Clinical Medicine 2014 Vol 14, No 4: 428–31

Genome-wide association studies: the good, the bad and the ugly

Author: TM Frayling[A]

Since 2007 genome-wide association studies (GWASs) have identified hundreds of common genetic variants (usually single nucleotide polymorphisms [SNPs]) associated with common diseases and traits. This information is catalogued online (see www.ebi.ac.uk/fgpt/gwas for an interactive diagram). To date, there are few examples where GWAS findings have translated into useful tests that can help individual patients. Instead the common variants identified by GWASs have provided a much-needed first step in improving understanding of common disease aetiology. In this short review, I discuss some of the advantages of GWAS approaches, including the biological insights arising from GWAS, and some of the limitations. Examples I focus on include type 2 diabetes and related metabolic diseases.

What is a genome-wide association study?

GWASs typically involve the direct genotyping of several hundred thousand SNPs in hundreds to thousands of DNA samples using microarray technology. After stringent quality control procedures, each variant is analysed against the trait of interest. Typically researchers will collaborate and combine data from studies with the same disease or trait available. As more studies have performed genome-wide genotyping in the last 6–7 years, GWASs using individuals of European ancestry have reached, for example, more than 250,000 individuals for height and BMI,[1,2] 32,000 Europeans for type 2 diabetes,[3] and 190,000 individuals for lipid levels.[4] Similar GWAS efforts have been performed in other major ethnic groups, eg for type 2 diabetes; the latest studies include 7,000 east Asian,[5] 5,500 south Asian,[6] and most recently 3,800 Latin American[7] and 6,000 Japanese[8] type 2 diabetes cases.

In contrast to the candidate gene and linkage study era before 2007, where many findings in common disease genetics proved to be false positives, the vast majority of associations identified by GWASs are extremely robust statistically and are reproducible in additional studies. Most findings have been reproduced in different ethnic groups, with some differences in allele frequency, eg the variants identified as associated with lipid levels in Europeans are also associated with lipid levels in Asians and, to a lesser extent, Africans.[9] In contrast to the linkage-based approaches, such as the use of affected sib pairs, GWAS findings identify small, tractable regions of the genome where, for most loci, the likely target gene is one of only a small handful of genes – typically 2–4, although some loci may contain 0, and some 20, genes. Most of these variants lie at the opposite end of the frequency and penetrance spectrum compared with mutations that cause monogenic diseases. Typically the frequency of a risk allele is >5%, meaning that more than 1 in 10 people will often carry the disease risk allele. As even common diseases typically affect only 1–5% of the population, the vast majority of risk allele carriers are unaffected and, because common diseases are polygenic, many affected individuals do not carry any one risk allele. Although in reality there will be a continuous spectrum between allele frequency and penetrance, most genetic findings have been at the two extremes; Table 1 lists some of the key differences between common risk variants and monogenic mutations. Two related striking features of discoveries from GWASs have emerged: first is the revelation of just how polygenic common diseases are; despite the identification of 70 variants associated with type 2 diabetes, 185 with lipid levels, 40 with heart disease and 180 with adult height, the vast majority of heritability remains unexplained. Second is how small the effect sizes are and how this has been reflected in the sample sizes required to detect associations, eg most of the 180 known variants associated with height required a GWAS sample size of 133,000 individuals and replication in an additional 50,000.[2]

Genome-wide association study findings in metabolic disease

These common variant associations currently have limited relevance to clinical decision-making. There are few examples where GWAS findings have proven useful to individual patients. The vast majority of variants identified by GWASs are common and of subtle effect. With the exception of the HLA alleles in autoimmune and inflammatory diseases, and some pharmacogenetic effects, the risks (usually expressed as odds ratios) associated with common alleles are <2.0, and for continuous traits such as body mass index (BMI) usually <0.1 standard deviation (SD), eg the most strongly associated variants associated with type 2 diabetes (the variant in *TCF7L2*[10]) and coronary heart disease (the variant near *CDKN2A/B*[11]) confer risks, expressed as odds ratios of approximately 1.4 per risk allele. The variants most strongly associated with BMI and height do so with per allele effects of 0.1 SD (approximately 0.4 kg/m^2) and 1 cm respectively.[12] Most of the variants confer much smaller

Author: [A]professor of human genetics, Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, Exeter, UK

Table 1. Key differences between known common and rare disease variants.

	Rare, monogenic disease	Common disease
Genetic architecture	Monogenic by definition	Polygenic
Penetrance of risk alleles	High	Low
Risk allele frequency	Low	Typically >5 %
Location	Usually coding	Usually non-coding
Methods to detect causal allele	Usually family based	Large case–control or population-based studies

effects than these examples. Many studies have combined information from multiple associated SNPs and shown that these explain more of the phenotype.[9,13,14] However, even when combined, the variants rarely provided sufficient statistical power to offer any predictive value, eg when considering the sensitivity and specificity of information from common genetic variants, the 40 strongest type 2 diabetes variants have a receiver operator curve (ROC) area under the curve (AUC) value of 0.63,[13] where 0.5 is the same as flipping a coin and 0.8 is considered clinically useful. Age at menopause is another common trait where predictive value could be very useful for women planning families, because fertility starts to decline sharply 10 years before the menopause and women are tending to have their families later. However, the four common variants most strongly associated with age at menopause and, in one case, reducing the age at menopause by 1 year per allele still do not provide useful predictive power, with an ROC AUC of 0.60 for early menopause (<45 years).[15]

There are some common diseases where directly genotyping sets of common variants could be useful to individual patients. Multiple common variants may prove useful in identifying individuals with coeliac disease[16] and different types of rheumatoid arthritis.[17] Over the next few years, it will be worth noting how additional discoveries in the inflammatory and autoimmune diseases, where genetic effects tend to be stronger, help clinicians and their patients.

What are the advantages of genome-wide association studies?

If there is no direct benefit to individual patients, what have GWASs achieved? To further complicate this question, the common variants identified by GWASs do not necessarily represent the causal allele, and do not usually identify which gene is involved at any locus. Many tens of common variants are often strongly correlated due to linkage disequilibrium, and this means that we cannot usually say which one is causal out of the many that have been inherited together as part of the same ancestral piece of DNA (an example is given in Fig 1). Furthermore these clusters of associated common SNPs often occur in non-coding regions of the genome, deep in introns, outside genes or sometimes overlapping many genes. Dissecting which causal allele and which gene are the target of that causal allele has proven difficult. However, several lines of evidence strongly suggest that the causal gene is usually one of the two or three closest genes, eg the regions of the genome identified by a GWAS as associated with type 2 diabetes are enriched for monogenic diabetes genes such as *HNF1A*, *HNF1B* and *PPARG*, and small non-coding regions of the genome (enhancers) critical for islet-specific gene expression.[18] Regions of the genome identified by GWASs as associated with height and altered lipid levels are enriched for monogenic genes, where mutations cause severe changes in height[2] or lipid levels[19] respectively (Table 2).

Despite the limitations of GWASs, the robustness of the associations has provided an important first step in understanding biology – the genome-wide, unbiased nature of the genetic mapping efforts means that many new unexpected genes and loci have been highlighted. Here I outline several examples in metabolic disease.

The 'fat gene': *FTO* or *IRX3*?

In 2007 one of the first GWASs for type 2 diabetes discovered a common variant in an intron of the gene *FTO* which was associated with BMI and obesity.[20–22] Carriers of two copies of the minor allele (16% of the population, the allele frequency being 40%) were on average 0.4 kg/m^2 larger. Studies in children showed that this association was primarily driven by differences in adiposity.[20] Since then hundreds of studies have tried to understand the role of the FTO protein in body weight regulation, with a combination of evidence from its expression in the brain[23] to mouse transgenic studies,[24,25] providing perhaps the most robust evidence that the obesity risk allele may operate through a gain-of-function mechanism on *FTO*. However, very recently, a new study, using a combination of human and mouse data, provides a case that the next-door gene, *IRX3*, is also involved

Key points

- Disease and trait altering variants discovered by genome-wide association studies (GWAS) tend to be common and of low penetrance
- Loci identified by GWAS are enriched for monogenic genes relevant to the disease or trait
- GWAS identifies variants, not genes, but the causal genes are likely to be very close in most regions
- Unlike candidate gene and linkage-based studies GWAS approaches have provided very robust, reproducible findings
- For most diseases, we likely need to discover more variants before GWAS findings can be translated into benefits to individual patients

KEY WORDS: Genetic variants, single nucleotide polymorphisms, genome-wide association studies ■

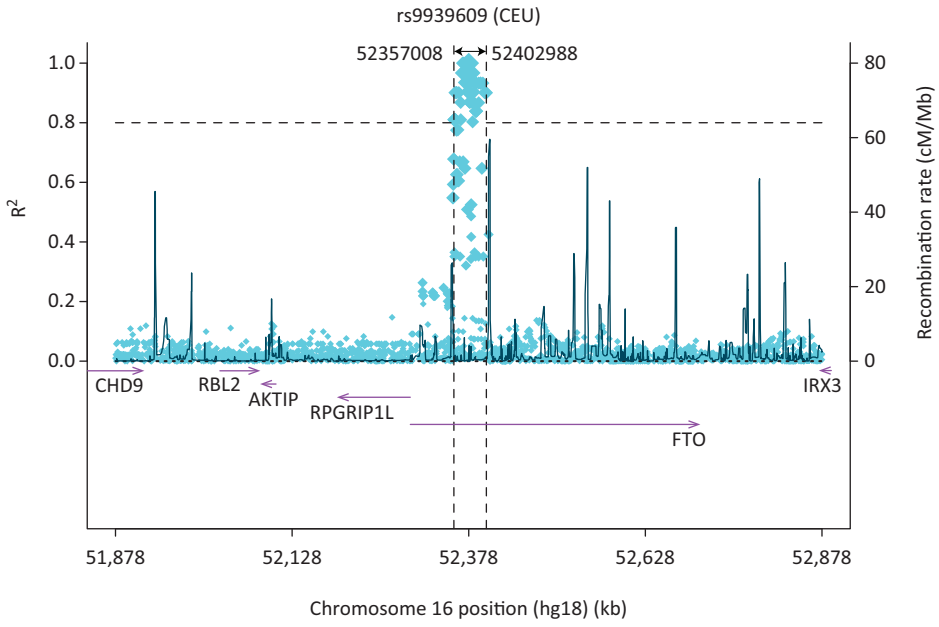

Fig 1. The region on chromosome 16 containing the *FTO* and *IRX* genes, the single nucleotide polymorphism (SNP) most strongly associated with obesity and body mass index (BMI) (rs9939609) and the correlations with other SNPs in the region caused by linkage disequilibrium.

in body weight regulation and may be the target gene of the obesity-associated SNPs.[26] Further studies are needed to establish which, or whether both, of *FTO* and *IRX3* are the causal genes. Either way, molecular studies of both genes have revealed more biological evidence for their role in body weight regulation, where previously there was none.

Genetic evidence for a metabolically obese–normal weight phenotype

Findings from GWASs have provided insight into the role of adiposity and metabolic traits in disease. These findings include that of a variant near the *IRS1* gene. Although it is not known if the variant operates through *IRS1*, it is a strong candidate given the role of *IRS1* in insulin signalling downstream to the insulin receptor. The variant was discovered through a GWAS of body fat percentage, and intriguingly the allele that predisposed to increased body fat percentage was associated with improved metabolic health, in the form of reduced circulating triglycerides, raised high-density lipoprotein (HDL), raised adiponectin levels[27] and reduced risk of type 2 diabetes.[28] The pattern of associations observed with the *IRS1* allele provides genetic evidence for the so-called 'metabolically obese–normal weight' phenotype (sometimes turned around and referred to as the 'metabolically normal–obese phenotype').[29] Researchers have hypothesised that individuals less able to store fat subcutaneously, due to reduced adipocyte differentiation or plasticity, may be thinner, but metabolically less healthy than others because they will store more fat viscerally – in the liver in particular.

Common variants identified by genome-wide association studies may modify monogenic phenotypes

Several studies have examined the effects of carrying multiple common, subtle effect alleles on monogenic disease traits. The penetrance of monogenic mutations in *BRCA1* and *HNF1A*[30] is increased in the presence of multiple common alleles predisposing to breast cancer and type 2 diabetes, respectively, eg women with a *BRCA1* mutation and carrying more common breast cancer risk alleles than average (based on alleles from seven SNPs) are at greater risk of developing breast cancer than *BRCA1* carriers with relatively few common risk alleles.[31,32] More studies may result in clinical use of this information.

Summary

Before 2007 the number of common genetic variants reproducibly associated with common diseases and traits was fewer than 20. There are now many hundreds of variants reliably associated with all types of diseases and traits, from male pattern baldness to height to common disease predisposition, including metabolic disease, autoimmune disease and germline predisposition to cancer.

Despite this success at identifying variants, the GWAS findings are not generally clinically useful to individual patients. Instead they represent a first step towards improved understanding of disease aetiology. ■

References

1 Speliotes EK, Willer CJ, Berndt SI *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 210;42:937–48.
2 Lango Allen H, Estrada K, Lettre G *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 210;467:832–8.
3 Morris AP, Voight BF, Teslovich TM *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 2012;44:981–90.
4 Willer CJ, Schmidt EM, Sengupta S *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;45:1274–83.
5 Cho YS, Chen CH, Hu C *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet* 2012;44:67–72.

Table 2. Allelic spectrum of known genes: examples of genes and loci where rare monogenic mutations and nearby common SNPs influence a very similar disease or trait.

Gene	Polygenic trait or disease	Monogenic disease
HNF1A	Type 2 diabetes	Maturity-onset diabetes of the young
HNF1B	Type 2 diabetes	Maturity-onset diabetes of the young
PPARG	Type 2 diabetes	Lipodystrophy with diabetes
Glucokinase	Stable hyperglycaemia	Stable hyperglycaemia
MC4R	Obesity/BMI	Severe obesity
POMC	Obesity/BMI	Severe obesity
ABCA1	HDL levels	Tangier's disease (low HDL)
LDLR	LDL levels	Familial hypercholesterolemia
Patched 1	Height	Gorlin's syndrome
HMGA2	Height	Overgrowth and lipoma
IHH	Height	Brachydactyly type 1A, acrocapitofemoral dysplasia
GDF5	Height	Chondrodysplasia (Hunter–Thompson type)

BMI = body mass index; HDL = high-density lipoprotein; LDL = low-density lipoprotein; SNP = single nucleotide polymorphism.

6 Kooner JS, Saleheen D, Sim X *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* 2011;43:984–9.
7 Williams AL, Jacobs SBR, Moreno-Macías H *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* 2014;506:97–101.
8 Hara K, Fujita H, Johnson TA *et al.* Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum Mol Genet* 2014;23:239–46.
9 Teslovich TM, Musunuru K, Smith AV *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010;466:707–13.
10 Grant SF, Thorleifsson G, Reynisdottir I *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 2006;38;320–3.
11 Samani NJ, Erdmann J, Hall AS *et al.* Genomewide association analysis of coronary artery disease. *N Engl J Med* 2007;357:443–53.
12 Weedon MN, Lettre G, Freathy RM *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nat Genet* 2007;39:1245–50.
13 Lango H, Palmer CN, Morris AD *et al.* Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* 2008;57:3129–35.
14 Paternoster L, Howe LD, Tilling K *et al.* Adult height variants affect birth length and growth rate in children. *Hum Mol Genet* 2011;20:4069–75.
15 Murray A, Bennett CE, Perry JR *et al.* Common genetic variants are significant risk factors for early menopause: results from the Breakthrough Generations Study. *Hum Mol Genet* 2011;20;186–92.
16 Abraham G, Tye-Din JA, Bhalala OG *et al.* Accurate and robust genomic prediction of celiac disease using statistical learning. *PLOS Genet* 2014;10:e1004137.
17 Han B, Diogo D, Eyre S *et al.* Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity. *Am J Hum Genet* 2014;944:522–32.
18 Pasquali L, Gaulton KJ, Rodríguez-Seguí SA *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* 2014;46:136–43.
19 Teslovich TM, Musunuru K, Smith AV *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010;466:707–13.
20 Frayling TM, Timpson NJ, Weedon MN *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007;316:889–94.
21 Dina C, Meyre D, Gallina S *et al.* Variation in *FTO* contributes to childhood obesity and severe adult obesity. *Nat Genet* 2007;39:724–6.
22 Scuteri A, Sanna S, Chen W *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLOS Genet* 2007;3:e115.
23 Gerken T, Girard CA, Tung YC *et al.* The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* 2007;318:1469–72.
24 Church C, Moir L, McMurray F *et al.* Overexpression of Fto leads to increased food intake and results in obesity. *Nat Genet* 2010;42: 1086–92.
25 McMurray F, Church CD, Larder R *et al.* Adult onset global loss of the fto gene alters body composition and metabolism in the mouse. *PLOS Genet* 2013;9:e1003166.
26 Smemo S, Tena JJ, Kim KH *et al.* Obesity-associated variants within FTO form long-range functional connections with *IRX3*. *Nature* 2014;507:371–5.
27 Kilpeläinen TO, Zillikens MC, Stančákova A *et al.* Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nat Genet* 2011;43:753–60.
28 Rung J, Cauchi S, Albrechtsen A *et al.* Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat Genet* 2009;41:1110–15.
29 Kramer CK, Zinman B, Retnakaran R. Are metabolically healthy overweight and obesity benign conditions?: A systematic review and meta-analysis. *Ann Intern Med* 2013;159:758–69.
30 Lango Allen H, Johansson S, Ellard S *et al.* Polygenic risk variants for type 2 diabetes susceptibility modify age at diagnosis in monogenic HNF1A diabetes. *Diabetes* 2010;59:266–71.
31 Mulligan AM, Couch FJ, Barrowdale D *et al.* Common breast cancer susceptibility alleles are associated with tumour subtypes in *BRCA1* and *BRCA2* mutation carriers: results from the Consortium of Investigators of Modifiers of *BRCA1/2*. *Breast Cancer Res* 2011;13:R110.
32 Mavaddat N, Peock S, Frost D *et al.* Cancer risks for *BRCA1* and *BRCA2* mutation carriers: results from prospective analysis of EMBRACE. *J Natl Cancer Inst* 2013;105:812–22.

Address for corresponding author: Dr TM Frayling, Genetics of Complex Traits, RILD Building, Royal Devon & Exeter NHS Foundation Trust and University of Exeter, Barrack Road, Exeter EX2 5DW.
Email: T.M.Frayling@exeter.ac.uk