European Journal of Human Genetics (2004) 12, 875–876
© 2004 Nature Publishing Group All rights reserved 1018-4813/04 $30.00
www.nature.com/ejhg

NEWS AND COMMENTARY

Deep genomics in shallow times: the finished sequence of human chromosomes 13 and 19

European Journal of Human Genetics (2004) **12,** 875–876. doi:10.1038/sj.ejhg.5201254

In the time between the draft human genome publication in 2001 and the present total of 11 finished chromosomes,[1] we have entered the era of shallow genome sequencing. Although the human and mouse genome projects sequenced each base at least seven times (7 × coverage), Craig Venter's poodle warranted only 1.5 × coverage,[2] and now the US National Human Genome Research Institute is considering proposals to sequence the genomes of perhaps a dozen mammals at 2 × coverage.[3] These data are intended for comparative genomics, to shed light on novel functional regions conserved across mammals, and would not be a good basis for the subsequent laboratory work required to validate/refute the predicted functions. The laboratory work to examine the functional variants of a novel gene or regulatory element requires a high-quality sequence in long, contiguous stretches. The ongoing search across the human genome for the small and often rare mutations involved in disease also depends on uninterrupted stretches of sequence. The finished sequences of chromosome 13 and 19 provide the deep coverage sequence required for these endeavors.

In 2001, the International Human Genome Sequencing Consortium (IHGSC) announced the initial sequencing of the human genome,[4] and casual observers might have been forgiven for assuming that the sequences of all human chromosomes were now known. In fact, the assembled draft human genome sequence was estimated to cover 90% of the gene-containing euchromatic regions (though only 28% was sequenced to an accuracy of 99.99%), and included around 150 000 gaps. Since then, the IHGSC has undertaken the arduous task of 'finishing': producing a genome sequence covering 99% of the euchromatic regions sequenced to an accuracy of 99.99%. On the 14th of April 2003, the IHGSC announced that this target had been reached, leaving less than 400 persistent gaps where highly repetitive sequences evaded current sequencing technology. A steady trickle of Nature papers has marked the emergence of each finished chromosome sequence, along with the annotation describing its notable features. Two recent papers chart the functional topography of chromosome 13,[5] the most gene-poor autosome, and chromosome 19,[6] which has the highest gene density of all human chromosomes. Together, they demonstrate the insights and novel analyses made possible by finished chromosome sequences.

Chromosome 13 is the largest human acrocentric chromosome and, in common with the other acrocentric chromosomes, its short arm is composed of heterochromatin rich in repeated sequences, including ribosomal RNA gene arrays. The long arm is euchromatic, contains most or all of the protein-coding genes of this chromosome and was the focus for the sequencing project. In total, 95.5 Mb of chromosome 13 was sequenced by The Wellcome Trust Sanger Institute, stretching from the repetitive, pericentromeric region at one end to the beginning of the telomeric repeats at the other.

The finished chromosome 13 sequence formed the basis for global analyses of the distribution of genes, repeats, GC content, single-nucleotide polymorphisms, and recombination events.[5] The most important aspect of the sequence annotation is the delineation of protein-coding gene structures. This is done by manual curation of the outputs from a battery of computational analyses: gene prediction algorithms and sequence similarity searches. This comprehensive and high-quality annotation is all freely accessible from the VEGA database.[7] This database is the destination for all manually curated annotation produced by The Human and Vertebrate Annotation (HAVANA) group annotators at The Wellcome Trust Sanger Institute.

The sequence was found to contain 633 protein-coding genes and 296 pseudogenes; it was estimated that more than 95.4% of the protein-coding genes present on chromosome 13 are within this set, on the basis of comparison with other vertebrate genome sequences. In addition, 105 putative noncoding RNA genes were found. These are remarkably low numbers for such a large volume of sequence data and show that chromosome 13 contains fewer genes than on chromosome 22, which is less than half as long. Overall, chromosome 13 has the lowest gene density (6.5 genes per Mb) among sequenced human autosomes, and contains a central 'gene desert' region of 38 Mb, where the gene density drops to only 3.1 genes per Mb. (The genome average is around 10 genes per Mb.) In contrast, the most gene-rich regions are at either end of the long arm of this chromosome.

Comparisons to other sequenced genomes show that even within the central gene desert there are numerous regions conserved, over half a billion years, between human and fish. These intriguing regions, believed to have structural or regulatory functions, will form the basis for some interesting experimental biology. Future work will also focus on the known disease susceptibility regions that occupy chromosome 13. There are 48 human conditions linked to genes on chromosome 13 in the Online Mendelian Inheritance in Man (OMIM) database,[8] and the biomedical benefits of the chromosome 13 sequence are already becoming evident, with an early phase of the sequencing leading to characterization of the breast cancer type-2 gene BRCA2.

The 55.8 Mb of human chromosome 19, the most gene-rich of all human chromosomes, was completed by The US Joint Genome Institute and Stanford University.[6] Although representing only about 2% of the human genome chromosome 19 contains 1461 protein-coding genes, 321 pseudogenes, and 14 RNA genes, leading to a protein-coding gene density of 26 per Mb – more than double the genome-wide average. As with the chromosome 13 publication, the authors have produced high-quality, manually curated annotation that is freely available via the JGI website.[9] (It is worth noting that the main sites for browsing human genome annotation also maintain an ever-expanding description of our chromosomes' landscapes: Ensembl,[10] UCSC Human Genome Browser,[11] and NCBI Map Viewer.[12]) Gene density is not the only unusual thing about chromosome 19, it also contains an unusually high density of repeat sequences: nearly 55% of the chromosome as compared to a genome average of 45%.

The landscape of chromosome 19 points to a dynamic evolutionary history, involving high levels of duplication and rapid evolutionary change. Nearly a quarter of the genes found belong to large tandemly arranged gene families that traverse more than 25% of the chromosome length. Included among these families are several of great medical significance such as those with immunoglobulin-like domains, which function in the immune response to infection, and the rapidly evolving cytochrome *P*450 subfamily II genes involved in the metabolism of steroid hormones, carcinogens, and other substances. As with chromosome 13, many segments of noncoding conservation were found with species as distant from us as fish. Of particular note was the 5 Mb closest to the centromere on the long arm of chromosome 19, which contains the highest density of noncoding conservation with both mouse and fish. This relatively gene-poor region would therefore seem to be rich in regulatory potential.

These preliminary investigations of chromosomes 13 and 19 raise more questions than they answer, but they also demonstrate that finished sequence can be a rich source of clues to the many mysteries of human biology and evolution. The essentially complete version of the human euchromatic genome sequence, thought by many to be impossible at the outset of the Human Genome Project, marks a remarkable achievement in the history of science. The end products of this project, the finished chromosome sequences, signal a new beginning in 21st century biomedical research.

Colin AM Semple is at the MRC Human Genetics Unit,
Western General Hospital, Edinburgh EH4 2XU.
E-mail: colins@hgu.mrc.ac.uk

1 Nature guide to the human genome, http://www.nature.com/nature/focus/humangenome/.
2 Kirkness EF *et al*: *Science* 2003; **301**: 1898–1903.
3 Pennisi E: *Science* 2004; **304**: 1227.
4 Lander ES *et al*: *Nature* 2001; **409**: 860–921.
5 Dunham A *et al*: *Nature* 2004; **428**: 522–528.
6 Grimwood J *et al*: *Nature* 2004; **428**: 529–535.
7 The Vertebrate Genome Annotation (VEGA) database, http://vega.sanger.ac.uk.
8 The online Mendelian Inheritance in Man (OMIM) database, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = OMIM.
9 Joint Genome Institute Human Chromosome 19, http://genome.jgi-psf.org/Chr19ncbi34/Chr19ncbi34.home.html.
10 Ensembl Genome Browser, http://www.ensembl.org/.
11 University of California at Santa Cruz (UCSC) Human Genome Browser, http://genome.ucsc.edu/cgi-bin/hgGateway.
12 NCBI Map Viewer, http://www.ncbi.nlm.nih.gov/mapview/.