Meeting report

The biology of genomes: sequence gives way to function

David A Hinds

Address: Perlegen Sciences, 2021 Stierlin Court, Mountain View, CA 94043, USA. E-mail: dhinds@perlegen.com

Published: 30 August 2005

Genome ***Biology*** 2005, **6:**342 (doi:10.1186/gb-2005-6-9-342)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/9/342

A report on the meeting 'The Biology of Genomes', Cold Spring Harbor, USA, 11-15 May 2005.

This year's Cold Spring Harbor meeting 'The Biology of Genomes' presented an eye-opening snapshot of how far the field has come since the sequencing of the human genome was completed. A common theme of the meeting was the integration of new whole-genome variation and comparative genomic datasets with high-throughput experimental methods to address biological questions that might previously have been answerable only on the scale of a single locus, if at all. Speakers discussed the systematic identification of functional sequence elements, detection of relationships between these elements and biological endpoints, and reconstruction of their evolutionary histories. Many of these whole-genome analyses raise questions of how to deal with incomplete coverage and nonrandom sampling, how to evaluate significance when many statistical tests are performed, and how to properly account for both false-positive and false-negative results in very large and complex datasets. It seems clear that much work remains to be done to sort out these statistical issues.

New genomics datasets: deeper and broader

The Encyclopedia of DNA Elements (ENCODE) project has the goal of characterizing all human functional genetic elements, initially in 44 selected intervals representing about 1% of the genome. Eric Green (National Human Genome Research Institute (NHGRI), Bethesda, USA) presented an overview of the project. Many ENCODE annotations are now available through the University of California at Santa Cruz (UCSC) Genome Browser [http://genome.ucsc.edu/encode] and other public repositories. Green sketched out a proposed follow-on project that aims to sequence functional elements in the ENCODE regions in 400 healthy individuals as a pilot for the use of sequencing data in clinical settings.

The progress made by the GENCODE consortium [http://genome.imim.es/gencode], a subgroup of the ENCODE project, towards characterizing all protein-coding elements in the ENCODE regions, was discussed by Roderic Guigó (Institut Municipal d'Investigació Mèdica, Barcelona, Spain). This project has produced more detailed annotations of known human genes in the ENCODE regions, and computationally predicted genes are being experimentally verified. This has revealed a complex landscape of alternatively spliced transcripts where a one-to-one mapping from gene to protein is more the exception than the rule. While protein-coding genes seem to be fairly well represented in current annotations, ongoing work by GENCODE will focus on improving the coverage of classes of genes that may be under-represented and difficult to predict with current methods. These include for example, short or intronless genes, genes with unusual codon composition, and genes with no known orthologs in other species.

The international HapMap project, which aims to characterize relationships between polymorphisms in the human genome, has recently completed its Phase I map of more than 1 million single-nucleotide polymorphisms (SNPs) genotyped across a panel of 269 individuals. Tom Hudson (McGill University, Montreal, Canada) gave a status report on the project. Ten ENCODE regions were selected for deep resequencing and SNP discovery, and these were genotyped in the HapMap panel at a much higher density of one SNP per 250 base-pairs (bp; data available at the organization's website [http://www.hapmap.org] and the dbSNP database [http://www.ncbi.nlm.nih.gov/SNP]). In Phase II of the project, to be completed by October of this year, Perlegen Sciences will attempt to genotype an additional 4.7 million SNPs.

Recent enhancements to the UCSC Genome Browser [http://genome.ucsc.edu] were described by Jim Kent (University of California at Santa Cruz, USA), including a

new view that displays pairwise linkage-disequilibrium data for SNPs genotyped by the HapMap project. The browser now includes genome assemblies for more than 25 organisms, and many new tracks are derived from comparative genomic analyses. Multiple genome alignments and comparative methods have been used to improve gene predictions.

Genetic variation in gene expression

Several speakers presented the results of work aimed at identifying genetic factors that explain differences in gene expression. Norbert Hubner (Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Germany) described work with a panel of recombinant inbred (RI) strains derived from the Brown Norway and the spontaneously hypertensive rat, a model system for insulin resistance, and presented the results of linkage analyses with gene expression in several tissues. In many cases, his team identified loci, often acting in *cis*, that substantially affected gene-expression levels. In RI mice, Peter Little (University of New South Wales, Sydney, Australia) has found that most of the genetic variation in gene-expression levels appeared to be tissue specific, while Chris Cotsapas (also at the University of New South Wales) reported the identification of several loci in RI mice that coordinated the expression of large groups of other genes.

Several groups have taken advantage of the dense SNP genotyping data from the HapMap project to identify genetic variants that affect gene expression. Vivian Cheung (University of Pennsylvania, Philadelphia, USA) described whole-genome linkage analysis with expression levels in lymphoblastoid cell lines from parent-parent-child trios in the Centre d'Etude du Polymorphisme Humain (CEPH) database which have been genotyped by the HapMap project. She reported the finding of a set of genes with strong linkage evidence for *cis* regulation, and the use of data from the HapMap project to identify nearby SNPs associated with the expression phenotype. In some cases, these associations were strong enough to be detected even in a whole-genome scan using the complete HapMap SNP dataset. Matthew Forrest and Barbara Stranger (Wellcome Trust Sanger Institute, Hinxton, UK) both described similar experiments in which expression levels of genes in the ENCODE regions were determined in lymphoblastoid cell lines derived from CEPH individuals, using Illumina bead-based expression arrays. Forrest and Stranger then identified associations with SNP genotypes located both in *cis* and in *trans*, using HapMap project data.

Evolutionary constraint and natural selection

The relatively recent emergence of extensive high-quality sequence data from a variety of mammals and other vertebrates has promoted the development of comparative methods for reconstructing the evolutionary history of modern genomes. David Haussler (University of California at Santa Cruz, USA) and George Asimenos (Stanford University, Stanford, USA) each described work on the reconstruction of genomic sequences of a common mammalian ancestor, through multiple alignments of available mammalian genomes in various stages of completion. Michele Clamp (Broad Institute, Cambridge, USA) described progress towards completion of low-redundancy sequencing of an additional 16 mammalian genomes, which will substantially improve the annotation of evolutionarily conserved sequences.

Conserved noncoding sequences are of considerable interest because they are likely to indicate important regulatory regions subject to functional constraint. Emmanouil Dermitzakis (Wellcome Trust Sanger Institute, Hinxton, UK) reported the identification of 418 noncoding sequences conserved across other mammals but with evidence of accelerated divergence in humans. In SNP data from the HapMap project, these regions were over-represented in human-specific alleles occurring at high frequency compared with other conserved noncoding sequences. Manolis Kellis (Massachusetts Institute of Technology, Cambridge, USA) described a comparative genomic approach to the identification of conserved functional elements in promoters and 3′ untranslated regions. Aligned human, mouse, rat, and dog sequences were searched for conserved patterns, which were then clustered into a limited set of common motifs. The method successfully identifies many known transcriptional regulatory motifs, and many of the identified 3′ UTR motifs could be shown to be targets of known or novel predicted microRNAs (miRNAs).

Other groups have investigated the use of human SNP variation data to detect signatures of selection. Stephen Schaffner (Broad Institute, Cambridge, USA) and colleagues have undertaken an analysis of the HapMap project data to identify regions with signatures of positive selection, such as an excess of rare alleles and low heterozygosity, or extended high-frequency haplotypes. He reported that outliers on these measures included some known targets of selection (lactase, selected for lactate tolerance; and beta-globin, selected for malaria resistance). Most identified loci, however, were in regions of unknown function or at least no known selective role. Peter Donnelly (University of Oxford, UK) described the construction of a fine-scale map of recombination rates and recombination hotspots based on Perlegen genotype data. The picture that emerges is one in which large-scale structure is mostly determined by genomic context and evolves slowly, but fine-scale structure is not well predicted by sequence features and is poorly conserved. This recombination map is being integrated with evidence of recent shared ancestry of extended haplotypes in the HapMap data to identify adaptive evolution, as reported by Gil McVean (University of Oxford, UK).

Structural polymorphism in the human genome

Looking at more extensive chromosomal rearrangements in human genomes, Evan Eichler (University of Washington, Seattle, USA) described the use of fosmid paired-end sequence data to identify 297 structural polymorphisms (insertions, deletions, and inversions) relative to the human reference genome sequence. These variants were significantly over-represented in duplicated chromosomal segments, and Eichler suggested that some of these sites might have experienced recurrent rearrangement. A targeted search for sequence copy-number polymorphisms (CNPs) in segmental duplications using array comparative genomic hybridization (array-CGH) was reported by Andrew Sharp (University of Washington, Seattle, USA). A set of 123 CNPs has been identified, and these were particularly over-represented in regions identified as potential rearrangement hotspots on the basis of comparative genomics. Duc-Quang Nguyen (University of Oxford, UK) has examined the distribution of recently identified CNPs in the human genome, using data from the Genome Variation Database [http://projects.tcag.ca/variation]. He reported that CNPs tend to contain more simple repeats, but also tend to be relatively gene-rich. Certain Gene Ontology terms related to environmental responses (that is, extracellular proteins, olfactory receptors, and acquired and innate immunity) were significantly over-represented in genes in CNPs.

Kelly Frazer (Perlegen Sciences, Mountain View, USA) described the detection of 99 intermediate-length deletions (80 bp to 8 kb) in microarray data originally collected for SNP discovery. Linkage disequilibrium between these deletions and nearby SNPs was essentially indistinguishable from linkage disequilibrium around SNPs with comparable ascertainment, indicating that these sites do not represent recurrent hotspots of variation. Frazer estimated that a few thousand such polymorphisms may exist across the genome. Jonathan Sebat (Cold Spring Harbor Laboratory, Cold Spring Harbor, USA) described a study of CNPs in CEPH parent-parent-child trios from the HapMap project, using representational oligonucleotide microarray analysis (ROMA). In this method, a reduced-complexity subset of the genome is selected using a restriction digest and hybridized to microarrays of probes designed to bind the expected digestion products. CNPs were associated with concentrations of non-Mendelian inheritance, deviations from genotype frequencies consistent with Hardy-Weinberg equilibrium, and loss of heterozygosity in the HapMap data.

Genetic architecture of complex traits

The extended allelic transmission disequilibrium test (EATDT), a new approach to the analysis of whole-genome association data, was described by David Cutler (Johns Hopkins University, Baltimore, USA). The EATDT tests for association of a phenotype with haplotypes as well as with SNPs, using permutation tests to evaluate significance. While many more tests are performed compared to a SNP-wise analysis, the method is more powerful because the permutation procedure preserves the correlation structure of the SNP data. Cutler also showed that the use of haplotypes constructed from common SNPs yielded reasonable power to detect rare variants not explicitly selected for genotyping.

Results of large-scale association studies with complex phenotypes are just beginning to emerge. Dan Arking (Johns Hopkins University, Baltimore, USA) reported a whole-genome association study of the so-called Q-T interval of an electrocardiogram using the Affymetrix 100K SNP genotyping platform. In a case-control design using individuals with extreme high and low Q-T interval, the SNPs with strongest evidence for association were selected for genotyping in additional samples. One SNP in *CAPON*, a gene involved in nitric oxide signaling with a role in cardiac contractility, showed good evidence for association in a second set of samples. David Reich (Harvard Medical School, Cambridge, USA) described results of admixture mapping studies of prostate cancer and multiple sclerosis. Admixture mapping uses patterns of extended linkage disequilibrium in a recently mixed population to locate genetic determinants that are differentially distributed across the ancestral founding populations. The method promises a substantially reduced amount of genotyping to complete a whole-genome scan compared with traditional association studies. While the prostate cancer screen was unsuccessful, several promising candidate regions were identified in the multiple sclerosis screen.

In a keynote talk, Tom Gingeras (Affymetrix, Santa Clara, USA) summarized recent work that aims to characterize the human transcriptome more completely using high-resolution oligonucleotide tiling arrays. He presented data from multiple cell types for 10 chromosomes tiled with probes every 5 bp. The data reveal a complex universe of transcribed fragments that confirm and extend our understanding of coding transcripts, but with a discouraging number of transcripts of unknown function (TUFs), many of which are not polyadenylated. If anything, the data reveal the limitations of our current understand of the transcriptome. Aravinda Chakravarti (Johns Hopkins University, Baltimore, USA) discussed recent progress towards characterizing the genetic architectures of complex traits, using Q-T interval and Hirschsprung's disease as examples. Drawing connections between elucidating the genetic basis of complex traits and understanding the history of the human species as represented in our DNA, Chakravarti ended on a philosophical note, pointing out that we have a responsibility to demonstrate the relevance of genetics to everyday life. That should become an easier proposition as we get better at making connections between complex multifactorial phenotypes and their genetic underpinnings.