

# WHOLE GENOME ASSOCIATION STUDIES IN CANCER GENETICS

Over the coming years our expanding knowledge of cancer genetics will have a major impact on our ability to predict an individual's level of risk of developing cancer, detect and diagnose cancer early, and select treatments which are most likely to be effective.



**Richard Houlston**  
**MD PhD FRCP FRCPATH**  
Richard Houlston is Professor of Molecular and Population Genetics and Team Leader of the Molecular and Population Genetics Team in the Section of Cancer Genetics at The Institute of Cancer Research.

## Genetics and cancer

In the year 2000, the Department of Health published The NHS Cancer Plan which set out the first ever comprehensive strategy to tackle the disease. The plan states: "Advances in genetics will lead to a greater understanding of inherited susceptibility to cancer. The relative influence of genes on cancer development is variable and ranges from situations where genetic factors predominate and are highly predictive of disease development, to others where they play only a minor role in modifying the effect of environmental exposure to toxic substances."

Cancer susceptibility is determined by exposure to environmental factors and inheritance of genetic factors, which act either alone or in combination to influence likelihood of disease. Most common cancers such as breast, colorectal and prostate cancer show familial clustering with relatives of patients being at a two-fold increased risk of cancer at the same site.

■ The higher rate of most cancers in identical twins versus non-identical twins or siblings provides strong evidence that most familial clustering is a consequence of inherited genetic variation rather than lifestyle or environmental risk factors. ■

After several decades of research, a great deal has been learned regarding the nature of inherited susceptibility to cancer. Much of this knowledge has been restricted to rare families carrying genetic variants conferring substantive cancer risks, for example occurrence of breast cancer

in families caused by inherited mutations in the *BRCA2* gene. However, such high-risk genetic variants are generally rare in the population and therefore make a restricted contribution to the overall disease incidence. Data on patterns of familial occurrence of cancer, excluding cases due to known high-risk genes, indicate that most genetic susceptibility to common cancers results from the combined effects of many genetic variants, some of which will be common, and each of which have a modest effect individually.

## Association studies

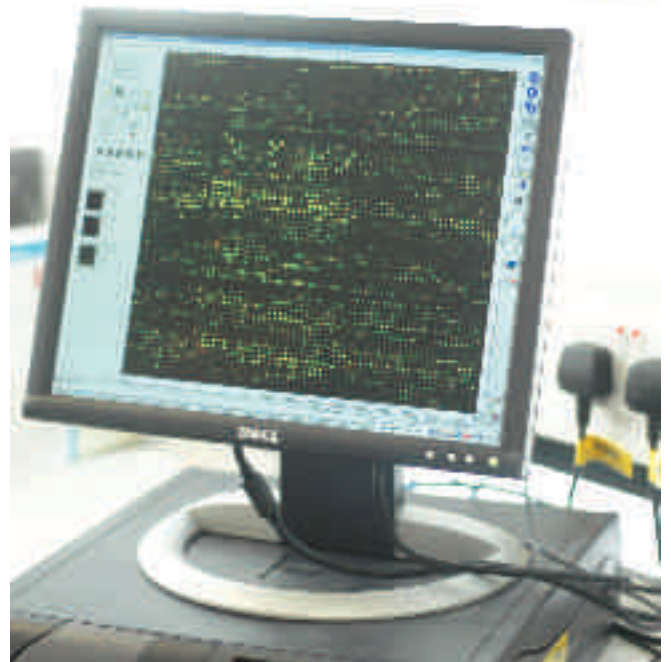
For a specific location in the human genome, allelic association is present when the distribution of genotypes differs between individuals with cancer and those unaffected. Such an association provides evidence that the locus of the genome under study, or a neighbouring region, is related to disease susceptibility. The aim of a whole genome association study is to detect such variants, searching in detail at many thousands of loci throughout the genome. Such studies are based on the 'common variant - common disease' hypothesis. Genetic variants arising a long time ago may have become common in the present population at frequencies ranging from a few per cent upwards. Some of these variants may predispose to common diseases, and combinations of these variants are proposed to underlie differences in disease susceptibility. Association mapping will be a powerful tool for mapping such loci with moderate effects.

Three factors are therefore motivating researchers to search for common modest risk variants associated with risk of developing cancer:

- (I) Risk variants can illuminate novel and important aspects of cancer development;
- (II) Common variants are likely to be important from a public health perspective since they make a more substantive contribution to disease prevalence;
- (III) Common modest risk variants are becoming easier to identify.

## Single nucleotide polymorphisms (SNPs)

Association studies are generally based on SNPs which are the most common form of sequence variation in the genome and consist of a variant at a single nucleotide in the genome. Following the sequencing of the human genome, large-scale harvests of SNPs have been conducted and there are currently over 10 million documented SNPs. These collectively account for over 90% of sequence variation between individuals. The functional effects of an SNP are dependent on the location and nature of the sequence change in DNA. Those occurring in genes are generally thought most



likely to confer functional effects, however SNPs occurring outside genes may influence expression of proteins.

Direct evidence that polymorphic variation defined by SNPs can influence an individual's risk of cancer is increasing. Examples include the relationship between SNPs mapping to the folate metabolism gene *MTHFR* and risk of colorectal cancer, and between variants of the gene *CASPASE8* and risk of breast cancer.

■ In addition to influencing cancer risk, SNPs may modify environmental exposures or the effects of drugs. For example, variants of *UGT1A6* have been reported to modify the protective effect of aspirin on colorectal adenoma risk. ■

## Haplotype blocks

Adjacent SNPs in the same chromosomal region are not inherited independently but as a combination of alleles that form haplotype blocks. Generally the closer two SNPs are on a chromosome, the more likely they are to be inherited together. This concept is captured by linkage disequilibrium (LD) which is defined as the non-random correlation between alleles at a pair of neighbouring SNPs and which underlies the principle of gene mapping by association analysis. Linkage disequilibrium between a marker allele and a disease susceptibility allele will result in both alleles being inherited together over many generations, thus the same marker allele will be detected

in affected individuals in multiple, apparently unrelated, families. Recombination between the marker and disease susceptibility allele will eventually dissipate the association (as can further mutational events) with the rate of decay being primarily dependent on the distance between the two alleles and the number of generations that has passed. The slowness of this decay, however, makes allelic association a useful tool. Additionally, the complexity of analysing a number of different SNPs within a particular gene or locus can be significantly reduced if there is strong LD between them, since the genotype of all the SNPs within the haplotype block can be inferred by the genotyping of only one or a few marker SNPs or tagging SNPs.

▣ Linkage disequilibrium can thus be further exploited in association studies by using tagging SNPs to reduce the number of SNPs that require genotyping, significantly lowering laboratory costs. ▣

Patterns of LD between SNPs have been characterised allowing subsets of tagging SNPs to be selected that, through LD with other variants, capture a large proportion of the common sequence variation in the human genome. This approach is unbiased and does not depend upon prior knowledge of function or presumptive involvement of any gene in disease causation. Moreover, it avoids the possibility of missing the identification of important variants in hitherto unstudied genes.

### Advancing research

Until recently, owing to financial and technical constraints, association studies on a genome-wide basis could not be contemplated; analyses were not surprisingly restricted to assessing the relationship between a limited number of defined variants and disease risk. Recent technological developments have led to the setting-up of analytical platforms which permit the simultaneous scoring of large numbers of SNPs and which are financially realistic.

While few genome-wide association studies of complex traits have been completed to date, the strategy has already yielded at least three examples of susceptibility loci, including one of prostate cancer, which appear to be robust. The relationship between patient genotype and aetiology of cancer is thus now open for further exploration.

The identification of genes associated with cancer predisposition and determination of their contribution to disease incidence is, however, contingent on having DNA samples from large systematic series of cancer patients.

▣ Researchers at The Institute have established large-scale collections of DNA for the most common cancers, typically of over 5,000 samples, to aid the discovery of genes associated with specific cancers. ▣

Supported by Cancer Research UK, major initiatives are currently underway in the research groups of Professor Richard Houlston, Professor Nazneen Rahman and Dr Rosalind Eeles in the Section of Cancer Genetics at The Institute to search for common predisposition genes for colorectal, breast and prostate cancer. For colorectal cancer, whole-genome genotyping has involved over 500,000 SNPs being genotyped in 2,000 colorectal patients and 2,000 controls. Those SNPs which show an association with colorectal cancer risk are being validated in a second phase, by typing 60,000 SNPs in over 10,000 samples. Phase 1 of the analysis has been completed with promising results and Phase 2 is underway. Phase 1 of the prostate cancer study is currently underway. Contingent on associations being established, the major challenges will be the fine mapping of regions in the genome displaying association and identification of causal variants. At The Institute we are well placed to expedite these analyses and examine gene-environment and gene-gene interactions.

Our expanding knowledge of the human genome coupled with recent significant advances in technologies such as high throughput genotyping are heralding a new dawn in the search for and identification of inherited cancer susceptibility. The new cancer genes are likely to have wide implications for public health in terms of allowing individuals at risk to be better identified and targeted screening provided.