# Genome Wide Association Studies and Common Complex Disease

Catherine Heeney *Usher Institute, University of Edinburgh*

**Contents**



## 1. Introduction

In some cases, a single genetic change is responsible for debilitating and potentially fatal conditions. Better known examples include Huntington's disease, cystic fibrosis and phenylketonuria. Finding these genetic changes in order to intervene in the inheritance of these genetic mutations was the main business of disease genetics research until late in the twentieth century. Yet, though numerous and often catastrophic for those affected, these diseases are rare in the overall population. Over the course of their lifetime, most people will be touched not by these conditions but by common ones such as diabetes, heart disease and chronic diseases of the bowel. For these conditions, the genetic contribution has been found to be much more complex. More than one gene is involved and each of the individual genes involved when considered separately contribute in a minor way to the overall risk of developing the condition. Therefore, it is necessary to identify numerous genes before beginning to explain their contribution to these diseases.

Between 2004 and 2007 a consortium of scientists funded by the Wellcome Trust attempted to address this problem using a relatively untried method termed the Genome Wide Association Study (GWAS). This is a type of case control study involving a comparison between the DNA of those with and those without a particular condition (Brookes 1999). Using this method, the Wellcome Trust Case Control Consortium (WTCCC) set out to identify the various genes involved in a number of common complex diseases.

The results from this project appeared in the journal Nature with the title "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls" (WTCCC 2007). The WTCCC found robust correlations between genetic mutations and common complex diseases including heart disease, psychiatric illnesses, and diabetes types 1 and 2. The WTCCC was one of numerous GWA studies published in 2007 and with more than 4000 citations has been profoundly influential (Manolio 2017; see EBI 2014). The study was a catalyst to widespread uptake of the GWAS method. In the subsequent decade it was widely used for determining the multiple genes, also known as polygenes, involved in common disease.

**Please cite as**: Heeney, Catherine (2019) Genome Wide Association Studies and Common Complex Disease. Genomics in Context, edited by James Lowe, published September 2019.

Traditionally, the search for disease genes had been carried out using the candidate gene approach. This meant selecting a gene or a number of genes and trying to establish a connection to the disease or condition of interest. However, in common complex disease the relationship between individual genes and disease is often a slight one, which contributes to the difficulty of deciding on suitable candidates for testing. Using the candidate gene approach for common complex disease genes is, therefore, a similar process to finding a needle in a haystack. In contrast, GWAS is a hypothesis-free method of searching for links between diseases and genetic changes (Hunter et al 2008). A GWAS in effect performs a search across the entire genome looking for statistical signals, which indicate that there is a relationship between a given gene and a given disease.

## 2. Finding genes for common complex disease

Techniques to detect genetic inheritance or the transmission of genes were developed and refined in extended programmes of experimental research with non-human organisms – famously, fruit flies – during the first half of the twentieth century (Kohler 1994). Throughout the twentieth century genetics research on the genetics of human disease dealt almost exclusively with monogenic disease, or "classic Mendelian recessive or dominant inheritance attributable to a single gene locus" (Lander and Schork 1994: 265). Methods focussed upon cases of inheritance of conditions through generations of a family or closely related populations. For example, Victor McKusick, a clinical geneticist and his team based at Harvard University spent the 1960s and 1970s conducting research with the reproductively isolated Amish community in the United States (Lindee 2005). This population manifested high numbers of cases of Ellis-Van Creveld syndrome, a rare condition that results in shortened limbs, extra fingers and toes and in some cases serious heart defects. McKusick and his team used family pedigrees to produce visual representations of cross-generational patterns of transmission of the condition.

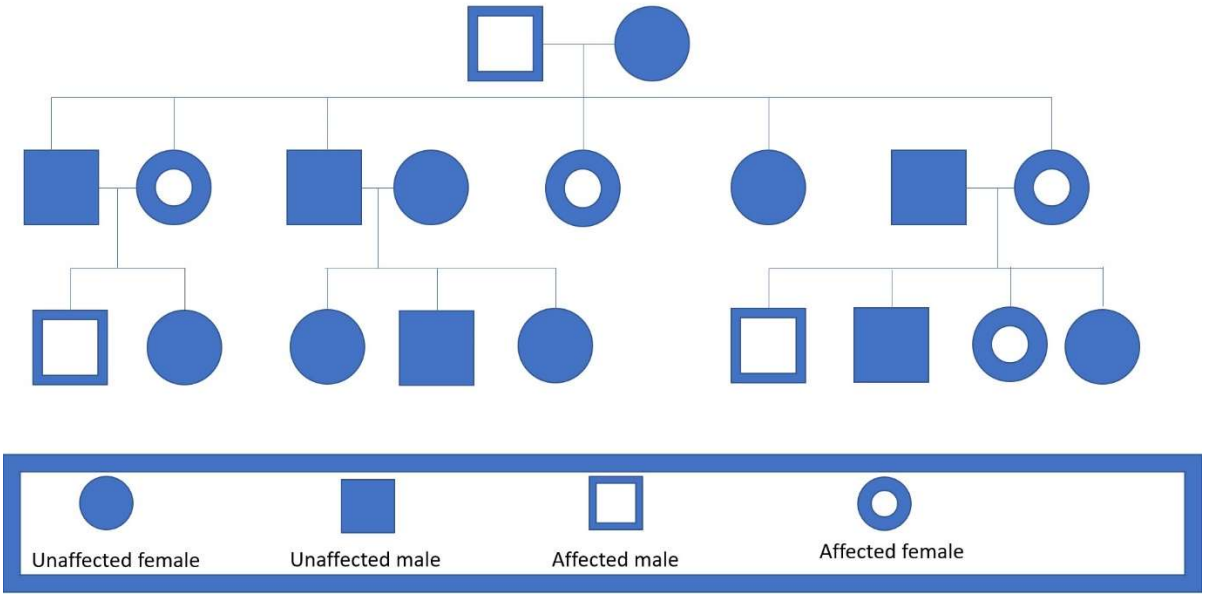


*Figure 1 - Illustration of a family pedigree diagram. Produced by Catherine Heeney.*

In the early 1980s, disease genetics acquired a new set of tools for tracing disease genes. Genetic markers allowed observation of the transmission of a gene at the level of DNA and (Botstein et al 1980). This technique used Restriction Fragment Length Polymorphisms (RFLPs), which is a genetic biomarker, to denote a place where the DNA sequence can vary between individuals of the same species. Alongside family pedigrees to map genetic locations to particular conditions, a method known as linkage mapping. RFLP linkage mapping involved geneticists searching within pre-specified blocks of DNA for the disease-causing genetic polymorphism (see Botstein and Risch 2003).

## 3. The search for common complex disease genes

Both the particular challenges of pinpointing complex genetic contributions as well as the potential benefits were familiar to those working on genetic factors involved in immune response. By the end of the 1950s, clinical researchers and immune biologists were aware of a cluster of polymorphic antigens (molecules that bind to antibodies, immune system proteins) on human leucocyte cells, that were implicated in the body's immune response (Thorsby 2009). By the 1980s, researchers agreed that genes found in a region of chromosome 6 controlled these antigens. The Human Leucocyte Antigen region or HLA at least partially regulated response to organ transplantation and blood transfusion but also autoimmunity. Autoimmune response plays a part in a number of common complex diseases. A 1967 study of Hodgkins' disease employed a tissue typing method developed by members of the HLA research community to find an association with an HLA gene, later named as HLA B5 (Bodmer 1990; Terasaki 1990).

In a 1970s study, leading HLA researcher Paul Terasaki compared patients with Ankylosing Spondylitis, a relatively common disease, which causes painful inflammation of the spine and joints, with matched but unrelated controls. Terasaki and colleagues found conclusively that the HLA-27 antigen was more prevalent in those with the condition (Brewerton et al 1973; Schlosstien 1973). HLA-27, moreover, provided an early and rare example of the use of polygenes in the diagnosis of disease (Terasaki 1990). Throughout the 1970s, research focussed on the relationship between HLA variants and juvenile diabetes, psoriasis and diabetes mellitus, with researchers concluding "that this area was a scientific gold mine" (Dausset 1990: 8). However, by the 1980s it was clear that whilst HLA genes made up a substantial part of the heritable component of a number of common conditions, they were insufficient to explain them fully. Therefore, the search for polygenes needed to expand beyond chromosome 6 to the other areas of the human genome (Risch 1987). At this point, the problems of finding polygenes involved in common complex disease came into sharp focus.

## 4. Changing methods

The linkage method had been successful in mapping "hundreds of simple monogenetic traits" (Lander and Schork 1994: 2036). However, outside of the HLA despite a notable exception in inflammatory bowel disease, throughout the 1990s linkage mapping methods were failing to detect genes implicated in common complex disease. This failure was manifested in multiple studies, which produced finds that could not be replicated in follow up studies (see Hugot 2001; Noble et al 1996). The search for polygenes and the problems associated with finding them soon raised the question of methods. Compared to many of the HLA genes, most polygenes individually account for a very modest effect upon disease development. By the early 90s, opinion began to coalesce around the

view that both linkage methods and association studies comparing small numbers of cases and controls were unsuited to finding these complex genetic causes.

One reason given for the unsuitability of linkage for finding new polygenes was its statistical inefficiency (Risch and Merikangas 1996). Whilst large numbers would be required to ensure that the relationship between a gene and common disease was not missed, linkage studies made things more complicated due to their reliance on recruiting families. Linkage needed related individuals to map the genes, which were transmitted between affected family members. The inheritance of common complex disease genes was not detectable by this method as they did not exhibit the clear modes of inheritance or co-segregation of gene and disease observable in Mendelian conditions. Still, one theory offered hope of a solution to this problem. The common disease common variant hypothesis postulated that whilst many of the genes involved in common diseases such as diabetes would have small effects in particular individuals they would nevertheless be common in the population (Reich and Lander 2001). Finding these small effects posed new statistical problems. The weak signal from polygenes needed to be amplified by large numbers of individual cases.

## 5. The promise of extending the health benefits of genetics

Despite these uncertainties, by the time of the Human Genome Project, finding interventions for common complex disease was a rhetorical goal. Francis Collins, then a leader of the Human Genome Project and John Bell the co-founder of the Wellcome Trust Centre for Human Genetics at Oxford University, talked of increasing knowledge about the genome founding interventions for common complex diseases (Collins 1999; Bell 1998). This promissory stance prompted scepticism from within the genetics community. In a highly critical article, genetic epidemiologist Neil Holtzman and health psychologist Teresa Marteau (2000) argued that even if the contributions of numerous individual genes to common complex disease were established, they would not provide any useful basis for clinical intervention. Genetic epidemiologists raised the prevention paradox, which meant that when viewed retrospectively the bulk of disease cases come from the part of the population at low to moderate risk (Clayton and McKeigue 2001). Nevertheless, in the years leading up to the completion of the draft human genome sequence in 2001, high profile figures in the field of genetics and genomics posed these difficulties as a challenge rather than an insurmountable barrier. These problems were increasingly dealt with by a focus on the limitations of existing methods for gene discovery and the search for alternatives.

In reality, the failure to replicate results in both linkage and association methods for polygenes and disease from the 1970s into the early 2000s made them useless as a basis for health intervention and was a real cause for concern (Cardon and Bell 2001; Schork et al 2000).

## 6. Strategies to replicate studies of common complex disease genes

The constraints of recruiting multiple relatives and inadequate sample size were blamed for the lack of replication (Risch and Merikangas 1996; Daly and Day 2001; Cardon and Bell 2001; Schork et al 2000). Originators of the Transmission Disequilibrium Test (TDT) attempted to tackle the issue of recruitment in the early 1990s. The TDT employs a case-control design to compare the genetic polymorphisms of affected individuals with those of their parents. It was originally used to look for non-HLA genetic mutations for insulin-dependent diabetes mellitus (Spielman et al 1993).

This improved the ability to recruit cases, as it required only a case and their parents rather than large families. It also had the advantage that both cases and controls would be from the same ethnic population. This avoided the problem of population structure, failure to account for which had produced spurious findings of relationships between genes and disease in so-called outbred populations (Spielman et al 1993). TDT was, therefore, a test of association that geneticists would potentially be comfortable with, as it both accounted for heredity and countered the dangers of hidden population structure. The TDT plays a significant part in the history of GWAS as this was the test for 'association' cited by Risch and Merikangas in their 1996 paper, which was the first to articulate the concept of the 'Genome Wide Association Study' (Tabery 2015). Later GWAS, such as that carried out by the Wellcome Trust Case Control Consortium, used unrelated cases and controls. Nevertheless, TDT provided an important methodological bridge from the use of family studies to the carrying out of association in outbred populations.

## 7. Sample size mattered

In 1999, the Wellcome Trust convened an expert working group to discuss the Human Genome Project's potential impact for health. One of the claims made at the meeting was that whilst the genome sequence would provide a helpfully precise map of the location and order of genes, large-scale studies would still be required to convert the genome sequence into reliable associations between genes and common complex disease. The reason given for this by one of the attendees was that the relationship between a specific genetic marker and a specific disease would only be present in a small sub-section of the overall study population. Moreover, the increased risk attributable to having the genetic variant is small (Campbell et al 2007). The initial study population needed, therefore, to be sufficiently large to allow for this subdivision without loss of the statistical power necessary to establish correlations (Wellcome Trust 1999).

Several thousands of research subjects would be necessary to provide such statistical power. Despite the improvements attributable to the TDT recruiting individuals and their parents was still not optimal in terms of attaining sufficient sample size. A better solution would be to conduct research in so-called outbred (not necessarily related) populations. The drawback was that where genetic differences and similarities within a population are unknown, comparisons between random individuals might contain bias. This bias results from smaller populations of different ethnic ancestry being present in the overall study population (Spielman et al 1993). Many geneticists worried that ignoring the presence of ethnic populations would lead to muddled results.

## 8. Creating the conditions for working in outbred populations

The HGP sequenced Single Nucleotide Polymorphisms (SNPs), single DNA base variants across the genome. One of the stated aims of the HapMap project that begun in 2002 after the HGP was to provide information about the different genetic variants present in particular ethnic populations, thus forming a basis to differentiate individuals based upon their DNA. HapMap thereby addressed the need to know the ethnic makeup of a study population in advance. Through linkage disequilibrium maps, HapMap established genetic similarities and differences for different ethnic groupings, which could be used to identify the presence of these groups in outbred populations. HapMap also promoted skills needed to deal with comparison of multiple SNPs (which had become the biomarker of choice following the HGP) in outbred populations (Reich and Lander 2001; Cardon and Palmer 2005). Epidemiologists and biostatisticians were already adept at controlling for

confounding variables in large, heterogeneous populations. A number of genetic epidemiologists argued that a focus exclusively on genetic factors, to the exclusion of environmental factors, would make an intractable problem into a tractable one. There was little mileage in "modelling the joint effects of genotype and environment" (Clayton and Mckiegue 2001: 1359). Association studies are comparisons based on DNA; they do not take into account environmental factors but instead focus entirely on the genetic factors involved in disease.

## 9. The SNP chip

As well as statistical theory and methods, a tiny quartz platform with attached strands of DNA contributed both to how and when the GWAS era began. By the late 1990s, several companies were involved in developing this microarray chip technology to support simultaneous testing for numerous biomarkers (Chakravarti 1999). The platform used the SNP biomarkers, which vary in a more limited way than predecessors such as the RFLP. SNPs usually have only two possible variants, one of which is common in the population and the other much less so. This simple structure meant they were suitable for use in GWAS, which requires an automated approach due to the volume of genetic variants involved. The development of microarray technology enabled the simultaneous comparison of many thousands of SNPs. The WTCCC became a celebrated early example of a GWAS, however the preliminary applications for the project submitted to the Wellcome Trust around 2004 did not suggest a full-scale GWAS but a more limited association study. This was because the technological capacity for such a thing did not yet exist.

[caption id="attachment_241" align="aligncenter" width="625"]

*Figure 2 - Technician preparing a SNP chip. Photograph by Daniel Sone. Royalty-free image obtained from: https://commons.wikimedia.org/wiki/File:Preparing_genotyping_arrays.jpg*

The biotechnology company Affymetrix produced chips supporting the processing of 500,000 genetic markers, just in time for the WTCCC project to carry out a GWAS (WTCCC 2007; Interview data from Making Genomic Medicine Project). The Affymetrix chip used in WTCCC made use of the linkage disequilibrium maps produced by the HapMap project to include so-called 'tagSNPs', which differentiated ethnic populations. Affymetrix had been working closely with scientists involved in HapMap and the WTCCC to harness developments in the field of genomics for use in research on common complex diseases. The chip had several glitches, however, which the WTCCC Analysis Group had to grapple with. One outcome of this was a new algorithm for detecting gene variants. This was named 'Chiamo' which means 'I Call' in Italian, a reference to the Italian heritage of one of the people who created it (WTCCC 2007; Interview data).

## 10. The GWAS era

Like the TDT and other more limited association studies before it, the GWAS was a case control method comparing the DNA of those with a condition to those without. As noted above, the WTCCC's methodological significance is evident both from the number of citations of the 2007 Nature paper. Former WTCCC chair Peter Donnelly stated in 2007 that the project had "shown conclusively that the new approach of analysing a large subset of genetic variants in large samples of patients and healthy individuals works" (WTCCC.org 2007). Statistical power calculations were an important discussion for genetics as the field transitioned to the use of outbred populations for GWAS. However, subsequently it seems uncontroversial that the small relationships between genes and disease could be found if data from very large numbers of individuals was available. Following the WTCCC the field of genetics seems to have entered an ever more data hungry path with large consortia pooling data resources from numerous existing studies (Burton et al 2009). The WTCCC by addressing many of the problems of determining genes for common complex disease was a new methodological era for disease genetics. However, it was soon clear that the relative risk scores based on GWAS results did not provide ground for clinical intervention (McCarthy et al 2008; Clayton 2009). Instead, it constituted another step towards that horizon.

## 11. Discussion

Despite a clear policy interest and continued scientific investment in genomics, many questions, problems and controversies remain in translating genomic and genetic science to the clinic. The ability of GWAS to help in predicting the outcomes for individual patients was questioned before, during and after large-scale GWAS were performed (Clayton 2009). Nevertheless, there were statements in the literature relating to the creation of genomic profiles for disease risk using a range of SNPs (Wray et al 2007). The results from GWAS have been argued to be impactful in indirect ways via the indication of biological pathways for example, or by allowing a risk score to be built up using several different genetic variants simultaneously (Torkamani, Wineinger and Topol 2018; Wray et al 2007). In 2007 there was great optimism that the WTCCC had laid the foundation stone for a suite of methodologies designed to tackle common disease.

The community puzzled over the issue of missing heritability: despite covering much of the genome a large part of the non-environmental component of common complex disease was still not explained. The WTCCC success also varied with diseases. Some of the autoimmune diseases showed some significant hits, especially in the HLA region, whereas others, such as hypertension, did not (Kruglyak 2008). However, GWAS made the problem of finding the genes for common complex

disease “doable”, creating an unmistakable bandwagon effect (Fujimura 1987). Arguably, the smallness of the effect on common disease of the relevant genetic mutations was what had created the conditions for GWAS. That the results from GWAS could not be directly used due to the small relative risks they revealed suggests a certain circularity to this story.

## References


Bell, John (1998) The new genetics in clinical practice, *British Medical Journal*, Volume 316 Number 7131, pages 618–620.

Botstein, David, Raymond L. White, Mark Skolnick and Ronald W. Davis (1980) Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms, *American Journal of Human Genetics*, Volume 32 Number 3, pages 314–331.

Botstein, David and Neil Risch (2003) Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease, *Nature Genetics*, Volume 33, pages 228–237.

Boyle, Evan A., Yang I. Li and Jonathan K. Pritchard (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic, *Cell*, Volume 169 Number 7, pages 1177–1186.

Brookes, Anthony J. (1999) The essence of SNPs, *Gene*, Volume 234, pages 177–186.

Chakravarti, Aravinda (2001) Single nucleotide polymorphisms…to a future of genetic medicine, *Nature*, Volume 409, pages 822–823.

Clayton, David and Paul McKeigue (2001) Epidemiological methods for studying genes and environmental factors in complex diseases, *The Lancet*, Volume 358 Number 9290, pages 1356–1360.

Clayton, David G. (2009) Prediction and Interaction in Complex Disease Genetics: Experience in Type 1 Diabetes, *PLoS Genetics*, Volume 5 Number 7, e1000540.

Collins, Francis S. (1999) Medical And Societal Consequences Of The Human Genome Project, *The New England Journal of Medicine*, Volume 341 Number 1, 28–37.

Dausset, Jean (1990) 'The HLA Adventure,' in: Paul Teraskai (editor) *History of HLA: Ten Recollections*, UCLA Tissue Typing Laboratory.

David, Chella S. (1997) The mystery of HLA-B27 and disease, *Immunogenetics*, Volume 46 Number 1, pages 73–77.

European Bioinformatics Institute (EBI) (2017) GWAS Catalog.

Fujimura, Joan H. (1987) Constructing 'Do-Able' Problems in Cancer Research: Articulating Alignment, *Social Studies of Science*, Volume 17 Number 2, pages 257–293.

Hugot Jean-Pierre *et al* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease, *Nature*, Volume 411 Number 6837, pages 599–603.

Hunter, David J., David Altshuler and Daniel J. Rader (2008) From Darwin's Finches to Canaries in the Coal Mine — Mining the Genome for New Biology, *New England Journal of Medicine*, Volume 358 Number 26, pages 2760–2763.

Lander, Eric S. and Nicholas J. Schork (1994) Genetic Dissection of Complex Traits, *Science*, Volume 265, pages 2037–2048.

Lindee, M. Susan (2005) *Moments of Truth in Genetic Medicine*, The Johns Hopkins University Press.

Manolio Teri (2007) 'Introduction, Goals of Symposium,' presentation given at 'Genome-Wide Association Studies for the Rest of Us: Adding Genome-Wide Association to Population Studies,' National Human Genome Research Institute, Boston, Massachusetts, June 2007.

Manolio Teri A. (2017) A decade of shared genomic associations, *Nature*, Volume 546, pages 360–361.

Noble, Janelle A. et al (1996) The role of HLA class II genes in insulin-dependent diabetes mellitus: Molecular analysis of 180 Caucasian, multiplex families, *American Journal of Human Genetics*, Volume 59 Number 5, pages 1134–1148.

Risch, Neil (1987) Assessing the role of HLA-linked and unlinked determinants of disease, *American Journal of Human Genetics*, Volume 40 Number 1, pages 1–14.

Risch, Neil and Kathleen Merikangas (1996) The Future of Genetic Studies of Complex Human Diseases, *Science*, Volume 273 Number 5281, pages 1516–1517.

Spielman, Richard S., Ralph E. McGinnis and Warren J. Ewens (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics*, Volume 52 Number 3, pages 506–516.

Thorsby, Erik (2009) A Short History of HLA, *Tissue Antigens*, Volume 74 Number 2, pages 101–116.

Todd, John A., John I. Bell and Hugh O. McDevitt (1987) HLA-DQβ gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus, *Nature*, Volume 329, pages 599–604.

Torkamani, Ali, Nathan E. Wineinger and Eric J. Topol (2018) The personal and clinical utility of polygenic risk scores, *Nature Reviews Genetics*, Volume 19 Number 9, pages 581–590.

Walsh, Emily C. *et al* (2003) An Integrated Haplotype Map of the Human Major Histocompatibility Complex, *American Journal of Human Genetics*, Volume 73, pages 580–590.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, Volume 447, pages 661–678.

---

**Published online**: September 2019

**Lead reviewer**: Ann Bruce

**Also participated in review process**: Miguel García-Sancho, James Lowe and Steve Sturdy