A Review of Genome-Wide Association Studies

David M. Budden
Department of Electrical and Electronic Engineering
The University of Melbourne, Australia
Parkville, Victoria 3052
Email: dbudden@student.unimelb.edu.au

I. GENOME-WIDE ASSOCIATION STUDY

A. Monogenic versus Complex Disorders

It is widely accepted that the phenotypic traits of an individual, whether they be individualising physical characteristics or susceptibility to a particular disease, are heavily influenced by their genome. In the case of Mendelian (monogenic) disorders, such as haemophilia and muscular dystrophy, there is a strong causal relation between disease and single nucleotide polymorphisms (SNPs), which are documented mutations of a single nucleotide (T,C,A,G) relative to some reference genome and occurring at a frequency of at least 1% throughout the species population [1]. SNPs occur within the human genome with a frequency of approximately one SNP per every three hundred bases, with the majority of SNPs being bialleic (exhibiting only two alleles, or number of alternative forms of the same gene or genetic locus). Unfortunately the majority of disorders are not monogenic, but rather individuals exhibit a susceptibility that is a complex and undiscovered function of the expression and corresponding interaction of many genes, as the effects of variations in protein biochemistry occur across the large signalling networks that control cellular behaviour. To complicate matters further, these networks are heavily influenced by many environmental factors.

![Image](https://example.com/image)

Fig. 1. Contribution of various factors toward both a) monogenic and b) complex disorders. The presence of a SNP (dark blue) is the primary cause of a Mendelian disorder, whereas complex disorders are influenced by the presence of several genetic variants, in conjunction with environmental and stochastic factors. Image reproduced from [1].

B. Motivation

The idea underlying a genome-wide association study (GWAS) is relatively simple: to classify a population into two groups based on the presence of lack of a certain physical trait or disorder, and by considering each of the 10 million known SNPs, attempt to identify those which are statistically correlated to that binary class. Although this methodology is only able to identify correlations and not causal mechanisms, knowing that the expression of a particular gene is associated with a particular disorder is a critical step in the development of effective detection, treatment and prevention strategies. Although conceptually straightforward, attempting to identify the presence or absence of 10 million SNPs within an individual’s genome is an impractically expensive and time-consuming process, even given recent advancements in next-generation sequencing technologies. This approach is statistically difficult due to the very high SNP-to-sample ratio, and is unable to capture structural genetic variants such as insertions, deletions and inversions [1].

C. Haplotypes and Linkage Disequilibrium

Fortunately, the probability of an individual exhibiting any two pairs of SNPs is not uniform; SNPs that are located close to one another (relative to the primary DNA structure) are intuitively more likely to be inherited together. Furthermore, meiotic recombination does not occur at random, but is rather localised about recombination “hot spots” [2]; it follows that two polymorphic sites are more likely to be associated if they are not separated by recombination hot spot than those that are. The resultant associated alleles at adjacent chromosomal loci are known as a haplotype, and the corresponding non-random association between two alleles (whether they be present on the same chromosome or otherwise) is measured in terms of linkage disequilibrium [3], which is most commonly reported as a D, D' or r2 score. Given two loci with alleles A, a and B, b, the haplotype frequencies are given in Table I.

<table>
<thead>
<tr>
<th>TABLE I. HAPLOTYP FREQUENCIES x FOR TWO CHROMOSOMAL LOCI WITH ALLELES A, a AND B, b.</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>x11</td>
</tr>
</tbody>
</table>

When no linkage disequilibrium is present, it would be expected that the probabilities of A and B are independent (i.e. P(A, B) = P(A)P(B)). This is often not the case, and linkage disequilibrium may be measured in terms of the difference in probabilities

\[ D = P(A, B) - P(A)P(B), \quad P(A, B) = x_{11}, \]
or the \( r^2 \) score

\[
    r^2 = \frac{D^2}{P(A)P(a)P(B)P(b)}.
\]

An example of a visual representation of linkage disequilibrium between SNPs is demonstrated in Fig. 2, where the diagonal intersection between any two SNPs (plotted along the x-axis) is colour-coded by \( r^2 \) value. In the presence of nonzero linkage disequilibrium, Table I may be rewritten in terms of \( D \), as demonstrated in Table II.

### Table II. Haplotype Frequencies for Two Chromosomal Loci

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>a</th>
<th>B</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(A)P(B) - D )</td>
<td>( P(a)P(b) - D )</td>
<td>( P(a)P(B) - D )</td>
<td>( P(A)P(b) - D )</td>
<td></td>
</tr>
</tbody>
</table>

Linkage disequilibrium and the presence of haplotypes are what make GWAS practically feasible; rather than testing for all 10 million SNPs, the presence of large sets of associated alleles may be statistically inferred by testing for the presence of a considerably smaller set of SNP “tags” [3]. For example, sequencing and genotyping studies have demonstrated that as few as 500,000 or 1 million SNPs, in European and African populations respectively, can serve as proxies for approximately 80% of common SNPs [4]. This due to the fact that haplotypes frequently extend across as many as 30,000 to 50,000 base pairs [5].

The idea that the majority of variation within the human genome could be captured by genotyping less than 1 million tag SNPs highlighted the need for precise mapping of linkage disequilibrium (or a “human haplotype map”), such that the most representative set of tags could be correctly identified [1].

**II. The International HapMap Project**

The goal of the International HapMap Project involved the development of a map of common patterns of DNA sequence variation in the human genome by an international consortium, by determining the genotypes of more than one million sequence variants, their frequencies and the degree of association between them [6]. DNA samples from 270 genetically diverse individuals were examined in two phases. During Phase I, 1 million SNPs were genotyped, and the results were published in 2005 [7]. In Phase II, 3 million SNPs were genotyped, and the results published in 2007 [8]. Phase II initially aimed to genotype 4.4 million distinct SNPs, but 1.3 million proved to be too difficult to type (due to the secondary structure of the genome) or were shown not to be polymorphic in any of the examined populations [1].

By precise mapping of linkage disequilibrium across the human genome, the International HapMap Project has facilitated feasible GWA studies by allowing the number of genotyped SNPs to be reduced by a full order of magnitude, while still capturing the majority of variant information available. In addition, HapMap has assisted in the development of new analytical techniques for reducing false associations, adjusting for ancestral differences among individuals and groups and assessing claims of replication of genotype-phenotype associations [1].

**III. Sequencing Technologies**

Despite the fact that International HapMap Project has successfully reduced the number SNPs that require genotyping for a GWAS, determining the presence or absence of the remaining set of 500,000 to 1 million proxy SNPs would not be feasible without modern, high-throughput sequencing technologies. Before the introduction of next-generation sequencing technologies (such as RNA-seq), DNA microarrays (such as the Affymetrix GeneChip and Agilent SurePrint) were the primary method of measuring the abundance of selected mRNA transcripts. This technology relies on the hybridisation of cDNA (DNA resulting from the hybridisation of an mRNA transcript, corresponding with a subset of the original DNA with intronic regions removed) or cRNA (transcribed cDNA product) against an array of short, immobilised probes of known nucleotide sequence, and involves the following steps [10]:

- **Sample preparation and labelling**: RNA is extracted from cells or tissues, and the corresponding cDNA product labelled with a fluorescent marker.
- **Hybridisation**: A solution of the labelled cDNA is applied to the array, forming duplexes with the probes on the surface of the microarray.
- **Washing**: The microarray is washed to remove unbound (or weakly bound, in the case of cross-hybridisation) cDNA.
- **Image acquisition**: The microarray chip is placed in a scanner, in which each fluorescent marker is excited by...
a laser. As these fluorescent markers correspond with labelled cDNA bound in duplexes with the microarray probes, the resultant fluorescent intensity (captured by the scanner) is an indication of the presence and proportion of each considered transcript.

- **Image Processing**: The image acquired during the previous step is processed, such that a numeric value is produced for the intensity of each individual probe on the microarray.

Moving beyond microarray technology, which can only measure the abundance of known, short DNA sequences, so-called next-generation sequencing provides a high-throughput method that is capable of identifying previously unknown sequences [11]. Conceptually, next-generation sequencing involves the fragmentation of cDNA by digestion with enzymes, before the resultant fragments are read by a sequencing machine such as the Solexa from Illumina, or SOLiD from Applied Biosystems. The sequenced fragments are then “stitched together” using string alignment against a reference genome, with computational genomics algorithms such as Smith-Waterman, BLAST (Basic Local Alignment Search Tool) or BLAT (BLAST-Like Alignment Tool).

### TABLE III. EVALUATION OF MICROARRAY AND NEXT-GENERATION SEQUENCING TECHNOLOGIES BASED ON A NUMBER OF KEY FACTORS.

<table>
<thead>
<tr>
<th>Throughput</th>
<th>Coverage</th>
<th>Cost per sample</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microarrays</td>
<td>High</td>
<td>Low for commercial Higher for custom</td>
<td>Can only detect known sequences Large base of published data and best-practice information</td>
</tr>
<tr>
<td>Next-generation sequencing</td>
<td>High</td>
<td>High</td>
<td>Less published data and best-practice information More effective for detection of low-abundance transcripts Can detect unknown sequences</td>
</tr>
</tbody>
</table>

### IV. INTERPRETATION OF GWAS DATA

#### A. Pre-Processing of SNP Data

As explained above, GWA studies are made possible by a combination of genome-wide haplotype mapping and high-throughput sequencing technologies. Unfortunately, both of these processes introduce potential sources of error. In the case of haplotype mapping, it is vital that the ethnicity of the individual being genotyped be known *a priori*, as it has been empirically demonstrated that the distribution of linkage disequilibrium across the genome differs significantly between individuals of European, Asian or African descent (and comparatively little within each of these populations [1]). For this purpose, companies such as Illumina produce SNP chips (microarrays containing pairs of allele probes for hundreds of thousands of well-known proxy SNPs) specific to a particular ethnicity.

Identifying and correcting for errors resulting from high-throughput sequencing technology is a complex and multistage process. In the case of microarray SNP chips, the following issues are prevalent:

- **Low quality samples**: This may be caused by incorrect handling of the physical microarray chip, or artifacts introduced during the manufacturing process. When a SNP probe does not reach the intensity threshold required to pass the detection filter or risk allele score requirements [12], a “NoCall” value is returned. The proportion of probes not returning a “NoCall” value, known as the *call rate* of the microarray, is often utilised to remove the need for human inspection in identifying this samples; an overall genotyping call rate less than 90% (or disproportionate between cases and controls) causes the sample to be disregarded. Low quality samples may also be identified by considering the Hardy-Weinberg principle, which states that allele and genotype frequencies in a population will remain constant between generations in the absence of evolutionary influences (such as non-random mating [13]). Concretely, given \( P(A) = p \), \( P(B) = 1 - p \), a significant divergence from the following equilibrium expectations would cause the sample to be rejected

\[
P(AA) = p^2, \quad P(AB) = 2p(1-p), \quad P(BB) = (1-p)^2
\]

A practical example of pre-processing for low quality samples by call rate and Hardy-Weinberg is presented in Sec. V, which describes a GWAS attempting to associate SNPs with postmenopausal breast cancer. A total of 8,706 SNPs (1.57% of the total proxy SNPs genotyped) were disregarded for failing to meet a 90% call rate threshold, and a further 28,710 (5.57%) were found to deviate from the Hardy-Weinberg proportions with a \( p \)-value of 0.005.

- **Relatedness of samples**: Undetected relatedness between samples can lead to excessive sharing of SNPs, which may further result in spurious false positives of genotype-phenotype association. Such relatedness may be identified by calculating the probability of two individuals sharing \( n \) alleles that are identical by descent (identical copies of the same ancestral allele), where \( n = 0, 1 \) or 2. In the example of two siblings, \( P(IBD = 0) = P(IBD = 2) = 0.25 \), \( P(IBD = 1) = 0.5 \), whereas for parent and child \( P(IBD = 0) = P(IBD = 2) = 0, P(IBD = 1) = 1 \) [14].

- **Population stratification**: Differences in allele frequency between cases and controls may be the direct result of systematic ancestral differences, and such population stratification can introduce false positives during the GWAS process. Principle component analysis (a mathematical procedure that applies an orthogonal transformation to convert a multivariate set of observations into a set of linearly uncorrelated variables, possibly of lower dimensionality [15]) may be applied to explicitly model these ancestral differences to minimise these spurious associations [16]. This process (implemented in EIGENSTRAT [16]) and similar applications of principle components analysis (such as that implemented in the PLINK software package [17]) have also been demonstrated as effective in minimising the impact of batch effects, such as those arising from sample collection, platform effects, DNA extraction methodologies and the unique practices of individual labs and research.
B. Post-Processing of SNP Data

Even after SNP data has been thoroughly pre-processed, there remains a high probability of emergent spurious associations from the sequencing output. These are often removed by human supervision and post-hoc processing. Specifically, the SNP cluster plots are inspected for indications of error resulting from copy number variation (insertion or deletion of nucleotides resulting from heritability or de novo mutation), which may then be verified by conducting array comparative genomic hybridisation-based studies [18]. Similar to the case of the International HapMap Project, an international consortium (the Copy Number Variation Project) was established with the intent of characterising as many copy number variants (CNVs) as possible, allowing the discovery of direct associations between CNVs and complex diseases. Spurious genotype-phenotype associations may be further identified by considering the biological significance and implications of the finding, by considering information known regarding gene expression profiles and the analysis of biological pathways [19].

Although post-hoc processing and the injection of biological system context is useful in the identification of many false positive SNP associations, much of the contextual information is still unknown, and therefore the process does not implement a perfect filter. Fortunately, GWA studies involve such large quantities of data that significant statistical adjustments and normalisation may be applied to the results of multiple tests over the same initial data.

In statistics, the null hypothesis refers to the default position that there is no relationship between two measured phenomena, such as the association between a SNP and a given disorder. Corresponding with this notion is the p-value, which represents the probability of obtaining the observed test statistics given the assumption that the null hypothesis is true [20]. It follows from this definition that, the stronger the correlation between SNP and disorder, the smaller the resultant p-value. A p-value of $5 \times 10^{-8}$ is generally accepted as the upper bound for genome-wide significance, below which the null hypothesis is said to be rejected. This threshold is derived by dividing the typical statistical threshold for rejecting the null hypothesis ($p = 0.05$) by the SNP dataset size, with the selection of such a stringent threshold necessary for the prevention of spurious associations when assessing hundreds of thousands of proxy SNPs. An example of experimentally determined p-values is presented in Table IV, representing the association between six SNPs and postmenopausal breast cancer.

Finally, batch effects are corrected by the process of quantile normalisation. Due to the significantly large size of the SNP dataset, it follows that the p-values for each SNP (for two experiments utilising the same DNA samples) follow the same distribution (i.e. all of the quantiles are equal), and that any significant difference between distributions be the result of experimental artifacts. By conducting several experiments (from different labs, using different microarray scanners, etc.), the reference distribution $F(x)$ (to which the others are normalised) may be obtained by averaging each quantile across the entire batch of analysed chips, following an algorithm introduced by Bolstad et al. [21]. The resultant quantile normalisation method is simply a specific case of the transformation $x_i' = F^{-1}(G(x_i))$, where $G(x)$ is estimated by the empirical distribution of each microarray. These quantile distributions are commonly visualised using Q-Q (quantile-quantile) plots, as demonstrated in Fig. 3, with the remaining divergence from the $y = x$ axis presumed to be representative of genuine biological variation.

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>p-value</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs15031126</td>
<td>0.000004</td>
<td></td>
</tr>
<tr>
<td>rs12505080</td>
<td>0.0000081</td>
<td></td>
</tr>
<tr>
<td>rs17157903</td>
<td>0.0000088</td>
<td>RELN</td>
</tr>
<tr>
<td>rs1219648</td>
<td>0.000012</td>
<td>FGRF2</td>
</tr>
<tr>
<td>rs7696175</td>
<td>0.000013</td>
<td>TLR1, TLR6</td>
</tr>
<tr>
<td>rs2420946</td>
<td>0.000015</td>
<td>FGRF2</td>
</tr>
</tbody>
</table>

Fig. 3. Example of a normalised Q-Q plot, demonstrating the effect of statistically significant SNP association (blue) relative to the null case (red). Image reproduced from [22].

C. Visualisation of Results

1) Manhattan Plot: A Manhattan plot is a succinct variety of a scatter plot that may be used to represent SNP association likelihood data, with the x-axis representing the genomic coordinates (binned by chromosome number) and the y-axis representing the negative logarithm of the p-value associated with each SNP (for example, a p-value of $10^{-10}$ will map to a y-axis value of 10). The threshold value of $5 \times 10^{-8}$ is often indicated with a horizontal line, to readily highlight each SNP that rejects the null hypothesis. An example of a Manhattan plot for a GWAS is illustrated in Fig. 4.

2) Forest Plot: A forest plot is used to graphically represent a meta-analysis of the results of randomised, controlled trials. In the case of a GWAS, its purpose is to compare, contrast and combine results from different studies to identify patterns,
V. Case Study and Evaluation

A. Association of FGFR2 Haplotypes with Breast Cancer Susceptibility

As one of the many examples of GWA studies revealing statistically significant association between SNPs and disease, research published in 2007 by Hunter et al. reports an association between alleles in FGFR2 haplotypes and susceptibility to sporadic postmenopausal breast cancer [9]. The GWAS was conducted by genotyping 528,173 SNPs in 2,287 postmenopausal white women (1,145 suffering invasive breast cancer, and 1,142 controls), and was successful in identifying four SNPs (in intron 2 of FGFR2) with a high level of association, with the most strongly associated SNP (rs1219648) exhibiting a \( p \)-score of \( 1.1 \times 10^{-10} \). These results were confirmed by three further replication studies, which disregarded four initially well-scoring SNPs at other chromosomal loci.

As an example of the pre and post-processing methodologies described above, this particular publication reported a number quality control statistics:

- **Assessment of SNP call rates:** A total of 8,706 SNPs (1.57% of the total proxy SNPs genotyped) were disregarded due to failing to meet the 90% call rate threshold, as described in Sec. IV-A.
- **Deviations from Hardy-Weinberg proportions:** All collected genotype data was assessed for deviation from the Hardy-Weinberg proportions, as defined in Sec. IV-A. Although 28,710 SNPs (5.57%) deviated with a \( p \)-value of 0.005 and 2,843 (0.55%) with a \( p \)-value of 0.001, none were removed from further analysis as they were determined to have equal negative effects among both cases and controls.
- **Assessment of population stratification:** Two independent sets of 7,050 and 7,061 SNPs (exhibiting very low linkage disequilibrium, as described in Sec.I-C) were analysed to determine whether any samples belonged to individuals of mixed origin (greater than 15% Asian or African ancestry, based on HapMap data [8]), resulting in the removal of 3 cases and 1 control. Principle components analysis (specifically the EIGENSTRAT implementation [16]) identified four significant principle components, of which only the first three were retained for further analysis.

The major results of this association study are graphically summarised in Fig. 4 and 2, with the experimentally determined \( p \)-values for six chromosomal loci presented in Table IV. Found to be the most significantly associated locus, FGFR2 is a tumor suppressor gene that is known to be amplified and over-expressed in breast cancer [24], [25]. Despite the identification of these statistically associated risk loci, the actual causal variant and mechanisms at these loci remain unknown, and the authors of the study admit that further research is necessary to determine these factors (beyond the scope of what is possible given only the GWAS framework). [9].
B. Criticisms of GWAS

DESPITE having identified more than 400 genetic variants responsible for a variety of Mendelian/simple diseases and other phenotypic traits, the methodology underlying a GWAS still draws a great deal of criticism [18]. As a simple motivating example, several studies (going back a century) have demonstrated that height is up to 90% heritable; although GWA studies looking at in excess of 30,000 people have identified dozens of associated variants, they have been able to attribute for little more than 5% of this observed heritability [26]. This relatively poor performance can be attributed to fundamental limitations in the GWAS methodology. As described earlier, a GWAS experiment tends to examine a smaller set of proxy SNPs, which have been experimentally demonstrated to be strongly representative of the much larger set of approximately 1 million known SNPs. Although this drastically reduces the financial and temporal cost of the experiment, it follows that a single proxy SNP represents a considerably large block of genetic material. As an example, if two people share some variant at one of these key locations, both may be scored as having the same version of any height-related gene in that area, despite the fact that only one person exhibits the rare mutation primarily associated with height [18]. Extending this idea to the case where tens of thousands of individuals are genotyped for this SNP, the averaged effect of this rare mutation will be considerably diluted.

GWAS has traditionally proven effective in the identification of genetic variants with a high penetrance, meaning that the majority of individuals carrying that mutated gene exhibit the phenotypic trait of interest. Unfortunately, it is now believed that many common diseases are more influenced by either rare, moderately penetrant or common, weakly penetrant variants, in which case the GWAS is unable to statistically link genotype to phenotype without considerably larger datasets [18]. Although many such arguments for this observed “missing heritability” have been proposed, one corresponds particularly strongly with modern understanding of biological systems; genes do not act independently, but rather exhibit epistasis: a phenomena in which the expression of one gene depends on the presence of one or more “modifier genes”, in that these modifier genes are able to promote or suppress expression of the other.

The problems introduced by epistasis in biological systems can be captured by a simple example: two genes may each add one centimetre to height, but together could add five. The ability for genes to “work together” (as in this example), or for one gene to mask the effect of another, are factors that are not incorporated into or well handled by the GWAS framework [18]. It has been demonstrated that expression within yeast genes is typically controlled by more than five genetic variants [27], and recent research has shown how the parameters of well-validated dynamic models of complex phenotypic traits can yield stronger genome-wide associations than polymorphisms for eight calcium-transient phenotypes in heart cells [19].

OVER the last decade, there has been considerable effort invested into the expansion of knowledge regarding nonlinear systems dynamic models of biological pathways and processes, with model repositories such as http://www.cellml.org and http://biomodels.net established to facilitate their exchange and reuse. Independent of the success stories or criticisms of GWA studies, it is now clear that such models are necessary for the full utilisation of phenotypic data, and to bridge the expanding gap between our ability to characterise genotype, and our relative inability to apply this information to the characterisation of phenotype [28]. Regarding the future improvement of GWA studies for discovering new associations between genotype and susceptibility to complex diseases, Dermitzakis and Clark stated that “a major breakthrough will be to predict and interpret the effect of mutational and biochemical changes in human cells and understand how this signal is transmitted spatially (among tissues) and temporally (spanning development)” [29]. Recent research has supported this opinion by demonstrated that much of the “missing heritability” of complex diseases may be attributed to functional interactions of multiple genetic variants [18], [19], [27], highlighting the importance of the future integration of GWAS findings into the wider context of biological processes and environment.

REFERENCES


