

## **Estimating Disease Risk with Genetic Screening**

**Helen O'Neill**

**I**t is more than a year since the complete sequence of the human genome was publicly announced (International Human Genome Sequencing Consortium, 2001). The impact of this milestone on medical research is already apparent. Scientists are geared up to adopt new experimental approaches in the race to combat human disease. New information flowing from the human genome project should fuel a revolution in medical research with wide-ranging implications for human health, longevity and employability. There are very good reasons to collect and use all of this information for the benefit of mankind, to improve health care and sustain life. It remains for society to dictate the level of protection placed on personal genetic information. Checks and balances will need to be implemented sooner rather than later. However, the time may come when the genetic blueprint of an individual will be viewed no differently to any other personal information. The potential to predict disease risk by genetic screening is considered here along with the health and medical benefits which should flow from new approaches to medical research fuelled by completion of the human genome project.

### **Genetics Today**

The science of genetics originated in the mid-1800s with the experiments of Gregor Mendel who investigated the inheritance of characteristics in pea plants. While the classical genetics experiments of the early 20<sup>th</sup> century identified many inherited traits, it was the discovery of the DNA double helix in 1953 by Watson and Crick which led to identification of DNA as the carrier of genetic information. All somatic cells in an individual carry genes from both the father and mother and these are inherited *en bloc* on chromosomes. The human cell has 23 different chromosome pairs including an XX pair of sex chromosomes in females and an XY pair in the male. The approximately 30,000 genes carried by the 46 chromosomes represent the human genome.

Each chromosome is made up of a string of four different nucleotides or bases, making up a four-letter code referred to as the sequence. For a given gene, the nucleotide sequence forms the genetic code, represented by triplets of bases. Genes are aligned along this strand, delineated by regions of coding rather than non-coding DNA. Each gene is first transcribed into an RNA intermediate form and then the base triplets are translated into amino acids. The amino acids are

---

*Helen O'Neill is a Reader in the School of Biochemistry and Molecular Biology, Faculty of Science, The Australian National University.*

joined together in the sequence dictated by the RNA, producing a protein molecule that performs a distinct structural or physiological role in the cell.

Different proteins have different cellular functions. Each different type of cell produces a distinct range of proteins providing all the necessary structural and metabolic functions required by the cell. Different types of cells or cells in different stages of development have a distinct pattern of gene expression and produce a unique profile of proteins. There are many examples of this division of function between cells. For example, B lymphocytes, which represent one type of white blood cell type, produce antibody molecules which are released into blood where they destroy blood-borne pathogens. In the pancreas, islet cells are specialised for production of the insulin protein. Any gene mutation which leads to reduced levels of insulin can precipitate development of diabetes.

The nucleotide sequence of a given gene can also vary slightly between individuals. This gives rise to allelic variants of the gene. One example of this is the ABO blood group system, where individuals carry either the A, B or O allele of a gene encoding a molecule called the glycosyltransferase enzyme. This molecule functions differently in individuals of different A, B, AB and O blood types, leading to the addition of variant carbohydrate side chains on a cell surface molecule expressed by red blood cells. The presence of multiple alleles, or 'polymorphism', in a population leads to the existence of variant forms of the one protein, most of which will be functionally indistinguishable and retain normal function. Some allelic variants result in proteins which function less well in some circumstances than others. If this protein is critical to cell function it can place an individual at risk of disease.

Some genes have large numbers of alleles. One example is the HLA antigen system encoded by the major histocompatibility gene complex important in tissue matching for transplantation. There are many HLA alleles and some have been associated with disease states like ankylosing spondylitis and hereditary hemochromatosis. HLA genes can be sequenced to determine which alleles are carried by members of a multigenerational family. This information can then be used to determine the pattern of inheritance and the penetrance of the particular disease.

Similar screening methods are used to determine the pattern of inheritance of mutant genes. Any single base change or deletion in a gene can lead to mutant forms which can be inherited and precipitate disease in some family members. Genetic screening to determine sequence changes in a gene of interest can determine the mode of inheritance of the mutant gene within the family of an affected individual. More drastic mutations can involve chromosomal aberrations such as polyploidy (more than two copies of the same chromosome as seen in Down's syndrome) or translocations between different chromosomes. Once a gene mutation has been linked to a particular disease, standard diagnostic tests based on gene sequencing can be devised to detect both normal and mutant forms of the gene. Genetic screening can then be extended to the wider population. In a similar vein, genetic screening can involve chromosomal or cytogenetic analysis to detect chromosomal abnormalities associated with disease risk. Genetic

screening to determine disease risk is a straightforward task for diseases linked to known genes or known chromosomal aberrations.

### **The Nature of Inherited Disease**

Monogenic diseases are caused by mutations in single genes leading to production of abnormal proteins affecting a vital cell function. Individuals who inherit the mutant form of the gene have a high risk of disease. The nucleotide sequence of a gene can be used to quickly identify individuals at risk of inheriting the disease or trait. Single gene defects precipitating phenotypic changes or disease can be easily traced in pedigree studies. There are four main patterns of inheritance, autosomal recessive, autosomal dominant, X-linked recessive and X-linked dominant. Autosomal recessive diseases require a gene copy to be inherited from both parents in order to see expression of the disease, while autosomal dominant diseases require only one copy. X-linked recessive diseases, like red-green colour-blindness, are usually only seen in males, whereas X-linked dominant diseases are found equally in both sexes. More than 1000 human disorders have a genetic component or are inherited to some degree, but most of these are extremely rare affecting between one in  $10^4$  to  $10^5$  people. Very few (~100) of these are due to a single gene defect.

The majority of diseases with a genetic component are complex diseases involving somewhere between 10 and 30 different genes. The issue of genetic screening for detection of disease is not so straightforward. Most disease states are complex or multigenic, involving sets of interacting genes. Many are also multifactorial, having both genetic and environmental cause. Diabetes is an example of a complex disease. It is genetically determined, with both genetic and environmental components and additive small effects contributed by multiple genes. Even common diseases like allergy, heart disease, depression and asthma show some evidence of genetic susceptibility but require environmental triggers to precipitate disease.

Propensity to develop some forms of cancer also has a complex genetic and environmental basis. Similarly, capacity to resist infection by a range of pathogens including viruses, bacteria and parasites is genetically controlled. Another complication is that some disorders cannot easily be attributed to a genetic cause. They can also have variable expressivity giving a range of severity from mild to extreme. In some cases the same disease phenotype (or appearance) results from different gene effects. In other cases, alternate allelic forms of the same gene can lead to different phenotypes. It can be argued that most diseases are so complex that 'they push the current array of molecular genetic analysis to their limits' (Schork, 1997). Despite our increased capacity for genetic analysis, it is not yet possible to develop genetic screening tests for most of these complex diseases.

The immune response plays an overriding role in many disease states. This is the defence system that monitors the body and removes damaged or infected cells. Many genetic effects related to complex disease map to genes that control the

immune response (Fahrer *et al.*, 2001). It is well known that an effective immune response can alter the course of disease. It occurs in response to most disease states including infection, cancer, autoimmune disease and injury and there is increasing evidence that genes controlling immune responsiveness have a major influence on the progress and resolution of most diseases. For example, one of the recently mapped genes for diabetes incidence in mice is *IL-12B* which encodes a soluble factor involved in regulation of immune responses involving T lymphocytes (Morahan *et al.*, 2001). The immune response involves networks of cells and secreted molecules produced by cells as part of the defence against pathogens. It is vitally important to work out which genes are involved in immune response development since a defect in any one immune response gene could be manifest in multiple disease states. Of particular interest are networks of regulatory genes in humans which control the immune response. This type of study has enormous potential for development of new drugs which moderate immune function and disease resolution.

### Gene Discovery

Family studies can be readily used to determine whether diseases are inherited and have a genetic component. However, the association or linkage of disease with particular genes segregating in that family is a much more complicated process. The endpoint of a successful linkage analysis is the discovery of a specific gene(s) associated with the disease and determination of the specific function of the protein(s) encoded by the gene(s). New technology developed in the past two decades has increased the capacity for gene discovery.

One tool now commonly used by geneticists is the library of SNPs or single nucleotide polymorphisms that have been located throughout the human genome. These represent known sites of natural sequence variation in DNA that can be detected by sequencing DNA across those regions in different individuals. A genome-wide map of 1.42 million SNPs has now been aligned with the human genome sequence (International SNP Map Working Group, 2001). The nature of SNPs makes them particularly suitable for high-throughout genotyping or allele identification so that linkage of disease states with SNPs can be easily calculated. With the complete sequence of the human genome the ability to link disease states with SNPs on known chromosomes has increased enormously. These will be used increasingly to map disease states to specific regions of the genome and to particular genes.

The complete sequence of the human genome has also been aligned and integrated with BAC clones (bacterial artificial chromosomes) representing regions of chromosomes that have been linked to many disease genes and human traits. (International Human Genome Mapping Consortium, 2001). This clone-based map will be used increasingly by geneticists to obtain rapid and accurate identification of genes linked to disease. BAC clones will be used to isolate genes of interest for further study. Once candidate genes are defined it will then be

important to define the products of those genes and to develop strategies to alter their aberrant effects.

While the sequencing of the human genome has led to the identification of many genes, the function of most genes remains unknown. One of the major challenges facing geneticists is the characterisation of the proteins produced by all genes in the human genome. It will also be important to work out how genes interact and regulate each other. Complex regulatory networks of genes have been described which involve a number of regulatory proteins influencing expression of sets of genes (Wyrick and Young, 2002). These networks are very important in development and act to govern gene expression as the organism acquires new form and function through its life cycle. An understanding of the interactions involved in complex regulatory networks represents a very challenging area of genetics.

### **Functional Genomics**

Various experimental approaches can be taken to assess gene function. The pursuit of information on how genes are expressed and interact to give the phenotype or appearance of an individual has been called 'functional genomics' or 'phenomics'. For complex diseases like diabetes, functional genomic studies aim to determine all genes contributing to the diabetic phenotype and to assess quantitative and regulatory aspects of gene expression. This same approach will be needed to assess gene function and expression in all diseases including different types of cancers.

One current activity of widespread interest is the gene profiling of cancers and diseased tissues. This involves analysis of the expression level of all genes within affected cancer cells. This analysis utilises small glass slides called micro-arrays which are dotted with an ordered pattern of probes. These probes represent short sequences of DNA called oligonucleotides representing short sequences of known genes. Micro-arrays are exposed to labeled RNA prepared from tissue. The binding or hybridization of a labelled RNA molecule to particular oligonucleotide probes will indicate gene expression within the tissue of interest.

Micro-arrays containing myriads of oligonucleotide probes specific for genes of interest allow simultaneous quantitation of RNA related to genes expressed in tissues or cancers from affected individuals. Custom-made micro-arrays can be designed to measure expression of genes related to infection with various pathogens. Micro-array data can also be used in cluster analysis to identify coincident gene expression and to map out gene networks under the control of common regulators.

The use of animal models is essential in order to proceed to the next step of linking genes with particular disease states and gene functions. This can involve 'reverse genetics' technology and the production of genetically engineered mouse strains. Transgenic mice are developed by transfer of new genes into the DNA or genome of a mouse. From this altered mouse a new inbred strain can be developed to study the effect of expression of the introduced gene on the

phenotype or characteristics of the animal. Similarly, 'knockout mice' are produced by mutating genes in order to study the effect of loss of gene expression on the phenotype of the animal. Other studies can include introduction of marked genes into isolated cells in culture using DNA vectors which facilitate new gene expression inside cells. The effect of introduced genes which are expressed within cells can be used to make predictions about gene function.

While these approaches give valuable information, 'forward genetic strategies' in which random genes are mutated offer a faster approach to identification of gene function. A number of large-scale mouse mutagenesis projects are now being conducted to develop a range of mouse mutants that can be used to study the effect of gene loss on animal development and disease states.

This approach is essential to add function to genes in the mouse genome that are linked to disease. Since many genes in mice have homologous or equivalent genes in humans it will be a straightforward progression to translate information obtained in mouse mutants to the human genome and to define a function for genes important in human disease states. In this procedure, male mice are mutagenised to produce single nucleotide mutations in genes with a view to breeding multiple mutant mouse strains. Mice showing phenotypes or diseases of importance are selected for further breeding, pedigree analysis and mapping of their mutated genes by linkage with SNPs. The complete sequence of the mouse genome is available, along with overlapping sequence and linkage maps for all the mouse chromosomes.

Using these tools, it will be possible to identify candidate genes controlling disease and to make predictions about the loss function of mutated genes and the nature of the proteins responsible for producing the disease state. For example, mice have been reported which are characteristically obese, diabetic, immunodeficient or which develop cancers at an early age (Justice, 2000). Many linkage studies to identify candidate genes are underway.

From the sequence of the human genome, we can now predict how much of the DNA contributes to protein production. In the human cell only about 1% of DNA is actually coding DNA which gives rise to proteins. The remainder of the genome consists of intergenic and intragenic non-coding DNA, most of which has unknown or unmapped function in the cell. There is increasing interest in the RNA transcribed from non-coding regions. Regulatory RNA produced from intergenic regions is now thought to influence the production of proteins (Mattick and Gagen, 2001), although the mechanisms by which this occurs are not well understood. This type of information removes the emphasis on proteins as the controlling elements of cell differentiation. The enormous potential for regulatory effects due to interactions between RNA and between RNA and proteins or DNA is almost beyond comprehension. It represents an area of genomics that extends beyond the scope of current molecular biology.

## **Contribution of Genomics to Human Health**

The human genome project will almost certainly increase our capacity for improved genetic screening to determine carriers of genes linked to serious disease. Two new technologies appear to offer the greatest potential for identification of susceptibility genes. These are SNP typing and the use of micro-arrays coated with oligonucleotides representing the variant forms of genes. Custom-micro-arrays could be used increasingly to identify individual genes directly affected by known mutations. They could also be used to determine gene expression changes resulting from those mutations. The development of a comprehensive list of disease genes and definition of their function will link genetics with medicine and precipitate new approaches to medical research.

As more genes are mapped and their functions defined, the capacity to develop drugs and therapeutic strategies to halt or reverse disease states should increase. Once the protein product of a gene is known, new drugs can be designed to specifically block protein function. The field of pharmacogenomics, involving development of drugs based on genes that determine disease, will expand. It may also be possible to develop individualised treatment based on the disease genes carried by an individual. For example, individuals could be screened for genes determining sensitivity to certain drugs. This could allow more accurate determination of drug dose and type so that therapy for cancer can be more accurately applied resulting in more effective treatment and fewer side-effects.

One important area will be the definition of the genes determining susceptibility to infection. Pathogens are probably the greatest threat to human health and longevity. There will be increased information on genes determining immune response potential and capacity to overcome infection. New immunoregulatory molecules will be defined and exploited for their potential use in immunotherapy. It is expected that some genes determining immune response to infection will be associated with a number of different diseases. Along with the discovery of new genes controlling the immune response will come increased opportunity to design vaccines to induce immunity and prevent infection. Preventative medicine may become available to individuals genetically more susceptible to certain types of infection.

Continual investigation will be needed to gauge the impact of infection on the human genome since the genetic basis of disease susceptibility can change as new pathogenic variants arise due to mutation of rapidly dividing micro-organisms. While most vaccines will be directed against infectious diseases, the potential exists to develop vaccines against other complex diseases such as autoimmune diseases and cancers. Another important area of investigation will be the impact of infectious agents on the manifestation of diseases with a known genetic component. For example, previous infection is thought to be associated with the development of some diseases. Associations have been noted between rotavirus infection and development of diabetes, retrovirus infection with multiple sclerosis and Epstein Barr virus with chronic fatigue syndrome.

The mutant forms of many genes are already known to contribute to cancer. The human genome sequence can be used to identify genes similar to or representing partial homologues of known cancer genes. Another approach is to fingerprint cancers to identify their gene expression profile. Using micro-array technology, it is possible to produce gene expression profiles for each of the different types of cancers (Alizadeh and Staudt, 2000). By comparing profiles of cancer cells with the normal non-cancerous cell counterpart, genes specifically expressed in the cancer can be identified. Mutations in gene products and copy number differences can be detected along with genes in the downstream pathways responsible for the defective phenotype. The new Cancer Genome Anatomy Project (Strausberg *et al.*, 2001) aims to accelerate the field of cancer genomics by producing an online database of all gene expression data related to cancers.

### **Concluding Remarks**

The capacity to perform rapid and multiple genetic screening tests will increase with the availability of the human genome sequence and associated linkage maps. Screening tests can be easily designed to detect small mutational differences between alleles in different individuals and to make predictions about disease risk associated with different genes and their allelic and mutant forms.

However, despite enormous technical advances, the very complexity of diseases with a genetic component remains a major inhibiting factor for development of routine genetic screening strategies to determine risk for the majority of diseases. Another important issue is that genetic information alone does not have predictive value for disease. Environment, lifestyle and infectious disease also impact on the manifestation of genetically determined disease. The main conclusion from this paper is that while it is now easy to establish diagnostic genetic screening tests, these will not tell us much about risk for the majority of diseases which are complex and for which there is significant environmental dependency.

Genetic testing will not be predictive for multigenic, multifactorial diseases until scientists understand the function of all genes and map out the interacting networks between those genes. The problem facing scientists relates to the sheer number of genes and environmental factors impacting on disease expression. The full impact of the human genome project will not be felt for decades. 'The Human Genome Project will not be completed ...until the functions of all human genes have been determined' (Sutherland, p.594).

Understandably, people are concerned about issues related to genetic screening and the impact it could have on their privacy, health and financial situation. The problem with genetic data relates to its familial nature. People will remain sensitive about the need to disclose personal genetic information to insurers. The laws controlling privacy of genetic information will need to be protective but adaptable to meet the needs of individuals, researchers and insurance companies. Adaptability will be important because the development of knowledge may well accelerate with time. Information disclosure may become a

lesser problem as the functions of more genes are defined. Institutions offering life and medical insurance will need to consider how all of this new information will affect the type of policies they can offer to different risk classes at different premiums.

One prediction is that multiple genes will be associated with risk for most diseases. In the long run it may be possible to assign a value of risk to each gene linked to a particular disease. People may be faced with the decision to undergo genetic screening to assess their disease risk and to make lifestyle choices, or to undergo a variety of medical treatments. Longevity should increase with the availability of new preventative and therapeutic medicines (Sutherland, 2000). The nature of health and life insurance products could be quite different from those available today. People will still want insurance but the industry will need to adjust to meet consumer needs.

Actuaries will need to keep abreast of new developments in genetics in order to price insurance and make it available to most individuals. Once a risk is assigned to every gene in an individual for every known disease, then it is hard to see how life insurance as we know it will exist. The issue for actuaries and insurers may well be development of new products that guard against unforeseen factors like major environmental shifts or the occurrence of catastrophic genetic change. These are events that cannot be predicted given current knowledge. This process of change will depend on the accessibility of information under regulated circumstances.

## References

- Alizadeh, A. and L. Staudt (2000), 'Genomic-scale Gene Expression Profiling of Normal and Malignant Immune Cells', *Current Opinion in Immunology* 12:219-25.
- Fahrer, A., F., Bazan, P. Papathanasiou, K. Nelms and C. Goodnow (2001), 'A Genomic View of Immunology', *Nature* 409:836-8.
- International Human Genome Sequencing Consortium (2001), 'Initial Sequencing and Analysis of the Human Genome', *Nature* 409:860-921.
- International Human Genome Mapping Consortium. (2001), 'A Physical Map of the Human Genome', *Nature* 409:934-41.
- International SNP Map Working Group (2001), 'A Map of Human Genome Sequence Variation Containing 1.42 million Single Nucleotide Polymorphisms', *Nature* 409:928-33.
- Justice, M. (2000), 'Capitalizing on Large-scale Mouse Mutagenesis Screens', *Nature Reviews Genetics* 1:109-15.
- Mattick, J. and M. Gagen (2001), 'The Evolution of Controlled Multitasked Gene Networks: the Role of Introns and Other Noncoding RNAs in the Development of Complex Organisms', *Molecular Biology and Evolution* 18:1611-30.

Morahan, G., D. Huang, S. Ymer, M. Cancilla, K. Stephen, P. Dabadghao, G. Werther, B. Tait, L. Harrison and P. Colman. (2001), 'Linkage Disequilibrium of a Type 1 Diabetes Susceptibility Locus with a Regulatory IL-12B allele', *Nature Genetics* 27:131-2.

Schork, N. (1997), 'Genetics of Complex Disease', *American Journal of Respiratory Critical Care Medicine* 156:S103-9.

Strausberg, R., S Greenhut, L. Grouse, C. Schaefer and K. Buetow. (2001), 'In Silico Analysis of Cancer Through the Cancer Genome Anatomy Project', *Trends in Cell Biology* 11:s66-s71.

Sutherland, G. (2000), 'Cracking the Code: Just How Long Can we Live?', *Medical Journal of Australia* 173:594-6.

Wyrick, J. and R. Young, (2002), 'Deciphering Gene Expression Regulatory Networks', *Current Opinion in Genetics and Development* 12:130-6.

*The author gratefully acknowledges helpful comments by an anonymous referee. This paper was stimulated by a recent conference on Genetics and Financial Services organised by the Centre for Actuarial Research at the ANU's School of Finance and Applied Statistics, the Institute of Actuaries and the Securities Institute.*