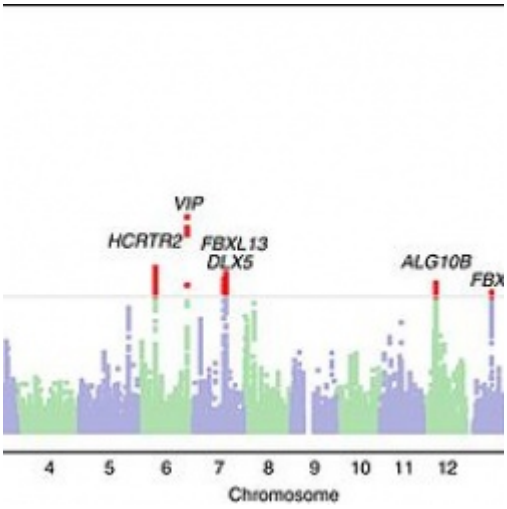
Published on *American Council on Science and Health* (https://www.acsh.org)

Genome-Wide Association Studies - ACSH Explains

By Chuck Dinerstein — July 27, 2018



Courtesy of Youna Hu, Alena Shmygelska, David Tran, Nicholas Eriksson, Joyce Y. Tung & David A. Hinds [1]

In the quest to separate nature from nurture, scientists seeking to understand the contribution of genetics have more tools. A recent paper in Nature Genetics, Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals, is one example, and I have written about several others here and here [2]. Genome-wide Association Studies or GWAS are increasing in the scientific literature and beginning to leak into the mainstream media. But how do these studies work, are they useful, how can I tell the good from the bad? Here's a guide for those of us, like me, who were in school a bit after Mendel but before CRISPR.

Introduction

Think of the human genome as a very lengthy book. As we repeatedly type the book, we make mistakes often in the same place and of the same form, like hitting the g instead of the t. The genetic equivalent of those typographical errors are single-nucleotide polymorphisms or SNPs, substituting one base pair for another. SNPs have many uses in science, but for GWAS studies they represent place markers on the genome providing a roadmap for the genome, marking an individual's genetic variations. A GWAS looks at genetic variations characterized by SNPs to see "whether on ***average***, [they are] associated with higher or lower levels of some outcome." [1] Note the use of the word averages, at heart GWAS are statistical analysis and reveal association, not causation.

These studies share a common underlying hypothesis, common disease-common variant. In brief,

when looking at a population, common diseases will have common disease genetic variants. Again, we need to carefully understand our words, these are population studies and tells us little about an individual. GWAS are scientific explorations, providing insight on where to dig deeper, where discovery is to be made; they are not a form or forum for “eugenic evidence.”

Creating the GWAS

GWAS studies begin by carefully defining a phenotypical outcome. In some instances, this is easy because of readily available quantitative measures, for example, height or blood levels of LDL. In other cases, there are no clear quantitative guidelines, and the researchers need to define the metrics of the phenotype. In the educational attainment study, years in school was the measure; GWAS of behavioral outcomes will more commonly have researcher specific definitions of outcome. The phenotype’s definition, whether quantitative or qualitative, is the first limitation on the scope of the study. The phenotypic characterization will be more problematic when the holy grail, the use of electronic medical records (EMR) to define a population are used. EMRs have no “data dictionary” and physicians may use varying definitions to describe the same problem, e.g., does hypertension begin at 130 or 140 mmHg systolic pressure?

With phenotype in hand, researchers work with well described quality controlled genomic data for a large population. Without reliable genetic data, there are no reliable results. Typically in a GWAS, part of the population’s genetic data is used to create, through statistics, a genetic profile associated with their phenotypic outcome. The remaining members of the population are used to test the profile’s performance. The populations involved have to be large, like the 1.1 million in the educational attainment study, because the profile’s predictive value increases as the population sampled grows, and you still need a good size test population for the actual research.

Researchers begin with each SNP, determining whether it is statistically associated with the outcome. But a single SNP contributes little on its own, genetic effects are primarily due to lots of SNPs. So they iterate over all the statistically “significant” SNPs until they have fashioned a stable profile. Statistically combining thousands of SNPs each with their own p-value of 0.05 is increasingly unable to separate true from false. So GWAS studies correct for this by significantly lowering the accepted p-value in the hopes of reducing the false positives. Typically a correction is made to the acceptable p-value which rather than 0.05 is, in a typical GWAS study, 0.0000001. [2]

The profile of genetic variation, sometimes further refined and termed a polygenic score, is then used to predict the outcomes for the remaining population, the test set, uncovering the genetic effect. Let’s take a moment to unpack that sentence. Predict conjures up describing the future with some significant certainty; a weatherman predicts the future and does so more reliability than a GWAS study. In the context of a GWAS, predict refers to seeing an association in the test set, not in the future and only for their defined population. And prediction is merely a statistical threshold; the genetic profile can be weak or strong. Scientists use the term “effect size” to describe the magnitude of the profile’s predictive abilities. But do not let the word effect fool you, GWAS talk about the prediction and strength of an association, they do not explain causation.

Sources of errors

The initial errors may lie in the definition of the phenotypical outcome or the quality of the genomic

data used in the analysis. SNPs are mile markers and SNPs are often found in similar regions of the genome. The presence of multiple SNPs in a limited area, when treated statistically may contribute more noise than signal to the genetic profile or polygenic score. More importantly, many genetic variations vary in a systematic way across environments. An environment in this usage can refer to the population being studied, and in the case of behavioral outcomes, like educational attainment, the social context of the behavior. Environment, particularly the population researched, is the second great limitation on the scope of these studies.

Limitations

- GWAS are population studies, while based on individual genomic variation, their only strength is in the aggregate. They do not tell anything meaningful about the individual.
- GWAS findings are limited to the population under study and the definition of the phenotype; they are not readily generalizable to other environments or phenotype definitions. Because the effect of genetic variation may be indirect, acting through alternate known or unknown pathways, a different setting or phenotype may result in greater or lesser effects.
- GWAS phenotypes matter, greater effect size is found when the outcome reflects quantified physical and chemical attributes. Behavioral outcomes show far less effect size. In the educational attainment study, the effect size was 11 to 20%, for studies of height effect size can be as high as 40%.
- GWAS describe the effects of multiple genetic variants, each making a tiny contribution. There is rarely a single gene involved.
- GWAS are exploratory studies; they provide clues of where to look, not proof.

Final thought

The researchers in the education attainment paper and that includes academics and scientists working for 23andMe make a point the media fails to mention.

“We recognize that returning individual genomic “results” can be a fun way to engage people in research and other projects and to stoke their interest in, and educate them about, genomics. But it is important that participants/users understand that these individual results are not meaningful predictions and should be regarded essentially as entertainment. Failure to make this point clearly risks sowing confusion and undermining trust in genetics research.”

GWAS are a powerful tool helping us understand the relationship between nature and nurture, but as with any tool, it has both limitations and moments of great utility. Hopefully, you can navigate between the two and understand for yourself what the science is telling us.

[1] Other markers of genetic variations besides SNPs can be used, but the methodology remains the same.

[2] This particular formula is called the Bonferroni correction and lowers the p-value based upon how many statistical tests were performed; p-value/number of statistical tests conducted. There are other formulas used to correct for multiple statistical testing, but a GWAS has to use some form of correction to be statistically valid.

Sources: The original spark for this explainer comes from Gene Discovery and Polygenic Prediction from a genome-wide association study of educational attainment in 1.1 million individuals Nature Genetics DOI: 10.1038/s-41588-018-0147-3. It comes with an excellent explainer [3] of its own which helped me understand and translate an interesting study. I also made use of a chapter [4] on GWAS studies found on PLoS.

Links
[1] https://upload.wikimedia.org/wikipedia/commons/0/05/Manhattan_plot_of_the_GWAS_of_self-reporting_of_being_a_morning_person.jpg
[2] https://www.acsh.org/news/2018/07/02/genetics-or-lifestyle-where-cause-disease-13139
[3] https://www.thessgac.org/faqs
[4] http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822#s3