Meeting report

Genomics and the renaissance of generalism

Andrew JG Simpson

Address: Ludwig Institute for Cancer Research, São Paulo, Rua Professor Antonio Prudente, 01509-010 São Paulo, SP, Brazil.
E-mail: asimpson@node1.com.br

Published: 9 June 2000

Genome ***Biology*** 2000, **1(1)**:reports411.1–411.2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/1/reports/411

A meeting report of the sessions on human, eukaryotic and bacterial genome sequencing at the American Society for Microbiology and Institut Pasteur joint conference: Genomes 2000 International Conference on Microbial and Model Genomes, Paris, April 11-15, 2000

Genome projects are undertaken by individuals who want to know everything about an organism and have the drive and energy to pursue such an apparently unachievable goal. The Genomes 2000 conference brought together a remarkable cross-section of such individuals who collectively described many highlights of genome sequencing projects from around the world.

The complete sequence of the euchromatin comprising 120 Mb of the 180 Mb *Drosophila melanogaster* genome presented by Craig Venter (Celera Genomics) revealed that the method of choice for genome sequencing is now essentially a resolved issue. High-throughput capillary sequencing has resulted in the virtual elimination of tracking problems and the resulting misassignment of reads. Paired sequences from random DNA fragments of known sizes are assembled by massively powerful computational hardware (Celera has the world's largest civilian supercomputer facility to accomplish such feats) and permit rapid, complex genome sequencing and assembly. The availability of a significant proportion of the *Drosophila* genome derived from carefully mapped and individually sequenced large genome fragments cloned in bacterial artificial chromosomes (BACs), provided the crucial validation of the whole genome shotgun method presented by Venter. The availability of such ordered sequence for the human genome will likewise greatly enhance the credibility and accuracy of the shotgun version (sequenced but not yet assembled at the time of the conference). But it seems unlikely that future whole genome sequences will see significant contributions from ordered large genome fragments, as the shotgun approach becomes the norm.

One of the more interesting of current challenges in genome research is the unambiguous identification of genes in eukaryotic genomes. In the case of the *Drosophila* genome, it was reported that gene identification was undertaken using a combination of two software algorithms together with available cDNA sequences and expressed sequence tags (ESTs). These confirmed, at least in part, the presence of around 65% of the almost 13,000 genes predicted. Is a fly really so much simpler than a worm (which has around 18,000 genes), or are we missing something somewhere? And what do all these genes do? Around 40% of predicted *Drosophila* genes have had functions assigned to them, while 48% are totally new to science and 11% are similar to genes of unknown function in other organisms. Undoubtedly this wealth of ignorance will set the agenda for cell and molecular biologists (and genome sequencers in their spare time) for the coming decades.

According to Venter, the human genome compiled by shotgun sequencing at Celera will be assembled, annotated and published by the end of the year. Stephan Beck (The Sanger Centre) and Jean Weissenbach (Genoscope) reported that the publicly funded Human Genome Project has already finished sequencing one third of the human genome using a BAC-based approach, and that a rough draft of the entire genome will be available by mid-2000. Both Venter and Beck estimated the number of human genes to be in the order of 80,000, although the density of genes actually annotated in the finished chromosome 22 to date, reported by Beck (545 for a supposedly gene-rich 1% of the genome), would indicate that the real number is rather below this value. Are genes being missed or is the overall gene number being overestimated? Genome comparisons may help resolve this issue, as gene-rich regions are relatively well conserved in evolution, whereas other regions are not. Venter reported that Celera is already embarked on a mouse genome sequencing project with the aim of aiding human gene identification, as indeed is the publicly funded international sequencing consortium. In this context, Weissenbach reported the results of a compari-

son using genome sequences from a puffer fish species. He found some 88,000 conserved exons by comparison with human genome data, at a rate of two or three per gene, leading to an estimate of the number of human genes of only 30-35,000! It will be fun to find out eventually just how gene-rich (or gene-poor) we really are.

The *Arabidopsis thaliana* genome described by Marcel Salanoubout (Genoscope) and Samir Kaul (The Institute for Genomic Research), will be the first plant genome to be completely sequenced when it is finished later this year. The compact gene arrangement (one gene per 4 kb), short transcripts (average 2 kb) and small number of exons (average five) in this genome should aid gene identification. Nevertheless, both Salanoubout and Kaul reported that only some 35-40% of predicted genes have corresponding cDNA or EST sequences at present. A pervasive feature of plant genomes is the presence of extensive repetitions and duplications. Indeed, Kaul reported that even in the compact *Arabidopsis* genome, some 38% of the genome is duplicated at the nucleotide level. The smallest grass genome is that of rice (*Oryza sativa*), the sequencing of which was reported by Takuji Sasaki (National Institute of Agrobiological Resources, Japan). Although there has been enormous investment in the construction of detailed maps and libraries, actual sequence accrual has been relatively limited within the Rice Genome Research Program. Sasaki reported, however, that this was likely to be imminently greatly advanced by sequencing data contributed by the private sector.

Of smaller size, but no less interest, are the genomes of *Plasmodium falciparum* (Sharen Bowman, The Sanger Centre), *Dictyostelium discoidium* (Adam Kuspa, Baylor College of Medicine) and *Encephatlitozoon cuniculi* (Christian Vivarès, Université Blaise Pascal). In the highly AT-rich genome of the *Plasmodium*, genes are relatively easy to identify as they stand out as comparatively GC-rich islands, although very small exons do remain difficult to pick out. The *Dictyostelium* genome is only some 34 Mb in size and is thought to contain only 8,000 genes, despite this organism's free-living lifestyle and ability to undertake multicellular development. The *E. cuniculi* genome is extraordinarily small, consisting of 11 chromosomes of between 207 and 305 kb and totaling only 2.9 Mb. The genome represents the probable minimal eukaryotic genome and also provides a model for studying intracellular parasitism.

The method of choice for sequencing bacterial genomes, the shotgun, has been defined as such since the completion of the first genome sequence (*Haemophilus influenzae*). Michael Fonstein (Integrated Genetics Inc.), however, pointed out that a considerable amount of bacterial genome sequence has been generated from cosmid clones, including the basis of the completed *Xylella fastidiosa* genome that I presented. Claire Fraser (The Institute for Genomic Research) estimated that, in the next two to three years, the international genome sequencing efforts will generate more than 200 Mb of bacterial genome sequence containing around 200,000 predicted genes, some two to three times the estimated number of genes in the human genome. Because of the compact nature of bacterial genomes, and the absence of introns, gene identification is much easier in microbes than in eukaryotes. The remaining issue then is the definition of putative function. Of the approximately 40,000 bacterial genes in published genomes, only around one half have had a putative function assigned to them and around one quarter of all genes are unique to a single sequenced species. Again, it is clear that in the microbial world there is an enormous amount of biology yet to be understood.

Amongst the nine recently completed (or very nearly completed) genomes presented during the first two days of the conference, three were human pathogens (*Listeria monocytogenes, Streptococcus pyogenes* and *Streptococcus pneumoniae*) and four free-living microbes (*Streptomyces coelicolar, Pyrobaculum aerophilium, Lactococcus lactis* and *Rhodobacter capsulatus)*. The other two - *Xylella fastidiosa* and *Buchnera* sp. APS - represent the first plant pathogen and first obligate symbiont genome sequences to be completed, respectively. Despite the enormous variation of biological function and lifestyle exhibited by these organisms, the presentations underlined the consistencies of bacterial genomes (uniform gene density of one gene per kilobase, one half of the genes with assigned function), the uniform approach to their study (large-scale sequence generation, gap closure and annotation) and, interestingly, the similar intellectual approach taken by the protagonists. Which brings me, albeit belatedly, to the theme of this article. I have recently read laments of the demise of the generalist in biology. I can assure the reader that, in fact, the generalist is alive and well and sequencing genomes. When a complete genome sequence is obtained, one is confronted with an organism's complete biological complexity. Although individual gene systems may be of particular interest in particular organisms - for example, the antibiotic synthesis genes in *S. coelicor*, the proteolytic enzyme genes in *L. lactis* and the virulence genes in *X. fastidiosa* - all the whole-genome studies presented at the conference were notable in the way that they rapidly allowed an overview, based on myriad details too numerous to list, of each organism's metabolism, evolutionary history, lifestyle and structure. Although frustrated by the thousands of genes with unknown function, the delight of the global, all-embracing analysis of organisms was apparent.

A moment's reflection and a little short-term memory reveals how much biology has altered in the last five years. I cannot recall a congress held in 1995 that collected together the highlights of human, insect, parasite and microbial biology. But at Genomes 2000, each highlight was presented by a specialized generalist, and enjoyed by the new generation of generalists empowered by megabases of sequence and gigabytes of memory to draw together the underlying truths of life as revealed through genome sequencing.