotype blocks is increased recombination frequency.

One of the anticipated uses of SNPs and information about haplotype block structure is to improve the power of association studies for human genetic disease. K.F. outlined a strategy to tackle complex genetic traits using high-throughput methods of genotyping. Based on an analysis of a large number of individuals (around 1,000) for genetic variation and plasma concentration of low-density lipoprotein (LDL), the importance of well-characterized case-control samples and replica testing of pooled samples becomes very evident. Richard Lifton (Yale University, New Haven, USA) discussed the genetic determinants of hypertension and metabolic syndrome. He used families at extreme points of the phenotypic range - extreme hypotension and extreme hypertension - to identify 15 genes associated with this disease; 14 of these are involved in the renin-angiotension system, causing either increased or decreased Na^+ reabsorption. Lifton argued that this is the reason that drugs targeting salt absorption are superior in the treatment of hypertension. A similar success story was echoed by Stephen O'Brien (National Cancer Institute, Frederick, USA) in a detailed study of a cohort of more than 1,000 individuals with acquired immunodeficiency syndrome (AIDS) who have been clinically monitored for over 10 years. His research has identified and/or confirmed 15 genes (including the immune-cell surface molecules CD4, CD5, RANTES, HLA class I, and others) that affect infection with human immunodeficiency virus (HIV-1) and disease progression. Finally, David Hunter (Harvard Medical School, Boston, USA) discussed the role of gene-environment interactions in common disease. He described several examples where dietary and medical advice should be dependent on genotype information, including the finding that the *APOE4* allele, which has been linked to both Alzheimer's disease and hypertension, is more associated with cognitive decline in individuals with uncontrolled hypertension than in individuals with controlled hypertension.

Model organisms

In a joint session between the two concurrent meetings, a series of talks focused on how model organisms might be used to move from genotype to phenotype or function. Keith Davies (Paradigm Genetics, Research Triangle Park, North Carolina, USA) described an industrial-level high-throughput transgenic facility that focuses on the systematic collection of phenotype information from *Arabidopsis*. Genotype-phenotype correlation data for over 16,000 *Arabidopsis* genes (including previously unknown genes) is in progress. Using a transposon-tagging system to detect open reading frames (ORFs) in yeast, Michael Snyder (Yale University, New Haven, USA) described the systematic experimental verification of genes in the yeast genome. He emphasized the need to annotate genes, as well as transcription-factor binding sites, experimentally, and not to rely strictly on *in silico* analyses. In contrast, Eric Lander (Whitehead Institute, Cambridge, USA) showed the power of comparative whole-genome sequence analysis of yeast genomes to systematically identify genes, regulatory elements and processes of genome evolution. His analysis of four yeast genomes predicted 5,695 'real' genes. Similar analyses between multiple mammalian genomes are revealing many unexplained conserved elements and "the spectacular state of ignorance" in the area of functional genomics.

A particularly novel aspect of this meeting was the emphasis on inherited patterns of gene expression. Several studies have shown that natural genetic variation can cause significant differences in gene expression, suggesting that phenotypic variation can result not only from coding variation but also from regulatory variation that affects gene expression. To study the genetic architecture of natural variation in gene expression, Leonid Kruglyak (Fred Hutchinson Cancer Research Center, Seattle, USA) conducted a linkage analysis of genome-wide expression patterns in a cross between a laboratory and a wild strain of *Saccharomyces cerevisisae*. Over 1,500 genes were differentially expressed between the parental strains. These loci fell into two categories: *cis*-acting modulators of single genes (around 20%) and *trans*-acting modulators of many genes (around 80%). Surprisingly, analysis of the *trans*-acting loci by molecular function did not show an enrichment of transcription factors. Kevin White (Yale University School of Medicine, New Haven, USA) described the evolution of gene expression in *Drosophila*. He addressed the question, "If evolution was played several times under similar conditions would it repeat itself?" *Drosophila* from 27 inbred strains (with 20 males and 20 females of each strain) were divided into six populations. Three populations were raised in a hypoxic environment and three in a normoxic environment for several generations. When gene expression patterns were compared between strains in the two oxygen levels, 195 (53%) of the 368 genes that showed greater than a two-fold expression difference had evolved in all six populations, suggesting that some genes are more prone to change their expression levels, perhaps due to selective pressures.

Stephanie Monks (University of Washington/Rosetta Inpharmatics, Seattle, USA) discussed the genetics of gene

expression in mice. She established a genetic map of expression for 111 F2 mice resulting from a C57BL/J6 x DBA/2J cross. For each of the F2 mice, RNA was isolated from the liver and 23,574 genes were assayed for expression using arrays. The study demonstrated that the distribution of quantitative trait loci (QTLs) controlling gene expression is non-random in the genome. Justin Fay (Washington University, St Louis, USA) addressed the question of whether or not transcriptional variation has functional consequences. Nine isolates of *S. cerevisiae* were grown in rich media in the presence or absence of copper sulfate. Two strains with demonstrated resistance to copper sulfate showed a reduced growth rate, and two different strains produced rust colored colonies. Gene expression differences correlated with resistance were enriched for oxidative stress and the unfolded protein response, while those related to coloration were almost exclusively in the methionine/sulfur assimilation pathway.

In general, the pattern and nature of human genome sequence variation was the primary focus of the meeting, although due diligence was given to insights that could be gleaned from excellent studies from model organisms such as *Drosophila*, mouse, yeast and *Arabidopsis*. The scope of presentations was significantly more 'global' than in past meetings, due in part to the near-completion of the human genome, improvements in genotyping and the amount of analyzed sequence data now available. The wider range of topics appealed to a broader base of biologists. During the course of this meeting, a number of the 'usual' questions emerged. What is the contribution of rare versus common SNPs to the molecular basis for complex genetic disease? What density of SNP markers is sufficient for discriminating disease associations? Is there functional significance to haplotype block structures? How soon will geneticists routinely be able to resolve the genetic basis of complex disease? While there were no final answers to these and other questions, it was clear that significant advances are being made in these directions.