The human genome and the future of medicine

John S Mattick

ABSTRACT

- The draft human genome sequence (about 3 billion base pairs) was completed in 2001.
- Humans have fewer protein-coding genes than expected, and most of these are highly conserved among animals.
- Humans and other complex organisms produce massive amounts of non-coding RNAs, which may form another level of genetic output that controls differentiation and development.
- Aside from classical monogenic diseases and other differences caused by mutations and polymorphisms in protein-coding genes, much of the variation between individuals, including that which may affect our predispositions to common diseases, is probably due to differences in the non-coding regions of the genome (ie, the control architecture of the system).
- Within 10 years we can expect to see:
  - ➤ increased penetration of DNA diagnostic tests to assess risk of disease, to diagnose pathogens, to determine the best treatment regimens, and for individual identification;
  - ➤ a range of new pharmaceuticals as well as new gene and cell therapies to repair damage, to optimise health and to minimise future disease risk; and
  - ➤ medicine become increasingly personalised, with the knowledge of individual genetic make-up and lifestyle influences.

MJA 2003; 179: 212–216

THE COMPLETION of the draft sequence of the human genome in 2001 was one of the great milestones of science.[1,2] However, this event is important more for what it has begun rather than for what it has finished. Determining the sequence has laid the foundations for determining the complete set of proteins that are produced in the human (the "proteome"), but we do not know the function of most of these proteins. Even when a protein sequence allows a reasonably confident prediction of the biochemical action of the protein, such as a probable tyrosine kinase or serine protease, the role of these proteins in human physiology and development has not yet been determined. This will take years to sort out, and will often rely on information about equivalent proteins in other organisms, both vertebrate and invertebrate.

Yet, even this is just the start of unravelling the meaning of the information stored in our genome. We still have to ascertain how variation in these proteins, and especially the sequences that regulate their production, affect our physiological networks and our individual characteristics, including susceptibility to disease.

An explosion of genome projects

It is important to appreciate that the Human Genome Project is just the flagship of a fleet of studies to explore the molecular and genetic basis of life and its diversity, which will provide the scientific and technological "scaffolding" for understanding the human genome and human biology. An updated and comprehensive list of completed and in-progress genome projects can be obtained from the Genomes OnLine Database,[3] which also has excellent links to relevant publications and information.

The first complete genome sequence, that of *Haemophilus influenzae*, was published in 1995. By late 2002, the sequences of 125 organisms had been completed, including 110 prokaryotes and 15 eukaryotes (including mouse, puffer fish, fruit fly, mosquito, nematode, rice, mustard plant, malaria, and two yeasts). At least 580 other genome-sequencing projects (346 prokaryotes and 235 eukaryotes) are in progress, including those of the chimpanzee, cow, pig, dog, rat, chicken, toad, salmon, bee, tick, shrimp, corn, wheat, tomato, and eucalyptus. Current sequencing projects also include the genomes of many fungal, protozoan and invertebrate pathogens of humans and other mammals, including trypanosomes and helminths. Sooner rather than later, the genomes of all organisms and variants of scientific or practical interest will be sequenced. The information generated by these projects will allow us to understand the genetic programming and evolution of life in exquisite detail.

Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD.

John S Mattick, AO, PhD, FRCPA, Professor of Molecular Biology and Director.

Reprints will not be available from the author. Correspondence: Professor J S Mattick, Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072. j.mattick@imb.uq.edu.au

We have fewer protein-coding genes than expected

What we have learned so far from genome sequencing is how little we know about our genetic programming (Box 1). However, there have been some surprising observations. The first is that the number of protein-coding genes in humans is much lower than expected — it had been predicted that humans would have at least 100 000 genes (encoding different proteins), but this is not so. Humans have about 30 000 protein-coding genes, similar to other vertebrates,[1,2,4,5] although the repertoire of protein isoforms is greatly expanded by alternative splicing of pre-mRNA.[6] Nevertheless, only about 1.5% of our genome is occupied by protein-coding sequences, which raises the question of what function the rest of the genome has, and how the complex

1: Facts and figures on the human genome

- The human genome consists of two sets of chromosomes (22 autosomes plus XX for females and XY for males). The Y chromosome has few conventional genes. One of these is a key gene called *sry*, which switches embryonic development into the male pathway. It also has some genes required for spermatogenesis. The other genes required for male development are located on the autosomes.
- Each set of chromosomes contains more than three billion base pairs of DNA sequence (A, T, G and C).
- Humans have about 1 in 1000 (0.1%) base-sequence differences, a total of about six million per diploid set of chromosomes; these sequence differences underpin our individual differences. A highly variable subset of these differences is the basis of DNA “fingerprinting”. The vast majority of genetic differences occur outside of protein-coding sequences.
- Humans vary from chimpanzees by 1% in their DNA sequence, and have much the same set of proteins.
- Humans and other mammals have about 30 000 protein-coding genes, far less than expected, but also tens of thousands of non-coding RNA genes, far more than expected, whose functions are unknown.

developmental program that unfolds following fertilisation is encoded within it.

We have much the same set of proteins as mice

The second surprise is the amazing conservation of protein-coding genes between different organisms. Recognisable equivalents of 99% of the proteins in humans can be found in the mouse, and vice versa, and about 95% are held in common.[4] Many human proteins are also very similar in structure and function to (and in some cases interchangeable with) those found in invertebrates, and include proteins controlling eye development, body plan, and circadian rhythms.[7,8] Moreover, of the about six million (0.1%) DNA-sequence differences between individual humans (in their diploid genomes), only about 20 000 are in protein-coding sequences, and most of these encode silent changes, which do not affect the amino acid sequence of the protein.[2] Therefore, apart from mutations that render particular proteins non-functional (the cause of classical monogenic diseases like cystic fibrosis and thalassaemia), individual humans have much the same set of proteins, and humans and mice have a very similar set of proteins. Thus, differences in the proteins alone may not be sufficient to explain the differences between individuals and between species.

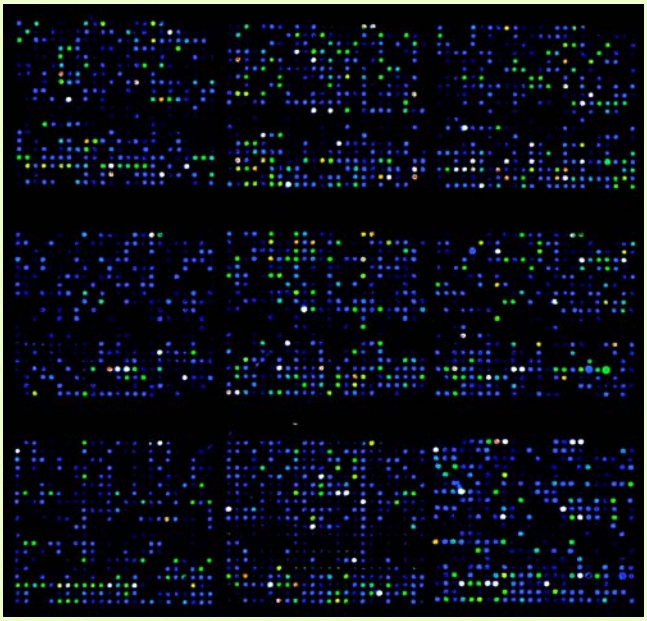
Human variation occurs in both protein coding and regulatory regions of the genome

Apart from polymorphisms in protein-coding sequences which may alter the structure and function of the protein in more subtle ways, most of what is termed “quantitative trait variation” — the small idiosyncratic differences in the way we look, and to some extent think, and which influence our physical and psychological characteristics — may be embedded in the remainder of the genome, and affect the regulatory architecture that controls the patterns and amounts of protein expression in different cells and tissues. Variation in regulatory sequences may also be a major source of our differential susceptibilities to infectious agents and to common diseases, such as diabetes, cancer, cardiovascular diseases and mental illness. Determining which variations (of which there are millions) affect these different susceptibilities and idiosyncrasies, and the interplay between each, will not be easy. In this respect, genetic epidemiology will become increasingly important, along with complementary studies in mice and other model organisms, to track down the genomic variation that underpins common diseases. Such studies will ultimately inform the development of new genetic diagnostic techniques to assess future risk, and to identify key points in physiological networks that might be targets for pharmaceutical intervention.

98% of the output of the human genome is RNA

The third and greatest surprise is the enormous amount of non-protein-coding RNA that is transcribed, and the extensive sequence conservation between the human and mouse genomes in introns (intervening sequences that interrupt protein-coding sequences) and “intergenic” regions.[4] These findings suggest that these regions have important, but not yet understood, functions. Introns comprise, on average, 95% of the primary sequence of protein-coding sequences in humans.[1,2] In addition, it appears likely that non-protein-coding RNAs represent at least half and perhaps as much as three-quarters of all transcripts in humans.[9,10] The transcriptional activity on human chromosomes 21 and 22 is at least an order of magnitude greater than expected from known protein-coding sequences,[11] and around 50% of all transcripts in the mouse do not contain substantial protein-coding sequences.[12] The expression of these non-coding RNAs is different in different cells, and a number of such RNAs have been implicated in diseases such as cancer, autism, spinocerebellar ataxia type 8, and cartilage–hair hypoplasia. Thus, we may have around 100 000 genes after all, but many, if not most, of them do not code for protein.[9] Another way of looking at this is that, although only about 1.5% of the human genome encodes protein, more than 50% of the sequences in the genome are actually expressed in a developmentally regulated manner.[10] Either the human genome is replete with useless transcription, or these RNAs are fulfilling unexpected functions. The strong possibility exists that this massive output of non-coding RNA sequences represents an internal communication mechanism to coordinate the complex suites of gene expression and programmed responses that underpin differentiation and development, and that this is a key part of the control architecture that enabled human complexity to evolve.[9,10] That is, most of our genome may be specifying the plans for the assembly, not just the components (proteins), of the system.

2: Gene expression during exposure of human endothelial cells to hypoxia



Pictured is a subset of a 20 000-gene array, with the colour temperature indicative of the level of transcripts binding to each spot. Illustration courtesy of Dr Sean Grimmond, Institute for Molecular Bioscience, University of Queensland.

What will be the impact of human genome science on medicine in 10 years' time?

Predicting the future is risky. Experience suggests that things will change faster than we expect, but usually not in the way that we expect. Some things expected to have happened quickly, such as the development of genetically engineered vaccines against malaria, have not yet eventuated, whereas other things have happened that we did not anticipate, such as the development of the polymerase chain reaction (PCR) in the late 1980s, which made DNA-based diagnostics a reality and revolutionised genetic engineering overnight.[13] Although the field of human genomics is in its infancy, there are nevertheless a number of trends emerging in terms of its impact on medicine and health.

DNA-based diagnostics

The technology of DNA-based diagnostics is already well advanced — it is now possible to place thousands (even millions) of DNA molecules on chips, where they can be read by combinations of advanced optics and electronics (Box 2).[11] This field is operating at the intersection of various sciences and technologies, including computing, advanced materials and nanotechnology, and is developing incredibly quickly. There are already many tests for genetic diseases and cancer predisposition, although these are still the province of specialised (usually hospital) laboratories rather than the doctor's surgery. The most probable route to broader implementation will be through the large pathology services, which are beginning to invest in this area as the number of validated tests expands, the databases and online services which support their interpretation improve, and patient and doctor awareness of them increases. To some extent, the rate of uptake of these tests will be affected by the availability of medical rebates, but patient demand (and/or litigation) will also drive this forward, as will the rapidly reducing cost of these tests.

Microbiological testing

Given the number of bacterial genomes that have been and are being sequenced, and the increasing trend towards partial or complete sequencing of pathogenic variants, it is likely that most microbiological testing will become DNA-based. There are problems to be faced, such as infections with mixed populations of bacteria, and DNA testing will not replace many of the current antibody-based tests. However, in most settings, DNA testing is likely to replace the standard microbiological laboratory plate test for identifying species and biotypes, including antibiotic resistances, and give much more precise information about the pathogenicity of the infecting agent.

Predictive genetic tests and pharmacogenomics

It will be some time before we see widespread use of human genetic diagnostics to assess disease susceptibility and future risk, except in the case of some inherited cancers (eg, those caused by mutations in *BRCA1* and *BRCA2*) and simple genetic diseases.[14] These generally involve low-incidence, high-penetrance genes, as opposed to testing for high-incidence, low-penetrance genetic variations, which may underlie many common diseases. Already, however, genetic tests have been used widely and effectively to reduce the incidence of some genetic diseases (notably thalassaemia) in some communities, although this raises ethical issues on which there are different views in society. My own perspective is that people affected should be permitted to make their own informed choices. This means that the profession of genetic counsellor will grow. Physicians will also need to improve their understanding of and access to human molecular genetic information, and ethical guidelines for its use and disclosure.

Advances in DNA diagnostic technologies using single-cell PCR should make non-invasive prenatal genetic testing (using fetal cells from the maternal circulation or from Pap smears) a reality within a few years, and lead to widespread prenatal screening for common genetic disorders, including chromosomal abnormalities and familial diseases.[15] This will avoid at-risk mothers having to make the difficult choice whether to have an invasive test that may threaten the fetus, as well as allowing low-risk mothers to take the tests safely, with obvious benefits.

There will also be tests that will assess an individual's likely response to different drugs. This new field of "pharmacogenomics" will allow physicians to avoid prescribing particular drugs to patients who may be at risk for a negative reaction, as well as to decide what drug and what particular dosage regimen might best suit an individual.[16,17]

The main limitation on the deployment of broader DNA-based tests to assess the genetic susceptibility of individuals to diseases like cancers, cardiovascular disease and stroke is knowledge, in particular how different (combinations of) genetic variations might affect a person's predisposition to any particular condition. This knowledge will come in time, but it may well be much longer than a decade before this widely penetrates the healthcare system. Nevertheless, ultimately we will all use our genetic "scorecard" to identify possible problems and to develop personalised strategies to reduce the risk of such problems, including lifestyle modification, pharmaceutical intervention and other therapies.

Personal DNA identification

DNA "fingerprinting", which assays a small but quite variable subset of human genetic variation, has become a powerful tool for individual identification in forensic medicine and in areas such as the identification of victims of major disasters. Although there are some societal concerns about the use of this technology, it seems more than likely that (subject to appropriate controls) our DNA signature will become our personal identity card in the future.[18]

New pharmaceuticals

Genomic information has become the major driving force in drug discovery. The identification of important genes affecting human development and disease will provide the basis for developing an entirely new generation of pharmaceuticals via genetically engineered proteins. It will also enable identification of new targets for drugs, which may then be developed by rational (computer-aided) drug design or by screening natural and combinatorial chemical libraries.[19,20]

Several genetically engineered protein pharmaceuticals are already in widespread use, including human insulin, erythropoietin, growth hormone, and hepatitis B vaccine. Computer-aided design was used by the Australian company Biota to develop the antiviral drug zanamivir (Relenza). In 2000, it was estimated that 40% of all new pharmaceuticals reaching the market were derived from genomic science, and many more are in the pipeline.[19-22] These are the beginnings of a new era of pharmaceutical development which will increase the range and sophistication of pharmaceuticals to intervene to correct abnormal physiology and to prevent disease, as well as to enhance aspects of lifestyle.

Gene therapy

Gene therapy is, in principle, an attractive option for dealing with serious monogenic diseases like cystic fibrosis, as well as for repairing acquired genetic lesions, such as those involved in cancer.[23,24] There have been some spectacular successes in some metabolic disorders where gene therapy can be carried out with haematopoietic cells *ex vivo*, with apparently permanent repair of the condition.[25] There have also been some adverse effects, such as the development of cancer in a treated patient, possibly as the result of the insertion of the new DNA next to a cancer-causing gene,[26] and the death of a patient in an earlier gene therapy trial.[23] Nevertheless, the main problem in developing gene therapy remains delivery of the DNA to large numbers of cells (and the relevant cells) *in vivo*, and it is difficult to predict when this might be achieved, or to specify the range of genetic conditions for which gene therapy might become feasible.

Living longer — and better

There is reason to expect that over the next decade we will begin to understand the molecular basis of ageing, which is clearly a genetically programmed process, even if it is accelerated by differing lifestyles. This may lead to new therapies, pharmaceutical or otherwise, to delay the onset of ageing. Genomics will have much to offer by identifying genes influencing ageing in humans and in model organisms, as well as identifying and enabling production of growth factors and hormones that may be used to "pretune" stem cells for transplantation or to alter the physiology of ageing *in vivo*.

With the new knowledge base that will flow from genomics, along with the new tools and therapies that will develop from it, medicine will continue to change from being the art of crisis management to the science of good health. Doctors will increasingly work with patients in the better knowledge of their particular genetic strengths and weaknesses, not just to treat disease, but to prevent it from occurring and, more particularly, to optimise the health and longevity of the individual. Good health is the most important lifestyle issue of all, and these developments will have a large and very positive impact on healthcare economics, as well as on the formal and informal productivity of our society, even if the proportion of discretionary income and of gross domestic product spent on healthcare continues to rise.

Change is inevitable and is accelerating

Lest the sceptics believe that some of these predictions are optimistic, consider the changes that occurred during the twentieth century. In 1901, no one knew that DNA contains our genetic inheritance, let alone dreamed that the complete sequence of the human genome would be deciphered in 2001. It has been said that "genetic prediction of individual risks of disease and responsiveness to drugs will reach the medical mainstream in the next decade or so. The development of designer drugs, based on a genomic approach to targeting molecular pathways that are disrupted in disease, will follow soon after".[16] The pace of acquisition of knowledge is accelerating and so will the pace of change. These generally positive developments will also bring problems, some of which are being addressed locally by the current inquiry into the Protection of Human Genetic Information being conducted by the Australian Law Reform Commission and the Australian Health Ethics Committee,[27] and a newly commissioned Inquiry into Gene Patenting.[28]

Underlying many of these particular issues are deeper ethical and philosophical tensions, as, on the one hand, some people are concerned about interfering with natural life processes ("playing God"), and, on the other hand, we all want to live longer and healthier lives. However, we should bear in mind that our ethical views evolve as quickly as the world around us, especially when we are faced with real choices rather than theoretical dangers. What will happen when we work out the genetic factors underpinning human intellectual and artistic potential is anybody's guess, especially if artificial intelligence has in the meantime become a reality through advances in computational science. To be around to find out would be nice.

Competing interests

None identified.

References

1. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001; 409: 860-921.
2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001; 291: 1304-1351.
3. GOLD: Genomes online database. Chicago: Integrated Genomics Inc. Available at: http://wit.integratedgenomics.com/GOLD/ (accessed Mar 2003).
4. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002; 420: 520-562.
5. Aparicio S, Chapman J, Stupka E, et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes. Science* 2002; 297: 1301-1310.
6. Graveley BR. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* 2001; 17: 100-107.
7. Duboule D, Wilkins AS. The evolution of 'bricolage'. *Trends Genet* 1998; 14: 54-59.
8. Rubin GM, Yandell MD, Wortman JR, et al. Comparative genomics of the eukaryotes. *Science* 2000; 287: 2204-2215.
9. Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep* 2001; 2: 986-991.
10. Mattick JS. Noncoding RNAs: a regulatory role? In: Nature Encyclopaedia of the Human Genome. London: Nature Publishing Group, 2003; in press.
11. Kapranov P, Cawley SE, Drenkow J, et al. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 2002; 296: 916-919.
12. Okazaki Y, Furuno M, Kasukawa T, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full length cDNAs. *Nature* 2002; 420: 563-573.
13. Ronai Z, Yakubovskaya M. PCR in clinical diagnosis. *J Clin Lab Anal* 1995; 9: 269-283.
14. Goodfellow PN. Impact of genomics on healthcare. Overview. *Br Med Bull* 1999; 55: 305-308.
15. Findlay I, Matthews PL, Mulcahy BK, Mitchelson K. Using MF-PCR to diagnose multiple defects from single cells: implications for PGD. *Mol Cell Endocrinol* 2001; 183 Suppl 1: S5-S12.
16. Collins FS, McKusick VA. Implications of the human genome project for medical science. *JAMA* 2001; 285: 540-544.
17. Zanders E. Impact of genomics on medicine. *Pharmacogenomics* 2002; 3: 443-446.
18. Williamson R, Duncan R. DNA testing for all. *Nature* 2002; 418: 585-586.
19. Dean PM, Zanders ED, Bailey DS. Industrial-scale, genomics-based drug design and discovery. *Trends Biotechnol* 2001; 19: 288-292.
20. Ward SJ. Impact of genomics in drug discovery. *Biotechniques* 2001; 31: 626-630.
21. Ji Y. The role of genomics in the discovery of novel targets for antibiotic therapy. *Pharmacogenomics* 2002; 3: 315-323.
22. Alaoui-Ismaili MH, Lomedico PT, Jindal S. Chemical genomics: discovery of disease genes and drugs. *Drug Discov Today* 2002; 7: 292-294.
23. Rubanyi GM. The future of human gene therapy. *Mol Aspects Med* 2001; 22: 113-142.
24. Bauerschmitz GJ, Lam JT, Kanerva A, et al. Treatment of ovarian cancer with a tropism modified oncolytic adenovirus. *Cancer Res* 2002; 62: 1266-1270.
25. Hacein-Bey-Abina S, Le Deist F, Carlier F, et al. Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N Engl J Med* 2002; 346: 1185-1193.
26. Marshall E. Clinical research. Gene therapy a suspect in leukemia-like disease. *Science* 2002; 298: 34-35.
27. Australian Law Reform Commission. Protection of human genetic information [ALRC inquiry website]. Available at: www.alrc.gov.au/inquiries/current/genetic/ (accessed Mar 2003).
28. Australian Law Reform Commission. Gene patenting [ALRC inquiry website]. Available at: www.alrc.gov.au/inquiries/current/patenting/ (accessed Mar 2003).

(Received 30 Oct 2002, accepted 21 Jan 2003) ❑