Proc. Assoc. Advmt. Anim. Breed. Genet. Vol 14

LESSONS FROM GENE DISCOVERY IN HUMANS

Parry Guilford

Cancer Genetics Laboratory, Department of Biochemistry, University of Otago, Dunedin.

SUMMARY
Enormous technological advances in human genome analysis have dramatically advanced the rate of discovery of disease susceptibility genes in humans. However, the success of gene localisation projects continues to rely on careful sample acquisition and phenotype description. Once a putative disease associated mutation has been identified, extensive validation is often required to prove the mutation is causative. If the mutation is not clearly inactivating, *in vitro* assays may be required to demonstrate the biological consequences of a mutation. This paper describes the key features of successful human gene linkage projects using a search for a gastric cancer susceptibility gene as an example.
Keywords: Gene, gastric cancer, susceptibility, mutation.

INTRODUCTION

The identification of genes in animals has a number of disadvantages compared to the same exercise in humans. In particular, resources for chromosomal localisation and the physical isolation of genes such as polymorphic markers, genomic libraries, physical maps and expressed sequence maps are relatively scarce. However, animal geneticists do have the significant advantage of being able to manipulate the study population by generating crosses and pedigrees of the desired size and complexity. Here, I highlight the differences between gene discovery in humans and animals using our isolation of a gene for inherited gastric cancer as an example.

IDENTIFICATION OF A GASTRIC CANCER SUSCEPTIBILITY GENE

Family A is a large New Zealand Maori family affected by a dramatic incidence of gastric cancer over several generations. About 25 people have died of the disease in the last 30 years. Examination of the family tree suggested a dominant inherited gene with incomplete penetrance was responsible for the cancer susceptibility.

To identify this putative gene, we first carried out a genetic linkage analysis to identify a conserved chromosomal fragment that was segregating with the disease in the family (Guilford *et al*. 1998). Accurate identification of individuals who have inherited the genotype of interest is critical to any linkage study. In the case of inherited cancer susceptibility, it is therefore necessary to be sure of the diagnosis, eliminate cases of sporadic disease (*i.e.* phenocopy) and be alert to the risk of more than one predisposing mutation segregating in the same family (genetic heterogeneity). In the gastric cancer study, the risk of genetic heterogeneity was only moderate due to the relative rarity of the inherited form of the disease. However, in other conditions such as inherited deafness, the risk of multiple susceptibility mutations segregating in the same family is extreme. It is the risk of phenocopy that poses the greatest threat to studies to localise cancer susceptibility genes due to the frequency of cancer in the general population. Two factors helped us minimise the risk of phenocopy.

Firstly, most gastric cancers in Family A were of a specific histological subtype (poorly differentiated), and the age of onset of the inherited disease was on average about 30 years younger than the age of onset in the general New Zealand population. Thus, individuals with other histological types of gastric cancer, or late onset disease, were regarded as "disease status unknown" for the linkage study.

The collection of DNA samples from affected people in cancer studies is rate limiting, largely because of the often short period of time from onset of the disease to death. Consequently, the sample collection is usually retrospective. In the gastric cancer study, the majority of samples from affected people came from paraffin-embedded archival tissue that had been stored by hospital pathology departments since the time of the person's diagnosis or surgery. In some cases, the archival tissues were collected from other surgical procedures unrelated to gastric cancer. For one individual, the only available source of DNA was a spot of blood on a Guthrie card taken for routine genetic testing when the person was one day old. Thus, the collection of samples from affected people was very time consuming, not only because of difficulty finding the samples, but also because of the need for consent from each patient's next of kin.

Rather than immediately doing a genome-wide scan for linkage with several hundred polymorphic markers spaced evenly across the entire genome, we used a candidate gene approach. This strategy searches for co-segregation of the disease with genetic markers that flank genes of interest. It is a limited approach because it requires considerable prior knowledge of the molecular biology of the disease and large amounts of guesswork. It is most appropriately applied when logistics limit the feasibility of a genome-wide scan. Nevertheless, despite the severe limitations of the candidate gene approach, it is advisable to eliminate major candidate genes from contention before embarking on any genome-wide linkage study. In the example of the gastric cancer study, the reliance on archival pathology material was a limiting factor. The amount of archived tissue available is often minimal (e.g. paraffin-embedded tissue samples) and its DNA tends to PCR-amplify inconsistently. Thus at the outset of the project, we were not in a position to do a genome-wide linkage study using hundreds of markers.

In the gastric cancer study we compiled a list of about 30 genes that were known to be mutated or abberantly regulated in gastro-intestinal cancer. Highly polymorphic microsatellite markers flanking these genes were selected from the Généthon set of 5000 markers (Dib *et al*. 1996). One of the candidates was the gene for the cell to cell adhesion molecule, E-cadherin (*CDH-1*). E-cadherin was chosen because it is mutated in the majority of poorly differentiated gastric cancers, but very rarely in the other histological forms (Berx *et al*. 1998). After excluding, several other candidates, linkage was found (with a maximum two-point LOD score of 5.0) to marker D16S752 which maps adjacent to *CDH-1* on chromosome 16q22.1. Other markers surrounding D16S752 were then genotyped to confirm a conserved haplotype was co-segregating with the disease. By identifying recombinants, the haplotyping also enabled the minimum region containing the gene to be identified.

To determine if mutation of *CDH-1* was indeed responsible for the cancer predisposition, and not another nearby gene, its 16 exons (Berx *et al*. 1995) were amplified by PCR and analysed by single stranded conformation polymorphism (SSCP). SSCP is a method for mutation detection that

identifies differences in DNA secondary structure that are associated with sequence variation. It is rapid, detects >90% of sequence changes, and requires only standard laboratory equipment. Analysis of several samples from affected and unaffected individuals by SSCP found a mobility difference between the two sample sets that was consistently observed in *CDH-1* exon 7. Direct sequencing of exon 7 showed that the SSCP shift was due to a G T change at the last nucleotide of exon 7 (position 1008).

MUTATION VALIDATION

Whether the identification of a DNA sequence change in a candidate gene represents the end of a study, or just a step on the way, largely depends on the nature of that sequence change. If the change has a clearly profound effect on gene function, then it is reasonable to assume the susceptibility gene has been found; if the change is more subtle a long period of mutation validation will be required. The exon 7 mutation resulted in an amino acid change in a calcium binding site in the E-cadherin protein. The amino acid sequence of the site is highly conserved, being identical in fruit fly, mouse and man. This conservation argues that any change in the sequence of the binding site is likely to be deleterious to the protein's function. However, the impact of any amino acid change can still be debated without direct experimental data showing altered function, particularly if the change involves amino acids of similar structure. The *CDH-1* exon 7 sequence change also overlapped the DNA consensus motif required for normal splicing of exons. Analysis of mRNA transcripts purified from patient's stomach mucosa showed that the change resulted in aberrant splicing between *CDH-1* exon 7 and 8 in the majority of transcripts from that gene. In addition, the 1008G T change was not observed in 100 unrelated, unaffected individuals screened by SSCP, adding further evidence that the 1008G T change was not simply a rare polymorphism segregating in family A by chance.

This data, although compelling, was still open to interpretation. Conclusive proof that a mutation is responsible for a disease generally requires either (i) direct biological data showing severely disrupted function or expression of the gene in mutation carriers or (ii) the identification of germline mutations in that gene in independent families. Since the initiation of the gastric cancer project we had collected samples from two further families suffering from familial gastric cancer of the poorly differentiated subtype. Sequencing of the *CDH-1* coding sequence from these samples identified two distinct germline mutations, a premature stop codon and an insertion mutation, which both clearly were capable of inactivating the gene (Guilford *et al*. 1998). We and other groups have now gone on to identify 24 different germline *CDH-1* mutations in 25 families of diverse ethnic origins. The phenotype caused by each of these 24 *CDH-1* mutations appears indistinguishable.

A search for a mutation in a gene needs to examine the entire coding sequence of all exons, the exon/intron splice sites, and if necessary, the genes non-coding regulatory sequences. It must also consider not only point mutations, but large sequence deletions or insertions.

In summary, the key features of successful gene discovery in humans are:

(i) accurate ascertainment of the pedigree and disease inheritance pattern; this is essential for both the linkage analysis and the need to minimise the risk of a second independent mutation segregating in the family.

(ii) clear, detailed definition of the phenotype; accurate clinical data and as many defining features of the disease as possible decrease the risk posed by phenocopy.
(iii) maximised number of affected samples; a minimum of approximately 12 affected samples is required for chromosomal localisation. Genotyping additional samples will decrease the size of the chromosomal interval that contains the gene. The smaller the interval, the fewer genes need to be screened for mutations.
(iv) access to additional affected families; mutation searching in related families potentially provides the simplest method for validation of a proposed susceptibility gene.

RECENT DEVELOPMENTS IN GENE IDENTIFICATION AND THE IMPACT OF THE HUMAN GENOME PROJECT

The release of the human genome sequence and a number of other genome initiatives have now dramatically simplified the task of isolating a gene responsible for a monogenic disorder. In human genome-wide linkage studies, progression from chromosomal localisation to gene identity has until recently first required the construction of a physical map of the chromosomal interval containing the gene. Physical maps portray clusters of defined, overlapping fragments of DNA. These fragments are of manageable size, and have been cloned into appropriate replication vectors to facilitate the analysis of their gene content. The release of the human genome sequence and integrated gene expression databases (e.g. dbEST http://www.ncbi.nlm.nih.gov/dbEST) has obviated the need to generate these physical DNA maps and also largely eliminated the need to identify the genes which map to the interval. Researchers can now go directly from chromosomal localisation to mutation detection. Mutation detection however remains laborious, particularly if a gene's intron/exon structure is complex. Recent developments in DNA fragment analysis (such as Transgenomic's WAVE® Analysis System) can however, speed the process of mutation screening.

Searching for novel genes that predispose to complex multifactorial diseases such as diabetes, psychiatric disorders, asthma and cardiovascular diseases, has until recently been largely impractical. In complex diseases, multiple genetic variants each contribute a small effect and different combinations of genes may result in a similar phenotype. Linkage analysis has limited power to detect such small effects. However, association studies, which compare large numbers of cases of disease with matched controls from the same population, have the potential to lead to the identification of genes contributing to complex diseases. To date, most association studies have been limited to candidate genes. This has been due to a lack of sufficient numbers of appropriate genetic markers and the absence of effective high throughput genotyping techniques. The highly polymorphic microsatellite markers that have been so successful for studies on monogenic disorders in families are too unstable (i.e. a high mutation rate) and too sparsely spread over the genome to be used in population studies. However, the last few years has seen the advent of extreme density maps consisting of markers known as single nucleotide polymorphisms (SNPs). SNPs are single nucleotide differences in sequence that are relatively stable (therefore suitable for identifying population differences but insufficiently polymorphic to be suited to family studies), extremely abundant and amenable to automated throughput. Because SNPs can occur in protein coding regions and gene regulatory sequences, they can also have functional consequences. Thus, their use in mapping can lead directly to the identification of causative mutations. The highest density SNP map currently under construction will eventually consist of 300,000 SNPs (Marshall 1999). Although very large

numbers of SNP markers are required for complex disease studies, (Collins *et al*. 1999) semi-automated technologies such as the Applied Biosystems 7900 real-time PCR system, which can analyse over 5,000 SNPs/day, and microarray-based techniques, which may soon be able to genotype 40,000 SNPs on a single microscope slide (Wang *et al*. 1998), provide the technical power to complete these studies.

CONCLUDING REMARKS

The technological advances made in the gene discovery process in humans will rapidly be adapted to other animal species. However, despite the changing technology, the need for observant animal breeders and scientists to identify and precisely define the trait of interest, and then establish a population of animals suitable for genetic analysis will not diminish. The technology will come, the true challenge will remain the ability to first observe the diversity.

REFERENCES

Berx, G., Becker, K.-F., Hofler, H., and Roy, F. v. (1998) *Human Mutation* **12**: 226.

Berx, G., Staes, K., Hengel, J. V., Molemans, F., Bussemakers, M. J. G., Bokhoven, A. V., and Roy, F. V. (1995) *Genomics* **26,** 281.

Collins, A., Lonjou, C., and Morton, N. E. (1999) *Proc. Natl. Acad. Sci. USA* **96**: 15173.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J., and Weissenbach, J. (1996) *Nature* **380**: 152.

Guilford, P., Hopkins, J., Harraway, J., McLeod, M., McLeod, N., Harawira, P., Taite, H., Scoular, R., Miller, A., and Reeve, A. E. (1998) *Nature* **392**: 402.

Marshall, E. (1999) *Science* **284**: 406.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lander, E. S., and *et al*. (1998) *Science* **280**: 1077.