Do we need more Genome Wide Association Studies?

by Stephan Menzel

Haematologica 2020 [Epub ahead of print]

Citation: Stephan Menzel. Do we need more Genome Wide Association Studies?
Haematologica. 2021; 106:xxx

Publisher's Disclaimer.
E-publishing ahead of print is increasingly important for the rapid dissemination of science. Haematologica is, therefore, E-publishing PDF files of an early version of manuscripts that have completed a regular peer review and have been accepted for publication. E-publishing of this PDF file has been approved by the authors. After having E-published Ahead of Print, manuscripts will then undergo technical and English editing, typesetting, proof correction and be presented for the authors' final approval; the final version of the manuscript will then appear in print on a regular issue of the journal. All legal disclaimers that apply to the journal also pertain to this production process.
Do we need more Genome Wide Association Studies?

Stephan Menzel

Red Cell Research Unit, King’s College London

Much of the individual biological traits we have, of what we look like, of our physical and mental abilities, of our risk to suffer from the non-communicable diseases that will ultimately end our lives, is encoded in the genetic ‘background’, consisting of millions of single-nucleotide polymorphisms (SNPs) and other common sequence variants that each have minute functional effects on regulatory sequences within our genome.

Genome-wide association studies (GWAS) are the tool of choice to make the connection between common-variant genotype data, collected either through genome sequencing or with genotyping arrays (‘chips’), and human phenotype. In its simplest form, GWAS compare the frequency for each of thousands or millions of common genetic variants between groups of patients and controls, thus identifying genetic risk factors for the diseases studied this way.

There are limits to what the traditional GWAS approach can achieve. Suffocating type-I error rates arising from the analysis of millions of genetic variants make it necessary to assemble very large groups of patients and controls, but even then, only the strongest genetic risk factors can be identified with meaningful certainty. Even so, finding this initial set of genetic factors has significantly enhanced our understanding of pathways leading to common disease or shaping health-relevant physiological traits. With the majority of disease risk factors still hidden, however, it is presently impossible to assemble enough genetic information to make clinically relevant predictions for any given person. The gambit for detecting additional disease-relevant genetic factors has been to the assemble ever larger subject cohorts, reaching hundreds of thousands of participants for some conditions. Still the majority of the disease-relevant genetic background remains untouchable (1), with thousands of weak-effect alleles still hidden, trapped in what is termed the ‘missing heritability’. General frustration with the GWAS approach is prevalent among researchers.

Corre et. al (2), on page … of this issue, report a study of the type that offers a way out of this trap. The authors present a quantitative-trait association study, comparing circulating levels of the hormone erythropoietin with genotype of a genome-wide SNP set. In contrast to the original case-control setup, such quantitative-trait GWAS offer crucial advantages. They allow ‘drilling down’ into the pathways underlying biological characters and disease pathogenesis, thereby reducing complexity and increasing the signal-to-noise ratio of genetic analysis. Quantitative-trait GWAS can utilise various large subject cohorts assembled for other studies, such as groups of patients or population samples, if the parameter of interest or related biological traits have been recorded. Loci and variants discovered in quantitative-trait studies can subsequently be evaluated with more complex traits, such as disease risk.

Several large GWAS with red blood cell traits have been conducted and the genes identified have contributed to our understanding of anaemia. This has been complemented with GWAS investigation of circulating erythropoietin levels, the main hormonal regulator of the system. Unsurprisingly, the set of genes detected overlap between the two approaches, e.g., *HB51L-MYB*, which is a quantitative-trait locus (QTL) for various red-blood cell traits (HbF%, MCV, MCH, RBC), has
also shown strong association with erythropoietin levels in a 2018 Dutch population study with 6,777 participants (Grote Beverborg et al. (3)). The present study of Corre et al., while smaller, has provided confirmation of HBS1L-MYB as an erythropoietin locus and the joint analysis of both cohorts has yielded a significance level of $p < 10^{-22}$. Heritability of erythropoietin levels was found to be higher than in another previous study (by Wang et al. (4)) and the set of genes detected is also somewhat different.

In quantitative-trait studies, heritability estimates and the spectrum of loci detected is fluid, and specific outcomes depend on peculiarities of subject recruitment, trait assay method, and measurement routines. However, with multiple cohorts available to study a given parameter and its related traits, a network of quantitative-trait studies can be built that, together with knowledge gained from lab-experimental studies, paints a picture of functional and genetic architecture of the investigated tissue system and any disease risk connected with it.

The most intriguing outcome of the present paper is the detection of a putative new QTL for erythropoietin levels on chromosome 15, with evidence for trait association ($p = 1.05 \times 10^{-7}$) just short of the acknowledged level of genome-wide statistical significance. Corre et al. have started to harness data from GWAS performed with blood cell parameters in an attempt to confirm the validity of this preliminary result: in the UK Biobank study variants at this locus were found associated with erythroid traits, e.g., with haemoglobin concentration and reticulocyte count at $p < 10^{-5}$, and it is not clear why Corre and colleagues have not presented a ‘look up’ of their new locus in the erythropoietin GWAS dataset of Grote Beverborg et al. Confirmation of initial, ‘suggestive’, findings in a set of related studies must be integral part to any QTL GWAS, thus harnessing the full power of this approach. Obtaining data for that from colleagues in the field is usually straightforward.

It will be fascinating to see how this story develops following publication in Haematologica. The possibility of uncovering a new mechanism regulating oxygen transport capacity through erythropoietin is tantalising. In general, present efforts to build large population cohorts of extensively phenotyped individuals with complementary genotype data (genome array or sequence) will generate increasingly powerful datasets allowing to decipher our genetic blueprint and help to fulfil the promise of genetics for the improvement of human health.

References