Genetics

The Human Genome Project: the next decade

R M Gardiner

Towards a molecular understanding of common childhood diseases

A draft version of the complete human genome sequence was published early in 2001. This was the culmination of both public and privately funded efforts initiated a decade ago. The new landscape of the genome contained several surprises, including the relatively small number of genes, 30–40 000, required to make a human. Attention has now shifted towards annotating the genome by assigning function to all the genes, and characterising human genetic variation manifested as single nucleotide polymorphisms (SNPs). The latter should allow the genetic basis of common disorders with “complex” inheritance to be elucidated.

Ten years ago I wrote a review for this journal entitled “The human genome: a prospect for paediatrics”.[1] In doing so a well known Goldwynism was ignored: “Never make predictions, especially about the future”. From today’s perspective the predictions in that article seem, however, rather cautious. The major goals of the Human Genome Project (HGP) which had just been initiated have been attained ahead of schedule and the molecular genetic analysis of rare human diseases continues to generate new biological insights of extraordinary depth. Yet for the paediatrician dealing with the daily round of childhood diseases the impact remains negligible, and it is true that our present knowledge of the human genome resembles a gigantic “parts” list in which at least half of the items have a catalogue number but no assigned function. So how has this project progressed so fast, where do we stand now, and what of the next one or two decades?

THE HUMAN GENOME PROJECT (HGP): THE PAST 10 YEARS

The HGP was launched in September 1990 with a projected completion date of 2005. The idea that sequencing the entire human genome might be a worthwhile endeavour arose in the mid 1980s, about a decade after Frederick Sanger and others introduced methods for sequencing DNA. This proposal sparked a fierce debate. Critics argued that it would be a mindless factory project siphoning research funds away from hypothesis driven research, that most of the sequence was “junk” of little biological interest, and that the sheer size of the human genome precluded its completion within a reasonable time frame without entirely new methodology.

The project was launched despite opposition and most of these fears have proved unfounded. Although completion of the sequence remained the ultimate goal, the project always encompassed wider aims and the creation of genetic and physical maps represented an essential preliminary to large scale sequencing. Most importantly it also included the sequencing of several model organisms and funding for the development of bioinformatics and research on the ethical, legal, and social implications of the project.

Genomes of increasing size and complexity were completed in rapid succession. The first genome of a living organism, *Haemophilus influenzae*, was sequenced in 1995, and those of baker’s yeast (*Saccharomyces cerevisiae*), the round worm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), and mustard weed (*Arabidopsis thaliana*) followed in rapid succession between 1996 and 2000. The last three of these have genomes of around 100 million base pairs, roughly equivalent to that of a medium sized human chromosome and just one thirtieth of the size of the entire human genome. Sequences of the two smallest human chromosomes were published in late 1999 and of course the complete “draft” of the human genome early in 2001. Competition between private and public sequencing efforts generated acrimony but spurred progress, and was a factor in the sequence being completed ahead of schedule.

THE LANDSCAPE OF THE HUMAN GENOME

The papers published simultaneously in two leading scientific journals[2,3] represent a milestone in biology and provided some fascinating new insights into the genetic blueprint of man. In this brief review, there is space to consider three of these: the architecture of the genome, the total number of human genes and their functions, and the comparisons which are now possible between human genes and those of other fully sequenced organisms.

Genome architecture: freeloaders and fossils

It was already known that protein coding sequences accounted for just 1.5% of the genome and that introns accounted for a further 24%. It is now clear that an unprecedented proportion, more than half, of the human eukaryotic genome consists of repeat sequences, the majority of which are so called transposable elements or transposons. It seems likely that most of these repeats are simply parasitic, selfish DNA elements, “freeloaders” that use the genome as a convenient host. Moreover, in humans most of these parasitic DNA repeats are very ancient and enfeebled. By contrast, the mouse genome has younger, actively reinserting sequences but they comprise a much smaller fraction of the genome. Is the human genome just lackadaisical about cleaning out these relics or do they serve some useful purpose? There is evidence that repeat sequences may have some positive effects, for example, in shaping the evolution of the genome and in creating new genes.

Chromosomal architecture: variation in gene density and transcriptional activity

It is now possible to generate a provisional “human transcriptome map” which reveals the gene expression profiles for any chromosomal region in various tissue types. This shows a striking tendency of highly expressed genes to cluster in specific chromosomal regions of high gene density. It is also apparent that whole chromosomes may differ in these respects. Chromosome 19 is packed with genes at an average of 23 per megabase, whereas chromosome 13 is gene poor at just five genes per megabase. Interestingly the three chromosomes responsible for most constitutional trisomies, 13, 18, and 21 all show low gene density and low gene expression. Presumably this accounts for the non-lethal effects of an extra copy of these chromosomes.

Gene number: not so many

The total number of human genes remains uncertain, but has been revised downwards from upper estimates of 150 000 to 30–40 000. Should this be regarded as a blow to our speciocentric

Abbreviations: HGP, human genome project; LD, linkage disequilibrium; SNP, single nucleotide polymorphism view of the biological world? This number of coding genes compares with 6000 for yeast, 13 000 for the fruit fly, 18 000 for the round worm, and 26 000 for a humble plant, the mustard weed.

It seems likely that the number will be revised upwards. Finding human genes is a difficult task, even with sequence in hand. Methods for gene prediction depend on looking for signatures of gene structure such as open reading frames, homologies to sequences of human genes, and evidence that a DNA sequence is expressed as messenger RNA. Long introns and rare transcripts make some genes difficult to detect and these may comprise the so called "dark matter" of undiscovered genes.

However, it is clear that the relation between gene number and biological complexity is not linear and the n value paradox may be more apparent than real. For example, taking a trivial mathematical model of biological complexity in which complexity is defined as the number of possible transcriptome states and a gene is either ON or OFF, a genome with n genes can encode 2^n states. On this basis, an extra 10 000 genes provides $2^{10\,000}$ extra states, a vast number, which certainly allows human beings to consider themselves superior to worms.

Gene complement and structure

It is also now possible to compare the nature and structure of human genes with those of other organisms with sequenced genomes. It is clear that most of our genes come from the distant evolutionary past. Genes involved in basic cellular functions, such as DNA replication and transcription, have evolved only once and remained fixed. Only about 10% of the protein families in our genome are specific to vertebrates. Human proteins are built from more domains and new combinations of domains, new architectures using old bricks. However, certain gene families do appear to have expanded in vertebrates. Some subserve vertebrate specific functions such as blood clotting or the acquired immune response. Others provide increased general capabilities such as genes for signalling, apoptosis, or control of gene transcription. The human genome is particularly rich in Zinc finger genes.

It is also clear that individual human genes encode a much wider repertoire of proteins, on average three or four, by mechanisms such as alternative splicing.

FUTURE PROSPECTS

If we fast forward to 2010, or even 2020, what changes can we expect in biology and medicine of special relevance to paediatricians? In the field of genomics the immediate challenges include completion of the human genome sequence, annotation of the genome, and characterisation of the pattern and extent of human genetic variation. At the present time the genomes of about three dozen organisms, most of those single cell microbes, have been completely sequenced. Model organisms next in line for complete sequencing include the mouse and zebrafish, and by 2020 it is anticipated that 1000 complete genomes will be in hand. In parallel however, interest is inevitably shifting already from genomics to proteomics.

In paediatrics the dissection of the molecular basis of rare Mendelian and chromosomal disorders will continue apace. More important perhaps, for the general paediatrician, is the prospect of understanding common early onset disorders with "complex" inheritance. These include asthma, type 1 diabetes mellitus, and the epilepsies, but also surgical abnormalities such as cleft lip and palate and pyloric stenosis, and the behavioural phenotypes of autism and attention deficit hyperactivity disorder. Of course, the leap from understanding to effective intervention is even more difficult, but at least a start will have been made. A selection of these themes are considered in more detail.

The human genome: annotation and characterising variation

So called annotation of the human genome remains a huge task. Not only must all the genes be identified, but functions must be assigned to the proteins they encode. As the final sequence is assembled the computer programs used to predict the presence of a gene will come closer to identifying a complete inventory, but their inherent limitations render this a difficult task as positional cloners searching through large genomic regions harbouring a disease gene already know.

Preliminary analysis of the predicted human protein coding genes has allowed functions to be tentatively assigned to about 50% of the putative gene products. Of 26 588 predicted human proteins, the most common molecular functions are transcription factors (1850, 6%) and nucleic acid enzymes (2308, 7.5%). Other highly represented functions include receptors, kinases, and hydrolases. There are 406 ion channels and 533 transporters. But there are 12 809 predicted proteins of unknown function. Some of these may represent false positive gene predictions but the rest remind us of how much remains to be discovered about the basic biology of man.

The second great task is the detailed characterisation of human genetic variation. On average, the genomes of two human individuals are 99.9% identical. The 0.1% which differs is what makes us individuals rather than clones. Most of the variation is represented by alterations at single nucleotides. A single nucleotide polymorphism, or SNP (pronounced SNiP), is defined as a single base pair in genomic DNA at which different alleles (bases) exist in normal individuals in some populations, with the minor allele frequency greater than 1%. The SNPs are likely to include the allelic variation that accounts for common disease traits (see below).

A massive effort is now in progress to characterise human genetic variation and create a SNP map of the human genome. A particular SNP may or may not influence the phenotype, depending on its nature and location. For example, a SNP in a coding region (cSNP) may alter an amino acid and a SNP in a regulatory region may alter gene expression. Most SNPs are however in introns or intergenic regions and are assumed to be "neutral" in evolutionary terms. Identified SNPs are then used to create "haplotype" maps which reflect the phenomenon of linkage disequilibrium (LD). LD is the non-random occurrence of specific alleles at adjacent loci. When a base change first occurs it does so on a particular chromosome with particular haplotypes—pattern of alleles at adjacent SNPs. Over time, meiotic recombination and other factors erode the haplotypes until linkage equilibrium, or random association is established.

The extent and pattern of LD within the human genome and across human genes is just beginning to emerge.[4] Empirical data indicate that extensive blocks of LD are present, at least in the Northern European population, which create haplotypes between 25 and 100 Kb in length. This is good news for the analysis of "complex" disease (see below) and probably reflects a recent bottleneck in human evolutionary history. Moreover, it appears likely that it will be possible to identify a small number of SNPs, perhaps half a dozen, in the average human gene, which will allow the main haplotypes of that gene present in a given population to be determined.

Model organisms: mouse and zebrafish

At the top of the list of model organism genomes to be sequenced comes the humble mouse. This small, furry creature shared a last common ancestor with humans about 100 million years ago. Its genome is similar in size to the human and the gene complement is similar. A host of human disease genes have orthologues in mouse, and identification of genes causing many of the many naturally occurring mutant phenotypes known has often helped in isolation of the corresponding human disease gene. Many extended chromosomal regions have maintained the same genes in the same order, so called conserved synteny.

Two new developments will build on the mouse genome sequence. Mutagenesis screens are underway to mutate many more mouse genes as a powerful strategy for assigning function. Secondly, it is likely that mouse models will prove extremely useful for unravelling the causes of disorders with “complex” inheritance.

Next in line is a small tropical fish, the zebrafish *Danio reria* which separated from humans 400 million years ago and promises to serve as the “canonical” vertebrate, especially for the investigation of development. It is the first vertebrate to prove tractable to large scale genetic screening of the kind used so successfully in fruit flies and worms. Developmental phenotypes are readily observed, thanks to its external development and transparency. Developmental programmes are highly conserved among vertebrates, and mutations in orthologous zebrafish genes have already provided models for human genetic disorders such as porphyria and Usher 1B syndrome. Like the mouse, large regions of human and zebrafish chromosomes show conserved synteny.

Analysis of disorders with “complex” inheritance

A number of common, important childhood onset diseases display familial clustering which is best explained by so called multifactorial inheritance: an interplay between several genes and environmental factors. These include type 1 diabetes mellitus, asthma, inflammatory bowel disease, the epilepsies, obesity, and behavioural disorders such as attention deficit hyperactivity disorder and autism.

In the past decade the spectacular successes seen in isolation of the genes for Mendelian diseases has not been matched by success in identification of susceptibility genes for disorders with such so called “complex inheritance”. In fact numerous genome wide linkage studies have frustratingly failed to find clear and replicable evidence for the location of the genes responsible for these traits. In any event, where good evidence for linkage has been found, the chromosomal region implicated has been extremely large.

Optimists believe this is about to change following the generation of so called SNP maps which can be used to find susceptibility loci by means of association rather than linkage. In reality success will depend on certain features of both the nature of human genetic variation and the genetic architecture of these common diseases.[5] As described above, recent data indicate that “blocks” of LD in the human genome are larger than some theoretical considerations had predicted. This means that genome wide searches can use fewer SNPs, although it creates a potential difficulty that if several sequence variants are inherited together (that is, are in LD) it will be less easy to spot the causal variant. Moreover, it appears that common variants of most genes can be characterised by a small number of SNP haplotypes, facilitating candidate gene association studies.

Of equal importance however, is the true genetic architecture of “complex” traits. A spectrum of possibilities exists concerning the number of loci, the magnitude of their individual effect on risk, their mode of interaction, and the number and population frequency of disease susceptibility alleles. The latter is particularly important. The so called “common disease–common variant” hypothesis suggests that each locus will harbour only a few susceptibility alleles, each at high frequency (for example, >5–10%) in the population. Recent calculations suggest that this will be the case,[6] but if it is not, association studies are doomed to failure. A large number of rare (<1%) alleles would be very difficult to detect using association analysis.

A recent application of the strategy of linkage analysis followed by association studies has resulted in the identification of a gene for Crohn’s disease, NOD2 in the pericentromic region of chromosome 16p.[7 8] If the underlying biology is favourable, this could be the first of many susceptibility genes for common diseases isolated during the next decade.

CONCLUSION

In the 1990 article I concluded that “the human genome today is as much a dark continent as Africa in the early nineteenth century. Its exploration is about to begin”. Phase 1 of that exploration has now been completed, but much of the excitement and exploitation remains in the future.

Arch Dis Child 2002;**86**:389–391

....

Author’s affiliation

R M Gardiner, Department of Paediatrics, Rayne Institute, UCL Medical School, London, UK

Correspondence to: Prof. R M Gardiner, Department of Paediatrics, Rayne Institute, UCL Medical School, 5 University Street, London WC1E 6JJ, UK; mark.gardiner@ucl.ac.uk

Accepted 14 December 2001

REFERENCES

1 **Gardiner RM**. The human genome: a prospect for paediatrics. *Arch Dis Child* 1990;**65**:457–61.

2 **The human genome**. *Nature* 2001;**409**:813–958.

3 **The human genome**. *Science* 2001;**291**:1145–1434.

4 **Daly MJ**, Rioux JD, Schaffner SF, *et al.* High-resolution haplotypes structure in the human genome. *Nat Genet* 2001;**29**:229–32.

5 **Wright AF**, Hastie ND. Complex genetic diseases: controversy over the Croesus code. *Genome Biology* 2001;**2**:2007.1–2007.8.

6 **Reich DE**, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001;**17**:502–10.

7 **Ogura Y**, Bonen DK, Inohara N, *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn’s disease. *Nature* 2001;**411**:603–6

8 **Hugot J-P**, Chamaillard M, Zouali H, *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn’s disease. *Nature* 2001;**411**:599–603.