J R Coll Physicians Edinb 2012; 42:139–42
http://dx.doi.org/10.4997/JRCPE.2012.211


# Genetic epidemiology

[1]JK Dunbar, [2]H Campbell
[1]*Specialty Registrar, Public Health Medicine, Public Health Directorate, NHS Tayside, Kings Cross Hospital, Dundee, Scotland, UK;*
[2]*Professor of Genetic Epidemiology and Public Health, Centre for Population Health Sciences and Institute of Genetics and Molecular Medicine, University of Edinburgh, Scotland, UK*

**ABSTRACT** The field of genetic epidemiology has advanced considerably over the past decade. The falling costs of genome-wide association studies have allowed the identification of many common genetic variants associated with risk of illness. This has advanced the understanding of disease pathogenesis but has not led to widespread changes in care. As these studies have become more common, a framework for reporting findings in the literature has been developed to ensure clinicians can accurately interpret the research. In future, whole genome genetic sequencing will likely lead to the identification of rarer variants more strongly associated with illness. Currently large resources for research, such as the UK Biobank project, are being developed which will allow detailed exploration of not only genes but the way their actions are controlled.

***Correspondence to** H Campbell*
***Public Health Sciences section, Centre for Population Health Sciences, The University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, Scotland***

***tel.** +44 (0)131 650 3218*
***e-mail** Harry.Campbell@ed.ac.uk*

**KEYWORDS** Genetics, epidemiology, genome-wide association study, genomics, epigenetics, research techniques

**DECLARATION OF INTERESTS** No conflict of interests declared.



## OVERVIEW

Genetic epidemiology has developed considerably in the past ten years both in terms of research methods and new discoveries giving insights into disease aetiology. In 2003 the Human Genome Project was completed, giving genetic researchers a full template of human DNA. Subsequently the International HapMap Project in 2005 provided researchers with tools to study common variants across the entire genome. These two major endeavours have made possible new approaches to studying genetic associations with disease. Prior to these advances, determining whether diseases were inherited and identifying genetic variants influencing human disease was restricted largely to the study of mutations in families or pedigrees (linkage analysis). This usually required medical data from related members of several pedigrees or family trees, typically in which there were several cases of the disease under study. A specific gene of interest was tested for in each individual to see if disease and gene were 'linked'. However, in the past five years it has been possible to study large numbers of genetic polymorphisms (with frequency in the population of >1%) across the entire genome. Polymorphisms are sections of DNA where the code differs from one person to another. Single nucleotide polymorphisms are those with a single letter of DNA code that is different. These studies typically look at between 100,000 and 2.5 million single nucleotide polymorphisms (or SNPs) in participants recruited in:

- large case-control studies for the study of genetic determinants of specific diseases; or
- large cohort studies for the study of genetic determinants of quantitative traits underlying disease, for example, biochemical and physiological parameters like serum cholesterol or high blood pressure.

These studies are commonly referred to as genome-wide association studies (GWAS). They make no assumptions regarding which variants may be associated with the disease or trait under study, leading to this research approach being called 'hypothesis free'. The falling costs of genotyping have led to many research centres across the world, in both public and private spheres, being able to undertake this research, and GWAS studies have been conducted on more than 200 conditions (as of January 2011).

## GENOME-WIDE ASSOCIATION STUDIES

The GWAS method has been very successful in identifying genetic variants that are associated with disease and disease traits. Over 700 genetic mutations have been shown to have statistically significant association with disease. These studies employ stringent statistical significance thresholds ($p < 5 \times 10^{-8}$) to account for multiple testing, correct for population stratification effects and require to be replicated in independent populations before being accepted for publication. They are not subject to the usual sources of confounding in case-control studies and so generally represent 'true positives'. This information has uncovered many new pathways underlying complex diseases and this new knowledge has great potential to identify novel 'drug

targets' and lead to new treatments. Furthermore, by understanding the genes that influence disease these studies may in future aid in predicting disease severity, response to treatment and outcomes. This knowledge could improve the efficacy and the cost-effectiveness of medical care. However, GWAS are an intermediate step in this process and it will take many years of subsequent research to realise this potential. An example of this was the recent identification of genetic mutations within a gene for urate transporter.[1] In this case several genetic variants associated with gout were found in and near the *SLC2A9* gene. When the gene was subsequently sequenced and identified it was found to be a novel kidney urate transporter whose dysfunction was strongly associated with the presence of gout in patients from a wide variety of populations.

Recent years have seen the price per scan of a human genome using single nucleotide polymorphism (SNP) arrays fall dramatically by several orders of magnitude. This in turn has permitted the conduct of very large collaborative studies with many tens of thousands of individuals across several populations being studied with the resultant power to detect common genetic variants with very small effects. The latest developments in genetic technology are now making whole genome sequencing possible at affordable costs (the aim is to make this possible for USD$1,000 per sample in the next few years). This will allow rare genetic variants, mutations and copy number variants to be studied as well in case-control and cohort studies.

Despite the increase in the number of studies showing significant association between genetic variants and disease, the applicability of this knowledge to clinical practice is limited. The effects of the common genetic variants that have been discovered are typically very small, with odds ratios below 1.5. A small number of genetic variants have been shown to strongly predict disease, such as variants that predict the likelihood of radiation therapy-induced second malignant neoplasms in patients with prior Hodgkin's lymphoma.[2] However, the vast majority of variants have small effects on risk and have not been found to be useful for disease risk prediction. It is expected that this will improve once information on rarer variants and copy number variants becomes available from sequencing studies in the next few years. It is expected that many rare genetic variants will be found to have much larger effects on disease risk (for example, odds ratios of 3–8). In addition, future information on gene–gene and gene–environment interactions from very large biobank studies should also improve future risk prediction. Individual differences in environmental changes and detailed phenotyping are crucial steps in trying to understand the complex relationship between genes and the environment. To this end the UK Biobank project has collected detailed data on over half a million adult participants, including personal data, medical history, a range of physiological measurements, data on exposures and data linkage to a wide range of health and social information.[3]

Future perspectives include research into epigenetic determinants of human disease. Epigenetics is the study of inherited factors that are not transmitted through the genetic code and can be influenced by environmental exposure; these factors can have an influence on susceptibility to disease by controlling gene expression. Cancer cell lines for example, have an increase in methylation of specific genetic sequences.[4] Given the potentially important relationship between epigenetic factors and disease, the US National Institutes of Health has stated that it will spend USD$190 million between 2008 and 2013 on the field of epigenetics.[5]

In response to the need for clear, consistent and accurate reporting of genetic epidemiological research findings, statements of best practice for reporting a genetic epidemiology study have been released. The two leading statements are STREGA (STrengthening the REporting of Genetic Associations) for reporting genetic association studies[6] and GRIPS (Genetic RIsk Prediction Studies) for reporting genetic risk prediction studies.[7] These two distinct statements give clear requirements for any author of a genetic epidemiology study on how best to disseminate clear critical information to enable readers to judge the validity of findings.

## CONCLUSION

In conclusion, genetic epidemiology has advanced extensively in the past 25 years from studies of single genes and the familial relationship of disease to large, complex studies with genome-wide arrays studying up to millions of genetic variants simultaneously. The advances in the technology have provided relatively cheap, fast and efficient means of determining subject genotype in studies and the resulting increase in research has uncovered much of interest. However core problems still remain, such as understanding the complex relationship between genes and environment. Future developments of large study populations such as the UK Biobank may be able to reveal more about what determines disease susceptibility and the gene-environment interaction. However this can only be achieved through vigorous, high quality research that is clearly presented to the wider research population.



*J R Coll Physicians Edinb* 2012; 42:139–42

## HIGHLIGHTS

- Genome-wide association studies (GWAS) have become a mainstay of the study of genetic epidemiology with several hundreds of studies undertaken on genetic variants to date.
- The widespread nature of these GWAS studies has demanded an improved quality of reporting so their significance can be reasonably assessed by clinicians.
- GWAS studies have identified numerous variants which are significantly associated with disease. However, few of these variants are able to strongly predict individual risk of disease.
- Variants of interest have helped develop understanding of the pathogenesis of disease, providing new potential therapeutic targets.
- Future developments will look at not only the genetic mutations but also the way their actions are controlled.

### *Further reading*

- **National Institutes of Health:** Talking glossary of genetic terms www.genome.gov/glossary/index.cfm
- **International HapMap Project:** http://hapmap.ncbi.nlm.nih.gov/

## REFERENCES


1 Vitart V, Rudan I, Hayward C et al. SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat Genet* 2008; 40:437–42. http://dx.doi.org/10.1038/ng.106

2 Best T, Li D, Skol AD et al. Variants at 6q21 implicate PRDM1 in the aetiology of therapy-induced second malignancies after Hodgkin's lymphoma. *Nat Med* 2011; 17:941–3. http://dx.doi.org/10.1038/nm.2407

3 UK Biobank [Internet]. Available from: http://www.ukbiobank.ac.uk/

4 Foley DL, Craig JM, Morley R et al. Prospects for epigenetic epidemiology. *Am J Epidemiol* 2009; 169:389–400. http://dx.doi.org/10.1093/aje/kwn380

5 US Department of Health and Human Services, NIH News, National Institutes of Health. *NIH announces funding for new epigenomics initiative* [Internet]. Maryland: NIH; 2008 [cited 2012 March 21]. Available from: http://www.nih.gov/news/health/sep2008/od-29.htm

6 Little J, Higgins JP, Ioannidis JP et al. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. *PLoS Med* 2009; 6:e1000022. http://dx.doi.org/10.1371/journal.pmed.1000022.

7 Janssens AC, Ioannidis JP, Bedrosian S et al. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. *Eur J Epidemiol* 2011; 26:313–37. http://dx.doi.org/10.1007/s10654-011-9551-z


EDUCATION

## SELF-ASSESSMENT QUESTIONS

**1. What has driven the large increase in genetic analyses over the past 15 years?** *Choose one answer.*

A. Strong public interest and subsequent private donations to charities focused on this research.
B. Massive increases in the provision of funding for genetic analysis.
C. Widespread increases in the capacity of universities to investigate the relationships between genes and disease.
D. Steadily decreasing cost of testing patients for genetic variants.
E. Pressure by political representatives.

**2. What is the UK Biobank?** *Choose one answer.*

A. A reference library of genetic information.
B. A repository of anonymised tissue samples to determine the natural incidence of genetic variants in the population.
C. A collection of data on patients which includes detailed medical history and a tissue sample to test for genetic variants.
D. A private initiative designed to find and develop patents on important and profitable genes.
E. A high street bank with an interest in biometric security.

**3. Why do genome-wide association studies (GWAS) require much higher p values to prove statistical significance?** *Choose one answer.*

A. Because associations are common.
B. To account for the multiple testing inherent in GWAS.
C. p stands for power and high p-values equate to greater power.
D. Associations with high p-values are more likely to be true associations.
E. GWA studies are likely to be inaccurate and this needs to be corrected for.

**4. What is epigenetics?** *Choose one answer.*

A. The study of the relationships between disease and environmental factors.
B. The study of disease and factors which are inherited and control the expression of genetic variants.
C. The study of genetics of epilepsy.
D. The study of the relationships between environmental and genetic factors.
E. The shorthand for genetic epidemiology.

**5. What was the main method of investigation used to determine the genetic aspects of disease (such as Mendelian disorders) before genome-wide association studies (GWAS)?** *Choose one answer.*

A. Linkage studies on large pedigrees (family trees).
B. Case-control studies.
C. Cohort studies.
D. Laboratory studies.
E. Case studies.

***This paper was originally published as part of the Public Health module in the RCPE Online Continuing Medical Education Programme. Online CME, including the answers to these questions, is available to Fellows and Members at: http://www.rcpe.ac.uk***

EDUCATION



*J R Coll Physicians Edinb* 2012; 42:139–42