Meeting report

The many uses of a genome sequence

Anna Sharman

Address: Genome Biology, 34-42 Cleveland Street, London, W1P 6LB, UK. E-mail: anna@genomebiology.com

Published: 30 May 2001

Genome ***Biology*** 2001, **2(6)**:reports4013.1–4013.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/6/reports/4013

A report on the Keystone Symposium on 'Human Genetics and Genomics', Breckenridge, Colorado, USA, 31 March to 6 April, 2001.

This symposium covered a wide range of topics, from new techniques for analysis of the human genome sequence to insights into complex diseases. The talks could be divided into those concentrating on processes downstream of transcription, those comparing different genomes, talks looking at the genetic basis of disease and others discussing the changes in society connected with genomics.

Proteins and RNA

Stephen Burley (Rockefeller University, New York, USA) described the consortium of New York labs that aim to solve the structure of a sufficiently large sample of proteins that all other proteins can be modeled by homology to a solved structure; he estimates that this means solving about 30,000 structures. Burley gave two examples of the impact that the solution of one structure can have. The consortium chose to solve the structure of the protein mevalonate diphosphate decarboxylase, one of the enzymes in the cholesterol biosynthesis pathway. The protein's fold turned out to be novel and, what's more, it could be used to model 120 other enzymes, including three other enzymes of the sterol biosynthesis pathway. The function of numerous hypothetical proteins could also be predicted from this analysis. The second protein from the cholesterol biosynthesis pathway to be solved, isopentenyl diphosphate isomerase, proved to define a new superfamily and its structure led to the modeling of the structures of 93 proteins. Burley stressed that these are not atypical examples - every new structure may have such a large impact. He did point out that their strategy, which focuses on globular domains that can be crystallized easily, will miss proteins with coiled coils, proteins with multiple transmembrane domains, and 'singletons', which do not share a fold with any other protein.

Dagmar Ringe (Brandeis University, Waltham, USA) showed some examples that act as cautionary tales for the structural genomics efforts. As part of such an effort, Bill Studier's group (Brookhaven National Laboratory, New York, USA) have solved the structure of an uncharacterized protein, YBL036. It proved to have a common fold, a TIM barrel, and to have some features in common with bacterial alanine racemase. The structure gave very few clues about the cellular function of the protein in this case. In another case, Ringe's group solved the structures of two aminotransferases (L-aspartate aminotransferase and D-amino acid aminotransferase), which they had assumed would be similar. In fact, the folds were completely different; these two enzymes have apparently evolved the same active site independently. Function can therefore not always predict structure. Ringe also gave examples of so-called 'moonlighting' proteins, which have more than one function - and which also show that structure cannot always tell you much about function.

Ruedi Aebersold (University of Washington, Seattle, USA) presented a method for identifying and quantifying individual proteins in complex mixtures. The technique involves cleavage with trypsin, separation by chromatography, and analysis by mass spectroscopy (MS). The mass spectrum of each peptide identifies the protein that it comes from. If two different isotope labels are added to two protein samples to be compared, the MS analysis can use the mass ratios to determine the relative concentrations of each peptide in the two samples. This system can quantify protein levels much more accurately than microarrays can quantitate DNA or RNA. Aebersold described the application of this technique to two questions: changes in membrane protein levels after treatment of mammalian cells with a phorbol ester, and the spectrum of phosphorylated proteins in a yeast cell grown on glucose. One surprise from the latter analysis was that almost all the enzymes of the glycolytic pathway are phosphorylated; only hexokinase was previously thought to be regulated by phosphorylation. It is clear that this method will be widely applicable.

Thomas Gingeras (Affymetrix, Santa Clara, USA) described the Affymetrix transcriptome project. Starting with the DiGeorge syndrome region on chromosome 22q11.2 and hoping to eventually cover the whole genome, they are using microarray chips to measure exactly which nucleotides are transcribed into RNA. The 25-nucleotide probes on the chip interrogate each base for 350 kilobases (kb) of this region of chromosome 22. (Every 30 bases will be interrogated for the whole-genome project.) When sample RNA is hybridized to the chip, the pattern of hybridization shows exactly which regions of the genome are transcribed into RNA in that sample. Some of the probes may recognize more than one position in the genome, but by combining the results from adjacent probes, the transcription signal can be separated from this noise. Gingeras and colleagues are also separating the cell nucleus from the cytoplasm to investigate RNAs that are present in one compartment or the other. The RNAs that they have found to be enriched in the nucleus include transposons, pseudogenes, repetitive elements and other transcripts lacking long open reading frames. The high level of transcription of repetitive elements - many with polyadenylated tails - was a big surprise.

Genome evolution

Many people are now comparing the human genome sequence with sequences from related animals. Svante Pääbo and colleagues (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany) are aiming to 'shotgun'-sequence 0.1% of the chimpanzee genome; the initial results from this project indicate that the average difference between chimp and human DNA is 1.3%. Edward Rubin (Lawrence Berkeley National Laboratory, Berkeley, USA) is starting a 'phylogenetic footprinting' study, sequencing one genomic region from a number of different primates. Gene Myers (Celera Genomics, Rockville, USA) said that the company has now completed its assembly of the mouse genome sequence and are using it to find regions conserved between mouse and human, which could be exons or regulatory regions. They are also using synteny between human and mouse to improve the assembly of both the human and the mouse sequences and to map the rearrangements that have occurred in the time since humans and mice diverged.

Rubin showed how powerful comparative sequence analysis can be. His group compared the sequence of a 1 megabase (Mb) region of human chromosome 5q31 with the syntenic region on mouse chromosome 11, looking for non-coding regions with over 70% sequence identity over more than 100 base pairs (bp); this strategy was chosen so that master regulatory elements (such as the globin locus control region) would be detected. They found 81 such regions, including some enhancers that were already known. Two thirds were also conserved in the dog genome, suggesting that they have important functions, although these might not be only in gene regulation but could also be in chromatin structure, chromosome pairing, or replication. Rubin and colleagues surveyed the promoters of nine genes in the 5q31 region for binding sites of the interleukin-regulating transcription factor GATA-3, which has been shown to play an active role in the transcription of most of the interleukin genes in the region; GATA-3 thus served as a positive control for the approach. He found 98 of these sites, spread across the promoters of all the genes. When these were compared with the mouse sequence, however, only the binding sites in the promoters of GATA-3-responsive genes were conserved.

Aravinda Chakravarti (Johns Hopkins University School of Medicine, Baltimore, USA) pointed out that before a mutation is fixed in the population, it must go through a stage of being a polymorphism. When there is selection keeping the sequence the same, polymorphisms will appear but will not be fixed. He has therefore looked for regions in the genome where there are many single-nucleotide polymorphisms (SNPs) in both chimp and human sequences but where the variations have not been fixed. Other regions that are also conserved but have low levels of polymorphism are probably the same simply because mutations have not arisen in them, not because they have an important function. This can reveal important regulatory regions in species that otherwise seem to be too closely related to give informative results.

To study variation among humans, Pääbo has sequenced a 10 kb region of the X chromosome (carefully selected because it is little affected by recombination or selection) in 70 individuals from all the major language groups of the world. He has found that, in a phylogenetic tree from these sequences, all the nine major branches included African sequences, whereas only three branches had sequences from Asia and Europe, supporting the hypothesis that the first humans spread over Africa and only a small group of them went on to colonize the rest of the world. This tree also shows that there was a period of rapidly increasing population size in human evolution, which is not seen in other primates.

The phylogenetic tree of primates has been controversial, but sequence data has led to a widely accepted tree in which chimps (and bonobos) are closest to humans, with gorillas next and orang-utans the most distantly related of the great apes. Blair Hedges (Penn State University, University Park, USA) has sequenced nine genes from orang-utan and gorilla and has used them, together with sequences available for human and chimps, to estimate the dates of divergence of primate species. Using very careful methods for calibrating such estimates with the fossil record, his group have put the human-chimp split at 5.7 million years ago, at a time when the Earth was rapidly drying up after a warm and humid period.

Evan Eichler (Case Western Reserve University, Cleveland, USA) has scanned the recently published human genome

sequence for duplicated regions - not the well-known repetitive elements, but sequences over 1 kb in size that share over 95% sequence identity. He has found that nearly 3.5% of the genome is duplicated. Eichler looked more carefully at one duplicated cassette that is present in 15 copies, all on chromosome 16. The one gene encoded by this cassette has a very unusual pattern of evolution: the exons mutate much faster than the introns, and silent substitutions are less common than substitutions that change the amino acid; both of these are clear signs of positive selection. It is clearly a functional gene, as it is expressed and splicing is conserved, but its function is unknown. Could the fact that it has been duplicated so many times be connected with this positive selection? More study of this fascinating example is needed.

The pattern of duplication found by Eichler and colleagues also has practical consequences for the 'finishing' of the draft human genome sequence. Duplications can mess up the assembly of shotgun-sequenced fragments, as two slightly different fragments can be assumed to be derived from the same, slightly polymorphic, region when they are in fact from different regions. In fact, some of the SNPs found by the genome project may in fact reflect differences between different versions of a duplication.

The genetic basis of complex diseases

Various speakers described methods for determining the gene(s) involved in complex human diseases. One popular method is whole-genome association using SNPs, which needs a large population. When families can be found in which more than one member has the disease, linkage analysis is simpler and fewer markers are needed.

In experiments that will help to speed up whole-genome association studies, Eric Lander's group (The Whitehead Institute, Cambridge, USA; presented by Francis Collins, National Human Genome Research Institute, NIH, Bethesda, USA) have found that linkage disequilibrium (LD) extends as far as 60 kb away from the average SNP, although this figure varies considerably between loci and also between human populations; for example, people of Northern European origin from Utah have much more LD than Nigerians. Lander has proposed that there may have been a bottleneck in the evolution of Northern Europeans about 20,000-60,000 years ago, such that effectively fewer than five individuals may have given rise to most of the modern gene pool. This hypothesis needs to be tested by looking at LD in other populations.

Huda Zoghbi (Baylor College of Medicine, Houston, USA) described her group's work on how long poly-glutamine repeats can cause neurodegenerative diseases such as Huntington's disease and spinocerebellar ataxias (SCAs). Nuclear inclusions of aggregated protein containing ubiquitin are seen both in patients and in some of the transgenic mouse strains that provide disease models. Using constructs expressing a version of ataxin-1 (the protein altered in one type of SCA) with a long poly-glutamine repeat fused to green fluorescent protein (GFP), Zoghbi's group have shown that these inclusions are reduced by inhibition of ubiquitin ligase in mice or by overexpression of two chaperone proteins in cells, and they are increased when proteasome function is inhibited in cells. These results led Zoghbi to propose a model in which proteins with long poly-glutamine tracts, being harmful to the cell for an unknown reason, are targeted by both chaperones (which increase solubility) and the ubiquitin-proteasome pathway (which start to degrade them). The amounts of protein overwhelm these pathways, so a large proportion of the aberrant protein aggregates in a 'holding' complex that contains components of the ubiquitin-proteasome pathway. This model implies that the nuclear inclusions seen in diseases are in fact a manifestation of the cell's protective response, rather than being (as others have suggested) the main cause of neuronal degeneration. Zoghbi's hypothesis suggests that enhancing chaperone action and targeting the ubiqitin pathway could provide promising therapies for poly-glutamine repeat diseases. Zoghbi's theory is supported by the work of Juan Botas (Baylor College of Medicine, Houston, USA), who has made a *Drosophila* model of SCA that shows the same features, including neuronal degeneration, nuclear inclusions, and behavioral abnormalities.

Adrian Hill (Oxford University, UK) described the search for genes that modify the susceptibility to three of the most prevalent infectious diseases: tuberculosis, leprosy and hepatitis B. A genome scan for leprosy-susceptibility genes, looking at about 400 microsatellite loci in 245 affected pairs of siblings with leprosy, came up with one significant locus on chromosome 10p13. The mannose receptor gene maps in this interval and is a candidate gene, though Hill and colleagues are still investigating whether it is really the important factor.

Of the other talks on diseases, two particularly interesting ones included the discovery of candidate genes that may lead to new insights into the mechanisms of disease. Maja Bucan (University of Pennsylvania, Philadelphia, USA) has used mouse mutagenesis to find genes involved in rhythms of sleep, rest and activity, which are frequently disrupted in bipolar (manic-depressive) disorder. One gene found was *Rab3a*, which is known to be important in synaptic transmission. Graeme Bell (University of Chicago, USA) has used linkage analysis to show that polymorphic variation in the gene encoding a calpain (calcium-regulated protease) affects susceptibility to type II diabetes in Mexican-Americans and Northern Europeans. It is not clear how calpains could be involved in diabetes, but Bell and colleagues are currently making animal models to investigate this.

One feature that became clear through all these talks is that complex diseases are the result of interaction between many

genes together with the environment and epigenetic effects. Klaus Lindpaintner (Hoffmann-La Roche, Basel, Switzerland) quoted Sir William Osler, who wrote in the classic medical textbook *The Principles and Practice of Medicine* in 1892, "If it were not for the great variability between individuals, Medicine might be a Science not an Art".

Genomic disease

James Lupski (Baylor College of Medicine, Houston, USA) discussed 'genomic disorders' - genome rearrangements that cause human diseases. He described Charcot-Marie-Tooth disease, which is due to a duplication, and hereditary neuropathy with liability to pressure palsies (HNPP), which results from the deletion of the same region as is duplicated in Charcot-Marie-Tooth disease, as well as other pairs of diseases with reciprocal deletions and duplications. A search for duplications reciprocal to known disease-causing deletions has come up with several new diseases with more subtle symptoms; one example is chromosome 17p11.2. When this region is deleted, the result is Smith-Magenis syndrome, which includes mental retardation and multiple congenital anomalies, whereas patients with this region duplicated have behavioral problems but no major organ system defects. Lupski pointed out that genomic disorders occur anew at 100 times the frequency of point mutations and occur at the same frequency in all populations; and they also tend to occur in the same places multiple times because they are initiated by repeated sequences. They are therefore of great significance for human health.

Andrew Feinberg (Johns Hopkins Medical School, Baltimore, USA), who entitled his talk "The epigenetics of genetics", used Beckwith-Wiedemann syndrome (BWS) as an example of a genetic disease caused by aberrant imprinting. He has found that many cancers have loss of imprinting and that, in the case of colon cancer, this is also found in the normal colon cells of the same patient, not just the tumor. It may thus be possible to use loss of imprinting to identify patients with cancer risk, and he is currently collaborating on a large clinical trial to test this hypothesis.

Genetics, genomics and society

Harold Varmus (Memorial Sloan Kettering Cancer Center, New York, USA) surveyed the ways in which the Human Genome Project has changed the culture of science. For example, it has given an increased opportunity to present biology to the public as something exciting; it has encouraged more long-term planning of science, and more non-hypothesis-driven research in large teams; it has encouraged more cooperation between scientists, both in the public and private sectors, with much more open sharing of new and published data and research tools; and it has created an explosion in demand for expertise in computing, for which the supply is insufficient.

Lindpaintner discussed the societal issues involved in genetic testing. He pointed out that every time we look at someone we are doing a genetic test for the presence or absence of the Y chromosome. This information can be used to discriminate, but the solution to this is not to prevent people from knowing your genotype but to fight discrimination in other ways. He also argued that medical and genetic information are not fundamentally different; many people are concerned about confidentiality of genetic test results, but can be happy to let others know about other factors that are far more likely to predispose them to disease, such as their age, their cholesterol level or the result of an X-ray. He proposed that, as part of an ongoing dialog with the public, education should be increased to help the acceptance of genetic testing, and described a CD-ROM available from the Roche Genetics Education Program [http://www.rochegenetics.com/CD_ROM/cd_rom.html] that is part of this attempt.

It became clear at this wide-ranging meeting that genomics and genetics are moving closer together rapidly, and that it won't be long before most human geneticists are using genomic methods and data in their research on individual genes. Meetings such as this Keystone Symposium should help promote this cooperation.