Meeting report

The meso-genomic era

Colin AM Semple, Martin S Taylor and Stephane Ballereau

Address: Department of Medical Sciences, Molecular Medicine Centre, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK.

Correspondence: Colin AM Semple. E-mail: Colin.Semple@ed.ac.uk

Published: 28 June 2001

Genome ***Biology*** 2001, **2(7)**:reports4015.1–4015.5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/7/reports/4015

A report from HGM2001, the sixth annual International Human Genome Meeting organized by The Human Genome Organisation (HUGO), Edinburgh, UK, 19-22 April 2001.

In his opening remarks, the president of HUGO, Lap Chee Tsui, included a plea to avoid the term 'post-genomic era'. One does not have to look hard to see why. According to the Genome Monitoring Table [http://www.ebi.ac.uk/genomes/mot/] hosted by the European Bioinformatics Institute (EBI) about 45% of the human genome is available in finished form, and the rest is available as fragmented 'draft' sequence. Also, accurate and comprehensive descriptions of the functions and often even the structures of the 30,000-40,000 predicted genes are, at the moment, rare. Added to this, the catalog of variations seen in the human genome is growing but far from complete. At the same time, the Entrez Genome pages [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome] at the National Center for Biotechnology Information (NCBI) say that there are currently sequences for 9 finished or draft eukaryotic, 40 bacterial and 10 archaean genomes, but a veritable avalanche of new genomes is already on its way. This new wave of data will extend genomics to new protozoan, fungal, plant and animal taxa, including eagerly awaited vertebrates such as zebrafish, pufferfish, the chimp and other apes. From several perspectives we are decidedly in the 'mid-genomic' era.

At this conference a good sampling of activities in the present 'meso-genomic' era was presented and several themes recurred. We carried out a superficial survey of the abstracts available at the HGM2001 website [http://hgm2001.hgu.mrc.ac.uk/] and found certain terms to be better represented than others (Figure 1). Many abstracts (134 out of the 543 online) dealt with studies of human variation and particularly of single-nucleotide polymorphisms (SNPs). Far fewer abstracts contained references to human diseases or disorders, reflecting the fact that much work on variation was concerned simply with the detection of SNPs and issues around data analysis. In such 'data-rich times' computational analysis is very much in evidence. Much of what we know about the function of human genes is inferred computationally and so, to rectify this, studies are underway to generate functional data using model organisms. Comparative genomics is rapidly coming of age and many studies, particularly those using the draft mouse genome, demonstrated its utility in sequence annotation. Fewer studies at this meeting concentrated on what precisely is transcribed and eventually translated from the genome, although important initiatives were described. It is also becoming evident that epigenetic mechanisms can be important in the journey from genotype to phenotype.

The finer details: genomic variation and complex traits

It is often quoted that all humans are around 99.9% genetically identical, but that still leaves millions of sites in the genome where we differ. The largest public repository for SNPs, dbSNP [http://www.ncbi.nlm.nih.gov/SNP/] at the NCBI, now contains more than 1.68 million of these variations. At the HUGO conference it became clear that the search is on for the variations that are important in identifying predisposition to disease and tracing our evolutionary history. A key question, which received a lot of attention, is the extent of linkage disequilibrium (LD) between SNPs. LD is a measure of the association between alleles; for example if no recombination has occurred between alleles they are in complete LD. The greater the distance such associations span, the more chance we have of finding SNPs that indicate predisposition to disease. A great deal depends on the ancient history of human populations. Some previous theoretical work, assuming constant population expansion, has suggested that LD should extend over only a few kilobases. It now appears, however, that the history of Northern European populations involved bottleneck events since migration

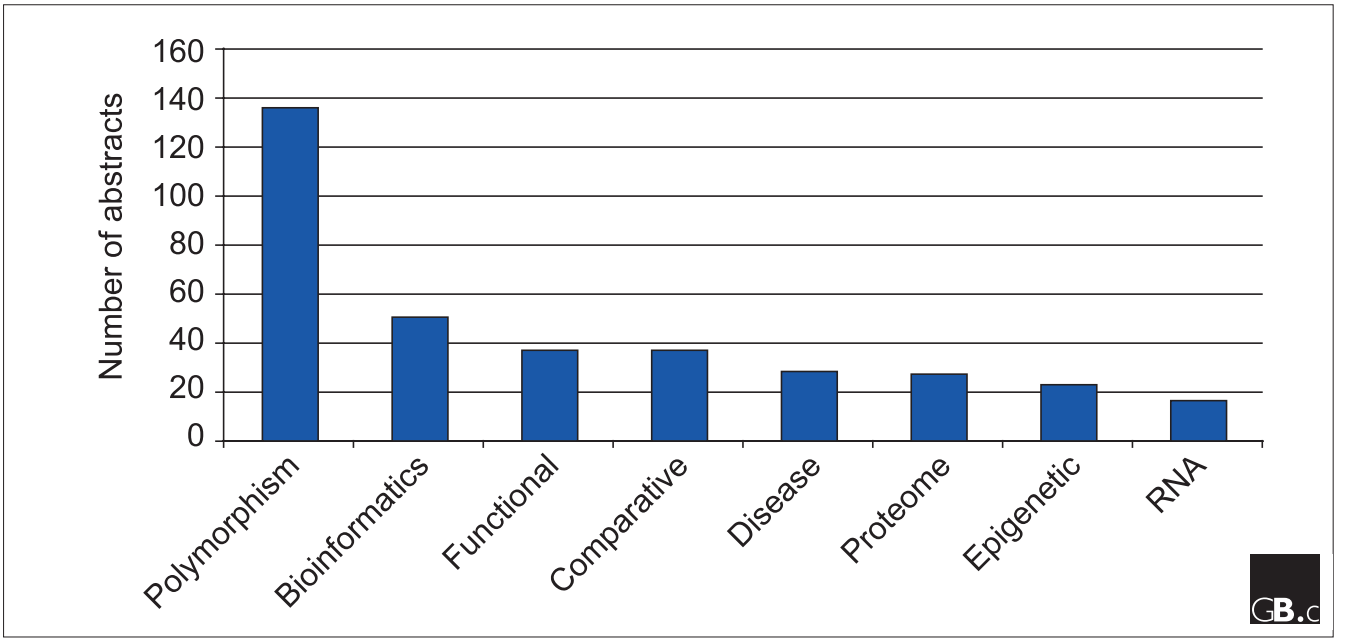

Figure 1
A survey of common terms in abstracts submitted to HGM2001. The categories on the horizontal axis were derived from searches for the presence of the following terms: polymorphism* or SNP*; bioinformatic* or comput*; function* and genom*; compar* and genom*; disease* or disorder*; proteom* or protein* and structur*; epigenet* or methylat* or heterochrom*; RNA and splic* or process*. Individual abstracts can appear in more than one category, so little can be concluded from the absolute heights of columns.

from Africa. Eric Lander (Whitehead Institute, Cambridge, USA) revealed that this has resulted in LD extending over some regions of around 120 kb (although great variation in LD was seen across the genome) and that a map defining perhaps 30,000 ancestral regions of LD can be constructed for such populations. In a Nigerian population, by contrast, LD extended over less than 10 kb and is more consistent with a simple model of population expansion. The results suggest that a representative set of only 100,000 SNPs could be effective for whole-genome association studies of disease in Northern European populations. Michael Olivier (Stanford Human Genome Center, USA) found LD extending over regions of chromosome 21 that were between 2 and 85 kb, but these regions were disrupted by segments of similar or larger size that showed no significant LD between SNPs. In addition, each segment was characterized by a restricted number of observed haplotypes, with the commonest haplotype found in over 50% of all chromosomes. These results illustrate the need to characterize the pattern of LD in a region of interest before association studies are undertaken. They also imply that initial localization of disease loci may be easier in populations with LD that extends over more than 100 kb (such as Northern European populations), but that finer mapping of such loci, to specific areas of genes, may be easier in populations with smaller blocks of LD.

Joseph Terwilliger (Columbia University, New York, USA) discussed the feasibility of identifying common susceptibility alleles that exert small effects. Much depends upon the relationships between three factors: the marker genotype, the susceptibility allele and the phenotype of interest. In association studies, we measure the correlation between the marker genotype and the phenotype. The success of this enterprise depends on two relationships: between the marker genotype and the susceptibility allele (measured by linkage and/or LD), and between the susceptibility allele and the phenotype. The latter relationship depends in turn on the capacity to infer the genetic factor from the phenotype (detectance). With complex traits such as predisposition to common diseases, however, there may be many different causal genotypes, making detectance low. In such cases, linkage and LD mapping may not give significant results. All this still leaves aside the question of gene-environment interactions, which can further complicate the observed patterns of disease.

A model association study was presented by Xavier Estevill (Duran i Rynals Hospital, Barcelona, Spain), who investigated susceptibility to panic and phobic disorders. His group identified an interstitial duplication on chromosome 15q24-26 that is significantly associated with panic disorder within families from a small Catalonian village. The duplication is estimated to be present in around 7% of the general population. The association was then replicated in a more diverse population of patients from elsewhere in Spain and followed up with studies of the expression and function of

genes within the duplication. One of these genes was overexpressed in transgenic mice and was found to cause an enhanced panic reaction in behavioral tests. In addition, the duplication was found in different forms within families, showed an absence of linkage with other 15q markers and exhibited mosaicism in patients' cells. It therefore appears to represent a novel, non-Mendelian disease mechanism. Allen Roses (GlaxoSmithKline, Triangle Park, USA) discussed successes in association studies based on SNPs. High-density disequilibrium mapping of SNPs has identified new susceptibility genes for psoriasis, migraine, diabetes mellitus (type 2) and Parkinson's disease within previously defined large linkage regions. He predicted that within the next few years clinical tests for drug efficacy or hypersensitivity will emerge using collections of SNPs that are associated with a given phenotype. The successes in identifying associations and the emerging LD patterns in the human genome have renewed optimism among those working on common diseases that show complex inheritance patterns.

The conference also included talks covering complex traits in species other than human. Complex genetic traits have been studied and manipulated in agriculture for thousands of years. The conventional breeding schemes of 'breeding the best to the best' have been incredibly successful, to the extent that most people would not think of eating - or indeed even recognizing - wild tomatoes. Dani Zamir (Hebrew University of Jerusalem, Israel) presented a model system for investigating quantitative trait loci (QTL) in tomato. Zamir and co-workers produced nearly isogenic lines (NILs) of tomato but included exotic chromosome segments from wild breeds of tomato, essentially using a back-cross strategy. Armed with NILs covering the entire genome, they found and mapped 23 QTLs for the brix (sugar) content of the fruit. Extension of the original strategy focused on one of the strongest QTLs, narrowing it to 484 bp within a gene for an apoplastic invertase, a key enzyme in sugar metabolism. Further work demonstrated that lower levels of mRNA from this gene are found in commercial cultivars than in the wild tomato. Not only is the tomato a good model for studying QTLs, it is also of substantial commercial significance: higher brix levels mean better ketchup.

Michel Goeorges (University of Liege, Belgium) described the fascinating inheritance pattern of "beautifully proportioned buttocks" (callipyge), in sheep. A ram with the desired trait was initially identified; through breeding of this ram and by using a least exclusion breakpoint mapping approach (that is, using recombination breakpoints to identify a common haplotype between individuals), a single, apparently autosomal dominant locus was pinpointed on chromosome 18 with a logarithm of odds (LOD) of 55. Classical genetic crosses were carried out and these revealed imprinting at the locus, with only those offspring that inherited a paternal callipyge allele expressing the phenotype. Back-crossing to produce sheep homozygous at the callipyge locus revealed that only heterozygous animals with a paternally derived allele expressed the phenotype, a phenomenon described as polar overdominance. Within the callipyge-determining locus four genes have been found, all of which are imprinted. It will be interesting to see the molecular explanation for this bizarre pattern of inheritance.

The same but different: comparative genomics

The human genome is unfinished and we still lack another complete vertebrate genome for comparison, but comparative genomics is already becoming popular. Jean Weissenbach (National Centre for Scientific Research (CNRS), Paris, France) described successes in comparisons between the human and more compact vertebrate genomes using the draft sequences from the two pufferfish species *Fugu rubripes* and *Tetraodon nigroviridis*. His group found such comparisons to be a valuable addition to the output of *ab initio* gene-prediction programs such as GenScan, for detecting novel exons and conserved, non-coding regions assumed to be important in regulation. By comparison, they had found human versus mouse comparisons to be more sensitive (detecting a larger number of exons) but also 'noisier' (generating many uninformative matches). Several studies were presented that exploit the draft mouse sequence to aid the annotation of the human genome. Lisa Stubbs (Lawrence Livermore National Laboratory, Livermore, USA) described a large-scale comparison between human chromosome 19 and 15 syntenic regions of the mouse genome adding up to 46 Mb of mouse sequence within 35 bacterial artificial chromosome (BAC) contigs. Combining these comparisons with other annotation methods (*ab initio* predictions and identifying matches to expressed sequences) gave a total of around 1,200 genes for human chromosome 19, with perhaps 50 genes identified by comparative genomics that had been missed by other methods. The mouse comparisons also identified many new candidate exons, 5′ untranslated regions (UTRs) and more than 4,000 non-coding conserved regions. Around 30% of the genes on human chromosome 19 are members of tandemly duplicated clusters, including vomeronasal receptors, olfactory receptors and zinc-finger genes. The regions containing clusters were found to have diverged considerably between human and mouse, varying in gene content, number and organization, in contrast to other regions.

Peter Holland (University of Reading, UK) has investigated the evolution of vertebrate genomes by studying the numbers of homeotic-gene clusters in vertebrates and amphioxus (the living invertebrate that is closest to the vertebrates). He postulates two episodes of tetraploidy, followed by selective gene loss in early chordate evolution near the time point at which the vertebrates emerged. Gazing further back in evolutionary time, Eugene Koonin (NCBI, Bethesda, USA) described using the combination of analyses of protein structure and comparative genomics to discover ancient protein domains originating in the last universal

common ancestor (LUCA) of all extant life forms. The main biological functions of the LUCA appear to have revolved around RNA metabolism and translation, whereas the DNA processing machinery is a later innovation. Koonin also offered insights into the way new domains evolved, by duplication and accretion, during eukaryotic evolution. Many eukaryote-specific domains are the result of explosive bursts of innovation; for example, exploiting the basic α helix in a variety of novel configurations such as the coiled coil.

It is clear that comparative genomics has thrown up a variety of new challenges for bioinformatics. Comparisons between multiple, long sequences that may only share small regions of limited homology is, in itself, a non-trivial problem. Burkhard Morgenstern (Munich Information Center for Protein Sequences (MIPS), Germany) presented a new version of the DIALIGN program that performs well in the identification of small regulatory regions in large non-coding DNA sequences. But although many conference participants emphasized the success of combinations of methods in gene prediction, we still lack software that combines intrinsic sequence properties (as detected by *ab initio* methods) with genomic-sequence comparisons and analyses of similarity to proteins or expressed sequence tags (ESTs). Tim Hubbard (Sanger Centre, Hinxton, UK) described how the Ensembl [http://www.ensembl.org] database, which provides annotation of the draft human genome, has incorporated comparisons to the draft mouse sequence. Hubbard and colleagues are now grappling with the prospect of incorporating comparisons between human and several other vertebrate genomes.

In spite of substantial conservation between species, there are clearly very significant differences in the genes and/or their regulation between species. One such difference between species is being investigated by Svante Pääbo (Max-Planck-Institute for Evolutionary Anthropology, Leipzig, Germany) using comparative genomics. His group focuses on the differences between humans and other, closely related species. Previously, the only well-characterized biochemical difference between humans and great apes was the level of hydroxylation of sialic acids (major components of cell surfaces in animals), which is of major importance for the success of xenotransplantation. Comparing the relative expression levels of 20,000 genes in blood, liver and brain between humans, chimpanzees and rhesus macaques has revealed that the rate of 'transcriptome' change has been similar between chimps and humans (using macaques as a standard) in blood and liver. In stark contrast, the rate of change of expression patterns in the brain is accelerated approximately three-fold in humans. A number of the most strikingly different expression patterns are now being followed up.

The wild frontier: from genotype to phenotype

With an established foothold in genomics and large-scale, public collections of microarray data on gene expression just around the corner, people have begun to peer over the horizon at the proteome. Most human genes have not been characterized functionally, so what we do know about them mostly comes from computational detection of homology. As Janet Thornton (University College London, UK) made clear, protein family members commonly exhibit some similarity in function, but there is remarkable functional variation within many families. By examining enzymes within 167 structural superfamilies from the CATH database [http://www.biochem.ucl.ac.uk/bsm/cath/CATH.html], which classifies groups of proteins by class, architecture, topology or homology, she estimated that only around 90 superfamilies contained members with conserved functions. In variable superfamilies, members often differed in substrate specificity. Thus, although structural homology can extend computational predictions beyond homology at the sequence level, there will always be many proteins for which no reliable predictions can be made. Thornton also discussed an analysis of proteins associated with disease garnered from the Online Mendelian Inheritance in Man (OMIM) [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM] database. By superimposing disease-associated mutations onto these protein structures she concluded that there was no general pattern in the positions of disease-causing mutations. Some occurred within regions, such as active sites, that would be predicted to alter function, but many others appear to relate to interactions with other proteins. It is widely accepted that a wider sampling of the protein-structural universe is needed to enhance our understanding of protein function and dysfunction. Tim Harris (Structural Genomix, San Diego, USA) discussed large-scale approaches to X-ray crystallography which might transform information obtained from the human sequence into novel three-dimensional protein structures. Initially, the focus is on structural families with pharmaceutical significance.

Traditionally, the development of new drugs involves testing novel compounds in disease models with little knowledge of the biochemical pathways or physiological systems underlying the disease. Because disruptions in the abundances and activities of proteins are the molecular basis of many diseases, proteomics promises to revolutionize drug discovery by allowing rational identification of drug targets. But the set of proteins present within a cell at any one time is the net outcome of many complex processes, including synthesis, degradation, modification and processing. Each protein can then participate in myriad interactions with other molecules in different places within the cell. In addition, protein abundances, cellular locations, activities and interactions may differ across tissues and between normal and disease states. A number of approaches to investigate this awe-inspiring complexity were presented. Ian Tomlinson (MRC Laboratory for Molecular Biology, Cambridge, UK) described recombinant-antibody technologies for use in large-scale studies of protein localization, quantification and interactions. The aim is to construct high-density antibody arrays

that are able to profile the expression of most human proteins in a given tissue within the next 3-10 years. Mark Vidal (Dana-Faber Cancer Institute, Boston, USA) outlined his approach to building a protein-interaction map for the *Caenorhabditis elegans* proteome. The idea is to amplify and clone every open reading frame in the genome for use in extensive yeast two-hybrid system assays. The resulting interaction map is to form part of an envisaged functional atlas for *C. elegans*. As Patricia Kuwabara (Sanger Centre, Hinxton, UK) pointed out, because around three quarters of human disease genes have *C. elegans* homologs, the worm could be a valuable source of functional annotation for the human genome. She outlined efforts to carry out comprehensive DNA microarray expression profiling as well as RNA-mediated interference (RNAi) the latter of which provides a rapid, sequence-specific method to abolish gene activity.

The mouse continues to be a favorite model organism for investigating gene function, and in this field investigations are also increasing in the form of screens and multiple studies carried out in parallel. Keats Nelms (John Curtin School of Medical Research, Canberra, Australia) and Sally Cross (MRC Human Genetics Unit, Edinburgh, UK) described the use of the mutagen N-ethyl-N-nitrosourea (ENU) for the generation of thousands of mice with random point mutations throughout their genome. All of the mutagenesis screens are picking up dominant, highly penetrant morphological and behavioral phenotypes. Each project is also following up specific groups of phenotypes: mutants affecting the immune system in the case of Keats Nelms and mutants affecting eye development and eye disease with Sally Cross. Whereas most ENU screens of mice are designed to identify only dominant mutations, the strategy employed by Nelms and co-workers generates pedigrees to uncover recessive mutations and intrinsically provides the basic resources to map the locus of the mutation by back-crossing. Not surprisingly, 90% of the mutations that have been found are recessive. As an alternative strategy for uncovering recessive ENU-induced mutations, Allan Bradley (Sanger Centre, Hinxton, UK) has been coordinating the development of mouse lines with *Cre-lox*-induced megabase deletions, which, when crossed with ENU-mutagenized mice would uncover recessive mutations in a locus-targeted manner. Other innovations from the Bradley lab include coat-color chromosome tags and megabase inversions, both of which have the potential to speed up the characterization of mutant mice.

The armored, three-spined stickleback (*Gasterosteus aculeatus*) is no stranger to the laboratory, having been well characterized in terms of morphology, color and behavior. Now David Kingsley (Stanford University, USA) and colleagues are using these fish to characterize the molecular genetic changes that are responsible for the morphological evolution of vertebrates. Geographically isolated populations of stickleback have undergone substantial adaptive radiation in lakes left in the wake of retreating glaciers around 10,000 years ago. The different species vary in size, shape, behavior, color and patterns of defensive armor. As most of these species have undergone parapatric divergence (as a result of population subdivision and reproductive isolation), isolating mechanisms are largely behavioral and can be overcome in the laboratory to produce fully viable F1 and F2 generations. With cDNA-library resources and a genome-wide linkage map in place, Kingsley is crossing species pairs and watching the divergent traits segregate in the offspring. It is immediately apparent that many of the segregating morphologies are caused by genes that have a major effect, rather than by many genes that have a relatively minor effect. These loci are being mapped and eventually the molecular basis of the trait will be uncovered. For one of the armor-patterning traits, it has already been found that one locus has independently changed twice to produce the same morphological adaptation.

Wendy Bickmore (MRC Human Genetics Unit, Edinburgh, UK) opened the workshop on chromosome structure and epigenetic mechanisms with the observation that it is often assumed that by combining the genome sequence with information about gene expression and the proteome we will understand inheritance. This is wrong, she said, because epigenetic processes such as histone modification and DNA methylation as well as chromatin-remodeling systems can all be important. A striking example was given by Mark Bailey (University of Glasgow, UK) who described his studies of the methyl CpG-binding protein 2 gene (MECP2). Mutations in this gene cause Rett syndrome, a severe X-linked, neurodevelopmental disorder that is the commonest cause of severe cognitive incapacity in the female population. MECP2 is believed to bind methylated CpG islands and mediate transcriptional repression. Bailey's work has identified genomic fragments to which MECP2 binds strongly. Intriguingly, many of these fragments contain Alu interspersed repeat elements, suggesting the possibility that Alu overexpression could be a cause of Rett syndrome. Not only the pioneering work going on in epigenetic mechanisms but also many other studies presented at the meeting showed that even when we finally have a functionally annotated human genome there will still be much to learn.