# **Concepts and relevance of genome-wide association studies**

ANDREAS SCHERER and G. BRYCE CHRISTENSEN

Dr Andreas Scherer is President and CEO of Golden Helix, Inc, a leading DNA analytics company. He holds a PhD in Computer Science from the University of Hagen, Germany. He is the author and co-author of over 20 international publications and has written books on project management, the internet and artificial intelligence.



He may be contacted at E-mail scherer@goldenhelix.com



Dr Bryce Christensen holds a PhD in Biomedical Informatics from the University of Utah in Salt Lake City, USA. Bryce has extensive professional experience with statistical analysis of whole-genome data from both microarray and next-generation sequencing platforms.

## ABSTRACT

Genome-wide association study (GWAS) technology has been a primary method for identifying the genes responsible for diseases and other traits for the past 10 years. Over 2,000 human GWAS reports now appear in the scientific journals. The technology is continuing to improve, and has recently become accessible to researchers studying a wide variety of animals, plants and model organisms. Here, we present an overview of GWAS concepts: the underlying biology, the origins of the method, and the primary components of a GWAS experiment.

Keywords: genome-wide association study, genetics, statistics, gene discovery

## 1. Introduction

The science of genome-wide association studies (GWAS) emerged about a decade ago as a powerful scientific tool to identify genes associated with the outward traits of an organism. GWAS has been developed as a primary method for the identification of disease susceptibility genes in humans. More recently, GWAS technology has experienced rapid growth in non-human applications,

particularly in the burgeoning field of agrigenomics<sup>1</sup>. High-throughput genotyping arrays are now available for numerous animals and crop species, enabling scientists and plant and animal breeders to improve breeding programs and food production through genetics. As more researchers are moving into this field, we would like to share some of our experience and give an overview of the key concepts underlying GWAS.

#### 2. The biology of GWAS

GWAS was initially developed to study the human genome. The human genome is a sequence of more than three billion DNA bases that can be represented by one of four letters: A, C, G or T. Much of the genome sequence is identical or highly conserved across the human population, but every person's genome is unique. A given person's genome sequence is likely to differ from the standard human reference genome at more than three million positions<sup>2</sup>. New mutations are introduced to the genome with every passing generation, and there are many old mutations that are now widely observed among all populations. These common mutations are generally called variants or polymorphisms.

The most common type of variants is the single-nucleotide polymorphism (SNP) which describes changes to an individual DNA base. The different forms of the same gene containing variable SNPs within the same site(s) are typically called alleles. GWAS methods are chiefly concerned with determining alleles associated with various SNPs in each study subject, and making statistical comparisons to identify SNPs or genes associated with a particular trait. If a certain allele is more common among individuals with disease than other healthy ones, this is interpreted as evidence that this allele or perhaps another nearby variant may cause the disease or at least increase the risk of disease.

Most SNPs result from one historical mutation event<sup>3</sup>. Because of this ancestry, each new allele is initially associated with the other alleles present on the particular chromosomal background where it arose. The specific set of alleles observed together on a single chromosome, or part of a chromosome, is called a haplotype. New haplotypes are formed by additional mutations or by chromosome recombination (also called crossing-over) during meiotic cell division. Haplotypes tend to be conserved, especially among individuals with recent shared ancestry (see Figure 1). Haplotype conservation is a very important factor for GWAS. The genetic variant that causes a particular trait may not be directly tested in the GWAS, but its signature may still be evident through the association of SNPs occurring within the same haplotype (see Figure 2). The nonrandom co-occurrence of alleles within a chromosome or haplotype is called linkage disequilibrium, or LD. The degree of LD in a population is shaped by selection, recombination rate, mutation rate, consanguinity and other factors.

60



**Figure 1** Haplotypes and recombination. This figure shows a small region of human chromosome 19. Genes and chromosome physical map coordinates are shown in the middle. The line tracing at the top shows the recombination rate determined from HapMap data – the peaks represent "hotspot" locations for meiotic recombination. The triangular plot in the lower section illustrates linkage disequilibrium (LD) patterns among SNPs in the region, with strong LD shown in red. LD measurements are based on genotypes from 649 individuals of European ancestry. Note that high LD is confined to regions of minimal historic recombination, and does not extend across the recombination hotspots.



**Figure 2** Genotypes and haplotypes. This illustration depicts four SNP loci in the genomes of three subjects. Each subject has two haplotypes, corresponding to the two copies of each chromosome typically present in human cells. Suppose that the C allele at SNP-2 causes a certain trait, but that SNP is not genotyped. The G allele at SNP-3 always occurs on the same haplotype with the causal allele, and if genotyped may serve as a proxy for the causal allele in GWAS tests. Further inspection shows that the causal allele always occurs on the A-C-G-A haplotype, and may also be detected using haplotype association testing.

#### 3. The origins of GWAS

GWAS became possible as the result of several scientific advances early in the 21st century. The completion of the Human Genome Project greatly improved our knowledge of the human genome and provided a much better context for the study of genetic variants<sup>4</sup>. The International HapMap project, which completed its first phase in 2005, conducted an unprecedented SNP discovery initiative and provided the first detailed human haplotype and LD maps<sup>5</sup>. These scientific efforts made it possible to identify relatively small numbers of SNPs capable of representing most of the common variation in the human genome. The GWAS era was born as biotechnology companies including Affymetrix, Illumina and Perlegen launched competing platforms to simultaneously genotype hundreds of thousands of SNPs.

The National Human Genome Research Institute and the European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog<sup>6</sup> recognise a 2005 analysis of age-related macular degeneration (AMD) as the first GWAS study. This study analysed about 100,000 SNPs in just 146 subjects, and identified the *cfh* gene as a major AMD risk factor<sup>7</sup>. Since then, GWAS has grown to produce hundreds of published reports each year. The volume of published human GWAS studies has plateaued in recent years, but the average size of the study cohorts continues to grow (Table 1). The largest GWAS studies today may include over 100,000 subjects.

	Primary GWAS cohort size			Replication cohort size, if used		
Year	No. of studies	Mean of genotyped subjects	Median of genotyped subjects	No. of studies	Mean of genotyped subjects	Median of genotyped subjects
2005	2	738	738	1	664	664
2006	8	862	821	5	3816	1584
2007	89	2454	1094	63	5957	2519
2008	147	5100	1983	114	9619	4981
2009	235	5748	1984	182	8060	3311
2010	330	7360	2383	223	10733	3835
2011	390	6881	2643	279	9390	3491
2012	382	7575	2662	256	9811	4000
2013	376	8708	2243	252	11276	3609

	Table 1	Growth	of GWAS
--	---------	--------	---------

62

This table shows the number of unique human GWAS papers published per year from 2005 to 2013 according to the NHGRI-EBI GWAS catalog, together with the mean and median number of genotyped subjects analysed. The number of those reports that included an independent replication cohort is also shown, together with the mean and median number of genotyped samples analysed in the replication stage.

#### 4. The practice of GWAS

GWAS studies can be designed to assess the genetic determinants of almost any qualitative or quantitative trait. Several issues must be considered in GWAS study design, including the selection of a genotyping platform, sample size and collection, statistical analysis plans, statistical power, correction for multiple testing and population structure.

Genotype data for GWAS are usually produced with microarray technology allowing the detection of polymorphisms within a population. Microarrays involve three basic principles:

- 1. The array contains immobilised allele-specific oligonucleotide probes, which are short pieces of synthesised DNA complementary to the sequence of the target DNA.
- 2. Fragmented nucleic acid sequences of the target, labelled with fluorescent dyes.
- 3. A detection system that records and interprets hybridisation signals measuring essentially genetic similarity.

There are many different microarrays or "chips" available for both human and non-human applications. Some chips are designed to test as many SNPs as practically possible-currently up to about five million. Some chips are specifically designed to test SNPs in coding regions of genes, which make up about 2% of the genome. Other chips may test relatively small numbers of SNPs that have been carefully selected to efficiently represent worldwide haplotype diversity. Some chips are designed for specific ethnic groups or may be enriched with SNPs from genes implicated in particular diseases. In selecting a genotyping chip, it is important to consider the goals of the current project, compatibility with data from past or planned future studies, and the budget available.



**Figure 3** Example Manhattan Plot. GWAS results are often visualised by plotting P-values on a logarithmic scale. The values are plotted in linear order based on the chromosomal locations of the SNPs. This type of figure is commonly called a "Manhattan Plot," alluding to its similarity with a city skyline. In the example above, the most significant SNP in the GWAS is on chromosome 6, with the highest  $-\log_{10}$  P-value of 6.68 in the plot.

The next endeavour required for an effective GWAS study is the collection and recording of the desired phenotype, which can be quantitative (integer or real-valued) or dichotomous (case–control). Quantitative traits can provide more statistical power to show a genetic effect, but the case–control study design can also be effective in identifying multiple genes associated with the phenotype. We can see examples of each in the literature<sup>8–10</sup>

The statistical analysis of genome-wide association can begin once samples have been collected and genotyped. The process begins with a thorough quality control analysis to confirm accuracy of the genotype data<sup>11</sup>. A statistical hypothesis test is performed for each SNP, with the null hypothesis of no association with the phenotype. There are a number of association tests available depending on which type of trait is being tested. Quantitative traits are generally analysed using linear regression approaches with the assumptions that the trait is normally distributed, variance within each group is the same, and the groups are independent. Popular analyses include ANOVA and GLM. Binary traits are commonly analysed using logistic regression, or tests such as a  $\chi^2$  or Fisher's Exact Test; logistic regression is popular because it allows adjustment for other covariates<sup>12</sup>. Specialised tests are available for study designs with family-based collection<sup>13</sup>.



**Figure 4** Statistical power in GWAS. Power was estimated using the PBAT<sup>22</sup> Power Calculator implemented in Golden Helix SNP and Variation Suite (SVS)<sup>23</sup>. Both figures show the statistical power to detect a true association for a dichotomous trait with significance level P < 5e-8, using an additive genetic model when the true mode of inheritance is also additive. Power is estimated using a simulation procedure for disease allele frequencies between 0.01 and 0.49. Power generally increases when the disease allele has higher frequency. The figure on the left shows the effect of increasing sample size when the effect size is held constant; OR1 (the odds ratio associated with having one copy of the disease allele versus no copies) is fixed at 1.5. The figure on the right shows the power difference to detect causal alleles with various effect sizes. The sample size in this figure is fixed at 1,000 cases and 1,000 controls.

64



**Figure 5** Principal Components Analysis. This figure shows the first (PC1) and second (PC2) principal components of the GWAS data for a group of samples with European ancestry. The samples are clearly stratified by ancestry and nationality. Samples are coloured according to ancestry and geography: CEU, Utah residents (CEPH) with northern and western European ancestry; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian population in Spain; TSI, Toscani in Italy.

Statistical power and multiple test correction are important and inseparable issues for GWAS. False positive associations are a great risk when testing large numbers of SNPs, so statistical evidence for association must be held to a high standard. The typical significance threshold used in human GWAS studies is *P*-value less than 5e-8, equivalent to a standard Bonferonni correction for one million independent tests<sup>14</sup>. Populations with greater genetic diversity, such as African populations, may require even greater stringency to determine that a test result is statistically significant. Very large sample sizes may be required to achieve such significance levels, especially for rare disease alleles and alleles with small effect sizes (Figure 4). Statistical power of GWAS is affected by many factors, some of which are beyond the investigator's control. These factors include: complexity of the genetic architecture of the phenotype, frequency and effect size of the disease allele, accuracy of phenotypic measurements and homogeneity of the phenotype, and LD relationships between causal variants and genotyped SNPs<sup>15</sup>.

Standard GWAS test statistics assume that all samples in the analysis are unrelated and selected from a uniform, random-mating population. Any departure from this assumption can cause unexpected results, especially in large study cohorts (groups of subjects encountering a certain event during a particular time period). For example, if individuals of a certain ethnicity are overrepresented in the control group of an experiment, the significance of test results throughout the genome may be consistently inflated due to the unique genetic background of that ethnic group. Principal components analysis (PCA) can be used to stratify subjects based on genomic similarity, and is often used to assess population stratification in GWAS cohorts (Figure 5). It is a common practice to adjust GWAS tests for principal components in order to account for the structure of the population. An alternative to PCA-based correction is to account for pairwise allele sharing among all study subjects using mixed linear model (MLM) regression<sup>16</sup>. MLM methods such as EMMAX<sup>17</sup> and GEMMA<sup>18</sup> effectively account for population structure in both human and agricultural populations.

#### 5. Beyond GWAS

GWAS is sometimes called a "hypothesis-generating" process<sup>19</sup>, as it is often the first step toward understanding the genetic architecture of traits. A successful GWAS will result in one or many SNPs found to be associated with the trait of interest. Researchers may then evaluate the functional consequences of each associated SNP, examine other variants in LD with that SNP, study the function of the gene where the SNP resides, and study the biological pathways in which the gene participates. Indeed, a great number of experiments may be required to fully understand the results of a GWAS. As the biology of the trait is elucidated, it may be possible to develop assays to test for disease risk or to improve disease treatment and prevention programs.

The first decade of GWAS provided many success stories, but debates continue about how to improve GWAS<sup>20</sup>. Many approaches have been proposed to increase statistical power, reduce false-negative rates, and incorporate biological context in GWAS results<sup>21</sup>. The coming years are likely to see continued innovations in both technology and analytic methods to make GWAS an even more effective and efficient method to study the underlying biology of diseases and other traits.

#### 6. References

66

- 1. Gurgul, A., Semik, S., Pawlina, K., *et al.* (2014) The application of genome-wide SNP genotyping methods in studies on livestock genomes. *J. Appl. Genet.*, **55**, 197–208.
- The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1092 human genomes. *Nature*, 491, 56–65.

- Francioli, L.C., Polak, P.P., Koren, A., *et al.* Genome of the Netherlands Consortium, van Duijn, C.M., Swertz, M., Wijmenga, C., *et al.* (2015) Genome-wide patterns and properties of *de novo* mutations in humans. *Nat. Genet.*, 47, 822–826.
- 4. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
- 5. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
- 6. Welter, D., MacArthur, J., Morales, J., *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001-6.
- Klein, R.J., Zeiss, C., Chew, E.Y., et al. (2005) Complement factor H polymorphism in agerelated macular degeneration. Science, 308, 385–389.
- Teslovich, T., Musunuru, K., Smith, A., *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466, 707–13.
- Habek, M., Brinar, V. and Borovecki, F. (2010) Genes associated with multiple sclerosis: 15 and counting. *Expert. Rev. Mol. Diagn.*, 10, 857–861.
- Bush, W. and Moore, J. (2012) Chapter 11: Genome-Wide association studies. *PLOS Comput. Biol.*, 8 (12), e1002822.
- Laurie, C.C., Doheny, K.F., Mirel, D.B., *et al.*; GENEVA Investigators. (2010) Quality control and quality assurance in genotypic data for genome-wise association studies. *Genet. Epidemiol.*, 34, 591–602.
- Clarke, G.M., Anderson, C.A., Petterson, F.H., *et al.* (2011) Basic statistical analysis in genetic case-control studies. *Nat. Protoc.*, 6, 121–133.
- Eu-ahsunthonwattana, J., Miller, E.N. and Fakiola, M., Wellcome Trust Case Control Consortium 2, Jeronimo, S.M.B., Blackwell, J.M., Cordell, H.J. (2014) Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLOS Genet.*, 10, e1004445.
- Pe'er, I., Yelensky, R., Altshuler, D. and Daly, M.J. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.*, 32, 381–385.
- Sham, P.C. and Purcell, S.M. (2014) Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.*, 15, 335–346.
- Vilhjálmsson, B.J. and Nordborg, M. (2013) The nature of confounding in genome-wide association studies. *Nat. Rev. Genet.*, 14, 1–2.
- Kang, H.M., Sul, J.H., Service, S.K., *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.
- Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, 44, 821–824.
- Stranger, B.E., Stahl, E.A. and Raj, T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187, 367–383.
- Manolio, T., Collins, F.S., Cox, N.J., et al. (2009) Finding the missing heritability of complex diseases. *Nature*, 461, 747–753.
- Marjoram, P., Zubair, A. and Nuzhdin, S.V. (2014) Post-GWAS: where next? More Samples, more SNPs or more biology? *Heredity*, 112, 79–88.
- Lange, C., DeMeo, D., Silverman, E.K., et al. (2004) PBAT: Tools for family-based association studies. Am. J. Hum. Genet., 74, 367–369.
- 23. Golden Helix Inc., Bozeman, Montana, USA.