BASIC SCIENCE FOR THE CHEST PHYSICIAN

Genome-wide association studies in lung disease

María Soler Artigas, Louise V Wain, Martin D Tobin

Departments of Health Sciences and Genetics, University of Leicester, University Road, Leicester, UK

Correspondence to
María Soler Artigas, 2nd Floor, Adrian Building, University of Leicester, Leicester LE1 7RH, UK; msa20@le.ac.uk

Received 1 July 2011
Accepted 18 July 2011
Published Online First
19 August 2011

ABSTRACT

Genome-wide association studies (GWAS) have provided new insights into the molecular mechanisms of lung function and lung diseases. GWAS, and studies that build upon their findings, will continue to provide evidence aimed at advancing understanding of lung disease. This paper summarises the key features of a GWAS, references some recent findings and discusses how the chest physician can interpret the validity and utility of future GWAS and related studies.

WHAT IS A GENOME-WIDE ASSOCIATION STUDY?

A genetic association study assesses the relation between genetic variation and a disease or a disease-related phenotype. Until about 2006, such studies usually included a handful of common sequence variants (single nucleotide polymorphisms, SNPs) in potentially relevant genes or regions. Although these candidate gene studies provided some robust associations, the choice of genes studied was limited by existing biological knowledge. Furthermore, the limited size and power of most candidate gene studies, coupled with liberal criteria for statistical significance, selective reporting and severe publication bias, led to a body of literature which is unreliable and difficult to interpret.

Improved cataloguing of SNPs in the human genome and an understanding of the inter-relationship between the SNPs through the HapMap project, together with improvements in genetic assays, made it possible to measure 300 000 or more SNPs spaced throughout the genome in each participant in case–control or cohort studies. These genome-wide association studies (GWAS) tend to focus on common SNPs (allele frequency >5%). As closely co-located common genetic variants tend to be inherited together over many generations, the genotype at one SNP can be strongly predictive of a nearby SNP (referred to as 'linkage disequilibrium', LD). Therefore, not all common variants need to be measured to have a fairly complete picture of common genetic variation across the genome. We can also exploit this concept to infer from GWAS data the genotypes at unmeasured SNPs—referred to as imputation—which is a valuable approach when we want to combine information across different studies, each of which has measured a different set of SNPs.

GWAS enable a hypothesis-free approach to studying association across the genome and are therefore much less prone to the reporting and publication biases which have plagued candidate gene studies. Crucially, they also have the potential to identify genes involved in biological pathways not previously connected to the disease under study.

INTERPRETATION OF A GENOME-WIDE ASSOCIATION STUDY

A GWAS typically tests association at about 2.5 million common genotyped and imputed SNPs in a first (discovery) stage and takes forward SNPs reaching a certain statistical threshold to a follow-up stage comprising independent samples. As an extremely large number of hypotheses are tested, genome-wide significance is typically defined as $p<5\times10^{-8}$ after meta-analysis of findings across discovery and follow-up stages.[1] At any SNP there are three possible genotypes representing 0, 1 or 2 copies of a particular allele. For example, a SNP which results in substitution of adenine (A) by cytosine (C) will result in three possible genotypes—AA, AC and CC corresponding to 0, 1 or 2 copies of the C allele. Association tests usually employ an additive genetic model which expresses effect size per copy of a particular allele. If dominant and recessive genetic models are also tested, then additional statistical corrections are needed to account for the greater number of hypotheses tested.

GWAS undertaken within a cohort, often to study quantitative traits, tend to be less susceptible to severe biases than case–control studies. Where cases and controls have been genotyped using different genotyping platforms, or have been genotyped in different laboratories or at different times, differential biases often result. Stricter quality control can actually make such biases worse, for example, by calling many homozygotes at a given SNP as missing in cases only. Approaches to filter rogue signals include comparing differences in missingness of genotypes between cases and controls and examining cluster plots of hybridisation intensities for each allele at a SNP to reveal possible genotyping errors. Evidence from further genotyping (validation) or from independent populations (follow-up and replication) can also help to distinguish between real and artefactual association signals.

A real association signal will tend to reflect the local characteristics of the genome, so if there are many nearby SNPs in LD, then the association signal may be manifest across all of these SNPs (figure 1A,B). While some association signals are localised, others include many genes and this may make it difficult to identify which gene is most likely to harbour one or more causal variants.

Confounding is a major concern in observational epidemiology studies, but the situation is somewhat different in genetic association studies. As

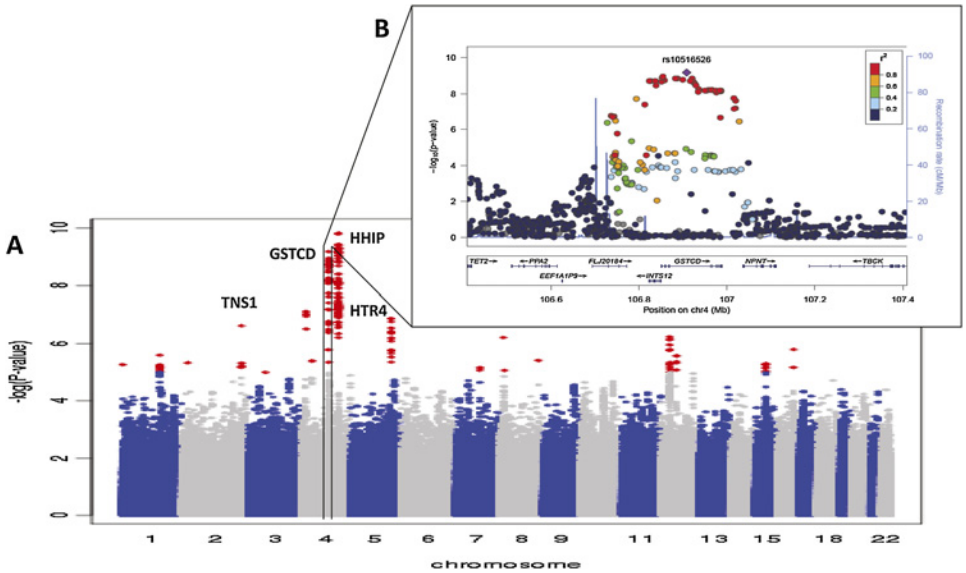

Figure 1 (A) Manhattan plot of the genome-wide association study (GWAS) of forced expiratory volume in 1 s (FEV_1) shown in the paper by Repapi *et al*.[5] Each of about 2.5 million single nucleotide polymorphisms (SNPs) is represented by a single dot. The x-axis shows the genomic location of the SNP and the y-axis its p value for association with FEV_1 (shown on a $-\log_{10}$ scale as very small numbers are hard to visualise). SNPs with p values $<10^{-5}$ (ie, $-\log_{10}$ p values >5) are highlighted in red. Association signals appear as towers (hence 'Manhattan plots') representing multiple SNPs that are highly correlated (in linkage disequilibrium, LD). (B) A region plot representing a 'zoomed-in' view of the signal at *GSTCD*, again with chromosomal location (x-axis) versus $-\log_{10}$ p value and each dot representing a single SNP. Additional information is also included on the region plot: (1) the SNP with the smallest p value is represented by a purple diamond; (2) other SNPs are coloured according to r^2 values ranging between 0 (no LD) and 1 (complete LD, perfect correlation); (3) shown in blue (right y-axis) is the genomic recombination rate: where recombination peaks, LD tends to diminish; (4) genes in the region are shown below the plot. In this case, the strongest association is seen in the *GSTCD* gene for a SNP named rs10516526. Each copy of the G allele at this SNP, which occurs on approximately 6% of chromosomes, is associated with an FEV_1 about 52 ml higher. The association is not limited to *GSTCD*, and the strength of association corresponds to the pattern of LD between the top SNP and surrounding SNPs.

genotypes are assigned at gamete formation by an essentially random process, lifestyle factors do not tend to confound genetic associations. Genetic association studies therefore need not adjust for covariates in a primary analysis. How does this affect our interpretation of the role of smoking, for example, which has a very large effect on respiratory phenotypes? A SNP which is causally associated with tobacco addiction will in turn increase susceptibility to chronic obstructive pulmonary disease (COPD), but in this case smoking is an intermediate on the causal pathway rather than a confounder. Therefore, it is preferable to adjust for smoking in a secondary analysis in order to assess whether smoking is indeed an intermediate on the causal pathway. In contrast, differences in ancestry can confound genetic associations. These ancestral differences are checked and accounted for using principal components approaches and 'genomic control'.[1]

Two key approaches have been employed to improve the power of GWAS to discover modest genetic effects. The first has been to recruit severe cases and/or refine the phenotype definition. A second approach has been to attain very large sample sizes, sometimes with some relaxation of phenotype definition, in order to discover variants that can be followed up subsequently in deeply phenotyped studies. Most of the discoveries to date relate to the latter approach. Meta-analysis of genome-wide data is often employed as a method to attain very large GWAS sample sizes. Note that meta-analysis of study level association test findings from a common analysis plan is very different from a meta-analysis of published findings. The former has been a very successful strategy for GWAS while the latter is susceptible to the severe biases present in the published literature.

LIMITATIONS OF GWAS

Unless very large and accompanied by large-scale follow-up studies, GWAS have little prospect of discovering novel associations that reach the stringent statistical thresholds of significance used to avoid false positive findings. All GWAS also require careful design and analysis to avoid the biases described above. Some robust findings from GWAS are referred to below. These explain a small proportion of the variance of the trait attributable to common sequence variation in the genome and leave unexplained variation that might be attributable to rare sequence variation or structural variation.

GWAS FINDINGS AND RESPIRATORY DISEASE

The study of genome-wide association in over 10 000 cases of asthma showed or confirmed association with *IL1RL1/IL18R1*, *HLA-DQ*, *IL33*, *SMAD3* and *IL2RB* and reported variants specific to childhood-onset asthma in *ORMDL3/GSDMB*.[2]

GWAS for COPD have reported associations near *HHIP* and in the region of *FAM13A* and the α-nicotinic acetylcholine receptors *CHRNA3* and *CHRNA5*.[3] Studies of quantitative lung function may improve understanding of susceptibility to, and severity of, COPD. To date, GWAS in over 20 000 individuals have reported associations with forced expiratory volume in 1 s (FEV_1) or FEV_1/forced vital capacity (FVC) in or near *TNS1*, *FAM13A*, *GSTCD/NPNT*, *HHIP*, *HTR4*, *ADAM19*, *AGER*, *GPR126*, *PTCH1*, *TSHD4* and *PID1*.[4] [5] A subset of these has already been shown to be associated with COPD and further reports of associations are likely as case−control studies of COPD attain larger sample sizes.

UTILITY OF GENOME-WIDE ASSOCIATION FINDINGS

GWAS have yielded novel findings for a range of common diseases and related traits. The majority of associated variants have a modest phenotypic effect (eg, RR of disease 1.1−1.2 or 1/10−1/20 SD for a quantitative trait). The major utility of such findings is therefore in what they tell us about the biological pathways involved in development and progression of disease. Indeed, GWAS have highlighted pathways that have not previously been implicated in respiratory disease. Such knowledge can inform the choice of molecular targets for the development of new drugs and, in some instances, will inform population-based preventive strategies for targeting the relevant molecular pathways.

FUTURE STUDIES

Evidence from other complex traits suggests that, with larger sample sizes, GWAS will reveal many more loci harbouring common sequence variants underlying respiratory diseases. Other studies will usefully focus on the role of rare sequence variants and structural genomic variants in respiratory disease. Understanding precisely how the molecular pathways highlighted by these genetic studies influence the development and progression of disease, and how they can be modified, provides a major challenge for the coming decade.

Funding This work was supported by the Medical Research Council (G0902313).

Competing interests None.

Contributors The literature review and drafting of the manuscript was wholly undertaken by the named authors.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES

1. **McCarthy MI,** Abecasis GR, Cardon LR, *et al*. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008;**9**:356—69.
2. **Moffatt MF,** Gut IG, Demenais F, *et al*. A large-scale, consortium-based genome-wide association study of asthma. *N Engl J Med* 2010;**363**:1211—21.
3. **Pillai SG,** Ge D, Zhu G, *et al*. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009;**5**:e1000421.
4. **Hancock DB,** Eijgelsheim M, Wilk JB, *et al*. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet* 2010;**42**:45—52.
5. **Repapi E,** Sayers I, Wain LV, *et al*. Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2010;**42**:36—44.

Pulmonary puzzle

A 67-year-old woman with fever, multiple lung opacities, visual impairment and acute respiratory failure

CASE PRESENTATION

A 67-year-old woman presented to the emergency room in acute respiratory failure. She had a 4-month history of dry cough, fever, intense fatigue, progressive dyspnoea and recent visual impairment. Her prior medical history included a mitral valve replacement in 2005 due to post-endocarditis mitral regurgitation. On examination, the patient was in hypoxemic acute distress. She was intubated and mechanically ventilated. A chest x-ray (figure 1A) just prior to endotraqueal intubation revealed bilateral air space disease and high-resolution CT of the chest (figure 1B) revealed irregular nodules, bilateral areas of conglomerate masses with air bronchograms of peribronchovascular orientation surrounded by ground glass opacities and no mediastinal lymphadenopathy. Laboratory findings indicated mild normocromic and normocytic anaemia, haemoglobin 10 g/dl, lymphopenia 168 cells/mm^3, C-reactive protein 157 mg/litre (0.0—3.0 mg/litre). Tests for HIV, serum cytoplasmic antineutrophil cytoplasmic antibody, antinuclear antibody and rheumatoid factor were negative. A trans-esophageal echocardiogram displayed a bioprosthetic mitral valve with mild stenosis, no evidence of endocarditis, and minimum regurgitation with negative blood cultures. MRI of the brain displayed bilateral optic neuritis.

An open lung biopsy was performed: angiocentric nodules comprised a mixture of lymphocytes with variable numbers of associated plasma cells and histiocytes (figure 2A). The lymphocytic infiltrate was mainly composed of small and intermediate size cells that were CD3 positive T lymphocytes. Scattered large atypical cells (figure 2B) were CD20 positive B lymphocytes. The polymorphic lymphoid infiltrate showed a distinctive predilection to infiltrate vessel walls (figure 2C), resulting in thickening of the lumens of affected vessels (figure 2D).

QUESTION

What is the diagnosis?

See page 280 for the answer

Leticia Barbosa Kawano-Dourado,[1] **Alexandre de Melo Kawassaki,**[1] **Vera Luiza Capelozzi,**[2] **Marcos Soares Tavares,**[1] **Carmen Sílvia Valente Barbas**[3]

[1]Pulmonary Division, Heart Institute (InCor), University of São Paulo Medical School, São Paulo, Brazil; [2]Pathology Division, University of São Paulo Medical School, São Paulo, Brazil; [3]Pulmonary Division, Heart Institute (InCor), University of São Paulo Medical School and Hospital Israelita Albert Einstein, São Paulo, Brazil

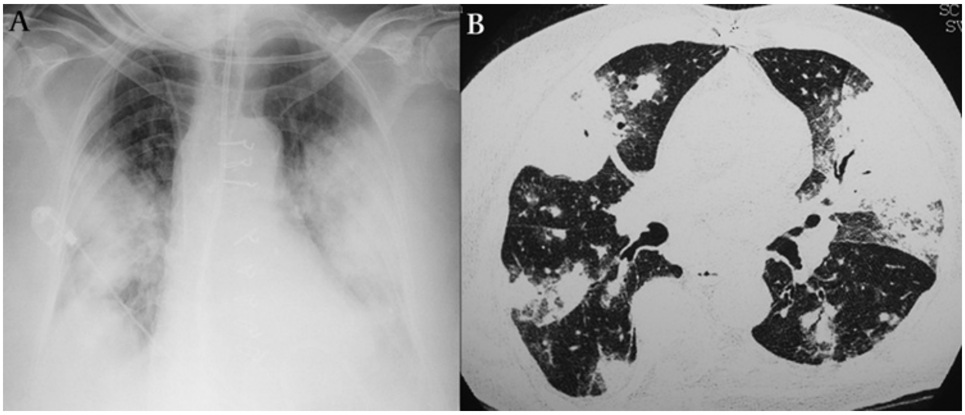

Figure 1 (A) Chest x-ray (anteroposterior view) just prior to endotracheal intubation revealed midline sternotomy metallic wires, bilateral mass-like opacities predominating in the middle and lower lung zones. (B) High-resolution CT of the chest revealed nodules, bilateral areas of conglomerate masses with air bronchograms of peribronchovascular distribution surrounded by ground glass opacities.