

Meeting report

Whose genome is next?

Danielle Kemmer* and Andrew Fraser†

Addresses: *Center for Genomics and Bioinformatics, Karolinska Institute, Berzelius väg 35, 17177 Stockholm, Sweden. Current address: Centre for Molecular Medicine and Therapeutics (CMMT), 950 West 28th Avenue, Vancouver, BC, V5Z 4H4, Canada. †Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

Correspondence: Andrew Fraser. E-mail: agf1ouk@yahoo.co.uk

Published: 28 November 2002

Genome Biology 2002, **3**(12):reports4037.1–4037.3

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2002/3/12/reports/4037>

© BioMed Central Ltd (Print ISSN 1465-6906; Online ISSN 1465-6914)

A report from the 14th Genome Sequencing and Analysis Conference, Boston, USA, 2-5 October 2002.

At the heart of the 14th Genome Sequencing and Analysis Conference, organized by The Institute for Genome Research (TIGR), were a number of ways to tackle a problem facing all avid readers - we know how to read, but which books should we choose? One might read Tolstoy and Flaubert until dawn and beyond, but neither wrote any books explaining how to make chocolate brownies. In short, reading sequence is like reading words, and now that we know how to do it, the question (given the current costs) is which genomes to read and why.

The meeting was opened by Barry Bloom (Harvard School for Public Health, Boston, USA), who gave a provocative reminder of how far we still have to go in our efforts to combat infectious diseases in both rich and poor countries - examples included the rapid rise in the developed world of drug-resistant strains of *Mycobacterium tuberculosis* and *Staphylococcus aureus* (60% of Japanese hospital cases are multiply drug resistant), and the incredibly fast rise of drug-resistant malaria strains in Africa. The hope, of course, is that genomic approaches will prove valuable in understanding pathogens and host-pathogen interactions and thus lead to the development of novel drugs and therapies, and this is already proving to be the case in the development of vaccines. In addition to the unquestionable humanitarian benefits of providing adequate healthcare for the whole global community (arguably a moral imperative), Bloom presented a possible model of healthcare as investment, arguing that the economic benefits arising from improved Third World health more than outweigh the capital investment required. As a stage-setting talk, it was ideal - the central message that

there is a huge amount more sequencing to be done was clear, as was the reminder that while scientific interest is important in determining future targets for sequencing, we have a duty to bear in mind the usefulness of our work for humanity at large, from Boston to Botswana.

More genomes, more species

The recurring question of what (or who) to sequence next is in many ways the natural counterpart to the comparative sequence analysis presentations that made up a substantial part of the meeting: would chimp, baboon, bonobo and orangutan sequence be 'more useful' than that of platypus, chicken, dog and snake, for example. The answer of course depends on the precise problem being addressed, a point made very clearly by Eddy Rubin (Lawrence Berkeley National Laboratory, Berkeley, USA). He illustrated two examples of comparative sequence analysis, the first in which human-murine comparison was perfectly sufficient to identify a conserved regulatory element in the interleukin gene cluster on human chromosome 5q31, and the second in which the alignment of multiple primate sequences was required to identify a novel apolipoprotein, apo(a). Thus, while some questions can be answered with what we already have, others will need a wider and deeper sampling of the vertebrate genome pool. Eric Green (National Human Genome Research Institute, Bethesda, USA) set out the current NIH plans for sequencing non-human genomes, including the genomes of chicken, dog, cow, several fish and, perhaps most interestingly, several marsupials.

Although most comparative analyses focused on either analysis of coding sequences or the identification of conserved regulatory elements, Victor Ambros (Dartmouth Medical School, Hanover, USA) presented a computational approach to identifying non-coding microRNA genes - which encode small

non-translated RNAs - in the nematode *Caenorhabditis elegans* by comparison of *C. elegans* sequence with that of its close relative *Caenorhabditis briggsae*. He estimates that there are around 150 such genes in *C. elegans*, and that about 10% of those identified have human counterparts, suggesting strongly that microRNAs are a widespread and common mechanism of gene regulation.

C. elegans-C. briggsae sequence comparisons also featured in the presentation by Andy Fire (Carnegie Institution of Washington, USA), which focused on some of the many ways in which hosts recognize their own genomes as distinct from foreign nucleic acids. For a long time it was unclear why transgenes become silenced in the *C. elegans* germline if they are inserted within the context of foreign genomic DNA (such as human sequence) rather than having coding regions inserted into *C. elegans* genomic DNA. Careful sequence analysis revealed that, unlike other available genomes, both the *C. elegans* and *C. briggsae* genomes have a remarkably regular phasing of AA/TT dinucleotides; Fire estimates that as much as 1-2% of each genome is devoted to this phasing. The observed phasing, which has a periodicity of 10 base-pairs, leads to the assembly of a DNA helix which has a markedly A/T-rich face; this is thought to be structurally different from a randomized genomic sequence. The significance of this phasing to the silencing of genes in the germline is confirmed by the finding that endogenous genes that are transcriptionally active in the germline are 'phased', whereas others are not. Thus, one way in which nematodes may distinguish self from non-self at the level of an entire chromosome may be a subtle structural difference in the DNA helical conformation brought about by extended periodicity in nucleotide sequence. Whether similar mechanisms are used by other eukaryotes is not yet known, but this phenomenon provides an intriguing example of the range of genetic mechanisms uncovered through the careful analysis of available genomic sequence.

Other highlights of the comparative sequence analysis talks included the presentation by Kelly Frazer (Perlegen Sciences Inc., Mountain View, USA) of a comparison of human and chimp chromosome 21 using high-density oligonucleotide arrays. Around 180 million 25mer oligonucleotides were used to sample the non-repetitive chromosome 21 sequences (making 22.5 Mb of the 33 Mb total), and the surprising finding was that around 9% of human chromosome 21 sequences are deleted when comparison is made between either human and chimp or human and baboon; furthermore, an amazingly large number, 35%, of the deletions are in genic regions, suggesting that small rearrangements and deletions may play a substantial role in genome evolution.

Finally, there were also presentations in which comparative analyses were used to reach beyond the identification of individual functional elements (coding or non-coding), to begin to describe entire gene networks. This approach was

illustrated by both Ed Marcotte (University of Texas, Austin, USA) and Peer Bork (European Molecular Biology Laboratory, Heidelberg, Germany). Both groups used gene-fusion data and phylogenetic profiles, along with genetic and physical interaction data, and microarray expression data, to assemble complex network models of gene interaction and function. Currently, such analyses are particularly useful in prokaryotes, for which far more genome sequences are available; but as more metazoan genomes (and systematically compiled functional data) become available, these computational 'systems' approaches look increasingly attractive.

In addition to the more general comparative analysis talks, presentations were also made of the completed public sequences for rat and mouse, by Richard Gibbs (Baylor College of Medicine, Houston, USA) and Kerstin Lindblad-Toh (Whitehead Institute for Biomedical Research, Cambridge, USA), respectively. Mouse-human comparisons of coding sequence suggest a lower human gene count than previous estimates - somewhere in the order of 28,000 genes, for those with sweepstake tickets - and a full 80% of human genes have a direct, single ortholog in mouse. There appear to be 25 clusters of mouse-specific genes, the great majority of which are involved in either reproduction (14 clusters) or immunity (5). Most interesting, perhaps, is the finding that while only 1.5% of the mouse genome is thought to be coding, models suggest that as much as 5% is under detectable selection. The 3.5% of selected non-coding sequence will clearly be fertile ground for future analysis.

Other fully sequenced genomes presented at the meeting included those of anthrax (Steven Salzberg, TIGR, Rockville, USA) and the sea-squirt *Ciona intestinalis* (Daniel Rokhsar, Department of Energy Joint Genome Institute, Walnut Creek, USA). Sequence analysis of this invertebrate chordate sheds light on aspects of vertebrate evolution: there is evidence for many 'vertebrate-specific' molecules (such as claudins and noelin), including several that are involved in the immune system (complement system components and Toll-like receptors). In addition, there was strong sequence evidence that *Ciona* has a cGMP-based light-sensing cascade very similar to that of vertebrates. Whether we sequence more vertebrates, more pathogens or more plants, what is absolutely clear is that the quantity of raw sequence data will continue to grow unabated in the foreseeable future, and the insights from comparative sequence analysis will grow accordingly.

How Y stops the rot

Although comparative analyses of multiple genomes appear to be the way of the future, there are still surprises hiding in individual genomes. This was perhaps best illustrated by David Page (Whitehead Institute for Biomedical Research) who presented a startling new model for Y chromosome evolution that arises from the newly available Y chromosome

sequence. The Y chromosome contains approximately 24 Mb of euchromatin (around 1% of the genome) and sequencing is now nearly complete for this stretch, which contains not only the extended region of close (greater than 99%) identity between the X and Y chromosomes but also the euchromatic portion of the so-called non-recombining region of Y (the NRY). Page suggests that the NRY be rechristened the MSY - the male-specific region of Y - for reasons that become obvious. About 30% of the euchromatin in the MSY is 'ampliconic': that is, it has more than 99% sequence identity to other MSY regions. But these 'repeats' are not low-complexity short strings of bases: rather, they are lengthy sequence duplications arranged as palindromes, with fully 25% of MSY euchromatin contained in eight palindromes that range in size from 36 to 1,500 kilobases. Furthermore, although the 18 single-copy MSY genes that have X homologs are all outside the ampliconic regions, all 72 testis-specific MSY genes are in amplicons. This clearly suggests that the maintenance of Y-specific gene integrity is closely associated with being in the amplicon palindromes, and led Page to the amazing model that Y-specific genes are maintained by Y-Y gene conversion events within the Y amplicons (or as one questioner paraphrased it "Y likes to have sex with itself"). Thus, the Y chromosome is not merely some rotting relic of an ancient autosome, but it is, rather, a specialized chromosome that maintains the integrity of its genes through unexpected intrachromosomal gene conversion events, estimated to be as frequent as 1-2 per generation. Clearly this model poses several major mechanistic questions, but as an illustration of the power of sequence analysis to shed light on complex biology it takes some beating.

One genome at a time, or a whole ecological niche?

As well as major efforts to sequence the complete genomes of many organisms, one of the most intriguing possibilities for current sequencing technologies is to generate essentially random sequence reads in such a way as to sample the genetic complexity of whole environments and in this way to take a 'sequence snapshot' of an ecological niche. This can be used not only to identify novel sequences belonging to hitherto unidentified species, but also to monitor how the genetic diversity of a particular environment changes over time. Sequencing could thus be a powerful way to check the ecological pulse of fragile or poorly understood environments. This approach was beautifully illustrated in the back-to-back talks of Edward DeLong (Monterey Bay Aquarium Research Institute, USA) and David Relman (Stanford University, USA).

DeLong used 'cultivation-independent' genomic approaches to study the vast spectrum of oceanic microbes, essentially using random reads of total isolated microbial populations from different local environments (for example, different ocean depths). Among other examples, he identified both

archaeobacteria and bacteria involved in anaerobic growth on methane as a carbon source; these organisms were often found to co-exist in structured syntropic aggregates. What was key, however, was that the bacteria identified had never been cultured in the lab, and such random-read approaches provide an excellent window through which to explore poorly understood organisms and ecologies. For Relman, the environment being sampled was the human body and the incredibly diverse range of microorganisms that live inside each of us. Relman estimates that of the cells present in the body only around 10% are human, with the remaining 90% being bacterial. More surprising, however, is the richness of strains. In just a single niche (for example, the subgingival cavity, 'between your teeth') one can find evidence for around 500 different species of bacteria, by using rRNA sequence clustering. Furthermore, over 50% of the rRNA sequences uncovered in this way had never been observed in any cultured bacterial strains, illustrating both how little we have yet sampled of possible genomes and also how sequence sampling of ecological niches can provide a means to investigate previously unanalyzed organisms.

In summary, the emerging picture from this meeting is that we have only scratched the surface of genome sequences and the insights that can be gained through sequence analysis. Whether these insights emerge from the detailed analysis of an individual genome, the comparison between the sequences of multiple genomes, or the sampling of the sequences present in whole ecological niches, sequencing and sequence analysis will play a major role in the way we approach biology for many years to come.