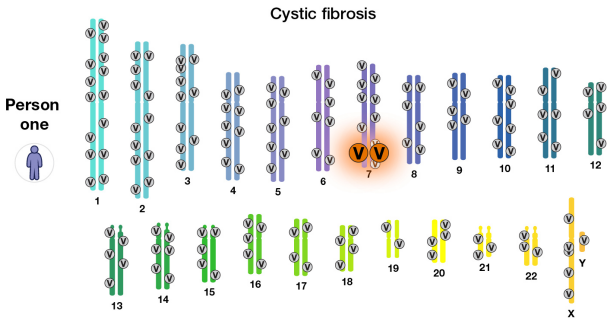
Mapping Genes to Complex Traits

Hae Kyung Im, PhD

THE UNIVERSITY OF CHICAGO

January 20, 2021

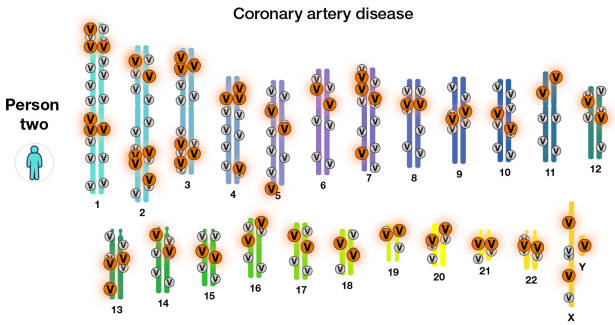
Example: Monogenic/Mendelian Disease (Rare)



https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores

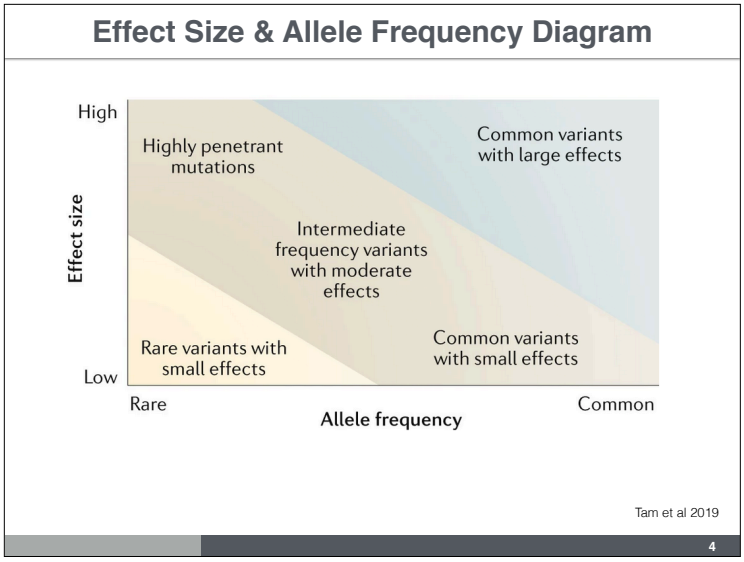
Let's agree in some common terminology. Monogenic or Mendelian diseases such as cystic fibrosis are caused by a loss of function mutation in a single gene. These diseases tend to be severe and because of selection the mutations tend to be rare.

Example: Complex Disease (Common)



https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores

Complex diseases are due to the accumulated effects of multiple and usually common variants.



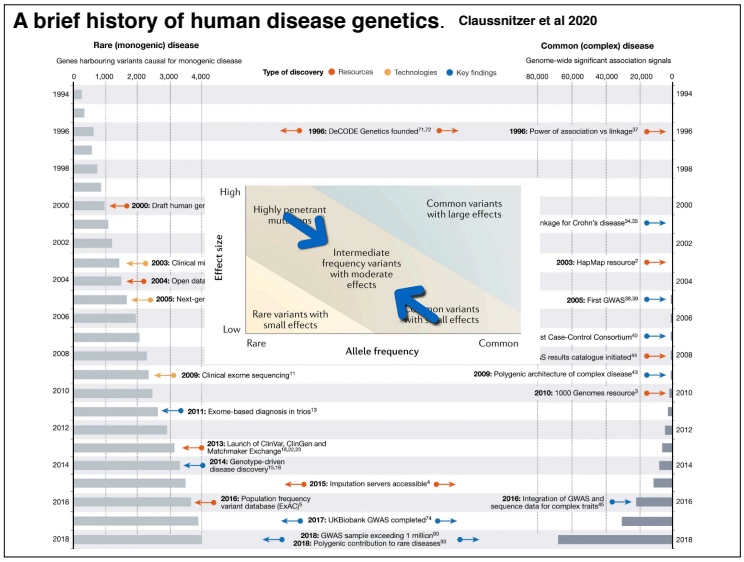
These two types of diseases and their causal variants can be placed in this diagram of effect size vs allele frequency.

The diagonal region of this diagram represents the allelic frequency and phenotypic effect spectrum where gene mapping typically occurs. Highly penetrant mutations (large effect size) tend to be rare and historically have been discovered via linkage. Examples are BRCA1 and BRCA2 (breast cancer), CFTR (cystic fibrosis).

Our focus will be in the lower regions of the diagonal band: common variants with small effects region and stretching towards the intermediate frequency variants with moderate effects, when sample size/power of the study permits. Association methods thrive in the lower portion of the diagonal.

Highly penetrant mutations have been mostly detected via pedigree-based linkage studies whereas GWAS are well powered to discover common variants with small effects.

Penetrance is the proportion of mutation carriers who manifest the disease. High penetrance is equivalent to large effect size but penetrance is a term used more in the rare disease context whereas effect size is more commonly used for common diseases.

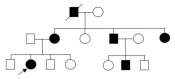


Even though in this course we will be focusing on common diseases and traits, as the field advances we are starting to erase the divide between Mendelian and common disease genetics.

I highly recommend to read this recent review of human disease genetics. Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., et al. (2020). A brief history of human disease genetics. Nature, 577(7789), 179–189. http://doi.org/10.1038/s41586-019-1879-7

Gene Mapping Methods

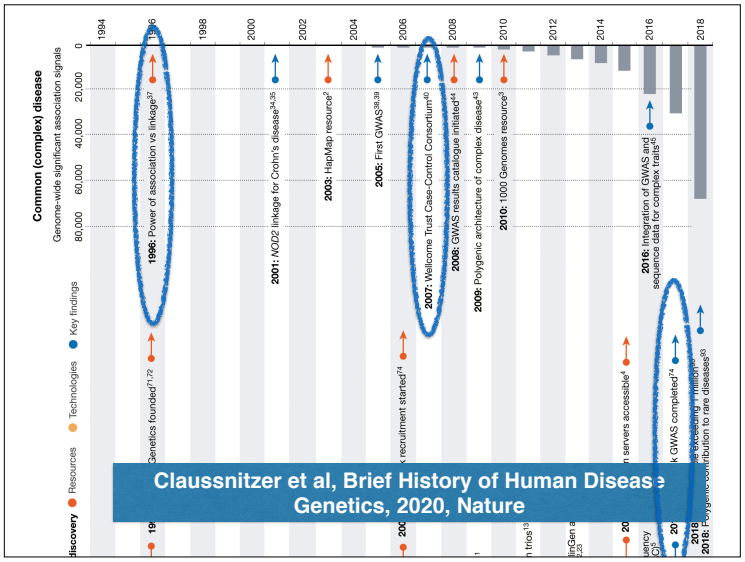
- Linkage analysis
  - popular before Human Genome Project's completion
  - based on co-transmission of genetic markers and disease genes
  - a few hundred markers can cover the whole genome
  - low resolution
  - pedigree/family based



- Association mapping
  - genetic markers in LD with disease genes
  - need a large number of markers (~ 1 million)
  - higher resolution
  - families not needed

Linkage and association are main approaches for mapping genes to diseases and other human traits. Before the completion of the Human Genome Project and the cost of genotyping more than a few hundred markers was prohibitive, linkage analysis was the most popular approach to identify disease loci. It is based on the co-transmission of genetic markers and disease genes. Advantages were that a few hundred markers were enough to cover the whole genome. But the downside was the low resolution and the fact that recruiting large families is more difficult than a large number of unrelated individuals.

Association methods are based on the LD between genetic markers and disease genes. For common variants (say, MAF>5%) about 1 million SNPs can tag most common variants. So that even if the causal variant is not measured but is common, a closely correlated (in LD) SNP can be detected. These are called tag SNP or index SNPs.



In 1996, Risch and Merikangas published a highly influential perspective paper that show the benefits of GWAS.
We have come a long way since the landmark publication of the WTCCC GWAS in 17,000 cases of 7 common diseases and controls.
Sample sizes continued to grow over the years, identifying increasing number of genomic loci associated with complex traits.
In 2006, recruitment for the UK Biobank project started.
Today we will look look at the UK Biobank GWAS performed half a million participants.

Review
Hypothesis Testing

Review: Hypothesis Testing

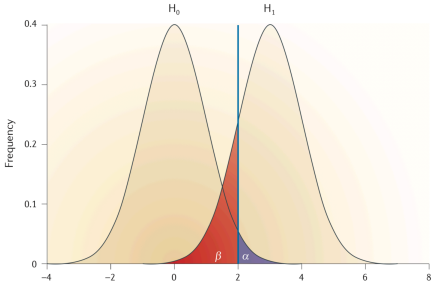
- **H_0** : null hypothesis
  - e.g. no difference in case-control allele frequencies
- **H_A** : alternative hypothesis
  - there is a difference, i.e. causal variant
- Test statistic **Z**
  - in many situations associated to a model
- Significance level
  - α = allowed type I error (reject null when null is true)
  - p-value = P(observed test statistic more extreme than threshold | H_0)

We reject the null hypothesis if the observed test statistic is more extreme than what we would expect could happen by chance

Before going into the techniques of association, let's review a few concepts.

We reject the null Ho if the observed test statistic is more extreme than a threshold, which is determined so that the probability of type I error stays below a pre-established significance level α. Usually, if p-values are smaller than the significance level, then the null hypothesis is rejected.

Review: Hypothesis Testing



α: type I error, probability of rejecting the null when the null is true

β: type II error, probability of not rejecting the null when the null is false

This figure shows the test statistics under the null hypothesis H0 and the alternative hypothesis Ha. The decision rule is to reject the null hypothesis if the test statistics is larger than a given threshold. Typically, we choose a significance level α -- the type I error that we are willing to accept-- and calculate the threshold above which the null hypothesis will be rejected. A typical significance level used in practice is α=0.05. But we will see that when we run multiple tests, things can go wrong very quickly, so that a much more stringent significance level is required.

They type II error is the probability of not rejecting the null when the alternative hypothesis is true. Power is the probability that we will reject the null when the alternative is true.

Regression Approach Single SNP

Regression Approach

$$Y = \mu + a \cdot \text{age} + \beta \cdot X + \epsilon$$

- Parameters β are estimated (using MLE, least squares, etc)
- Null hypothesis β = 0
- **Many types of traits can be treated with the same approach**
- **Can correct for covariates (age, sex, ethnicity)**
- Prediction

The evidence for association can be quantified using the Pearson Chi2 test of independence between case status and genotype using a table of counts. This only allows to test binary traits. To accommodate other types of phenotypes, such as continuous traits or counts, we can use a regression approach.

The phenotype, Y, is modeled as a constant term, effects of covariates (e.g. age, sex, ethnicity) and the genetic effect.

Advantages: we can correct for covariants, we can use for prediction.

Regression Approach for Quantitative Traits

Quantitative trait:

e.g. height, BMI, systolic blood pressure

$$Y \sim N(\mu, \sigma^2)$$
$$\mu = E(Y) = \beta_0 + \beta_1 X_1$$

Linear Regression

genotype: aa, aA, AA

X_1 : dosage = number of A alleles

Quantitative traits are typically modeled with normal errors and mean given by the a constant β0 and a genetic effect β1. X here indicates the number of minor alleles.

Regression Approach for Disease Traits

Binary trait:
e.g. disease status, hypertension

$$Y \sim \text{Bernoulli}(\pi)$$

$$\text{logit}(\pi) = \log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1$$

$P(Y = 1) = \pi$

$\text{odds} = \frac{\pi}{1-\pi}$

genotype: aa, aA, AA

X_1 : dosage = number of A alleles

Logistic Regression

Binary traits are typically modeled using logistic regression. Instead of the E(X), we use the log of the odds = log(π/ (1-π)) of being a case.

odds = prob / (1 - prob)
beta1 = log odds ratio

for an individual with X1 = 0, the log odds of having the disease is beta0
for an individual with X1 = 1, the log odds of having the disease is beta0 + beta1

therefore, beta1 is the log of the odds ratio

Recall that the log of a ratio is the difference of the logs:
log(A/B) = log(A) - log(B)

Regression Approach for Count Data

Count data:

e.g. number of reads that align to an exon

$$Y \sim \text{Poisson}(\lambda)$$

$$\log(\lambda) = \log(E(Y)) = \beta_0 + \beta_1 X_1$$

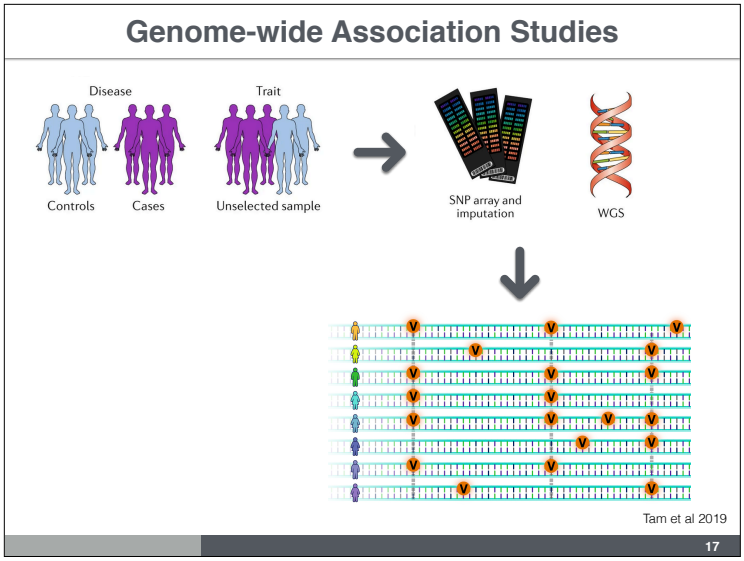
Poisson Loglinear Regression

genotype: aa, aA, AA

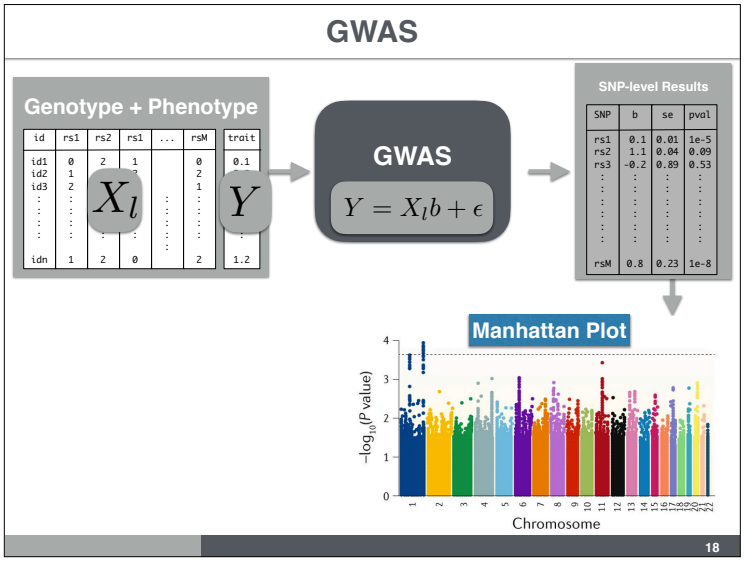
X_1 : dosage = number of A alleles

Count data can be modeled with a poisson log

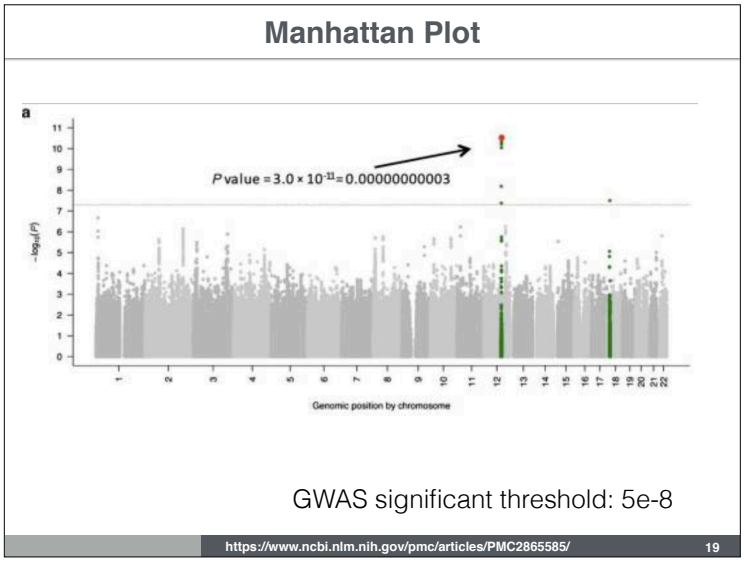
Genome-wide Association Studies



GWAS collect individuals (cases and controls for disease or a population sample of individuals for quantitative traits), measures the genotype of individuals in a set number of genomic locations (~1 million markers), and performs association test between the phenotype (case status or quantitative trait) and each of the genetic marker (typically SNPs, single nucleotide polymorphisms).



The genotype of the individuals is represented as a matrix X_l , the phenotype is represented as a vector Y. Single marker SNP association test is performed, leading to a table of SNP-level results with effect size, standard error of the effect size, and p-values.



Manhattan plots are used to visualize GWAS results. Points above -log(5. 10^{-8}) are called GWAS significant.

"a Manhattan plot (−log10[P] genome-wide association plot) of a genome-wide association study on systolic blood pressure in 29,136 individuals in Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE). The genome-wide significance level is set at 5 × 10−8 and plotted as the dotted line. Any single nucleotide polymorphism (SNP) within a region of 5 Mb containing a SNP reaching the genome-wide significance threshold is colored in green. The most significant SNP in this experiment is colored in red (rs2681492 in the ATP2B1 gene). The P value is indicated for demonstration. b Quantile-quantile (QQ) plot of the data shown in the Manhattan plot. c QQ plot of simulated data showing an early separation of the observed from the expected, suggesting

population stratification. (a and b adapted from Levy et al. [22••], with permission.)"
Ehret, Genome-Wide Association Studies: Contribution of Genomics to Understanding Blood Pressure and Essential Hypertension, 2011,

QQPlot

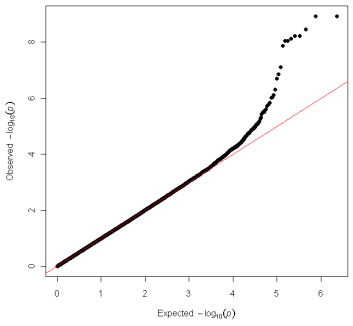
- A Q-Q plot is a useful tool to present the GWAS results and check for potential issues

X-axis: the expected –log(P-values) under the null hypothesis of no association. I.e., the negative log10 of a set of uniformly distributed p-values.

Y-axis: the observed –log(P-values).

Dots above the 45-degree line (upper right) deserve a closer look.

A QQ plot can also be used to check for population stratification (more later).



20

In addition to the Manhattan plot, qqplots is a useful visualization of GWAS results to detect possible issues with the analysis. This compares the observed distribution p-values with the expected distribution under the null hypothesis of no real relationship between genotype and phenotype. Recall that under the null hypothesis, p-values are distributed uniformly. So if we order the p-values under the null, they will be nearby 1/m, 2/m ,..., 1. This is the expected distribution.

In typical well-behaved GWAS, most points should line up at the identity line and a few at the right end depart from the identity line.

LD Allows Detecting Association even if Causal Variant Not Available

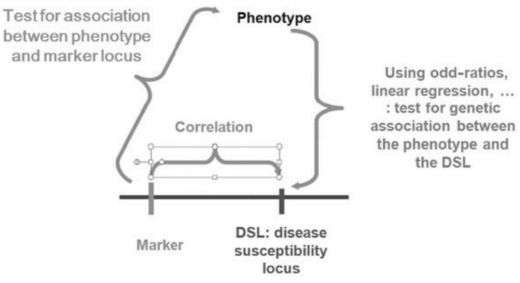


Fig. 5.4 Indirect association: Guilt by association

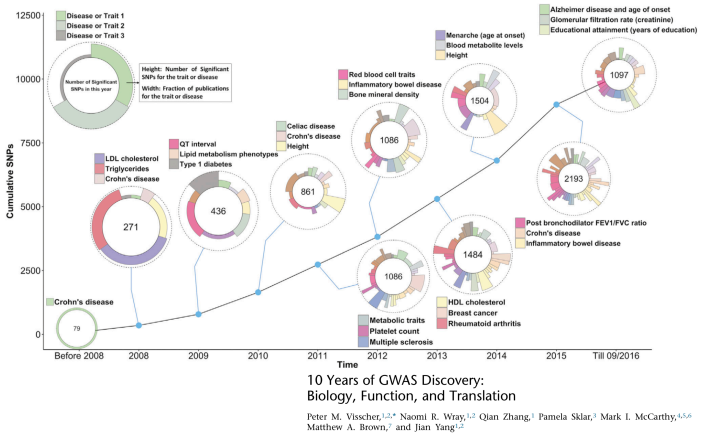
~ 1 Million SNPs can "tag" most common disease susceptibility loci

21

What happens when the causal variant is not genotyped? Doesn't GWAS miss it?

The reason why GWAS works well in identifying causal loci is LD. If the causal variant is common (say minor allele frequency > 5%), then they will be correlated with some marker in the genotyping chip. Top SNPs in GWAS loci, are called 'tag SNPs' or 'index SNPs'. We cannot know for sure which variant is causal from the association result but we can be confident that the causal variant is correlated with the top SNP (assuming the locus is not a false positive due to various possible confounding).

GWAS SNP-Trait Discovery Timeline

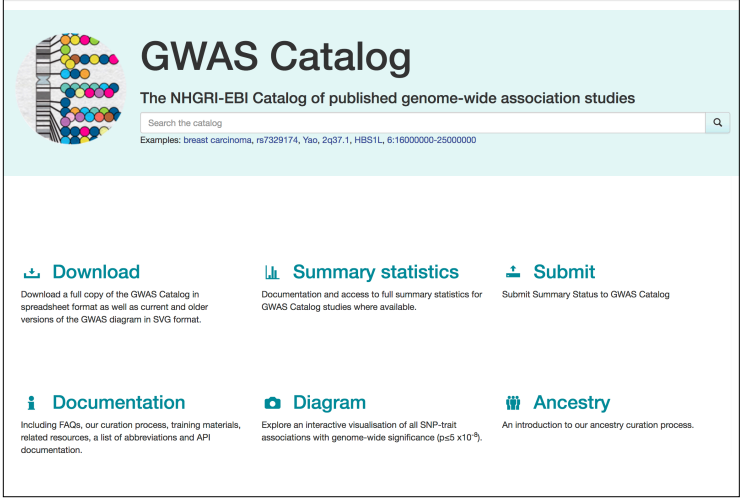


10 Years of GWAS Discovery:
Biology, Function, and Translation

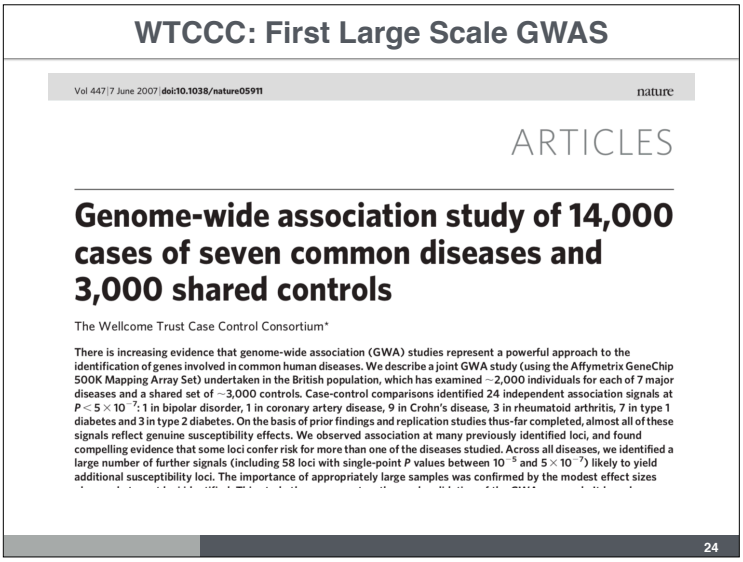
Peter M. Visscher,[1,2,*] Naomi R. Wray,[1,2] Qian Zhang,[3] Pamela Sklar,[3] Mark I. McCarthy,[4,5,6] Matthew A. Brown,[7] and Jian Yang[1,2]

GWAS have been so successful that thousands of them have been performed since 2005, with discoveries that grow continuously. As of 2020, more than 70K SNP-trait associations are reported in the GWAS catalog.

80,000+ SNP/Trait Associations



WTCCC: First Large Scale GWAS



Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

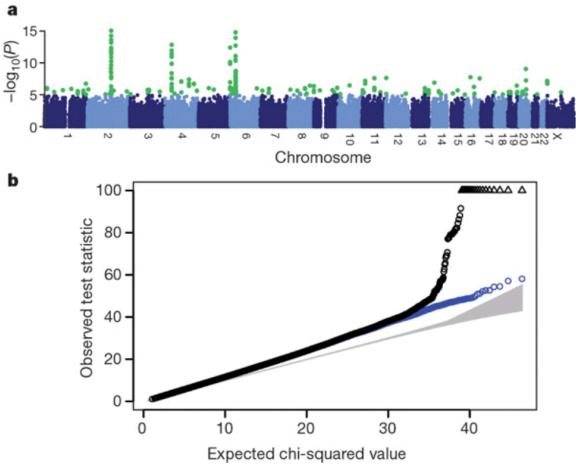
There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point P values between 10^{-5} and 5×10^{-7}) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes

the Wellcome Trust Case Control Consortium's GWAS is a landmark study of 7 common diseases, the first large scale GWAS performed.

WTCCC

- The Wellcome Trust Case Control Consortium (WTCCC)
- GWA studies of 2,000 cases and 3,000 shared controls for 7 diseases
- platform: Affymetrix 500K Set
- main paper published in 2007
- results and summaries freely available
- genotype data access granted to qualified investigators

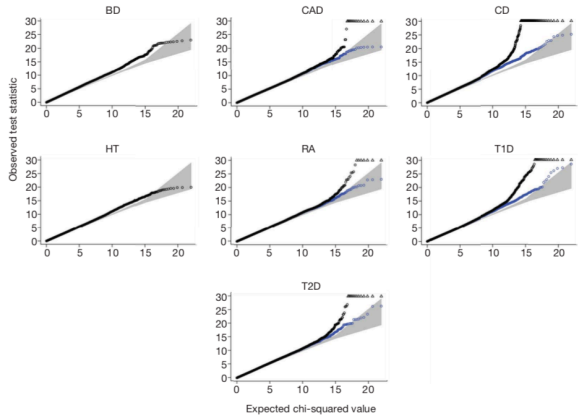
WTCCC: Population Structure



Here we see the manhattan plot and qqplot from WTCCC for the association between site (recruitment location) and genotype. Significant peaks are seen in chromosomes 1, 4, 6, and 20 indicating that there are significant differences in allele frequencies in these loci across sites.

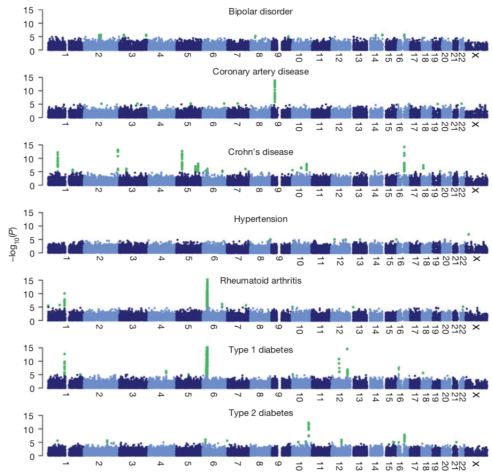
The departure from the identity line of most points, is an indication of population structure. We will get back to this concept later. Most variants show small frequency differences between sites, which does not pass GWAS significance (5e-8) due to their small effects.

QQplots for WTCCC diseases



These shows the qqplots for all 7 diseases from the WTCCC. Given a bit of departure from the identity line early on seems to indicate some population structurehere. At this time, methods for correcting for population structure were still under developed. In a later lecture, we will look into methods to account for population structure.

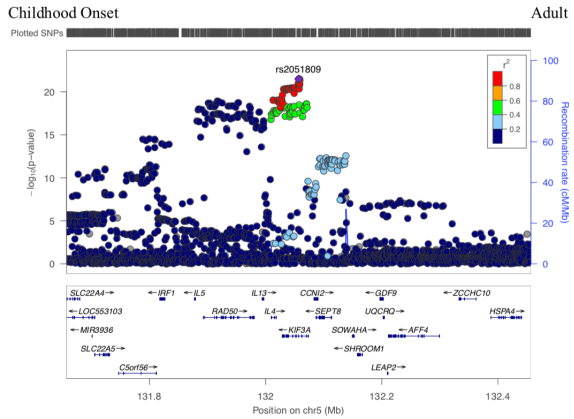
WTCCC: Manhattan Plots



28

Manhatthan plots for the 7 WTCCC diseases.

Zoom in with Regional Plots



29

We can zoom into the manhattan plot using locus zoom. These are staples of any GWAS paper at the moment. Variants in LD tend to have similar association p-values. This is a good sign. If we find variants that are significant by themselves, this may be a sign of genotyping issue.

References

- General intro to association (a bit old)
  - N. M. Laird and C. Lange, The fundamentals of modern statistical genetics. Chapters 7 and 9.
- First large scale GWAS
  - The Wellcome Trust Consortium "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls" (2007) Nature 447:661-678.
- Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., et al. (2020). A brief history of human disease genetics. Nature, 577(7789), 179–189. http://doi.org/10.1038/s41586-019-1879-7
- Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome- wide association studies. Nature Reviews Genetics, 1–18. http://doi.org/10.1038/s41576-019-0127-1

30