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

The Celera paper: sequencing by random shotgun cloning

ArticleInfo		
ArticleID	:	3977
ArticleDOI	:	10.1186/gb-spotlight-20010213-01
ArticleCitationID	:	spotlight-20010213-01
ArticleSequenceNumber	:	48
ArticleCategory	:	Research news
ArticleFirstPage	:	1
ArticleLastPage	:	2
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 16 Science, Venter *et al*. announce the sequencing of the euchromatic portion of the human genome by a whole-genome shotgun sequencing approach (*Science* 2001, **291:**1304-1350). The sequencing achievement was accomplished by Celera Genomics in nine months in a factory-scale project involving 300 automatic squencing machines (ABI PRISM 3700) producing 175,000 sequence-reads per day. The company generated 14.8 gigabases (Gb) of DNA sequence and combined data with the public GenBank) database to generate a 2.91 Gb consensus sequence (94% coverage) representing over eight-fold coverage of the genome.

Venter *et al*. constructed a series of high-quality plasmid libraries (with 2, 10 and 50 kilobase inserts) from five individual DNA donors of diverse ethnic origin. Data from paired sequence reads (averaging 543 basepairs (bp)) were pooled with GenBank data that had been shredded into 550 bp fragments. The authors used two different assembly strategies: whole-genome assembly (WGA) or a compartmentalized shotgun assembly process (CSA). The authors estimate that their final sequence includes 240 Mb that are not represented in the public database. The total 2.9 Gb consisted of 107,227 contigs, with an equal number of gaps (most of which are estimated at less than a kilobase). The achievement demonstrates the power of the shotgun approach and provides a mountain of data for analysis. Key points include:

-The authors developed an integrated evidence-based approach (called Otto) to predict and annotate genes. They provide a conservative estimate of 26,383 genes, and used less stringent predictions programs to extend this number to an upper limit of 39,114. This is approximately twice the number of genes in fruitflies.

-Over 40% of genes cannot be assigned molecular function. The genes that differ from other sequenced organisms include those implicated in acquired immune defence, neural development, cellular signalling, homeostasis and apoptosis. Combinatorial diversity is likely to contribute to human genetic complexity.

-Gene density correlated with high GC content, and CpG islands are enriched (from 2.5% to 40%) near gene start sites.

-Gene-poor 'deserts' account for about 20% of the genome. Gene-rich chromosomes - 17, 19 and 22 - have fewer deserts (about 12%), while gene-poor chromosomes - 4, 13, 18 and X - have more (27.5%).

-Recombination rates differ more between the body of the chromosome and the telomeres (4.99 difference) than between males and females (4.44).

-About 40% of the genome is made up of repeat sequences, whose frequency may correlate with gene density.

-There are 1,077 duplicated blocks covering 3,522 distinct genes. Retrotransposition (generating intronless paralogs and processed pseudogenes) affects many ribosomal proteins and elongation factors. There is also evidence for large duplications involving gene clusters.

-Analysis of small nucleotide polymorphisms (SNPs) revealed a rate of 1 per 1200 bp, very few of which (<1%) affect protein function. The authors mapped over 2 million SNPs onto their genome sequence, which varied in distribution density and correlated with GC content.

The potential of the high-throughput shotgun sequencing approach has been impressively validated - although Celera did also rely on the publicly available data. It remains to be seen how Celera will succeed in providing access to the scientific community and in its pledge to tackle proteomics.

References

1. *Science*, [http://www.sciencemag.org]

2. Shotgun sequencing of the human genome.

3. Celera Genomics , [http://www.celera.com]

4. ABI PRISM 3700 DNA Analyzer, [http://www.appliedbiosystems.com/ga/3700/]

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

6. *Science* human genome special issue, [http://www.sciencemag.org/genome2001]

7. A whole-genome assembly of Drosophila.