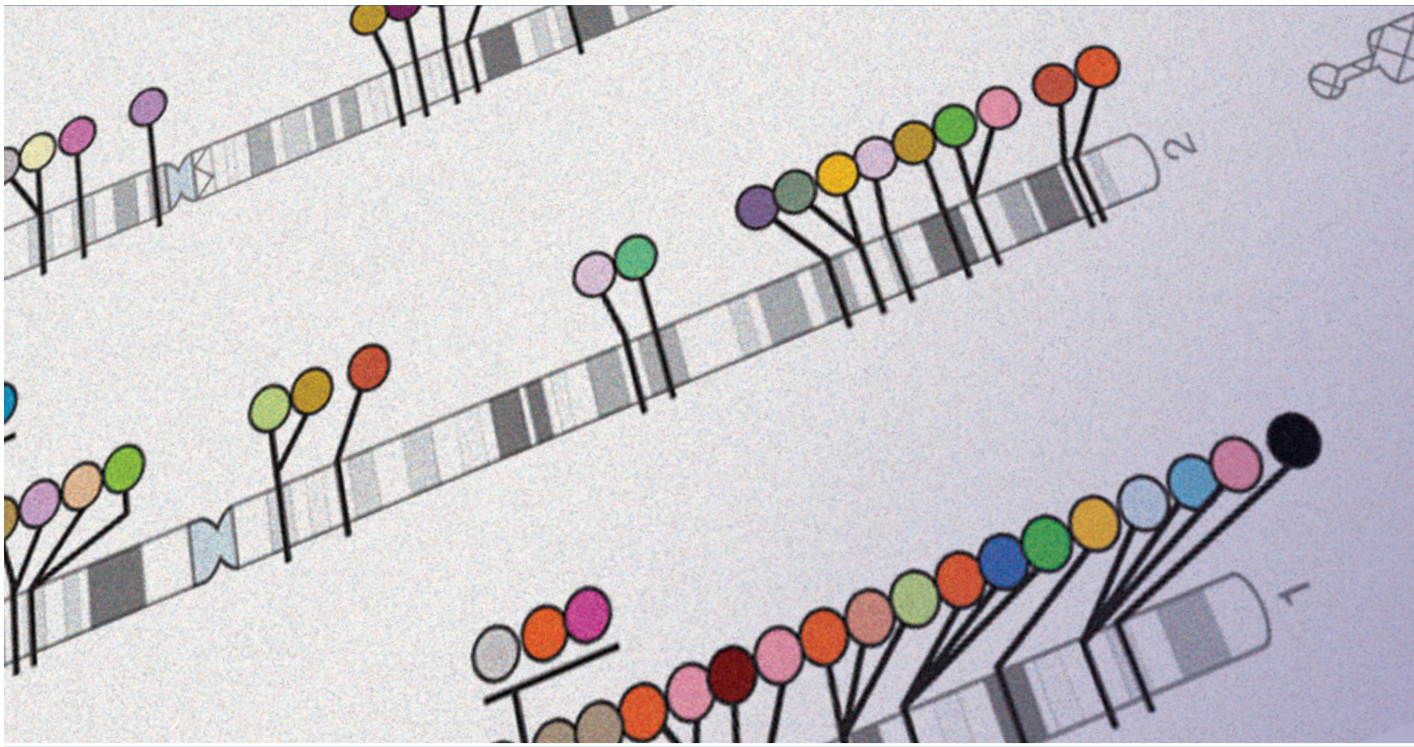
The Academy of Medical Sciences | FORUM



Genome-wide association studies: understanding the genetics of common disease

Symposium report

July 2009

The Academy of Medical Sciences

The Academy of Medical Sciences promotes advances in medical science and campaigns to ensure these are converted into healthcare benefits for society. Our Fellows are the UK's leading medical scientists from hospitals and general practice, academia, industry and the public service.
The Academy seeks to play a pivotal role in determining the future of medical science in the UK, and the benefits that society will enjoy in years to come. We champion the UK's strengths in medical science, promote careers and capacity building, encourage the implementation of new ideas and solutions – often through novel partnerships – and help to remove barriers to progress.

The Academy's FORUM with industry

The Academy's FORUM is an active network of scientists from industry and academia, with representation spanning the pharmaceutical, biotechnology and other health product sectors, as well as trade associations, Research Councils and other major charitable research funders. Through promoting interaction between these groups, the FORUM aims to take forward national discussions on scientific opportunities, technology trends and the associated strategic choices for healthcare and other life-science sectors.

The FORUM builds on what is already distinctive about the Academy: its impartiality and independence, its focus on research excellence across the spectrum of clinical and basic sciences and its commitment to interdisciplinary working.

Acknowledgements

This report provides a summary of the discussion at the FORUM symposium on 'Genome-wide association studies' held in October 2008. The Academy gratefully acknowledges the support of GlaxoSmithKline for this event. For further information please contact Dr Robert Frost, Manager, FORUM, robert.frost@acmedsci.ac.uk.

www.gsk.com

ISBN No: 978-1-903401-22-4

The Academy *of* Medical Sciences

Genome-wide association studies: understanding the genetics of common disease

Symposium report

July 2009

Contents

Summary

Genome-wide association (GWA) studies are a powerful new tool for deciphering the role of genetics in human biology and common disease. By analysing hundreds of thousands of genetic variants, and comparing individuals with a specific disease against carefully selected controls, the approach is, for the first time, identifying multiple genetic changes associated with common polygenic diseases. GWA studies have been made possible by detailed mapping of the genetic sequence and by technological advances that allow the simultaneous genome-wide comparison of these variations. In the last two years, the technique has been successfully applied in studies of diseases of major medical importance such as cancer, diabetes and coronary artery disease.

The search for functionally important genetic variants now lies at the heart of much biomedical research. Each variant that is robustly linked to a disease offers a possible route to understanding the underlying biological pathways and potentially to the development of new treatments. The construction of detailed 'molecular signatures' and the classification of molecular sub-types of specific conditions is informing a new taxonomy of disease. Increasing knowledge of molecular variation brings the prospect of stratifying human populations according to genotype, improving the design of clinical trials, and enhancing patient care. Opportunities to develop safer and more effective treatments through targeting a patient's underlying biology must be seized.

Success in identifying genetic variants that predispose to common diseases can also improve disease diagnosis and management. Individually, many of the common variants identified to date confer only a small risk of the disease, limiting the immediate utility of genetic profiling to predict individual disease susceptibility. However, by examining the patterns of variation across the genome it may become possible to identify subgroups at differing degrees of risk. This has the potential to impact on screening procedures for specific conditions and the targeting of preventative measures.

Despite the many successes and exciting potential of GWA studies, there is considerable scope to further capitalise on the opportunities and secure real benefits for healthcare. Fulfilling this promise will take time and will require input from scientists across academia and industry. Moving from a statistical indication that a gene variant or region of DNA is involved in a disease, to locating and identifying causal variants and the associated biological pathways, presents a significant challenge – one that can only be met by greater integration between three historically distinct approaches to disease causality: genetic mapping, epidemiology and studies of pathophysiological mechanisms.

Additional factors that contribute to disease must be identified through detailed re-sequencing of DNA regions of interest, and work to assess the role of epigenetics and other structural variations. In turn, knowledge of individual variants must be built on with improved methods to study the impact of gene-gene and gene-environment interactions. Success will be dependent on responsible data sharing amongst researchers in academia, industry and the NHS. Mechanisms to provide genotype and phenotype data to researchers need to be developed and incentives put in place to recognise advances in translation. Effective communication between researchers and clinicians from different disciplines will be crucial to progress.

Moving forward there is a need to:

- Identify additional factors that contribute to genetic variance, including the role of rare single nucleotide polymorphisms, copy number variants and epigenetics.
- Collect samples in diverse populations for multiple diseases. These collections

should have some commonality of clinical datasets, patient consent and data access arrangements if they are to have maximum impact.

- Provide researchers with appropriate access to high quality data from prospective studies, population-based samples such as Biobank UK and disease registries.
- Encourage input from both academia and industry and facilitate collaboration and sharing of information across research disciplines.
- Invest in bioinformatics and statistical *in silico* methods to interpret sequence data and develop tools for the assessment of gene-gene and gene-environment joint effects on clinical endpoints.
- Study differences in gene expression across diverse tissue types and develop improved *in vivo* and *in vitro* models in which human causal variants can be assessed.
- Translate the wave of genetic findings on common diseases into improved diagnostics, preventions and treatments.

The first wave of GWA studies has generated a flood of data and the knowledge gained has the potential to have a major impact on medical science and healthcare. We are only in the early stages of a process that will have a major impact on our understanding of health and disease. Substantial and continued investment will be needed to ensure that the UK maintains a leading international position in this exciting area and to translate new knowledge into benefits for patients.

Introduction

The last ten years have seen a rapid expansion in our understanding of human genetic variation. At the start of the millennium the focus was on identifying shared genetic material. The Human Genome Project mapped the entire chain of base pairs in human DNA and provided a reference sequence for the 99% of the genome that is common to all individuals. As the decade has progressed, the focus has shifted to exploring the genetic differences among individuals and increasing our understanding of how genetic changes contribute to phenotypic diversity.

Within the human genome are millions of sequence variations that vary in frequency and in the size of their effect on a given disease or trait. In single gene disorders, also described as monogenic diseases, a defect in a single gene can cause the condition. In contrast, the 'genetic architecture' of common diseases is more complex and can involve the interaction of multiple genetic and environmental factors.

Genome-wide association (GWA) studies represent a powerful new tool for deciphering the link between common genetic variation and disease. The approach simultaneously interrogates hundreds of thousands of sites across the genome where individuals differ from each other. By comparing differences among individuals with a specific disease and carefully selected controls, GWA studies have successfully identified variants associated with a range of common diseases and quantitative traits such as height,[1] lipids[2] and body mass index.[3] In 2008 alone, major GWA studies were published on: Alzheimer's disease, bipolar disorder, breast cancer, coronary artery disease, Crohn's disease, multiple sclerosis, rheumatoid arthritis, stroke and type 2 diabetes.[4] The identification of variants or genetic loci associated with particular diseases offers a route to understanding the underlying biological pathways and ultimately to informing the development of new therapies.

To showcase the latest findings from this research, the Academy of Medical Sciences held a one-day symposium on GWA studies. The symposium aimed to:

- Highlight the latest research findings from GWA studies.
- Consider methodological issues relating to GWA studies.
- Identify barriers to translating GWA findings.
- Identify areas where further action may be needed to more fully understand the genetic aetiology of common disease.

The symposium did not focus in detail on the ethical issues arising from GWA studies. A separate meeting on 'GWA and ethics' was organised by the Wellcome Trust in July 2008 and reviews of the issues raised by GWA studies and genetic research more broadly are available.[5]

The symposium included presentations from leading national and international experts and was co-chaired by Sir John Bell FRS PMedSci, President of the Academy of Medical Sciences, and Professor Lon Cardon FMedSci, Head of Genetics at GlaxoSmithKline. The meeting was grouped into three sessions: GWA studies and disease pathways; science and methodology; and commercial and clinical applications.

Speakers' presentations drew on examples from across a number of disease areas to illustrate the significance of GWA studies to

[1] Weedon MN, *et al* (2008). *Genome-wide association analysis identifies 20 loci that influence adult height.* Nature Genetics **40**, 575-583.
[2] Diabetes Genetics Initiative (2007). *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science **316**, 1331-1336.
[3] Cho YS, *et al* (2009). *A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits.* Nature Genetics **41**, 527-534.
[4] An updated list of published GWA studies can be found on the National Human Genome Research Institute's *Catalog of published genome-wide association studies*. http://www.genome.gov/gwastudies
[5] For example: Kaye J, *et al* (2009). *Data sharing in genomics – re-shaping scientific practice*. Nature Reviews Genetics **10**, 331-335.

date and the potential they offer for the future. The meeting concluded with a discussion of the steps needed to enhance the design and interpretation of GWA studies and to facilitate the translation of findings into commercial and clinical applications. A symposium programme and list of attendees are annexed.

The meeting was attended by around 50 invited researchers, industry representatives, clinicians, medical funders, stakeholders and policymakers, enabling perspectives to be shared on a range of topics. We are extremely grateful to the symposium speakers and attendees for their thoughtful presentations and remarks.

This report seeks to capture key themes and issues raised during the symposium and is intended for researchers, policymakers, research funders, industry and other stakeholders. Key areas covered by presentations and discussion at the symposium that are considered in this report are:

1. Studying the genetics of common disease
2. Genome-wide association studies
3. New insights into disease biology
4. Unravelling genetic variation
5. Epigenetics
6. Designing better studies
7. The translational journey
8. How to move forward

1. Studying the genetics of common disease

Over the course of the 20th Century a combination of theoretical insights, basic science research and clinical observation fuelled a growing understanding of the genetics of disease.[6] By studying the inheritance of a condition in generations of an affected family and utilising new molecular mapping techniques, the role of genetics in a rare, sub-group of diseases was revealed. It became clear that for a group of disorders, a variation in a single-gene can be sufficient to cause the condition. For these single-gene (or monogenic) conditions, the associated genetic variation is uncommon within the population but has a large effect. Knowledge of the mechanism by which genetic factors cause single-gene disorders has provided important information about basic pathophysiological processes.

In contrast to progress in understanding single-gene disorders, much less is understood about the genetics of more common diseases. The 'genetic architecture' of common diseases is more complex and involves the interaction of numerous genetic variants, as well as environment and behavioural factors. So far, most genes identified as involved in common disease have been discovered by virtue of their large effect and high penetrance – i.e. the chance of getting the disease for those people with the mutation is high. However, these discoveries relate only to relatively rare sub-forms of common disease. Examples include mutations in *BRCA1* and *BRCA2* which increase the risk of familial breast and ovarian cancer.

Highly penetrant mutations associated with common disease have a prevalence of only one in several hundred to several thousand people. So while these rare variants have a large effect they impact on only a small proportion of cases of the disease. In contrast, the effect of more common variants is more subtle. More than 50% of the population may carry a specific genetic variant but it may only confer a slight increase in risk of disease. The frequency of these variants means that in combination with other genetic factors they play an important role in a greater number of cases, but do not have strong predictive power individually. Many efforts are now under way to increase understanding of common human genetic variation.

There are a number of ways to categorise genetic variation. Three key aspects are:

- The mechanism of variation: DNA sequence variations, such as single-nucleotide polymorphisms (SNPs), are the most common change. While there are millions of common SNPs in the human genome, there are even more rare variants as new mutations arise every generation and many of these are not passed on or do not become highly variant throughout the population. Other forms of variation include larger changes to the DNA sequence, changes to DNA structure and differences in the number of copies of a gene.
- The frequency of the variation in the population: Common variants are broadly defined as genetic variants with a minor allele frequency (MAF) of at least one percent in the population, whereas rare variants have a MAF of less than 1%.
- The size of the risk conferred by a given variant: Effect sizes are measured using an 'odds ratio' - a measure of risk that compares the probability of disease occurrence with a risk allele, with the probability in a control group. In continuously variable traits such as lipid levels, sizes are measured by how much of the observed variability they can explain.

Within the human genome are millions of sequence variations that vary in frequency and in the size of their effect on a given disease or trait. Single nucleotide polymorphisms (SNPs) are the most common form of variant, arising due a single base substitution at a

[6] Guttmacher MD & Collins FS (2002). *Genomic medicine – a primer.* New England Journal of Medicine **347**, 1512-1520.

given genetic locus. Projects such as the International HapMap Project have been crucial to cataloguing and mapping the location of SNPs and now cover approximately 25-35% of the 9-10 million common SNPs across the genome.[7] This information has had a central role in making the study of the genetics of common disease a reality and has been integral to the development of GWA studies.

[7] The International HapMap Consortium (2007). *A second generation human haplotype map of over 3.1 million SNPs.* Nature **449**, 851-863.

2. Genome-wide association studies

Genetic studies of disease have traditionally fallen into two broad categories: family-based linkage studies and association studies with candidate genes. Family-based linkage studies proved effective in identifying genes of large effect in single gene diseases such as cystic fibrosis and Huntington's Disease. The approach has been less successful in studies of complex diseases due to the involvement of multiple genes and difficulties in successfully narrowing down the linkage signal to a specific gene. To study complex traits, researchers have used association studies which look for statistical correlation between a specific genetic variant and a disease. This technique carries the potential to identify genes that do not segregate clearly in families due to the complex interplay of other genes and environmental triggers. However, candidate-genes are selected on the basis of an *a priori* hypothesis about their role in disease meaning the approach can be restricted by how much is already known of the underlying disease biology.

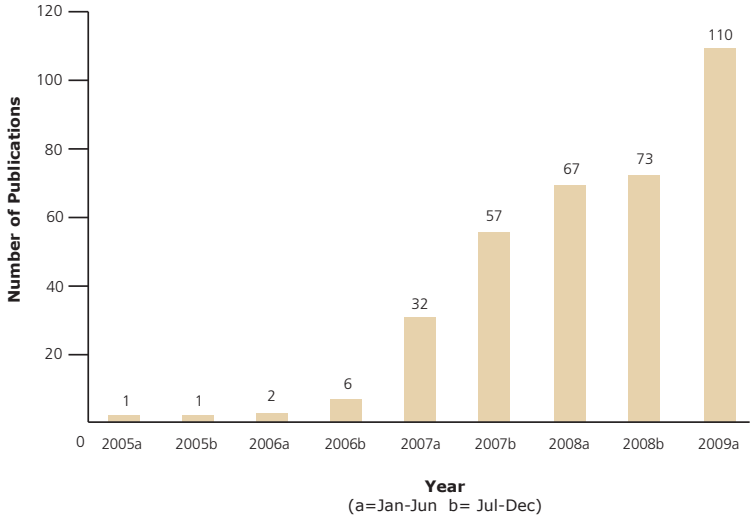
The genome-wide, non-hypothesis nature of GWA studies represents a powerful new tool. The approach has been made possible by more detailed information on the differences among individuals and improved technologies that allow the simultaneous analysis of hundreds and thousands of different positions or genetic loci. By scanning the genomes of large numbers of individuals and comparing differences among cases with a specific disease and carefully selected controls, it has been possible to identify genetic variations associated with common diseases.

A typical GWA study has four parts:[8]

- The selection of a large number of individuals with the disease or trait of interest and a suitable comparable group.
- DNA isolation and high quality genotyping.
- Statistical tests for associations between a genetic variant and the disease.
- Replication of identified associations in an independent population sample and further study of findings.

GWA studies have transformed the landscape of genetic research. As recently as 2004, few genetic variants were known to reproducibly influence common polygenic diseases. In the past three years, the number of published GWA studies has increased dramatically (see Figure 1), identifying hundreds of associations of common genetic variants with over 80 diseases and traits.[9] For the first time it has

Figure 1: Published GWA studies[10]



[8] Pearson TA & Manolio TA (2008). *How to interpret a genome-wide association study*. Journal of the American Medical Association **299(11)**, 1335-1344.
[9] An updated list of published GWA studies can be found http://www.genome.gov/gwastudies
[10] Source: National Human Genome Research Institute's *Catalog of published genome-wide association studies*. http://www.genome.gov/gwastudies (accessed June 2009).

been possible to begin to define 'molecular signatures' for complex diseases and start to decipher the link between genetic variation and common disease. Identifying and characterising the genetic variants associated with a given disease has important implications for understanding disease biology.

One of the disease traits for which the GWA approach has been most successful is type 2 diabetes (T2D). It is well established that multiple genetic, environmental and behavioural factors combine to cause T2D disease. However despite its growing global prevalence, the molecular mechanisms involved in the development of T2D are poorly understood and, despite numerous candidate genes and linkage studies, the field of T2D genetics had succeeded in identifying few genuine disease susceptibility loci. The advent of GWA studies has transformed the situation, leading to an expansion in the number of T2D loci to almost 20.[11]

In many cases these loci were previously unsuspected of playing a role in the genetic basis of T2D. While in combination these loci only account for a small proportion of the observed heritability, each associated variant is a potential new route to improved understanding of disease aetiology. Results from GWA studies have shown that genetic propensity to develop T2D seems to involve genes in several different pathways. The association of melatonin receptor 1B (MTNR1B) with T2D indicates the involvement of the circadian rhythm pathway in fasting glucose levels[12] and other research has established that common variants near the merlanocortin-4 receptor (MC4R) influence fat mass, weight and obesity risk at the population level.[13] These novel findings offer unique insights into the pathogenesis of T2D and, in the main, point towards pathways that affect pancreatic B-cell formation and function.[14] Although the associated variants exert modest to small effects on the risk of disease, this has no relationship to the potential importance of the underlying pathway and its applicability for therapeutic intervention.

The value of GWA scans in identifying common variants of small effect has been further demonstrated in studies of common cancers. A GWA study using four comparable sets of colorectal cancer (CRC) cases linked a variant that occurs in around 50% of the European population to increased CRC risk.[15] This research providing evidence for the existence of common CRC susceptibility alleles and supports the idea that variation in inherited risk of colorectal cancer is due to combinations of common, low-risk variants. By the middle of 2009, GWA studies had identified ten common genetic variants associated with colorectal cancer susceptibility, with several suggesting the involvement of components of the transforming growth factor beta signalling pathway.[16] GWA studies into predisposition to other common cancers tell a similar story, identifying multiple common variants of small effect.[17]

[11] McCarthy MI & Zeggini E (2009). *Genome-wide association studies in type 2 diabetes.* Current Diabetes Reports **9(2)**, 164-171.
[12] Prokopenko I, *et al* (2009). *Variants in MTNR1B influence fasting glucose levels.* Nature Genetics **41(1)**, 77-81.
[13] Loos R, *et al* (2008). *Common variants near MC4R are associated with fat mass, weight and the risk of obesity.* Nature Genetics **40(6)**, 768-775.
[14] Pascoe L, *et al* (2007). *Common variants of the novel type 2 diabetes genes CDKAL1 and HHEX/IDE are associated with decreased pancreatic B-cell function.* Diabetes **56**, 3301-3104.
[15] Tomlinson I, *et al* (2007). *A genome-wide association scan of tag SNPs identified susceptibility variant for colorectal cancer at 8q24.21.* Nature Genetics **39(8)**, 984-988.
[16] Tenesa A & Dunlop M (2009). *New insights into the aetiology of colorectal cancer from genome-wide association studies*. Nature Reviews Genetics **10**, 353-358.
[17] Easton DF & Eeles RA (2008). *Genome-wide association studies in cancer.* Human Molecular Genetics **17(2)**, 109-115.

3. New insights into disease biology

The impact of greater understanding of the molecular variation underpinning common diseases could be substantial. Identifying which genes are involved in a disease has the potential to provide new routes to understanding disease aetiology, but may also make it possible to: design more effective drugs and potentially reduce adverse drug reactions; identify population groups at increased risk of disease; and screen and diagnose disorders more effectively.

Success in identifying common genetic variants that predispose to common diseases has led to suggestions that these variants may be used to predict an individual's risk of disease. The major limitation for most complex traits is that the variants identified to date explain only a small proportion of variation in disease risk, limiting their prognostic and diagnostic potential. Returning to the example of T2D, despite the successes of GWA studies, the variants identified currently provide only about the same information on disease risk as traditional risk factors such as current weight and body mass index.

Rather than predicting an individual's risk of disease, results from GWA studies may have a role in predicting disease risk in population groups. By incorporating all the known variants associated with a disease it may be possible to identify sub-groups of the population at distinctly different levels of risk for that condition. Further consideration needs to be given to how this information should be used but depending on the magnitudes of risk involved and the appropriate cost-benefit calculations there is the potential to use this information to inform decisions around the targeting of screening and preventative approaches. Appropriately applied, robust GWA findings could be used, for example, to guide cancer risk profiling strategies and determine the size of the population that should be screened to identify a given proportion of cancer cases.

However, as previously highlighted, the greatest impact of GWA studies will be uncovering the biological pathways underlying polygenic diseases and traits. Even in psychiatry, where disorders can be difficult to measure and understanding of pathogenesis has been limited, early results are promising. By early 2009, GWA studies of subjects with attention-deficit hyperactivity disorder, autism, bipolar disorder, major depressive disorder and schizophrenia had all been completed.[18] These studies have shown that psychiatric disorders are amenable to the GWAS approach and offer the promise of greater understanding of the biology of these conditions.

The identification of genes in which variation appears to confer risk to both schizophrenia and bipolar disorder already challenges the assumption that these are completely distinct entities with separate underlying disease processes.[19] Further insights into disease pathogenesis will also emerge, for example, findings from completed GWA studies support a role for ANK3 (ankyrin G) and CACNA1C in bipolar disorder, suggesting that bipolar disorder is part of an ion channelopathy.[20] The identification and replication of common variation associated with autism is one further example of the impact of the GWA approach. The association of autism with a region on chromosome 5p14.1 appears to confirm the importance of CDH9/10, with research showing that CDH10 is highly expressed in fetal brain tissue, particularly in an area thought to influence speech and social interactions.[21]

[18] The Psychiatric GWAS Consortium Steering Committee (2009). *A framework for interpreting genome-wide association studies of psychiatric disorders.* Molecular Psychiatry **14**, 14-17.
[19] Hennah W, *et al* (2008). *DISC 1 association, heterogeneity and interplay in schizophrenia and bipolar disorder*. Molecular Psychiatry. Published online 4 March 2008.
[20] Ferreira MAR, *et al* (2008). *Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder.* Nature Genetics **40(9)**, 1056-1058.
[21] Ma D, *et al* (2009). *A genome-wide association study of autism reveals a common novel risk locus at 5p14.1*. Annals of Human Genetics **73**, 263-273.

Making associations between genetic variations and disease phenotypes is the first step towards developing interventions based on genetic information. This might involve identification of therapeutic targets within causal pathways or the discovery of new biomarkers, allowing improved monitoring of disease progression and treatment response. The ability to define a molecular taxonomy of common diseases and stratify populations according to genotype has the potential to:

- Make clinical trials more cost-effective and time-efficient by enrolling patients for whom the intervention is more precisely matched with their underlying biology.
- Classify diseases into sub-phenotypes based on genetic information, resulting in improved treatments and an expanded use of pharmacogenetics.

GWA studies are laying the groundwork for an era in which the current 'one size-fits-all' approach to medical care will give way to more targeted strategies. Completed studies have already proven successful in uncovering polymorphisms associated with individual differences in drug efficacy and safety. For example, a variant in the *SCLO1B1* gene has been identified as markedly increasing the risk of statin-induced myopathy, with researchers estimating that 60% of incident myopathy could be attributed to the variant.[22]

[22] SEARCH Collaborative Group (2008). *SCLOIBI variants and statin-induced myopathy – a genome wide study.* New England Journal of Medicine **359**, 789-799.

4. Unravelling genetic variation

To date, GWA studies have focused largely on understanding the pattern and nature of single-nucelotide differences within the human genome. Given the small effect sizes of the associated variants, increasing statistical power through data sharing and meta-analysis of studies has been a major feature of progress in identifying common variants (see Box 1). However, even for traits for which a large number of loci have been identified, only around 10% of the genetic variance can currently be accounted for. This raises the question of how the remainder of the genetic variation can be explained and identified?

Completed research shows that GWA studies conducted using sample sizes of around 2,000-5,000 individuals have sufficient statistical power to confidently identify common variants with an odds ratio of 1.5 or greater. It is likely that only a few, if any, common variants with modest to large effect sizes remain to be discovered for most complex traits investigated. Looking beyond common and rare SNPs, some of the missing heritability will be identified through examining other forms of genetic variation, including:

- Structural variants, including copy number variants, deletions and inversions of genetic material.
- Joint effects, including gene-gene and gene-environment interactions.
- Epigenetic modifications.

Recent studies that have identified larger polymorphisms emphasise the value of investing in more comprehensive and systematic studies of human structural genetic variation. It is estimated that structural variants underlie greater than 70% of the nucleotide bases that vary in humans, suggesting that these play an important role in phenotypic diversity among individuals. Studies have looked for associations between rare structural variants and autism and schizophrenia and have identified specific deletions involved in both of these diseases. For instance, recurrent deletions and duplications of a 600kb interval on chromosome 16 were found in multiple

Box 1 Collaboration in genome-wide association studies

Networks of collaborative GWA studies, involving multiple study samples and phenotypes, have been integral to demonstrating the power and potential of this approach:

The Wellcome Trust Case-Control Consortium (WTCCC) was able to demonstrate the effectiveness of a 'common control' design in which 3,000 UK controls were compared with 2,000 cases from each of seven different diseases. Established in 2005 and involving 50 research groups across the UK, the Consortium has identified new variants across the diseases studied.[23] The second phase of WTCCC, begun in April 2008, includes 15 collaborative studies and 12 independent studies totalling approximately 120,000 samples.

The Genetic Association Information Network (GAIN) is a public-private partnership involving six different studies with case-control or family trio designs. The Network includes four private sector partners: Pfizer, Affymetrix, Perlegen Sciences and Abbott; and one academic partner, the Broad Institute of MIT and Harvard. The GAIN policies promote broad freedom of information, by rapidly placing data in the public domain and by encouraging the initial genotype-phenotype associations to remained unrestricted by intellectual property claims.[24]

[23] Wellcome Trust Case Control Consortium (2007). *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature **447**, 661-678.
[24] The GAIN Collaborative Research Group (2007). *New models of collaboration in genome-wide association studies: the Genetic Association Information Network.* Nature Genetics **39(9)**, 1045-1051.

unrelated individuals with autism and have been estimated to account for 17% of cases.[25]

It is important that association studies involving structural variants are subjected to the same standards of quality control and replication that have been developed for SNP-based studies. It is a priority to catalogue the locations and frequencies of common structural variants and to empirically determine their linkage disequilibrium patterns across the genome. Copy number variation (i.e. individual differences in the number of copies of a particular gene or genomic region) is also likely to influence predisposition to some common diseases. Extensions of GWA studies to study copy number variation (CNV) have already resulted in discoveries of both de novo and inherited CNV that is associated with risk of common disease.[26]

[25] The International Schizophrenia Consortium (2008). *Rare chromosomal deletions and duplications increase risk of schizophrenia*. Nature **455**, 237-241.

[26] McCarroll S (2008). *Extending genome-wide association studies to copy-number variation*. Human Molecular Genetics **17(2)**, 135-142.

5. Epigenetics

Consideration of genomic variation must also include the role of epigenetic changes, modifications of the DNA or associated proteins, other than DNA sequence variation. Epigenetic changes include histone modification, positioning of histone variants, nucleosome remodelling and DNA methylation. These changes do not alter the underlying genomic sequence, but stably modify the DNA and chromatin proteins. Epigenetic processes are essential to normal development and are a key mechanism by which cells generate functional diversity.

The term 'epigenome' is used to describe the chromatin states that are found along the genome, defined for a given time and cell point. For a given genome there may be hundreds or thousands of epigenomes depending on the stability of the chromatin states. Recent years have seen the development of several strategies for genome-wide analysis of the epigenome and microarray and high-throughput technologies have been used to map chromatin modifications, cytosine methylation and non-coding RNAs across chromosomes and entire genomes.[27] High throughput application of chromatin immunoprecipitation (ChIP) is one way to study protein-DNA interaction and chromatin changes associated with gene expression.

Variation in chromatin states is highly abundant in experimental and natural populations and provides an important additional source of phenotypic variation. It is now known that there are over 28 million positions where methylation can vary (methylation variable positions or MVPs). There is a case for integrated (epi)genetic GWA studies which bring together classical sequence-based quantitative genetics and epigenome dynamics.[28] Initiatives such as the Alliance for the Human Epigenome and Disease (AHEAD), which aim to provide a high-resolution reference epigenome map, will be crucial to this goal.

[27] Down TA, *et al* (2008). *A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.* Nature Biotechnology **26(7)**, 779-785.
[28] Johannes F, Colot V & Jansen RC (2008). *Epigenome dynamics: a quantitiative genetics perspective.* Nature Reviews Genetics **9**, 883-889.

6. Designing better studies

Consideration should be given to optimising all components of the GWAS process: from the selection of case and control samples; the implementation, analysis and interpretation of studies; and the reporting of results.

The power of GWA studies can be increased by focusing on case participants who are more likely to have a genetic basis for their disease, such as early-onset cases or those with multiple affected relatives. Adopting a more stringent case definition can also guard against misclassification bias. For diseases that are difficult to diagnose reliably, ensuring that cases are truly affected (for example, by testing or imaging), will be important.

GWA studies to date have used various commercial genotyping platforms containing approximately 300,000 to 1 million common SNPs, excluding approximately 10-20% of common SNPs that are only partially tagged, or not tagged at all. A perfect tool would provide complete information at every variable point in the genome. In practice, current studies typically capture a high proportion of the information for around 65-80% of variant sites where the minor allele frequency is above 5%. Some regions of the genome are covered well, others less well, and low-frequency alleles are generally not interrogated with current study designs. Also, current panels were derived from small sets of reference samples and thus do not account well for populations with high genetic diversity.

Choices made for study design, conduct and analysis all potentially influence the magnitude and direction of results from GWA studies. Transparent reporting of results helps to address gaps in empirical evidence and to improve understanding of study design.

The Strengthening the Reporting of Genetic Association studies (STREGA) initiative builds on previous attempts to enhance the transparency or reporting, regardless of choices made during design, conduct, or analysis.[29]

In addition to optimising the case-control design that has dominated GWA studies to date, a better understanding of human genetic variation would be facilitated by:

- Transferring GWA study results to other populations.
- Resequencing to find rare variants.
- Expanding the number of cohort studies.

With rare exceptions, the GWA studies carried out so far have focused on populations of European ancestry for primary, high-throughput genotyping. However, the frequency of genetic variations differs among populations. Variants that are found to be associated with a particular trait or disease in any given population will often not be transferable for risk prediction in individuals from a different population. The discovery, using samples of East Asian origin, of diabetes susceptibility variants mapping to the *KCNQ1* gene highlights the importance of extending these studies to a wider range of populations.[30]

GWA studies have succeeded in finding common variants of relatively modest effect, almost always much less than a 2-fold increase in risk. However, if susceptibility alleles are rare and have even smaller effect sizes, then unrealistically large sample sizes are required to achieve convincing statistical support for a disease association. The GWA studies currently being conducted are therefore not able to capture the contribution made by rare variants to complex traits. Much remains to be determined about the relative contribution of

[29] Little J, *et al* (2009). *STrengthening the REporting of Genetic Association studies (STREGA) - an extension of the strengthening the reporting, of observational studies in epidemiology (STROBE) statement.* Journal of Clinical Epidemiology **62(6)**, 567-608.
[30] McCarthy MI (2008). *Casting a wider net for diabetes susceptibility* genes. Nature Genetics **40(9)**, 1039-1044.

rare variants to common complex traits and the ability to generate genome sequences of thousands of individuals in a cost-effective way will make the study of rare variants possible. Rapid cost-effective methods for sequencing entire genomes are needed to study the role of rare variants, including non-coding and structural variants.

The majority of existing studies have been case-control designs and therefore can provide only a snapshot assessment of the association of a genetic variants and a particular trait. The collection and analysis of carefully phenotyped prospective cohorts will enable researchers to study the natural progression of a disease.

Cohort studies involve collecting extensive baseline information in a large number of individuals who are then observed to assess the incidence of disease in subgroups defined by genetic variants. Although cohort studies are typically more expensive and take longer to conduct than case-control studies, they often include study participants who are more representative of the population from which they are drawn, and they typically include a vast array of health-related characteristics and exposures for which genetic associations can be sought.

7. The translational journey

The rapid growth in published GWA studies and large-scale initiatives such as the WTCCC and the HapMap Project have contributed to heightened expectations about the capacity of this research to generate tangible translational benefits. The initial wave of GWA studies has presented an unprecedented number of promising signals of association between genomic variants and complex traits. Each discovery serves as a potential starting point for future genetic and functional research. However, translation of these initial findings will take time and there is a need to validate and refine association signals, identify underlying causal variants and bridge the gap between association and mechanism.[31,32] Participants at the meeting identified a number of steps to accelerate translation of both the current findings and the anticipated future wave of data.

Identifying the causal variant

The task of moving from a confirmed association signal to the identification of the causal variant at a given locus is not straightforward.[33] Important insights can be gained from expression studies[34] and experiments are being conducted that simultaneously examine differential gene expression and genome-wide variation.[35] Publicly available expression quantitative trait locus (eQTL) data exist for a growing number of tissues. These data sets may be valuable tools for identifying whether any identified variants within the association signal have transcriptional effects. Overlap between the associated patterns with respect to disease and gene expression has the potential to highlight putative mechanisms and enable a targeted approach to resequencing and fine mapping.

It is hoped that advances in high-throughput resequencing technologies and the efforts of the 1000 Genome Project should enable progress in identifying causal variants. The 1000 Genomes Project is an international research consortium formed to create a more detailed map of biomedically relevant DNA variations at a resolution unmatched by current resources. The project involves sequencing the genomes of approximately 1200 people from around the world and receives major support from the Wellcome Trust Sanger Institute, the Beijing Genomics Institute Shenzhen and the National Human Genome Research Institute (NHGRI). Sequencing many human genomes, unselected with regard to phenotype, should provide a resource of variants to support deeper understanding of loci influencing human disease, and inform a next generation of association studies that explore rare and structural variants.

Deciphering the underlying biological mechanism

The biological pictures being revealed by GWA studies are still largely incomplete. Many of the associations identified by GWA studies do not involve previous candidate genes for a particular disease, and many associated markers are in genomic locations harbouring no known genes.

Identifying the functional basis of the link between a genomic sequence and a given complex trait presents a significant challenge – one that can only be met by greater integration between three historically distinct approaches to disease causality: genetic mapping, epidemiology and studies of pathophysiological mechanisms.

[31] McCarthy M, *et al* (2008). *Genome-wide association studies for complex traits: consensus, uncertainty and challenges*. Nature Reviews Genetics **9**, 356-368.
[32] Fugger L, Friese MA & Bell JI (2009). *From genes to function: the next challenge to understanding multiple sclerosis.* Nature Reviews Genetics **9**, 408-417.
[33] Ioannidis J, Thomas G & Daly M (2009). *Validating, augmenting and refining genome-wide association signals.* Nature Reviews Genetics **10**, 318-328.
[34] Nica AC & Dermitzakis ET (2008). *Using gene expression to investigate the genetic basis of complex disorders*. Human Molecular Genetics **17(2)**, 129-34.
[35] Cookson W, *et al* (2009). *Mapping complex disease traits with global gene expression.* Nature Reviews Genetics **10**, 184-192.

Each discovery of a biologically relevant locus is a first step in a translational journey. To move forward on this journey there is a need for:

- Informative functional and computational studies to move from gene identification to possible mechanisms that could guide translational progress.
- Relevant and functional assays for associated genes.
- Tractable animal models or highly relevant *in vitro* models in which human causal variants can be assessed.

Variation in gene expression is an important mechanism underlying susceptibility to complex disease. The simultaneous genome-wide assay of gene expression and genetic variation could provide immediate insight into a biological basis for disease associations identified through GWA studies, and help to identify networks of genes involved in disease pathogenesis. Expression data from densely genotyped human samples and covering diverse tissue types would aid researchers in their attempts to move from statistically associated variants to identifying the biological mechanisms underlying a disease. The first wave of GWA studies typically focused on individual SNPs, however, pathway-based approaches, which jointly consider multiple variants in interacting or related genes in the same pathway will become of increasing importance.[36]

[36] Wang K, *et al* (2009). *Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn's disease.* The American Journal of Human Genetics **84(3)**, 399-405.

8. How to move forward

The last few years have seen an explosion in the number of common genetic variants linked to complex traits. The GWAS approach has changed the landscape of human genetic research, linking new, often unexpected, genetic loci to a range of complex diseases. Technological advances in microarrays and high throughput genotyping have driven forward the field from testing one SNP at a time to the assessment of millions of SNPs per individual.

It is predicted that the pace of discovery will accelerate further as a result of second-generation GWA studies, follow on analyses and meta-analyses. The ability to identify predisposing or protective genetic factors has begun to provide novel insights into disease biology.

Despite the many successes, GWA studies present several challenges including: an unprecedented volume of data; difficulties explaining more than a small proportion of the genetic variation and identifying true disease pathways.

Looking forward, more still needs to be done to:

- Find additional loci that contribute to genetic variance, including beginning to decipher the impact of gene-gene and gene-environment interactions.
- Refine the location and phenotypic consequences of causal variants.
- Progress from known loci and variants to functional mechanisms.

Participants identified a number of key points:

- Identifying missing variation and building understanding of the allelic variation that underlies common disease will require:
    - Complete genome sequencing for many individuals with and with out a given disease.
    - Large samples in diverse populations for multiple diseases and traits.
    - Better methods to interrogate efficiently structural variation in large samples.
    - Improved annotation of variation across the genome, especially of non-coding regions.
    - Further assessment of the role of epigenetics (and other structural variants) in the inherited risk of disease.
- Collaboration between groups with large, well-defined sample sets has been a major feature of progress to date. Data needs to be shared across academia and industry to drive innovation and accelerate progress from genetic studies to the biological knowledge that can guide the development of predictive, preventative and therapeutic measures. Mechanisms to provide raw data to researchers need to be developed and incentives put in place to recognise advancements in translation.
- The first wave of GWAS studies has generated a flood of data; further studies will follow, looking in new diseases areas or seeking to replicate previous associations. Investment in bioinformatics is needed to put in place appropriately skilled individuals and the computation methods to interpret sequence data. This should include the tools for the comprehensive assessment of gene-gene and gene-environment joint effects.
- The majority of existing GWA studies have been based on case-control study designs and therefore can provide only a snapshot assessment of the association of a genetic variant and a particular trait. The collection and analysis of carefully phenotyped prospective cohorts is needed to study the natural progression of a disease and the interplay between genetic and environmental factors.
- Many of the best insights from GWA studies will identify difference at the cellular level. Unlocking molecular cell biology will require effective communication between researchers from different disciplines and clinicians.

- Harnessing the opportunities of genomic medicine in risk factor identification and disease prevention will require researchers' access to high quality data from prospective studies and disease registries. The potential for selective screening procedures and the stratification of patients based on molecular biology will also require continued education of patients and the general public on risk and benefit.

Underpinning all the factors described above is the need for input from both academia and industry and better collaboration and sharing of information across disciplines. The results from completed GWA studies are already providing novel insights into disease biology, with the promise of identifying new biological pathways and new drug targets. The last two years have seen exciting advances, however, we are in the early stages of a process that will have a major impact on medical sciences and health. Fulfilling the promise of GWA studies will require the coordinated input from scientists in academia and industry, research funders, regulators, policy makers and government.

Annex I symposium programme

Genome-wide association studies

Friday 31 October 2008, Wellcome Collection Conference Centre, London

09.00 **Welcome and introduction**
Professor Sir John Bell FRS PMedSci, President, Academy of Medical Sciences

09.10 **Session 1: GWA studies & disease pathways**
Chair: Professor John Todd FMedSci, University of Cambridge

- Type II diabetes - Professor Mark McCarthy FMedSci, University of Oxford
- Mental illness - Professor Nick Craddock, Cardiff University
- Cancer - Professor Ian Tomlinson FMedSci, University of Oxford

10.45 Discussion

11.00 Refreshment break

11.20 **Session 2: Science & methodology**
Chair: Dr Teri Manolio, National Human Genome Research Institute (US)

- Interpreting GWA studies - Professor David Goldstein, Duke University
- Towards integrated (epi)genetic GWA studies - Professor Stephan Beck FMedSci, University College London
- GWA research: implications for population and public health
  Professor George Davey-Smith FMedSci, Bristol University

12.55 Discussion

13.15 Lunch

14.05 **Session 3: Commercial and clinical applications**
Chair: Professor Martin Bobrow FRS FMedSci, University of Cambridge

- Impact of GWA studies on drug discovery.
  Professor Lon Cardon FMedSci, GlaxoSmithKline
- GWA studies and pharmacogenetics.
  Professor Rory Collins FMedSci, University of Oxford
- Genome variation, cancer and international consortia.
  Professor Tom Hudson, Ontario Institute of Cancer Research

15.40 Discussion

15.55 Refreshment break

16.15 **Session 4: Discussion - next steps?**

Chairs: Professor Sir John Bell FRS PMedSci & Professor Lon Cardon FMedSci

17.00 End

The Academy gratefully acknowledges the support of GlaxoSmithKline for this event

Annex II symposium delegates

Professor Tim Aitman FMedSci
Professor of Clinical & Molecular Genetics
Imperial College

Professor Stephan Beck FMedSci
Professor of Medical Genomics
University College London

Professor John Bell FRS PMedSci
President, Academy of Medical Sciences

Dr Ewan Birney
Senior Scientist
European Bioinformatics Institute

Professor Martin Bobrow FRS FMedSci
Emeritus Professor of Medical Genetics
University of Cambridge

Ms Sue Bolton
Health & Biotechnology Issues Team
Government Office for Science

Dr Laura Boothman
Policy Officer
Academy of Medical Sciences

Dr Susan Bull
Senior Researcher
Ethox Centre, University of Oxford

Professor Lon Cardon FMedSci
Senior Vice President, Genetics
GlaxoSmithKline

Professor Rory Collins FMedSci
Professor of Medicine & Epidemiology
University of Oxford

Professor Nick Craddock
Professor of Psychiatry
Cardiff University

Professor John Danesh
Professor of Epidemiology & Medicine
University of Cambridge

Professor George Davey-Smith FMedSci
Professor of Clinical Epidemiology
Bristol University

Sir Colin Dollery FMedSci
Senior Consultant
GlaxoSmithKline

Professor Peter Donnelly FRS FMedSci
Professor of Statistical Science
University of Oxford

Dr Audrey Duncanson
Science Portfolio Manager
Wellcome Trust

Dr Neil Ebenezer
Policy Manager, NHS Genetics
Department of Health

Dr Robin Fears
Senior Policy Advisor
Academy of Medical Sciences

Dr Robert Frost
Manager, FORUM with industry
Academy of Medical Sciences

Professor David Goldstein
Professor of Molecular Genetics & Microbiology
Duke University

Professor Tom Hudson
President and Scientific Director
Ontario Institute for Cancer Research

Dr Joanna Jenkinson
Programme Manager
Medical Research Council

Dr Zahid Latif
Bioscience, Medicines & Healthcare
Technology Strategy Board

Professor Mark McCarthy FMedSci
Professor of Diabetic Medicine
University of Oxford

Dr Teri Manolio
Director, Office of Population Genomics
National Human Genome Research Institute

Professor Patrick Maxell FMedSci
Head, Division of Medicine
University College London

Dr Helen Munn
Chief Executive
Academy of Medical Sciences

Lord Naren Patel FMedSci
Hon. Consultant, Obstetrician & Gynaecologist
University of Dundee

Ms Nicola Perrin
Senior Policy Advisor
Wellcome Trust

Dr Harald Schmidt
Assistant Director
Nuffield Council of Bioethics

Dr Simon Smith
Research Development Group
Astra Zeneca

Mr Laurie Smith,
Medical Science Policy, Manager
Academy of Medical Sciences

Dr Carol Symes
Senior Research Manager
Cancer Research UK

Professor John Todd FMedSci
Professor of Medical Genetics
University of Cambridge

Sir Mark Walport FMedSci
Director, Wellcome Trust

Professor Ian Tomlinson FMedSci
Professor of Molecular & Population Genomics
University of Oxford

Professor Peter Weissberg FMedSci
Medical Director
British Heart Foundation

Professor Andrew Wilkie FMedSci
Nuffield Professor of Pathology
University of Oxford

Academy of Medical Sciences
10 Carlton House Terrace
London, SW1Y 5AH

Tel: +44(0)20 7969 5288
Fax: +44(0)20 7969 5298

E-mail: info@acmedsci.ac.uk
Web: www.acmedsci.ac.uk