Meeting report

How many genes does it take to make a human being?

Pablo D Rabinowicz, Erik Vollbrecht and Bruce May

Address: Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA.
E-mail: rabinowi@cshl.org

Published: 28 July 2000

Genome ***Biology*** 2000, **1(2)**:reports4013.1–4013.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/2/reports/4013

A meeting report from the 13th Annual Cold Spring Harbor meeting on Genome Sequencing and Biology, May 10-14, 2000. Cold Spring Harbor, New York.

The 'C value paradox' - the lack of correlation between genome size and organism complexity - has been explained by differences in ploidy and in abundance of repetitive elements. This allowed the speculation that organism complexity might correlate with gene number, rather than genome size. Thus the human, one of the most complex organisms, is expected to harbor a large number of genes. With the current availability of partial or complete sequence of several model organisms' genomes as well as two thirds of the human genome, a more accurate estimate can now be made of the number of genes in each organism.

This was a recurring theme during the 13th Annual Cold Spring Harbor Meeting on Genome Sequencing and Biology. Early in the meeting, Gerald Rubin (HHMI and the University of California Berkeley) pointed out that the products encoded by the 13,600 genes found in *Drosophila* can be grouped into a 'core proteome' (number of distinct protein families in an organism) of about 8,000. These numbers are comparable to those in *Caenorhabditis elegans*, the other multicellular animal whose genome sequence has been completed, two-fold higher than those in *Saccharomyces cerevisiae*, and almost six times those in *Haemophilus influenzae*. Over the past few years, the accepted estimate of the number of human genes has been around 100,000. Now, however, it seems that although humans might have a larger number of individual genes than flies or worms, this number, too, may be less than previously estimated. Rubin also predicted that the human core proteome will be similar to that of *Drosophila* and *C. elegans*, and will contain few novel, human-specific protein domains. A much larger repertoire of particular proteins could be produced by alternative splicing and post-transcriptional and post-translational modifications.

Several talks in the human sequencing session supported Rubin's notion. André Rosenthal (Institute of Molecular Biotechnology, Germany) and Yoshiyuki Sakaki (RIKEN, Japan) reported that the finished sequence of chromosome 21 contains large gene-poor regions. This is in contrast to previous findings when analyzing chromosome 22, the first human chromosome to be completely sequenced and annotated. If gene-finding models are accurate and if the gene-rich and gene-poor chromosomes together reflect an average gene content for the genome, then the number of human genes would approach only 40,000. Hugues Roest Crollius from Genoscope (France) described the 'Exofish' bioinformatics tool [http://www.genoscope.cns.fr/exofish] that finds exons in human DNA by comparison with DNA from *Tetraodon nigroviridis* (pufferfish). Exofish assumes that sequences coding for proteins evolve more slowly than non-coding sequences, and hence that it is possible to identify genes by looking for sequences that have remained similar during evolution. Exofish predicted similar gene numbers in chromosomes 21 and 22 as have been identified by other methods, and predicted only 28,000 to 34,000 genes in total when applied to the working draft containing 42% of the human genome. The emerging debate was best reflected in a betting pool initiated by Ewan Birney (European Bioinformatics Institute, UK) to guess the number of genes discovered by 2003, when the highly accurate annotated sequence of the human genome is to be released. The statistics of this betting pool presented by Francis Collins (director of the National Human Genome Research Initiative) in his Keynote Lecture, showed a mean bet of around 50,000 genes and some experts bet as low as 27,000 or as high as 160,000. Interestingly, plants may climb to a higher level of 'complexity' when reviewed in this context, as the model

plant, *Arabidopsis* is predicted to contain about 25,000 genes.

An important and old question that must be answered in order to determine the exact number of genes in any organism is how to define a gene. Annotations for genes should greatly improve in the future, as a result of efforts to integrate expression data and information from the published literature. Martin Ringwald from the Jackson Laboratory and Steven Gullans from Harvard described databases containing expression information from mouse [http://www.informatics.jax.org] and human [http://www.hugeindex.org] tissues, respectively, which will be integrated with other genomic resources such as SWISS-PROT [http://www.expasy.ch/sprot/sprot-top.html] and GenBank [http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html]. Meanwhile, the ambitious BioKnowledge Library at Proteome, Inc. [http://www.proteome.com] is employing curators to comb the published literature and produce summaries to be used as annotations for genes. Taking this approach one step further, Hagit Shatkay from National Center for Biotechnology Information (USA) presented GenTheme, an algorithm to map genes to documents in PubMed and search for relationships between genes through their coincidences in published abstracts.In the absence of any stock-shaking announcement from Celera Genomics Inc., another main feature of the conference was *Drosophila*, whose genome sequence was recently completed by a collaboration between the Berkeley Drosophila Genome Project, Celera Genomics and others. Gene Myers from Celera described the controversial strategy of shotgun sequencing, used to almost complete the fly genome sequence. The strategy consists of breaking the genome into small (average 2 kilobase pairs, kbp), medium (10 kbp) and large (100-200 kbp) size fragments and obtaining reads from both ends of each clone. Before all the reads are assembled, those containing repetitive DNA sequences (which would complicate assembly) are separated from the dataset. To identify these repetitive sequences, all fragments are compared against each other to look for overlaps 40 bp long. A 40 bp overlap is considered to be not coincidental, and more than 6% mismatch in overlapping sequences is taken to indicate repetitive elements. The remaining low copy sequences are then assembled into contigs and the pairs of reads from medium and large clones are used to build scaffolds (sets of ordered contigs). The result of this strategy was a set of scaffolds, 25 of which contained more than 95% of the assembled sequence (116 Mbp). The 1,600 gaps that currently remain are being closed one-by-one using bacterial artificial chromosomes (BACs). The audience was eager to know how the strategy will work with the human genome, whose completion Celera recently announced. Without giving details, Myers answered that application to the human is essentially an issue of scale. For example, the computer memory required would be a few thousand gigabytes. Jeff Bailey of Case Western Reserve University (USA) was less optimistic, pointing out that the four-fold higher genomic duplication rate in human will add an extra factor of complexity. While Celera stand by its claims that it is possible to assemble the whole fly as well as human genomes by shotgun sequencing, Steven Scherer (Baylor College of Medicine) pointed out that the published assembly of the fly genome was obtained with significant aid from the BAC-based physical map produced at his institution. Obviously, the private sector will take advantage of the data produced by the public effort in the case of the human genome, where formal collaboration has not been possible. In this regard, a singularly helpful tool for shotgun assembly of the human genome is the physical map produced at the Genome Sequencing Center at Washington University in St. Louis (USA). John McPherson, from this center, presented an almost complete physical BAC map in which 97% of the contigs are anchored to chromosomes or chromosomal regions. McPherson is confident that the gaps in the map will soon be closed.

Single nucleotide polymorphisms (SNPs) are taking the stage for the post-genome era. Lincoln Stein (Cold Spring Harbor Laboratory) announced that more than 140,000 SNPs are available in the SNP consortium (TSC) database [http://snp.cshl.org/] and there are many more in other databases (for example HGBASE [http://hgbase.interactiva.de] and dbSNP [http://www.ncbi.nlm.nih.gov/SNP]). Nevertheless, a detailed analysis of TSC SNPs, presented by Pui-Yan Kwok, showed that - although almost 70% of the SNPs are informative - care should be taken when using these markers to avoid, for instance, using a SNP whose minor allele is rare. The present challenge is to genotype SNPs in an affordable way. Several alternatives are being explored. One of these techniques, called 'minisequencing', was presented by Ann-Christine Syvänen from Uppsala University, Sweden, for genotyping disease alleles in the Finnish population. The method consists of a DNA polymerase primer extension array using fluorescent nucleotide analogs corresponding to the polymorphic nucleotide. The reaction is performed in 384-well format on microscope slides and the genotype is determined with a fluorescence scanner. A variation on this theme was presented by Jingwen Chen from Glaxo Wellcome, Inc. In this case, allele-specific oligonucleotides, generated by a fluorescent or non-fluorescent single-nucleotide extension, are annealed to a complementary oligonucleotide linked to a differently fluorescent microsphere. The technique can be multiplexed by using a different fluorescent microsphere for each SNP. Flow cytometric analysis discriminates which SNP is being analyzed, according to the microsphere fluorescence, and which of the two alleles (fluorescent or non-fluorescent) is present in each sample. Mass spectrometry is frequently used for genotyping SNPs and has the advantage of being highly accurate. Ivo Gut (Centre National de Génotypage, France) presented a cheaper variant of this method that uses small reaction volumes. Because SNPs are the most frequent polymorphisms, they are being used in association studies of genetic

diseases. Nevertheless, a large proportion of false-positive associations is often observed. For example, Joel Hirschhorn (Whitehead Institute/MIT) could not replicate the results in 17 out of 18 SNPs previously associated with type 2 diabetes. It is encouraging, though, that by using new SNPs more associations could be identified.

Far from being over, genomic sequencing will go on - perhaps even more aggressively - during the 'post-genome' era. Probably one of the most auspicious perspectives for the near future in this regard was suggested by Collins. In his view, the human and other genome sequencing projects have generated expertise which, in aggregate, will allow the community to easily approach the simultaneous sequencing of a number of genomes, such as those of zebrafish and rat. We can expect further multicellular organisms to take the stage at coming meetings.