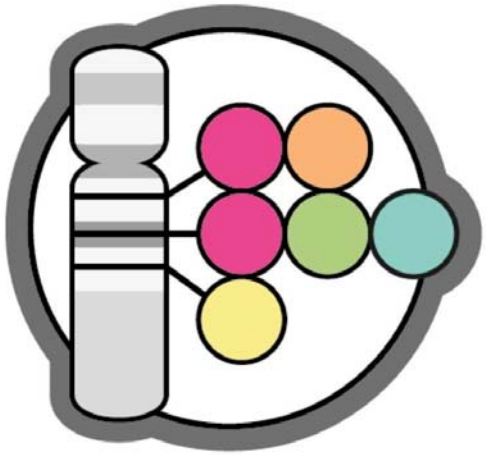
The GWAS Catalog



https://www.ebi.ac.uk/gwas/home

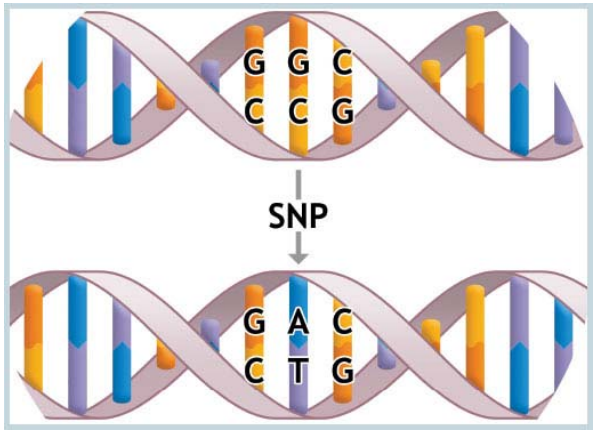
- Useful concepts
- The failure of linkage studies
- The common disease/common variant hypothesis
- What is a genome-wide association study (GWAS)?
  - The logic behind GWAS
  - GWAS ushered the personalized medicine era
  - Reasons for non-replication of GWAS
  - Has GWAS failed us or have we misunderstood GWAS?
- The GWAS catalog
- Practice

Concepts Underlying the Study Design

The modern unit of genetic variation is the single nucleotide polymorphism or SNP (single base-pair changes in the DNA sequence). In genetic studies, SNPs are used as markers of a genomic region and are the most abundant form of genetic variation in the human genome.

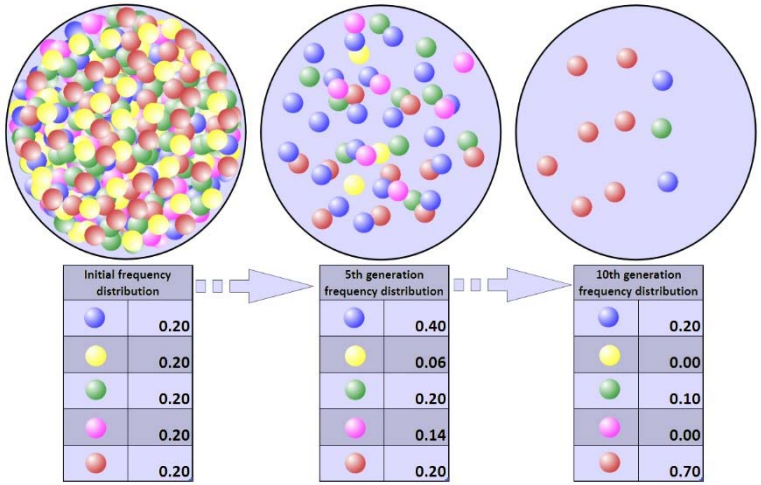
Most SNPs have are non-functional or have a minimal impact on biological systems. Some SNPs have functional consequences, causing amino acid changes, changes to mRNA transcript stability, and changes to transcription factor binding affinity.

Most SNPs have two alleles, meaning within a population there are two commonly occurring base-pair possibilities for a SNP location.



The frequency of a SNP is given in terms of the minor allele frequency (MAF) or the frequency of the less common allele. For example, a SNP with a minor allele frequency of 0.40 implies that 40% of a population has the *G* allele versus *A* the more common allele (the major allele), which is found in 60% of the population.

Random genetic drift responsible to changes in allele frequencies of most SNPs creating a confusion between causal and neutral alleles



Rare genetic disorders (e.g., cystic fibrosis) are caused due to extremely rare SNPs (=*mutations*) that induce a detrimental change to protein function.

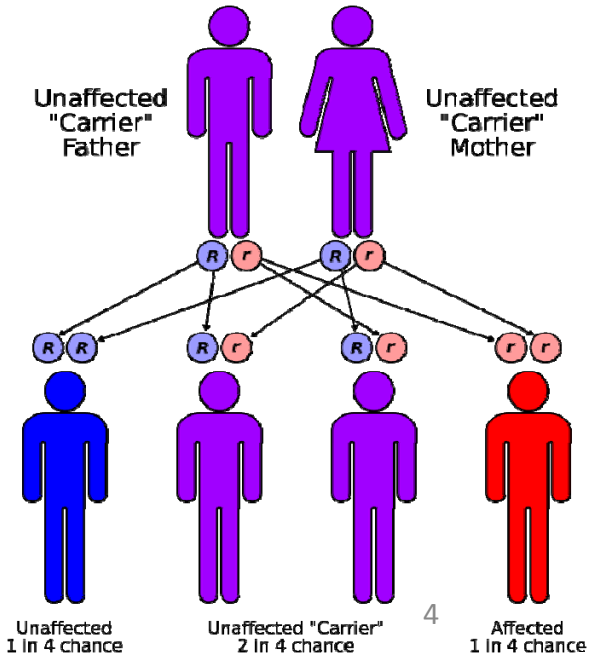
In the genetics literature, the term SNP is generally applied to *common* single base-pair changes, and the term mutation is applied to *rare* genetic variants.

The failure of linkage for common disorders

Rare disorders (e.g., CF) are caused by multiple mutations within the *CFTR* gene whose effect is so strong that CF follows an autosomal dominant inheritance pattern in families with the disorder.

The identification of *CFTR* mutations was achieved by genotyping affected, and examining how those genetic markers segregate with the disease across families. *Linkage analysis*, was successfully applied to identify genetic variants in other rare disorders like Huntington disease. Applied to more common disorders, linkage analysis has **not** fared well.

The *common disease/common variant* (CD/CV) hypothesis



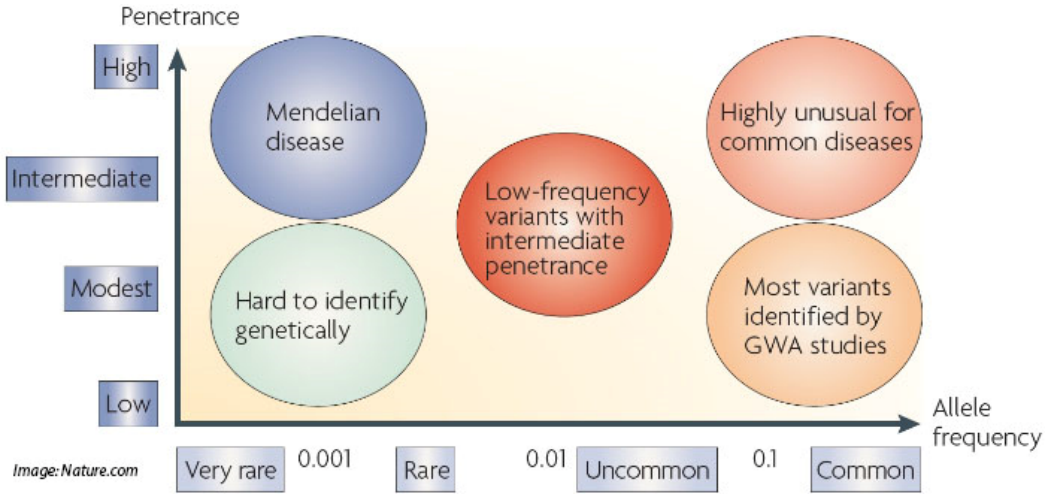
The common disease/common variant hypothesis

Common disorders are likely influenced by genetic variation common in the population. **This has ramifications of this for studying diseases.**

If common genetic variants influence a disease, the effect size (or penetrance) must be small, compare to rare disorders. **Here is why**:

- If a common a disease allele (MAF=0.4) inflicts the phenotype in 40% of the population then the population prevalence and MAF will be correlated.
- If that allele causes a small change in gene expression that alters risk for a disease by some small amount then the prevalence of the disease and the MAF would be slightly correlated.

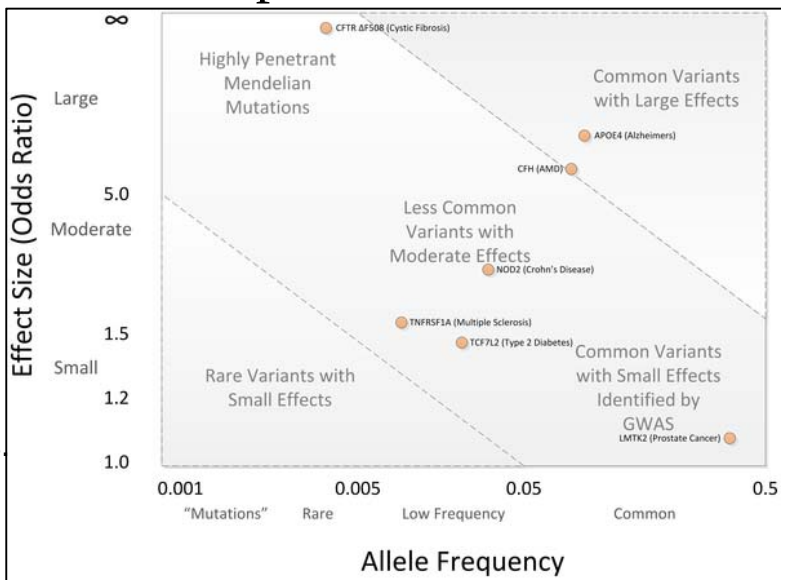
Therefore, common variants have low penetrance.



If common alleles have small genetic effects (=low penetrance), but common disorders show heritability (=inheritance in families), then multiple common alleles must influence disease susceptibility. **Here is why**:

- If a twin study estimates the heritability of a common disease in 40%, then 40% of the total variance in disease risk is due to genetic factors.
- If the allele of a SNP incurs only a small degree of disease risk, then that SNP explains only a small proportion of the total genetic variation. Therefore, the total genetic risk due to common genetic variation must be spread across genetic factors.

Therefore, family-based studies will be successful for complex disorders, compared to population-based studies.



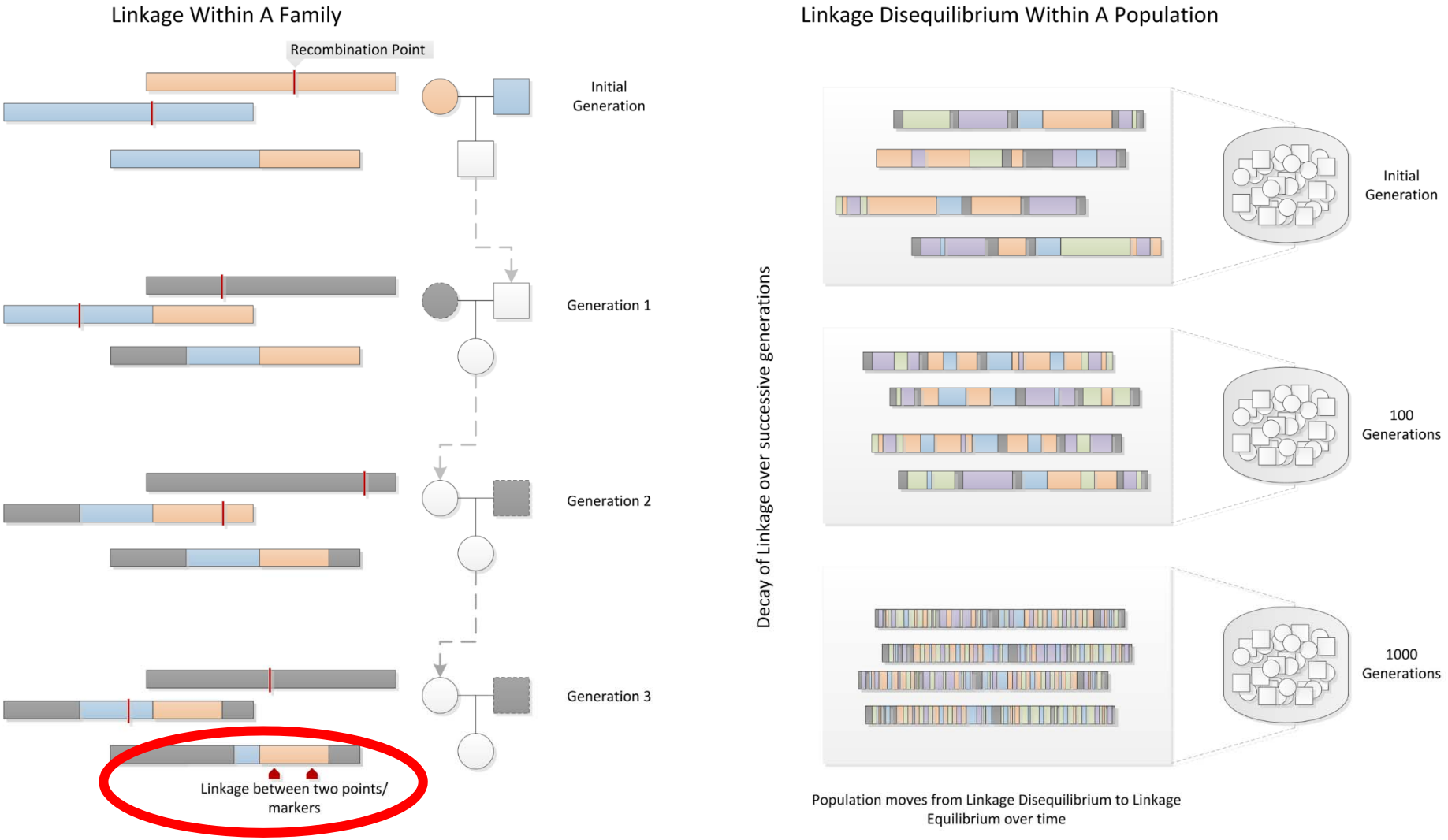
Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12): e1002822. doi:10.1371/journal.pcbi.1002822
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002822

What is a genome-wide association study (GWAS)?

A central goal of human genetics is to identify genetic risk factors for common, complex diseases (e.g., schizophrenia) and for rare Mendelian diseases (e.g., cystic fibrosis). Understanding the biological basis of genetic effects has an important role in developing new pharmacologic therapies.

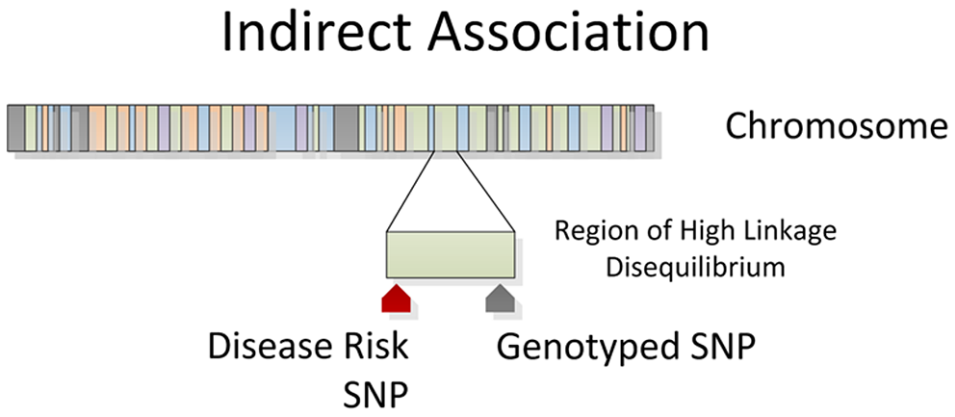
- There are many different technologies, study designs, and analytical tools for identifying genetic risk factors.
- GWAS analyzes DNA sequence variations throughout the genome to identify genetic risk factors for common diseases.
- GWAS's goals are to use genetic risk factors to make predictions about who is at risk and to identify the biological underpinnings of disease susceptibility for developing new prevention and treatment strategies.
- One of the early successes of GWAS was the identification of the *Complement Factor H* gene as a major risk factor for age-related macular degeneration (AMD).
- GWAS can be powerful tools for investigating the genetic architecture of **common** (single gene) diseases.

The logic behind GWAS

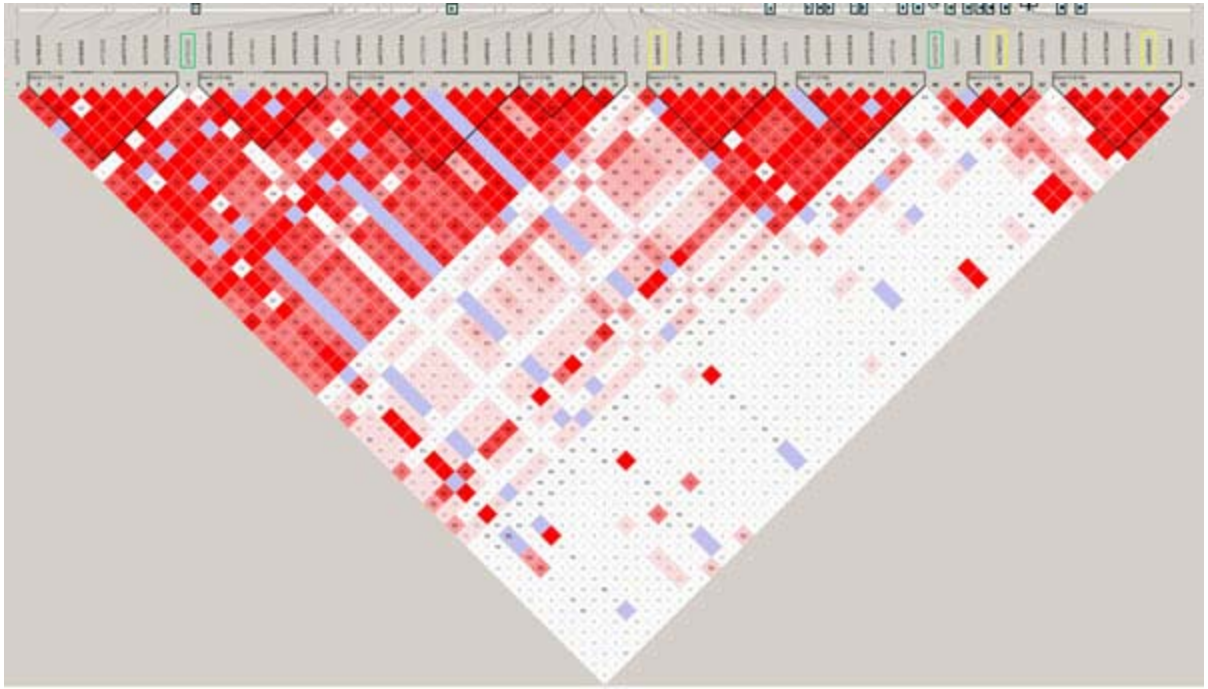


Within a family, linkage occurs when two genetic markers remain linked on a chromosome rather than being broken apart by recombination events during meiosis. **In a population**, contiguous stretches of founder chromosomes from the initial generation are sequentially reduced in size by recombination events.

Assumption. Genotyped SNPs often lie in a region of high linkage disequilibrium with an influential allele. The genotyped SNP will be statistically associated with disease as a surrogate for the disease SNP through an indirect association.

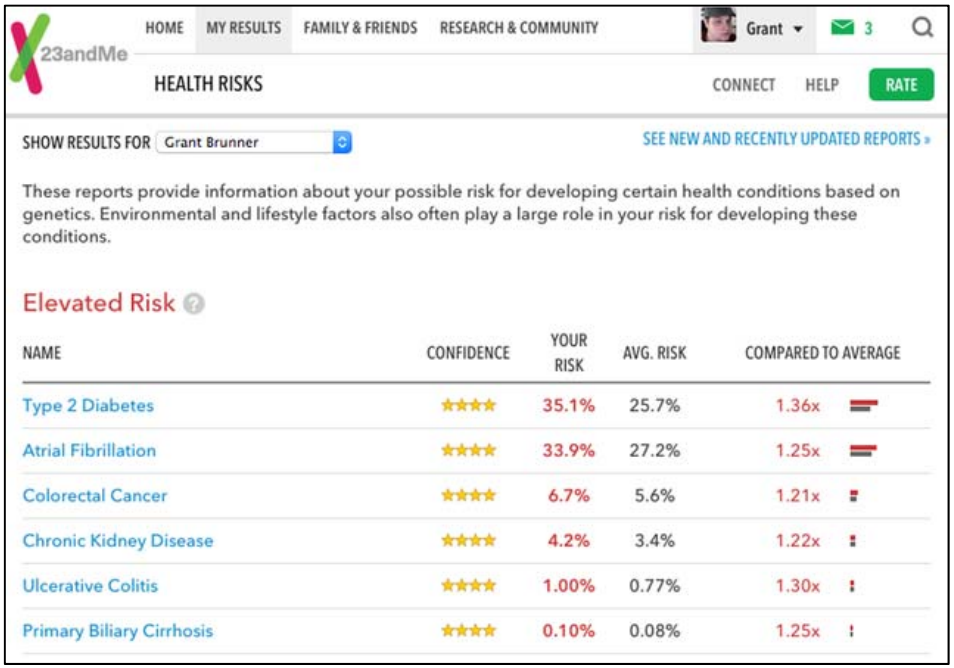


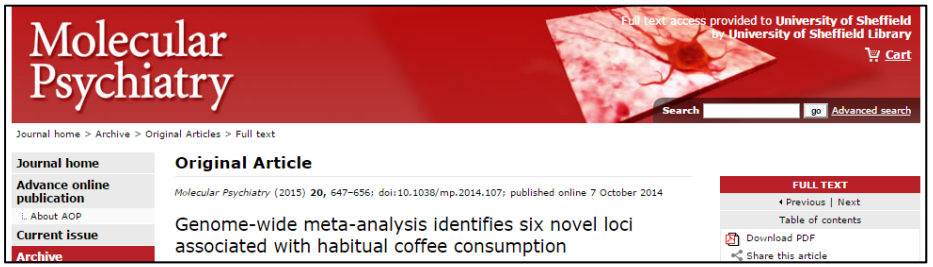
"Tag SNPs" allow us to predict the value of the SNPs that they tag.



GWAS ushered the personalized medicine era

The widespread availability of low-cost technology for measuring an individual's genetic background has been harnessed by businesses that market genetic testing directly to the consumer.





Replication?!?!?!?

For better or for worse, GWAS have paved the way for the era of personalized medicine and personal genetic testing.

Reasons for non-replication of GWAS

1. The original observation was a false positive due to sampling error.

2. The follow-up study had insufficient power.

3. The genotypic coding used in the initial study may not accurately reflect the true underlying association, leading to a loss of power (**correlation<>causation!**).

4. The variant may be a poor marker for the trait due to differences in linkage-disequilibrium structure between the studies.

5. Differences in design or trait definition may lead to inconsistencies.

6. The absence of an association in the subsequent studies may be due to true etiologic heterogeneity.

Has GWAS failed us or have we misunderstood GWAS?

Is the GWAS approach founded on a flawed assumption that genetics plays an important role in the risk for common diseases?

GWAS showed that 10%–50% of phenotypic variation is captured when all SNPs are considered simultaneously for a number of complex diseases and traits in support of pedigree studies suggest that a substantial proportion of variation in susceptibility for common disease is due to genetic factors.

Have GWASs been disappointing in not explaining more genetic variation in the population?

The aim of GWAS is to detect loci that are **associated** with complex traits. In some cases, the detection of such loci has led to the discovery of new biological knowledge about disease.

Have GWASs delivered meaningful biologically relevant knowledge or results of clinical or any other utility?

Yes, in some cases. No, in others cases.

Are GWAS results spurious?

Yes, in some cases. No, in others cases.

Visscher et al. (2012) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3257326/

The GWAS catalog

"The GWAS Catalog was founded by the NHGRI in 2008. The Catalog is a quality controlled, manually curated, literature-derived collection of all published genome-wide association studies assaying at least 100,000 SNPs and all SNP-trait associations with p-values < 1.0 x 10^{-5} *(Hindorff et al., 2009)*."

Practice

- What is an Odd Ratio (OR)?
- How many autism GWAS studies are in the catalog?
- Which SNP has the highest OR for autism?

- How many SNPs are associated with bipolar disorder?
- On which chromosomes are they found?

- Which gene is best associated with Coffee consumption?
- Where is it located?
- Is this association true for all worldwide individuals?
  How many SNPs were used in these study to infer association?

- Download the catalog and observe the data.
- How many SNPs are in the catalog?
- What is the typical number of markers in these studies?