J. Appl. Genet. 42(4), 2001, pp. 399-403

Invited Editorial

Human Genome Sequenced: achievements & shortcomings in big science

Nikolaus BLIN

Division of Molecular Genetics, University of Tübingen, Germany

With broad public attention (including a White House, Washington, D.C. press conference) the release of the almost complete human genome sequence was celebrated as a milestone in a key area of present scientific activities. In a multicentered and international program, equipped with substantial budgets, academic institutions laid down the basis for this "big science" approach about 10 years ago. The goal was to complete the human sequence by 2003-2005. Due to improved technology and to increasing commercial competition the project was finished ahead of time with the announcement of the databases in February 2001. But despite all the media hype and hefty arguments about gene patenting, the present sequence collections should be considered "work in progress" with many more details needed to be filled in. Interesting and surprising facts derived from the data and also manyfold gaps and problems of interpretation are discussed with examples presented below.

In a fulminant race academia vs. commerce the human genome was announced to have been sequenced in February this year, with both parties claiming victory (www.nature.com/genomics/human; www.scienceonline.org/feature/data/genomes). HUGO, the academic-based genome project achieved its goal by first creating a physical map of the genome. A series of overlapping fragments of about 100-200 kilobases were generated, fingerprinted and mapped, thus covering the entire genome. Each fragment was sequenced, then the sequence was overlaid onto the map scaffold and merged to reassemble the human genome.

Received: August 8, 2001.
Correspondence: N. BLIN, Univ. Tübingen, Inst. of Anthropology & Human Genetics, Div. of Molecular Genetics, Wilhelmstr 27, 72074 Tübingen, Germany, email: blin@uni-tuebingen.de

Due to an immense accumulation of primary data, major outlines of information were presented in print, all available details can be found in electronic data collections such as http://genome.cse.ucsc.edu; www.nhgri.nih.gov; www.sanger.ac.uk/HGP).

Indeed, the findings are astounding and some of them came rather unexpectedly. Moreover, next to disclosing many details within the human genome, the sequence data allow comparisons to be made to other sequenced species, thus elucidating evolutionary trees (LIU et al. 2001), the development of karyotypes and the interrelation of genomic structure and function.

Most human genetic variations occur as different nucleotides at single base positions – called single nucleotide polymorphisms, or SNPs. The latest map of nucleotide diversity across the human genome catalogues 1.42 million SNPs across the genome. On average, there is one SNP every 1.9 kilobases. Nucleotide diversity varies greatly across the genome, and the pattern of diversity varies for different populations.

More than half of the euchromatic genome is comprised of repeat sequences, with the vast majority (45%) accounted for by repeats derived from parasitic DNA, called 'transposable elements' or 'transposons'. The elements propagate by replicating themselves at one site in the genome and then by inserting the copy into another site. This degree of transposition came as a surprise since it is unprecedented in any other sequenced genome, compared with that of the fly (*Drosophila*) and the worm (*Caenorhabditis*). Duplications also appear to have had a significant role in genome evolution, with roughly 5% of the sequence arising from duplications of large blocks (of more than 10 kilobases) within and between chromosomes. Again, this finding represents a much more prevalent feature in man than in the fly, the worm or yeast (*Saccharomyces*). Duplications enable one copy of a gene to relocate to a new site, where it may take on a distinct physiological function. Highly homologous duplicated regions are likely to have contributed greatly to the expansion of gene families in humans, as can be aptly exemplified by the large olfactory receptor gene family, which comprises more than 1,000 members.

A rather controversial finding is the unexpectedly low number of genes, about 32,000. Here, both databases (the academic human genome project HGP and the commercial Celera) seem to agree quite well and many speculations have been presented to explain this astounding fact. However. depending on sequence data for a particular region, both versions can vary substantially. In the vicinity of anion channel gene (chromosome 3, clone RP11-219D15) Celera lists 84 genes, HGP expects 148 genes (K. JURKAT-ROTTS, pers. communication). In the meantime, the algorithms used for detecting genes among new sequences have been disputed and new numbers (up to 60,000) were generated from the official ge-

nome sequence. It seems it will take quite a while and many more functional data to settle this matter.

Achievements

- released 12.2.2001 (ahead of projected date)
- about 32.000 genes disclosed
- only 5% of the DNA represent genes
- many viral insertions
- maybe up to 200 genes derive directly from bacteria
- uneven chromosomal gene density (with many genes on e.g. chromosomes 17,19,22 and less on e.g. chromosomes 4,13,18)

Despite these achievements, a set of shortcomings of this "big science" approach needs to be discussed with examples given below.

Due to the logistics of the sequencing strategy, contigs may have remained incomplete and fragments may have been assembled in reverse order. We encountered the latter problem when mapping the *POLR2F* gene to 22q13 (PUSCH et al. 1996). Within 40 kb of the sequenced cosmid, the computer program fused exon 2 to exon 5/4/3, instead of exon 3/4/5. Our knowledge of the cDNA prevented this misalignment. In the meantime, several such problems were reported for both, HGP and Celera, databases. It has been estimated that over 100,000 gaps remain to be filled in, more than 10% of all sequences display assignment problems and that only 1/3 of all sequences reach 99.9% precision status.

Moreover, while euchromatic gene coding regions caught all the attention, heterochromatic parts of chromosomes remained quite out of focus. First of all, genes promised rich patent yields and then, repetitive territory is much more difficult to traverse. Ambiguities were promised to be resolved during the coming two years. Aptly, the present releases should be entitled "work in progress".

Probably, the hottest scientific debate arose from the released gene number, estimated at 32,000. This would not be much more than the expected 26,000 genes in *Arabidopsis*. While no linear increase in gene number in connection with the growing complexity of organisms should be expected, a valid question remained whether all genes were unambiguously detected within the human genome. A series of genetic features can disclose an easy register: sequence overlaps, genes within genes, interrupted genes, variable promoter use, alterations at polyA- and splice sites, editing. Therefore, an alternative application of additional algorithms allowed other investigators to come up with quite divergent gene

estimates, up to 60,000, well in line with many previous calculations. It presently becomes obvious that, no matter what the final count will be, human genetics will not continue to present to us simple, monocausal problems. In the long run, we will have to consider multiple facts such as population polymorphism, transcriptional and translational regulation, gene cascades, feedback loops and epigenetic effects, all playing a role in influencing hereditary traits, which will be multigenic or even multifactorial

The final – also debated and by no means resolved – problem touches the field of gene patenting. While proponents argue that financing the expensive genome research would become impossible without commercial support, which obviously depends on commercial rewards in form of gene patents, this has not been proven beyond doubt. Patents can also prevent access to molecular data and surely diminish free and open information exchange. Thus, for many years, HUGO opted for an open sequencing data policy (HUGO 1995). In their effort to secure many potential candidates for the gene market, Celera patented about 6,500 genes while publishing their genome data share (http://public.celera.com/genomics).

Shortcomings

- 5% of all sequences undisclosed
- about 100.000 gaps remaining
- possibly >10% wrongly assigned or oriented sequence fragments
- only 30% of all sequences with 99.9% precision
- correction of remaining ambiguities expected until year 2003
- 6500 genes patented (not freely available)
- gene number broadly disputed

While scientists, free-market-advocates, patent lawyers and venture capitalists will still debate on this issue for the coming years, one question remains: what does the general public think of the entire genomic business, what is the layman's awareness including expectations and fears? A Europe-wide poll (16,000 individuals in 16 EU countries) evaluated this question (GASKELL et al. 2000). While the majority supported gene technology in areas of medication, gene food was not accepted by more than 50% of the population. Although this is not the place to analyze the poll's results, a general acceptance of medical application also seems a positive voice for the HUGO program and its biomedical goals.

Table. Disposition of European Union individuals towards gene technology

+2 to –2	Useful	Accepted	With risk
Gene tests	1.3	0.8	0.1
Genetic engineering in medication	1.2	0.7	0.3
Cloning of human cells	0.9	0.3	0.5
Animal cloning	0.1	–0.2	0.5
Gene food	–0.1	–0.2	0.7

Points on a scale +2 (positive) to –2 (negative) assigned to 5 questions relating to gene technology (adapted from Nature Biotechnology 2001, 18: 935).

Acknowledgements: Travel support by the Polish/German exchange program (BMBF/KBN pol 99/031) is acknowledged. My lab colleagues Peter GÖTT, Carsten PUSCH and Tina SCHRÖDER deserve many thanks for numerous discussions.

LITERATURE and RELEVANT URL ADDRESSES

www.nature.com/genomics/human/

www.scienceonline.org/feature/data/genomes

www.nhgri.nih.gov

http://genome.cse.ucsc.edu

www.sanger.ac.uk/HGP

http://public.celera.com/index.cfm

www.gene.ucl.ac.uk/hugo

www.ebi.ac.uk

GASKELL et al. (2000). Biotechnology and the European public. Nat. Biotech. 18: 935-938.

www.hugo-international.org/hugo/statements.html

LIU et al. (2001). Molecular and morphological supertrees for Eutherian mammals. Science 291:1786-1789.

Nature (409(6822):813-958) February 15, 2001.

PUSCH C., WANG Z., ROE B., BLIN N. (1996). Genomic structure of RNA polymerase II small subunit (hRPB14.4.) locus (POLR2F) and mapping to 22q13.1 by sequence identity. Genomics 34: 440-442.

Science (291(5907):1145-1434) February 16, 2001.