Meeting report

More biology from the sequence

Martin S Taylor

Address: Medical Genetics Section, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK.
E-mail: martin.taylor@ed.ac.uk

Published: 31 July 2001

Genome ***Biology*** 2001, **2(8)**:reports4018.1–4018.5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/8/reports/4018

A report on the Cold Spring Harbor meeting on Genome Sequencing and Biology, Cold Spring Harbor, NY, USA, 9-13 May 2001.

We are now clearly within the genomic era. Large-scale sequencing centers are running efficiently and are able to churn out several fold coverage of a eukaryotic genome by shotgun sequencing in a few months. This is an impressive technical and logistical *tour de force*, but for genome biology as a whole it represents just the data-acquisition phase. The shift from a data-starved field with a need for better sequencing-related technologies to a field with an avalanche of data and the need to interpret it has been reflected over the past few years by the changing emphasis at the annual Cold Spring Harbor Genome Sequencing and Biology meeting. Predictably, finishing and annotating the human genome was a major discussion point this year. The other big themes were comparative genomics, highly parallel investigations of gene expression and function, and the variability of genomes.

When is finished finished?

Although not immediately obvious from the title, this meeting has traditionally had a substantial slant towards eukaryotes and particularly humans. Although that trend continued this year, the genomes of other organisms have also been sharing the center stage. Of particular note was the entertaining keynote speech by Paul Nurse (Imperial Cancer Research Fund, London, UK) announcing the completion of the *Schizosaccharomyces pombe* genome. Unusually but not uniquely for a eukaryotic genome, 'complete' means finished in this case; it is all there, from one end of each chromosome to the other, without gaps.

It is widely 'known' that the human genome is finished; it has been on many news reports, and in the newspapers there have even been heads of state patting their own backs for the achievement. It was, then, interesting to hear the progress being made in the actual finishing and assembly of the human genome. Ian Dunham (Sanger Centre, Hinxton, UK) reported the progress in filling the gaps in the human chromosome 22 sequence (the second human chromosome to be 'finished'). When published last year, there were 11 gaps in the chromosome 22 sequence; one of these gaps has now been closed, leading to the discovery of a further three genes. On the p (short) arm of chromosome 22, gaps in the sequence are a result of low-copy repeats that cause problems with cloning, clone selection and assembly. In the remaining gaps, clones that have been identified to cover the gap have internal deletions, so they still contain the gap; this supports the many anecdotal reports of a correlation between high GC content and 'unclonable' regions. Efforts are currently underway at the Sanger Centre to clone mouse sequences syntenic to the gaps in order to understand better the sequences that may be in the gaps and to provide resources for gap filling. Ian Dunham also reported on attempts to annotate the genes on chromosome 22 fully. Similarly, Nobuyoshi Shimiau (Keio University School of Medicine, Tokyo, Japan) described efforts to clone and sequence the full-length cDNAs for predicted genes on chromosome 21.

Currently the human genome is in fact in a draft state; this is equivalent to being given a huge, jumbled-up jigsaw with 10% of the pieces missing and most of the pieces looking very much like each other. Fortunately, there are several other sources of information available that can be used to help assemble the genome. Pre-eminent among these are the finger print maps of bacterial artificial chromosome (BAC) clones and paired end reads of clones. Other data, such as radiation-hybrid map positions, are also being used to enhance assemblies. Of the three human genome assemblies (University of California Santa Cruz (UCSC), Celera Genomics Inc. (Rockville, USA) and the National Center for Biotechnology Information (NCBI, NIH, Bethesda, USA)), Jim Kent

(UCSC) and David Lipman (NCBI) outlined improvements to their respective assemblies.

Kent reported two-to-threefold reduction in false contig joins by inferring fragment order and orientation from boundaries of the sequence with the cloning-vector backbone (the Sp6 and T7 ends) as well as several improvements to the UCSC Genome Browser. Both the NCBI and UCSC are also including cDNA-to-genomic sequence alignment to facilitate the ordering and orientation of sequence fragments.

The Ensembl project [http://www.ensembl.org/] for genome annotation is making use of the UCSC assembly and providing baseline annotation of the human genome. Ewan Birney (EMBL European Bioinformatics Institute (EBI), Hinxton, UK) presented the Ensembl project as the "Linux of genome biology", operating in a fully open culture with scientists free to distribute, develop or contribute to the project. Ensembl is not just a human genome annotation system and browser; it is now also being used for the mouse genome project and is under development for the rice and *Fugu* genome projects. On top of all of this, Ensembl is embracing the Distributed Annotation System (DAS [http://biodas.org/]) so that other biologists can contribute their annotation of the genome and compare annotation from multiple sources.

Copy after copy

Segmental duplications have clearly been important in shaping many, if not all, of the genomes of extant organisms. It is equally the case that such duplications are actively shaping the genomes of primates and causing serious problems for the finishing of the human genome. For example, on the human Y chromosome there are several copies of a 3 megabase (Mb) palindrome that are 99.97% identical (Richard Willson, Washington State University, St Louis, USA). Classical yeast artificial chromosome (YAC) and BAC mapping techniques and sequence assembly have typically failed or been misassembled in such duplicated regions. Willson described a "combat sequencing strategy" for the sequencing of a single Y-chromosome and a resolute policy of "take no SNPs" for sequence assembly - given that what look like single-nucleotide polymorphisms (SNPs) could in fact be errors of assembly in repetitive regions. It is extremely likely that just such a rigorous approach is going to be needed for the duplication-rich pericentromeric and peritelomeric regions of the genome.

Although as Evan Eichler (Case Western Reserve University, Cleveland, USA) said, the 'problem' regions of the genome containing low-copy-number repeats have been, in part, intentionally left until last and are likely to be severely underrepresented in the rough draft sequence, some of the most interesting and unexpected observations are being made in these regions. Willson reported that testis-specific genes located in the duplicated regions of the Y chromosome have been found to be deleted in some forms of male sterility, demonstrating either divergence of function or dose-dependence effects, even though almost identical copies of the genes exist in adjacent segments of sequence. Eichler described a 19-20 kilobase (kb) sequence that is restricted to human chromosome 16, with 15 copies interspersed along its length, sharing 97-99% identity between copies. What is truly fascinating about these duplications is the gene that they contain. For each of the duplicated regions, the gene's reading frame is maintained, splice sites are highly conserved, and introns appear to be selectively neutral, but puzzlingly, the protein sequence is not: amino-acid changes are 15 times more frequent than expected under neutral selection. There appears to be essentially no conserved region of the encoded protein, and amino-acid substitutions have been made that do not appear to reflect any constraint on physico-chemical properties. No homolog of these genes has been found in mouse, and within other primates these genes are also undergoing duplication and divergence within the encoded protein. The comparison with primates also demonstrated that, unlike the standard model of gene duplication and subsequent functional divergence of one copy, none of the gene copies is being conserved. Work is currently underway to understand the function of this gene family and the significance of its rapid evolutionary change.

Comparative genomics

There are several more eukaryotic genome sequences on the way: mouse, *Fugu*, zebrafish, rat, rice, dog and more, with further announcements of genome projects likely in the next few years as the genome centers start to look for new projects. Richard Mural (Celera Genomics Inc.) described the 5.5x whole-genome shotgun coverage of the mouse genome generated by Celera [http://www.celera.com/]. Using three strains of laboratory mouse (129X1/SvJ, A/J, and DBA/2), Celera have identified 2.7 million SNPs where sequence derived from separate strains overlaps and contains discrepancies. Mural also reported the amazingly high rate of SNPs found within strains, one in 10,000 nucleotides across the genome, although under cross-examination by Eric Lander (Whitehead Institute, Cambridge, USA), Mural admitted that many of these SNPs probably reflect sequencing errors.

As expected, Celera have been finding good correlation of synteny between the mouse and human genomes. An interesting general theme emerging from comparative analysis of the human and mouse genomes is the consistently smaller distance between anchored markers (such as orthologous exons) in the mouse than in the human; this finding was reported by both Mural and Lisa Stubbs (Department of Energy Joint Genome Institute, Oak Ridge, USA). In her comparison of human chromosome 19 and the 15 segments of homology to it found in the mouse genome, Stubbs has found that all synteny breakpoints are in tandemly duplicated

regions, such as clusters of odorant-receptor and zinc-finger genes. Through mouse-human genomic alignment, Stubbs and co-workers have found many candidate new exons for known genes and conserved 5′ and 3′ non-coding regions as well as approximately 30 new genes that, it is claimed, would be entirely missed by non-comparative prediction methods. In summary, 80% of known human exons on chromosome 19 have significant matches to the mouse sequence, single-copy genes are overwhelmingly conserved in the mouse (only three are convincingly not conserved), 37% of conserved sequence features are not associated with any gene features, and in total 5.4% of chromosome 19 shows significant conservation at syntenically conserved regions.

Kelly Frazer (Perlegen Sciences, Santa Clara, USA) and colleagues have started to investigate the global pattern of mammalian genome conservation, using BAC contigs of regions from mouse and dog syntenic to human chromosome 21 to hybridize oligonucleotide microarrays that represent about 74% of human chromosome 21 (130 million oligonucleotides). From this study, it appears that 1.6% of non-repetitive base pairs are conserved between human and mouse and 3.9% between human and dog. Interestingly, about 30% of the conserved sequence is located at least 10 kb from any known gene. There is also a trend for shorter conserved motifs to be overrepresented in non-coding regions. The big advantage with this technique is once you have the sequence and have synthesized the oligonucleotides for your region of interest in one genome - a huge financial hurdle - the comparative analysis of the region in many genomes becomes readily tractable.

The importance of sequence depth (the number of homologous regions for which sequence is available) in comparative genomic analysis is intuitively obvious, but because of the resources needed to produce good-quality sequence for the orthologous region of multiple metazoan organisms this has been produced for only a handful of well-studied loci such as the *Hox* gene clusters, the major histocompatibility complex (MHC) and the globin genes. The work of Frazer and colleagues provides the technological means to address this issue, potentially in a hugely parallel manner for many genomes.

The gold standard for comparative genomics is base-by-base sequence comparison, however. In a major effort to achieve this for five target regions corresponding to segments of human chromosome 7, the orthologous segments of 11 other vertebrate genomes are being sequenced. Jeff Touchman and James Thomas (both from the National Human Genome Research Institute (NHGRI), Bethesda, USA) described the pipeline for identification, contig assembly and sequencing of these regions from each of the following species: mouse, rat, pig, cow, dog, cat, baboon, chimpanzee, chicken, zebrafish and pufferfish. This is seen very much as a pilot project to evaluate which genome sequences will be the most appropriate to aid in the annotation of the human genome, to provide an understanding of vertebrate genome evolution and to provide a dataset to act as a catalyst for the development of much-needed bioinformatics tools for comparative genomics. As many of the genomes that are now being produced are likely to remain in the draft phase indefinitely, it will be interesting to see what proportion of the findings from this work would be discovered if only draft, rather than finished sequences were used. One preliminary finding from this work is particularly worthy of note: plotting the degree of sequence conservation against accepted evolutionary distances, the mouse and rat would appear to be significant outliers from the general trend being substantially more divergent than expected. This observation has implications for the interpretation of the mouse genome and its use in the comparative annotation of other genomes. Discussions after the presentation also highlighted a community-wide need for centralized information on the availability of BAC libraries and possibly even for coordination of library distribution.

There has been much publicity for the potential for comparative genomics in the prediction of gene structure. Ian Korf (Washington University, St Louis, USA) and Roderic Guigo (Genome Informatics Research Laboratory, Barcelona, Spain) presented their efforts to integrate cross-species sequence homology into *ab initio* gene-prediction methods. Korf's program, Twinscan [http://genes.cs.wustl.edu/], produces conservation profiles from BLASTN alignments, which are included as components of a gene prediction based on the program Genescan. Guigo's program (as yet unnamed) utilizes similar methods, but uses TBLASTX alignments integrated with the Geneid [http://www1.imim.es/software/geneid/] *ab initio* gene-prediction method. Both programs have been designed to work with shotgun mouse sequences for human gene prediction, and both groups concluded that further development of these approaches was necessary to achieve an optimally performing package.

What else is in my genome?

Matthew Meyerson (Dana-Farber Cancer Institute, Boston, USA) presented a strategy to use the mass of genomic and cDNA sequence data to uncover the existence of previously unknown pathogens. For example, computational subtraction of human cDNA sequences from the human genome should leave you with no cDNA sequences (assuming the human genome was complete). In reality that is obviously not the case: you get contamination with sequences from *Escherichia coli*, mouse, rat and other more exotic species. Meyerson's idea is that, as well as this erroneous annotation and *E. coli* sequence, human pathogens will be represented when libraries have intentionally or inadvertently been made from infected tissue. As a proof of principle, 7,073 expressed sequence tags (ESTs) derived from a human cervical carcinoma cell line were subtracted from the human genomic sequence and the *E. coli* genome. There remained 22 candidate sequences, of which ten amplified normal

human genomic DNA and two amplified only DNA from HeLa cells (the cell line used). Both of these sequences were from human papilloma virus 18, a common cause of cervical cancer. Turning their attention to the 3.3 million human ESTs, Meyerson and colleagues searched the set of ESTs from which genomic and *E. coli* sequences had been subtracted against known human pathogen sequences. They found, amongst other things, that 0.2% of ESTs derived from a human liver cDNA library are from the hepatitis B virus. Meyerson reported finding several other sequences in the databases that are probably microbial in origin, and these are being followed up.

Contaminants of the human genome were also the topic of a poster by Paul Kits and Greg Schuler (NCBI, Bethesda, USA) who are ridding the database of human genome sequence of foreign contamination, using a battery of screens against vectors, transposable elements, non-human repeat sequences and a host of other datasets. The majority of contamination was predictably found to be cloning vectors and *E. coli* genomic sequence. Bacterial transposons have also been efficient at invading cloned sequences, with at least 110 insertions by the IS10 element. Sequence from several other sequencing projects have also crept into the human genome. Interestingly, since results of this work started to be released, there has been a 60% decrease in the incidence of contamination from sequencing centers.

Functional genomics

The *Saccharomyces cerevisiae* Genome Deletion Consortium has generated a series of 20,000 yeast strains with precisely defined gene deletions covering 96.5% of the genome - an impressive feat. Each deletion strain is also uniquely identified by a 'molecular barcode'. Adam Deutschbauer (Department of Genetics, Stanford University, USA) and coworkers have been making good use of this great resource: using the molecular barcode on microarrays, they have measured the growth rate of nearly 5,000 homozygous diploid mutant yeast strains in parallel. In elegantly designed screens, more than 50 genes have been newly associated with germination and sporulation. This work also represents the first screen specifically for germination mutants in yeast and has more than doubled the number of genes known to be necessary for germination. Interestingly, this work has also shown that the majority of genes with a meiosis-specific transcription pattern are not necessary for efficient sporulation, at least under the test conditions.

RNA-mediated interference (RNAi) is an amazingly simple and powerful technology, particularly in the model organism *Caenorhabditis elegans*. The nematode is fed or injected with double-stranded RNA, which through some little understood 'black magic', probably involving targeted RNA degradation, effectively eliminates the expression of the target gene. The attractiveness of this system has led Andrew Fraser (Wellcome CRC Institute, Cambridge, UK) and colleagues to produce a library of bacterial clones expressing double-stranded RNA that can be used directly in RNAi experiments. So far, Fraser and colleagues have generated libraries for genes on chromosomes I, II and X, representing 41% of predicted and known genes. Screening only by visual examination of anatomy and behavior, about 20% of genes have given phenotypes. It is interesting, although not necessarily unexpected, that more highly conserved genes are more likely to produce a detectable phenotype in these screens. Unfortunately, RNAi appears less effective at generating neuronal phenotypes when RNAi and known null mutations for genes are compared, although it appears to work extremely well for all other tissues.

Marc Vidal (Dana-Farber Cancer Institute, Boston, USA) described the idea and progress of the *C. elegans* 'ORFeome' project. The basic idea is to generate a library containing every *C. elegans* open reading frame (ORF) in a vector that would allow the rapid and efficient movement of inserts between vectors. Such a resource would potentially allow all-versus-all protein-interaction studies, a basis for the expression of all proteins, and high-throughput protein structural studies. In generating this resource, a side effect is to effectively test the prediction of every gene in the genome and provide evidence for or against its predicted expression and exact genomic structure. The ORFs are cloned by PCR amplification from cDNA libraries using ORF-specific primers. Expression has been demonstrated for more than 17,300 genes; for 10,000 of these there was previously no EST or cDNA evidence of their existence.

Polymorphism

In contrast to sequencing technologies, there is clearly still a need for technology development in high-throughput genotyping and polymorphism detection. Sanya Tyagi (Public Health Research Institute, New York, USA) presented molecular beacon technology, in which oligonucleotides with fluorophores and quenchers result in a fluorescent signal when the oligonucleotide hybridizes to a complementary sequence. This technology is being used for detection of SNPs as well as *in situ* hybridization, and is being developed for several infectious-disease diagnostic uses. The other technologies presented included pyrosequencing, a four-enzyme sequencing system (Mostafa Ronaghi, Stanford DNA Sequencing and Technology Center, USA); an exo-proof-reading assay (Patrick Cahil, Genome Therapeutics Corp, Waltham, USA), which relies on the enzymatic release of the 3′-most nucleotide of an oligonucleotide and the associated fluorescent label if it is not complementary to the hybridized sequence; and the Invader assay (Michael Oliver, Stanford Human Genome Center, USA), which uses a structure-specific cleavase to distinguish sequence-induced structures when partially overlapping oligonucleotides are hybridized over the site of an SNP (the number of fluorescein molecules

released reflects the ratio of alleles at the SNP). All these technologies are being scaled up to high-throughput scales.

Microarray-based technology is emerging as the format of choice, independent of the actual genotyping or SNP-detection method used. A variation on the typical microarray scheme was presented by Mark Chee (Illumina Inc., San Diego, USA), in which arrays are randomly assembled on the specially pitted tips of optic fibers. Each pit contains a bead with bound oligonucleotides, 50,000 beads per array. Although the identity of each bead is unknown at the time of assembly, a series of specific hybridizations is carried out to uniquely identify each bead. Oligonucleotides bound to each bead have been used for hybridization, oligo-ligation (template-sequence-dependent ligation of oligonucleotides) and PCR-based methods of SNP genotyping. The arrays are readily clustered into higher-order arrays and have been used in 96-well-plate conformation for parallel processing of samples.

Linkage disequilibrium (LD) is the fundamental basis for all positional cloning efforts, yet its large-scale behavior across the human genome is little understood. In an elegant introduction to the topic, David Reich (Whitehead Institute, Cambridge, USA) described LD as "a string of alleles occurring together in a population more often than would be expected based on the product of their individual frequencies". The underlying assumption is that each of these strings of alleles (haplotypes) was derived from an ancestral chromosome. Using 19 'randomly selected' regions of the genome, and determining the haplotypes from 44 Utah individuals of Northern European descent, Reich and colleagues found the LD half-length (the physical distance at which significant *p*-values for LD occur >50% of the time) to be typically 60 kb but extending beyond 100 kb in places. The same pattern was also found for a Swedish population, but a strikingly distinct pattern was observed when Yoruba tribespeople (from Nigeria) were tested, with LD extending markedly less far than in populations of European origin. This is argued to provide evidence for a substantial founder effect in the European population, with as few as 50 founder individuals giving rise to all Northern Europeans. The major conclusion from this work is that LD gene mapping is practical with existing SNP datasets, but fine-structure mapping may require a more diverse population than Northern Europeans.

In a systematic investigation of LD over the long arm of chromosome 22, Elizabeth Dawson (Sanger Centre) and colleagues have selected SNPs and deletion/insertion polymorphisms every 7-15 kb along chromosome 22q and genotyped them in seven families from the Centre d'Etude du Polymorphisme Humaine (CEPH) genotype collection and 92 unrelated individuals. For the markers typed to date, the average spacing is 22 kb and the median is 14 kb. LD was found to be a generally good predictor of physical distance, although a plot of LD across the whole chromosome exhibits many pronounced peaks and troughs. The upper limit of LD detected in this study was approximately 300 kb. The beauty of this approach combined with finished genomic sequence is that physical distances are accurately known and genome features can be directly correlated with observations of LD. One striking observation has already been made: LD strongly correlates with gene content.

Directions for the future

For the immediate future, the most dramatic developments in eukaryotic genome biology are likely to be in comparative genomics, with the release of whole-genome shotgun sequence for several metazoans. The substantial challenge still remains to accurately annotate this sequence, decode much of the functional information and understand the evolutionary processes that have culminated in its existence.

The emphasis at this meeting has shifted away from the technical aspects of data acquisition in favor of interpretation. We have already had glimpses of the next shift, in which these data are being used to address the new and old questions of biology. As all good science should do, genome biology is throwing up more questions with every answer.