PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

The Human Genome Consortium paper: sequencing by collaborative mapping

ArticleInfo		
ArticleID	:	3978
ArticleDOI	:	10.1186/gb-spotlight-20010213-02
ArticleCitationID	:	spotlight-20010213-02
ArticleSequenceNumber	:	49
ArticleCategory	:	Research news
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 2001–02–13 OnlineDate : 2001–02–13
ArticleCopyright	:	BioMed Central Ltd2001
ArticleGrants	:	

ArticleContext	:	130592211

Jonathan B Weitzman
Email: jonathanweitzman@hotmail.com

In the February 15 Nature, the International Human Genome Sequencing Consortium announces the completion of the first draft of the sequence of the human genome. The publication is the achievement of the decade-long Human Genome Project based on open international collaboration (involving 20 groups) and rapid, unrestricted data release (via GenBank). The draft sequence covers 94% of the genome, which is estimated to be about 3.2 Gigabases (Gb), 25 times the size of any previously sequenced genome and 8 times the sum of all sequenced genomes. The 61-page report by Lander *et al.* (*Nature* 2001, **409:**860-921) details the sequencing strategy and analysis of the genomic landscape, repetitive structures, gene content and comparisons with other sequenced genomes.

The Human Genome Project was based on a 'hierarchical shotgun sequencing' approach involving the assembly of large-scale physical maps, followed by coordinated systematic sequencing of selected clones. Bacterial and P1-derived artificial chromosomes (BACs and PACs) were selected from eight large-insert libraries from anonymous human donors, and used to generate a genome-wide physical map (*Nature* 2001, **409:**934-941). Automated sequencing reached levels of over 1,000 nucleotides per second, 7 days a week. A total of 29,298 overlapping large-insert clones were assembled into 1,246 fingerprint clone contigs representing 7.5-fold genome coverage. Assembly of the sequence was achieved using PHRAP software assigning 'quality scores'. The total length of sequence amounts to 2.693 Gb and provides a treasure-trove for detailed genome analysis. Key findings include:

-The genome-wide average GC content is 41%, but regional variations confirm a correlation between GC content and gene density. There are 28,890 CpG islands in non-repetitive regions (an average 10.5 per Mb).

- Recombination rates increase as the length of the chromosome arm decreases, and higher rates are observed in distal chromosomal regions. This may reflect differences in chromosome accessibility during meiosis.

-Repeat sequences account for over half of the human genome and provide a "rich palaeontological record". The most common (45%) are transposon-derived repeats and as many as 47 human genes may be derived from transposons. Yet there is no evidence for DNA transposon activity in the human genome in the past 50 million years, demonstrating the slow rate of clearance of nonfunctional sequences.

- Analysis of noncoding RNAs reveals the unexpected observation that we have less tRNA genes than worms.

- The integrated gene index (IGI) provides a cautious estimate of the total gene number. There are 31,778 estimated proteins, with 14,882 from known genes and the rest from gene predictions. Corrected calculations estimate as few as 24,500 true genes, twice as many as nematode worms and fruitflies.

- The average coding sequences of human genes (1,340 bp) are similar to worms and flies, although they are spread out over larger regions (due to larger introns, on average 3.3 kb). About a third of the human genome is transcribed, with only 1.5% representing coding sequences.

- The integrated protein index (IPI) shows significant similarity with fly, worm and yeast proteomes, particularly for proteins involved in metabolism, DNA replication, transcription and translations, protein folding and degradation.

- Complexity of the human proteome appears to stem from large-scale protein innovation and extensive alternative splicing.

- Hundreds of genes resemble bacterial proteins not found in other eukaryotes, suggesting extensive horizontal transfer from prokaryotes. These may play roles in xenobiotic metabolism and stress responses.

- About 7% of protein families are 'vertebrate-specific', including proteins implicated in immune defence and the nervous system.

- Many gene families have undergone expansion in humans (e.g. 111 keratin genes and thousands of olfactory receptor genes).

Future work will focus on 'finishing' the draft. An immediate challenge is filling in the 145,514 sequencing gaps that currently account for about 80 Mb. Further cross-species comparisons (particularly with the mouse and pufferfish genomes) are likely to be very informative. The authors conclude "that the more we learn from the human genome, the more there is to explore."

References

1. *Nature*, [http://www.nature.com/]

2. Human Genome Project, [http://www.ornl.gov/hgmis]

3. International Human Genome Sequencing Consortium, Genome Hub , [http://www.nhgri.nih.gov/genome_hub.html]

4. GenBank, [http://www.ncbi.nlm.nih.gov/Genbank]

5. The Use of Human Subjects in Large-Scale DNA Sequencing, [http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/human_subjects.html]

6. Documentation for PHRAP, [http://bozeman.mbt.washington.edu/phrap.docs/phrap.html]

7. *Nature* genome gateway - human genome, [http://www.nature.com/genomics/human]

8. Genetic Information Research Institute, [http://www.girinst.org/~server/repbase.html]

9. The gene guessing game

10. Ensembl, [http://www.ensembl.org]

11. *Nature* genome gateway - papers - *C. elegans*, [http://www.nature.com/genomics/papers/c_elegans.html]

12. *Nature* genome gateway - papers - *D. melanogaster*, [http://www.nature.com/genomics/papers/drosophila.html]

13. Mouse genome informatics, [http://www.informatics.jax.org]

14. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence.