Genome-Wide Association Studies

—

Ryan Collins, Gerissa Fowler, Sean Gamberg, Josselyn Hudasek & Victoria Mackey

Introduction

- The next big advancement in the field of genetics after the Human Genome Project was Genome-wide association studies.
- This study works to associate genes with traits/diseases throughout the genome using SNP's to determine the location of these genes.
- These studies required many advancements, particularly financially, costing ~$250 million in the first 5 years globally.
- From this thousands of loci associated with hundreds of diseases have been found

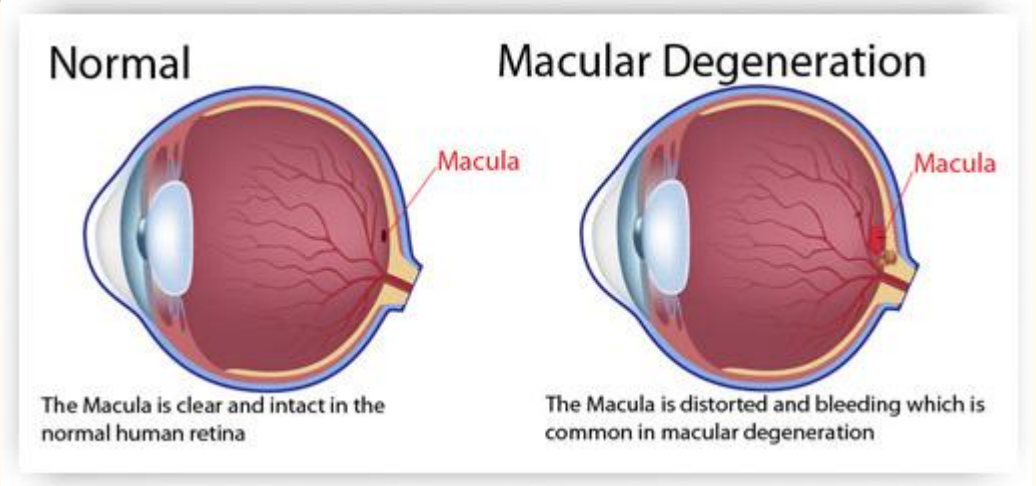
http://www.genome.gov/Glossary/index.cfm?id=91

Before GWAS

- A 1996 paper titled “The Future of Genetic Studies of Complex Human Disease” by Neil Risch and Kathleen Merikangas proposed that family linkage studies could only identify alleles that increased risk more than two fold.
- Genetic mapping at this time was restricted by the technology available, as scoring thousands of markers in the genome between thousands of individuals with and without a disease or trait was impossible without modern advancements.

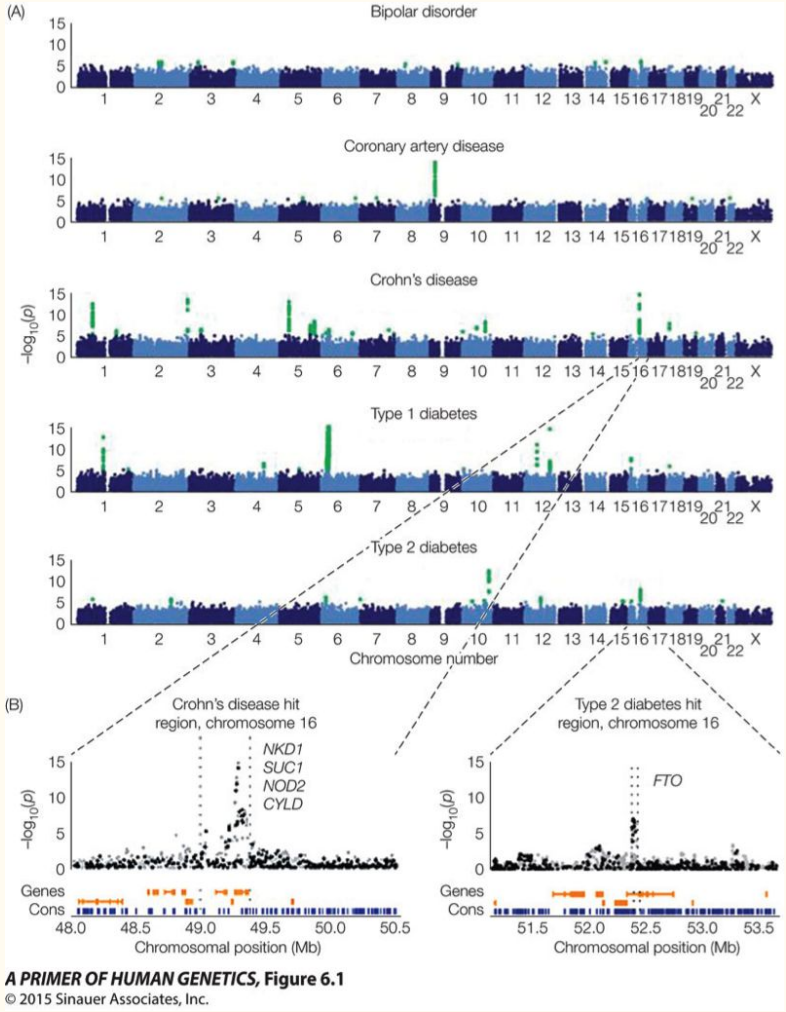
First GWAS

- Age related macular degeneration was the first positively associated disease which was found in 2005, using 96 Caucasian individuals with this disease and only 50 without.
- 100,000 SNP's were used in this study, which found a block of SNP's in the complement factor H gene (*CFH)* that showed a fourfold risk increase in heterozygotes and sevenfold in homozygotes.
- A subsequent study found another associated loci (*HTR1)* which has a risk factor increase of 11- fold in homozygotes.



Wellcome Trust Case Control Consortium

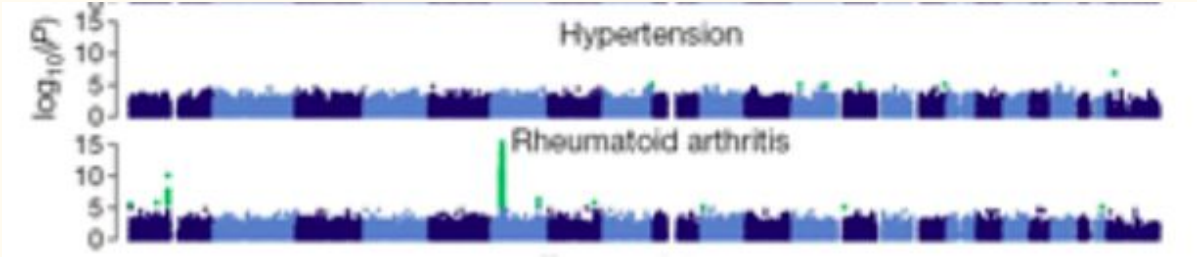
- Additional experiments were conducted that included:
  - fine mapping of entire genes
  - altered gene expression due to polymorphisms
  - Showing an ability to quickly map genetic factors without the use of family studies
- In 2007 the WTCCC conducted a study that found gene associations for seven different diseases finding 23 associations in total.



A PRIMER OF HUMAN GENETICS, Figure 6.1

- Gene associations for each disease:
  - bipolar disorder = 0
  - coronary artery disease = 1
  - crohns disease = 9
  - type 1 diabetes = 7
  - type 2 diabetes = 3

Figure 6.1 Results of the Initial WTCCC Study in 2007. (A) Manhattan plots for five of the seven diseases in the study. Each plot shows the significance of the association at each of 360,000 SNP sites ordered by position along the chromosomes, indicated as the negative logarithm (base 10) of the *p*-value: highly significant associations generate peaks (shown in green). (B) Fine-scale association mapping of two of the peaks shows that for each highly significant SNP, there tends to be a cluster of related associations due to linkage disequilibrium in the vicinity. Dotted vertical lines show the locations of recombination hotspots that set the limits of the haplotype block. Black points are actually genotyped, and gray ones are imputed statistically. Tracks beneath each plot show the locations of candidate genes in the vicinity and regions of sequence conservation across mammals. (After WTCCC 2007.)



- No genes were found for hypertension
- Four genes were found for rheumatoid arthritis, one of which is shared with type 1 diabetes

WTCCC

- A 18-fold increase in risk for type 1 diabetes was found in homozygotes at the HLA complex
- 21 of the associations (out of 23 found) showed a per allele risk of twofold or less
- This study failed to replicate five associations of candidate genes previously found. It did however outline the importance and significance of using GWAS to discover common and rare disease associations.
- As of 2016, 16696 unique SNP-trait associations have been cataloged on the GWAS cataloging website, which includes traits as well as diseases.

Basic Methodology

Basic Methodology – Purpose

- To determine if either allele at a given SNP is overrepresented in either cases or controls
  - Cases: individuals with a disease of interest
  - Controls: individuals without the disease of interest

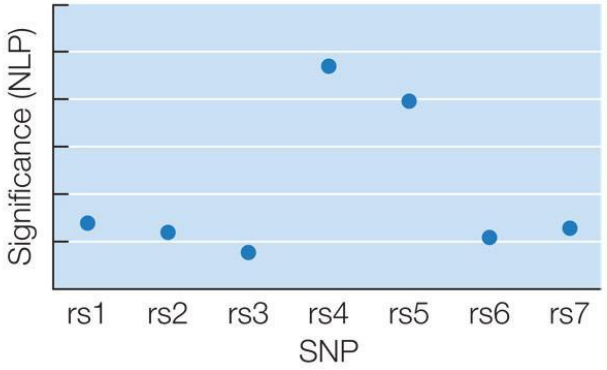
- SNPs that shows a stronger association than expected by chance are regarded as markers of candidate genes

Basic Methodology – Procedure

1. Gather sample
   - Usually at least 1000 cases and 1000 controls
2. DNA sample
3. Hybridize DNA to the array
4. Identify (“call”) genotypes
5. Impute additional SNPs
6. Perform statistical analysis
7. Interpret findings

Example of a Mendelian trait with complete penetrance

	SNP							
Individual	**rs1**	**rs2**	**rs3**	**rs4**	**rs5**	**rs6**	**rs7**	**Status**
ID001	A	T	C	C	A	G	A	Case
ID002	A	C	T	C	A	G	A	Case
ID003	A	C	C	T	G	–	A	Control
ID004	T	T	C	T	A	G	A	Control
ID005	A	T	C	T	G	G	A	Control
ID006	A	C	C	C	A	–	A	Case
ID007	T	C	T	T	G	–	A	Control
ID008	T	T	C	C	A	G	A	Case
ID009	T	C	T	C	A	–	G	Case
ID010	T	T	T	T	G	–	A	Control



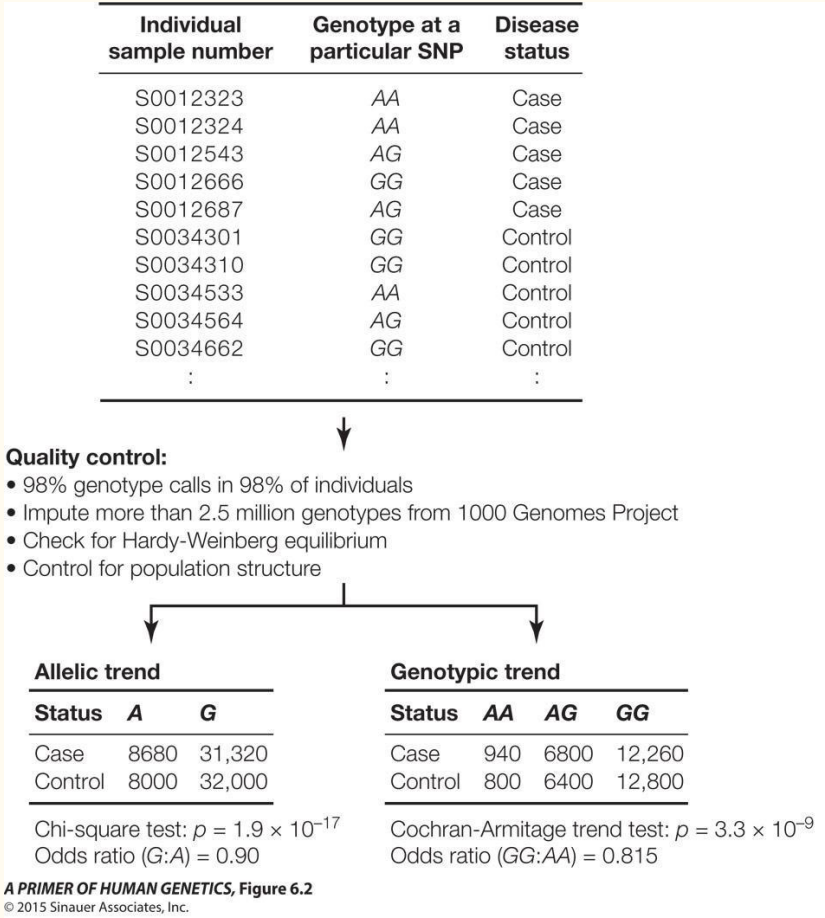
A PRIMER OF HUMAN GENETICS, Box 6.1

Power of a GWAS

- Function of:
  - Number of individuals of each class that are sampled
  - Effect size of the allele
  - Frequency of the allele
  - Variation within each genotype class

Association test

- Allelic trend test
  - Contrasts frequencies of the two alleles of a SNP in cases and controls
  - Chi-square test
- Genotypic trend test
  - Compares 3 genotype classes
  - Assuming heterozygotes have risk intermediate to the homozygotes
  - Cochran-Armitage trend test



Individual sample number	Genotype at a particular SNP	Disease status
S0012323	*AA*	Case
S0012324	*AA*	Case
S0012543	*AG*	Case
S0012666	*GG*	Case
S0012687	*AG*	Case
S0034301	*GG*	Control
S0034310	*GG*	Control
S0034533	*AA*	Control
S0034564	*AG*	Control
S0034662	*GG*	Control
:	:	:

Quality control:
- 98% genotype calls in 98% of individuals
- Impute more than 2.5 million genotypes from 1000 Genomes Project
- Check for Hardy-Weinberg equilibrium
- Control for population structure

Allelic trend

Status	*A*	*G*
Case	8680	31,320
Control	8000	32,000

Chi-square test: $p = 1.9 \times 10^{-17}$
Odds ratio (*G*:*A*) = 0.90

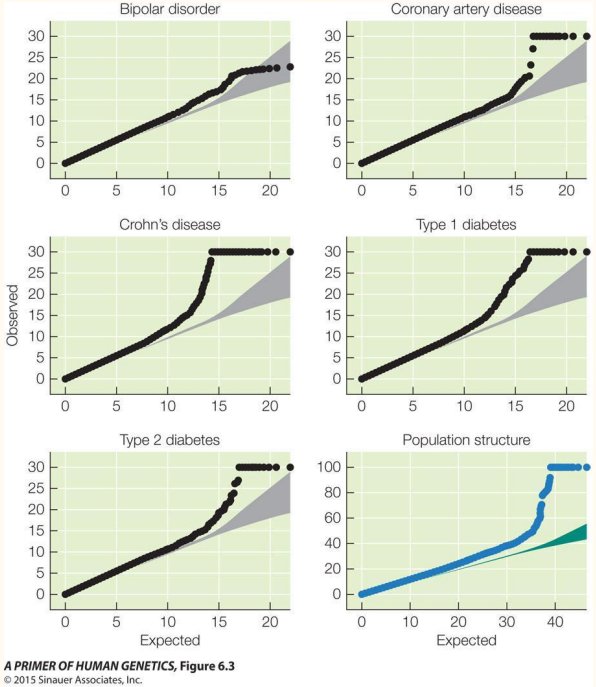
Genotypic trend

Status	*AA*	*AG*	*GG*
Case	940	6800	12,260
Control	800	6400	12,800

Cochran-Armitage trend test: $p = 3.3 \times 10^{-9}$
Odds ratio (*GG*:*AA*) = 0.815

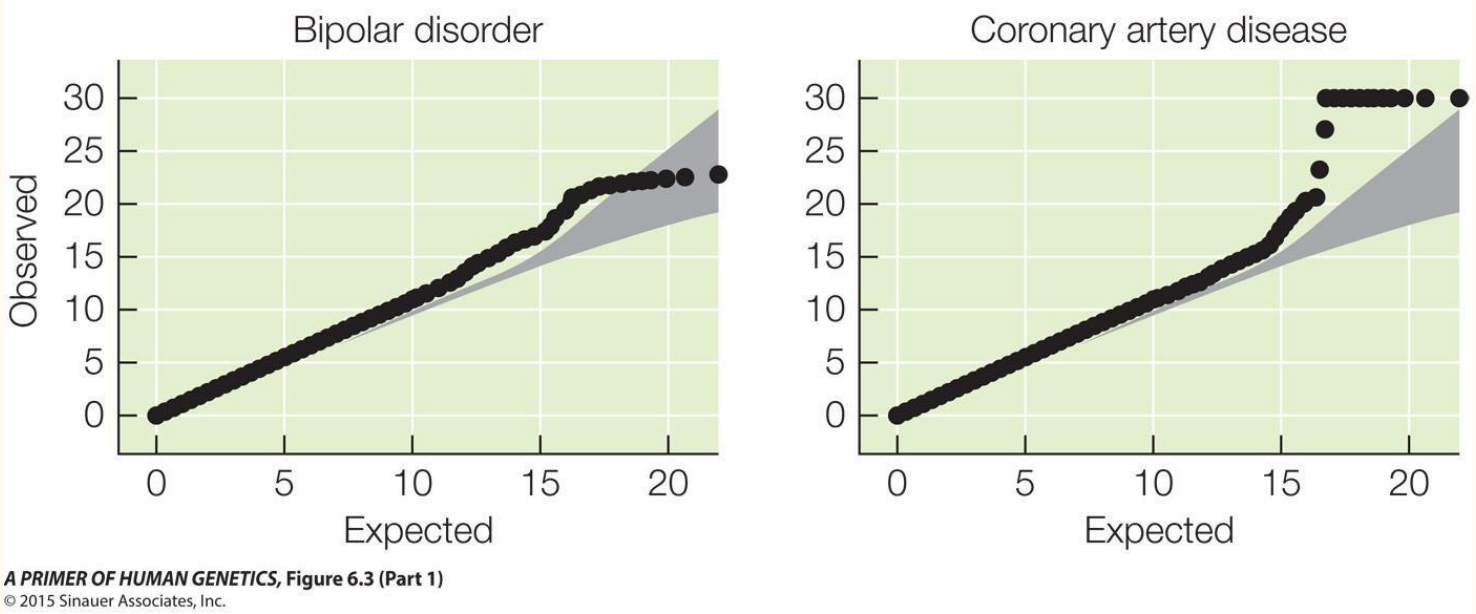
***A PRIMER OF HUMAN GENETICS*, Figure 6.2**

Q-Q Plots

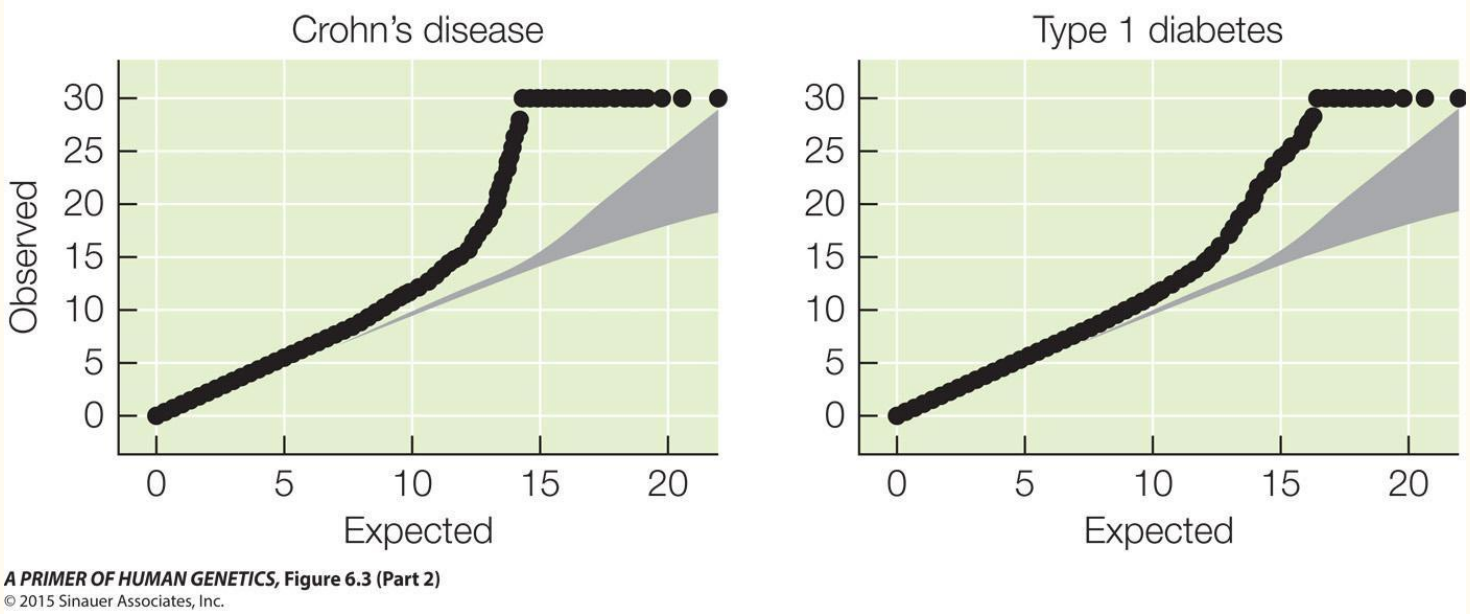


A PRIMER OF HUMAN GENETICS, Figure 6.3

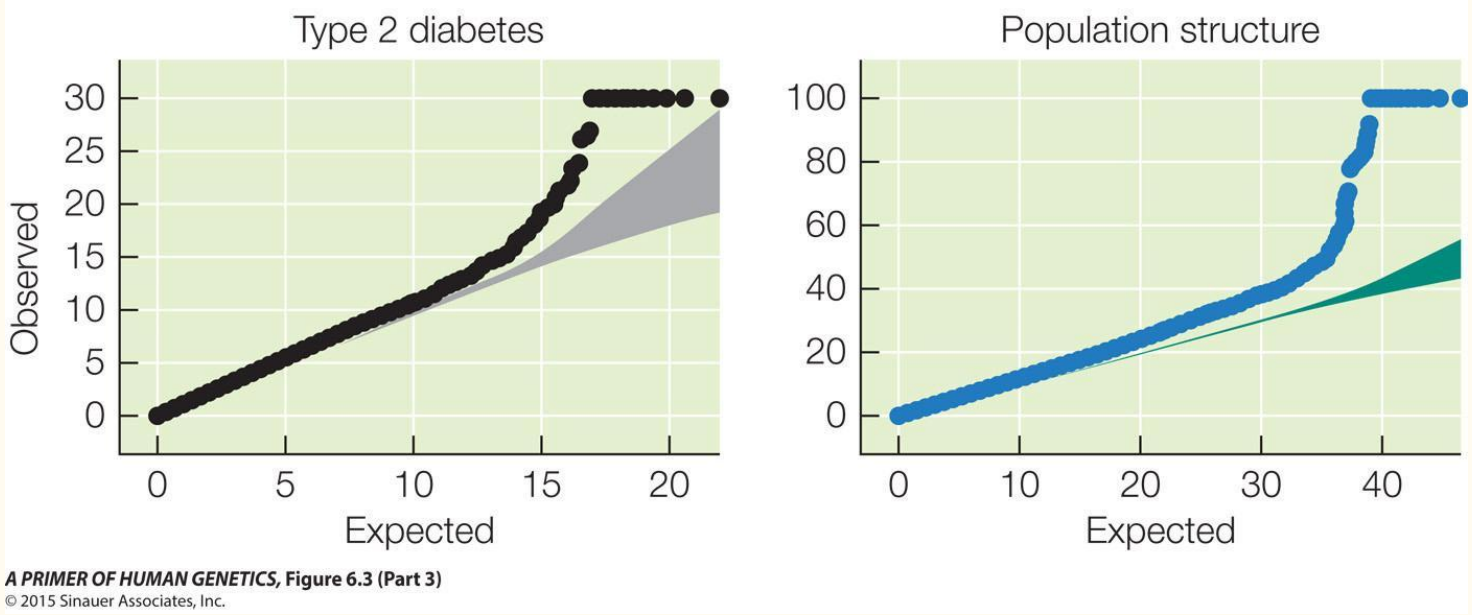
- Observed vs. Expected range of values on *y* and *x* axes, respectively
  - Chi-square or NLP values
- Gray area indicates the expected range of values under the null hypothesis
- Points well above the gray area correspond to GWAS “hits”



***A PRIMER OF HUMAN GENETICS*, Figure 6.3 (Part 1)**



***A PRIMER OF HUMAN GENETICS*, Figure 6.3 (Part 2)**



***A PRIMER OF HUMAN GENETICS*, Figure 6.3 (Part 3)**

What statistical threshold must be exceeded to conclude a SNP contributes to a trait?

- Generally, $p = 0.05$
  - If 20 tests are performed, chances are that 1 of them will yield a false positive
  - If 100 000 tests are performed, 5 000 type 1 errors will occur

- Bonferroni correction
  - Study-wide significance is adjusted to α divided by the number of tests

$$\frac{0.05}{1\,000\,000\ tests} = 5 \times 10^{-8}$$

Gold Standard

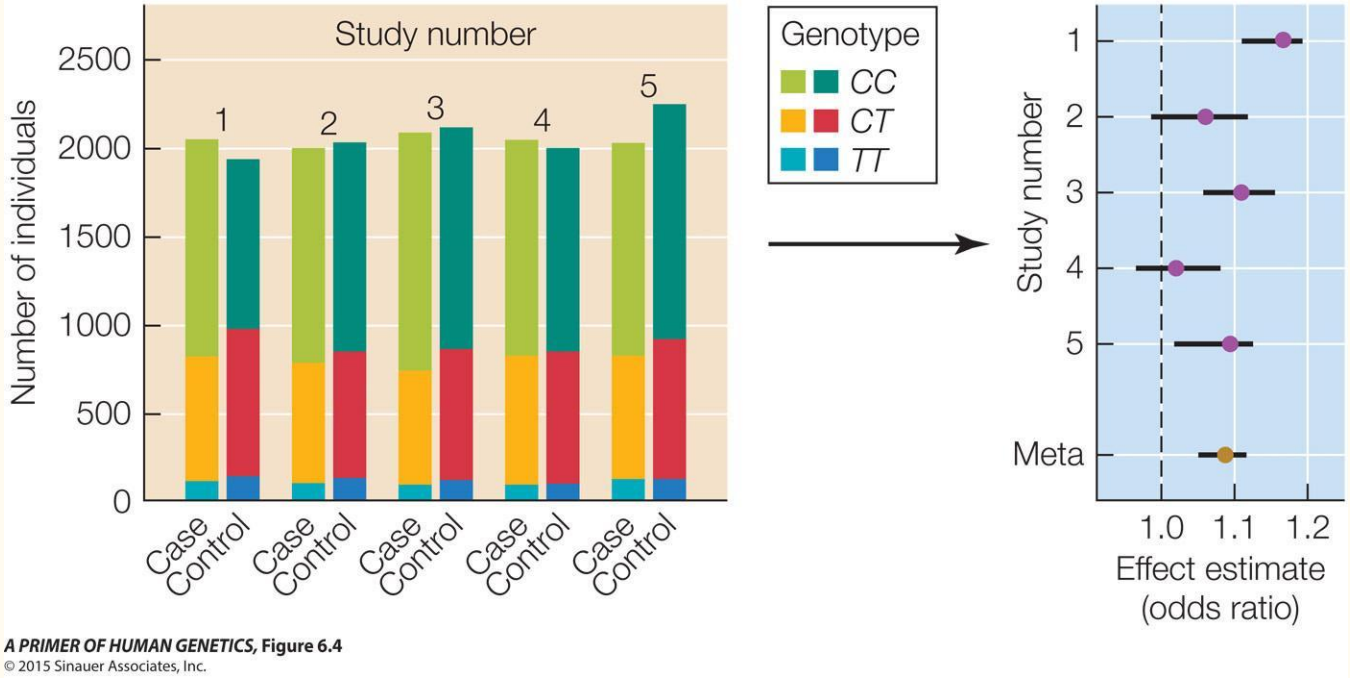
$p < 5 \times 10^{-8}$ with replication

- Not required that both phases exceed the threshold independently
  - Together, should show the effect in the same direction – combined *p*-value is smaller than that of either phase

Meta-analysis

- Combining *p*-values from multiple studies
  - Resulting *p*-value can be thought of as an average *p*-value weighted by the study sizes and boosted by the combined sample size
- Results can be visualized in a forest plot
- Usually, if the effect is in the same direction in multiple studies, it can be regarded as true
  - Even if the individual effects are small

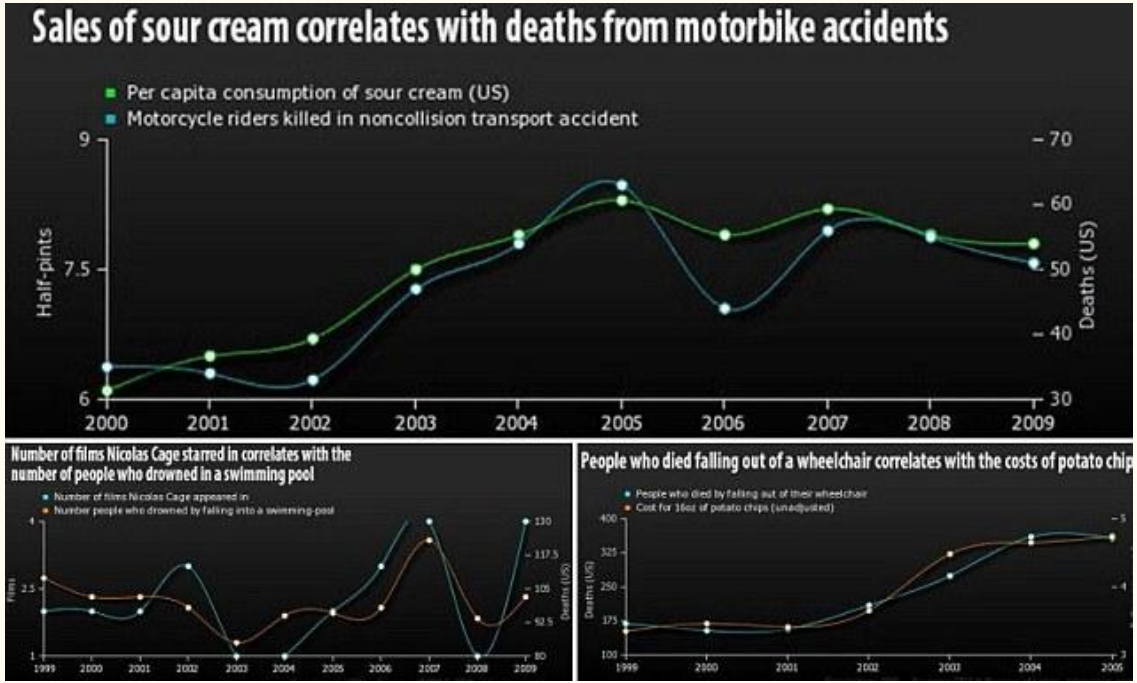
Meta-analysis converges on the best estimate of the true odds ratio



A PRIMER OF HUMAN GENETICS, Figure 6.4

Population Structure:

Misleading the analytically-challenged since the beginning of GWAS...



Hidden population structures in a study can present data with apparent genomic patterns that would not otherwise arise if the population sample had been picked more carefully

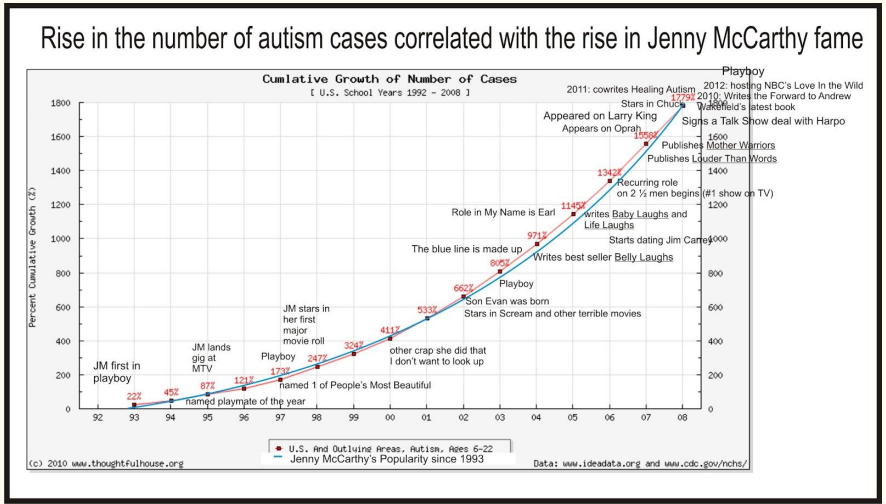
For Example...

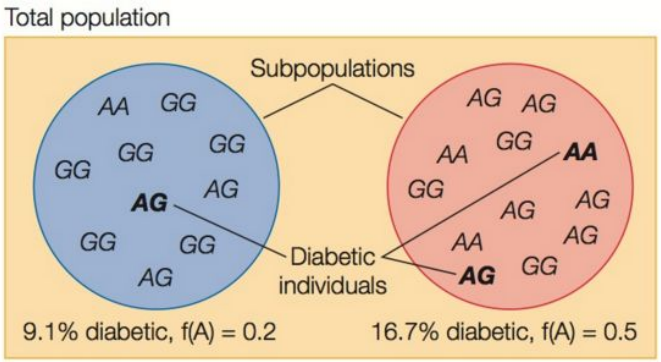
If you are looking for a gene connected to religious affiliations in a sample of Americans but that sample includes Americans of Irish or Iranian descent there will be an apparent correlation between Irish genes in catholics as well as Iranian genes in muslims.

False Correlation Due to Population Structure

Not only is population structure at work here, but cultural and environmental effects are as well.

This was a silly example but these effects can occur less obviously and do cause problems in GWAS that must be accounted for.





Blue subpopulation

	AA	*AG*	*GG*
Case	80	640	1280
Control	800	6400	12,800
Case/control	0.1	0.1	0.1

Red subpopulation

	AA	*AG*	*GG*
Case	200	400	200
Control	1000	2000	1000
Case/control	0.2	0.2	0.2

Total population

	AA	*AG*	*GG*
Case	280	1040	1480
Control	1800	8400	13,800
Case/control	0.155	0.124	0.107

Odds ratio (*A*:*G*) = 1.2 $p = 10^{-8}$

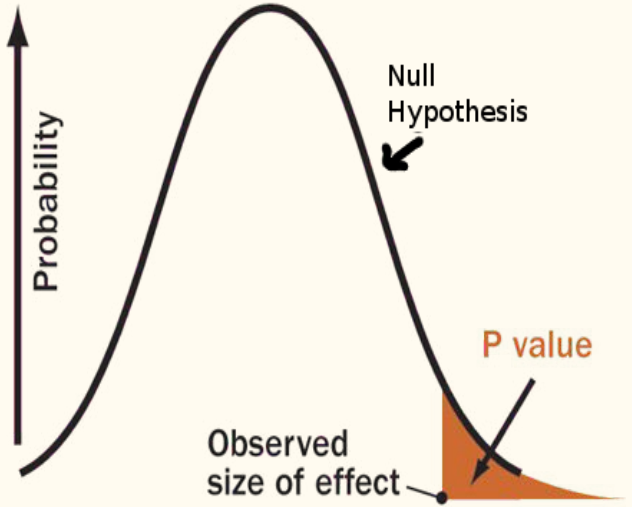
Figure 6.5 Population Structure. If a population consists of two subpopulations, which may be unrecognized, then despite the absence of any association in either subpopulation, there may appear to be a spurious correlation between genotype and disease risk. In this hypothetical example, 9.1% of the individuals in the blue population have diabetes irrespective of genotype, and there is a minor (A) allele frequency of 0.2; in the red population, 16.7% of the individuals have diabetes, and the A allele has a frequency of 0.5. In the total population (yellow), the A alleles are more likely to be observed in people with diabetes, which gives an odds ratio of 1.2 and genome-wide significance, yet the odds ratios in the two sub-populations are 1.0.

- No association between subpopulations makes it appear that there is a correlation between genotype and disease risk
- Blue pop'n with diabetes has the disease irrespective of genotype
- Red pop'n with diabetes has a minor allele frequency (A allele)
- Total population results in an odds ratio of 1.2 which is considered genome-wide significant, even though separately the sub-population odds are not significant

Genomic Control

- One of the simplest ways to correct for population structure
- An adjustment easily made to all p-values, downweighting by a factor called λ_{gc} so that the nulls may be closer to random expectation
- Those p-values that still remain significant indicate a true positive (detects the condition when the condition is present)

- When population structure is known, biases can be removed by manipulations in favor of equal population distribution as well as the formation of a control with $\lambda_{gc} = 1$



Genomic Control

- Typically, obtaining equal distributions requires removing individuals from the sample which results in a lower sample size
- The homogeneity of the allele frequencies within the populations tend to make up for this loss
- However, this is one of the main challenges in GWAS, as sample populations tend to be restricted to very specific groups.

- This specificity could require looking into past generations of relatives of the participant which can be difficult and time consuming in some cases.

Genomic Control

And even when it is not a small sample size more problems arise as we start to notice that it isn't enough, for example, to just ensure the relatives are from the Netherlands but we must also specify that they are from the Northern Netherlands.

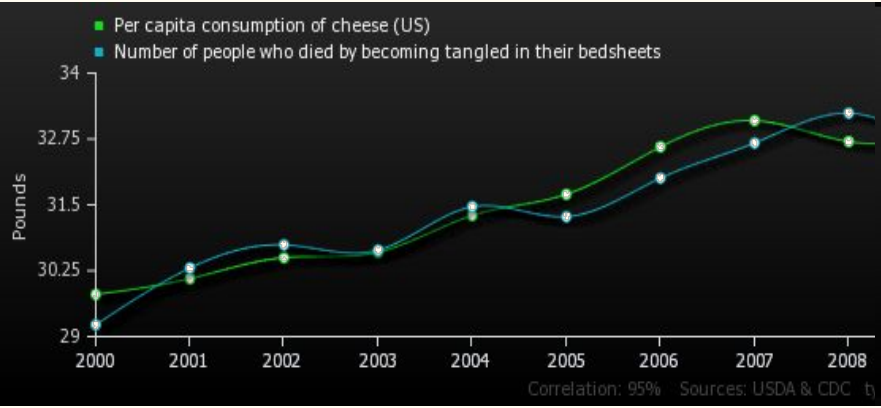
This leads to more work for distinguishing those borders that classify a subject as 'Northern' as well as increasing the difficulty for constructing appropriate controls.

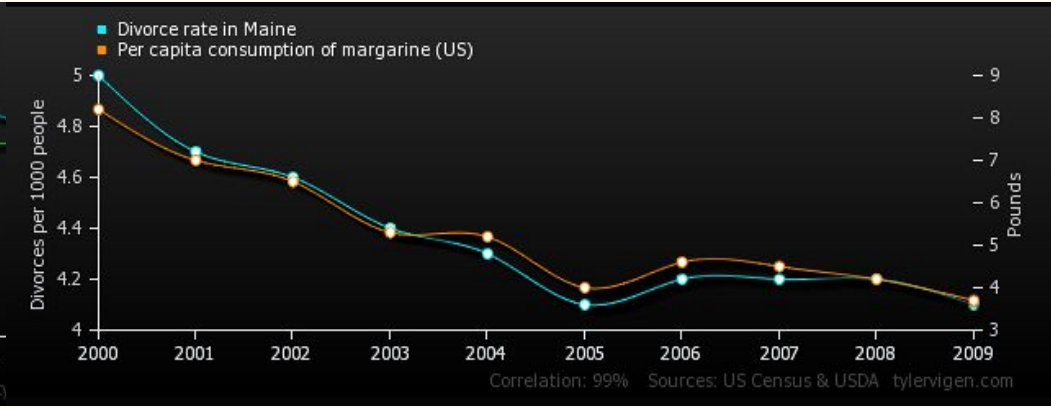
Measuring Population Structure & Making Estimates with Statistical Models

- **Principal Component Analysis (PCA)** [Chapter 3] and **Multi-Dimensional Scaling (MDS)** [displaying data under more than one dimension] can be used to reduce data by sifting through the genotypes to give more objective indications of population structure than collected self-reports
- **PCs** can be used to distinguish people according to their African, Asian or European Ancestry which also leads to the detection of smaller scale population structure
- If the PC values are used as continuous covariates in logistic regression models that involve the case versus control as the dependent variable it has been proven to remove the need for λ_{gc} adjustments

Measuring Population Structure & Making Estimates with Statistical Models

However, mixture-studies can also occur in which case the uniform λ_{gc} adjustments are still made to the continuous traits leading to analogous studies that are very effective at preventing false positive outcomes in result of population structure.

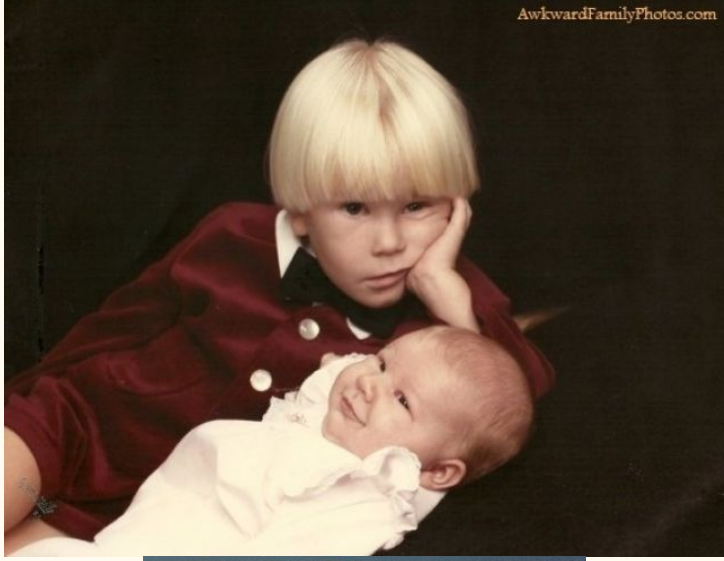




Family Specific GWAS

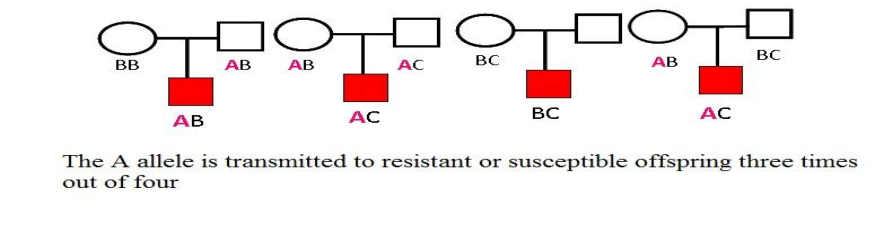
Population Structure is no longer an issue!

- Used for the enrichment of alleles in siblings affected by disease or transmission disequilibrium tests (TDTs)



TDTs (Transmission Disequilibrium Tests)

- No need for study controls
- Answer the question: “Are either of the alleles in a heterozygous parent more commonly transmitted than the other to an affected offspring “
- Example: if 1000 children are affected as well as both biological parents, 500 of which are heterozygous, it would be expected that under random assortment 250 of each allele should be transmitted, however if 300 transmissions of one allele was observed then it can be significantly associated with the disease
- Difficult to perform: obtaining biological samples from parents can be time consuming and costly

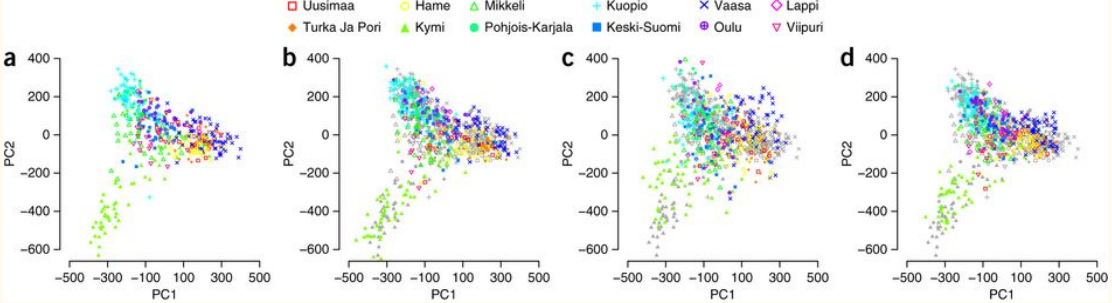


The A allele is transmitted to resistant or susceptible offspring three times out of four

Genome Wide Association Studies Success

- **For GWAS to be successful:** there must be strong linkage disequilibrium (LD)[non-random association of alleles at different loci] between tagged SNPs in the genotyping panel and the causal variants
- **Humans are ideal:** because the even spacing of 500,000 SNPs ensure that there is at least one SNP every 10kb in the genome

Fine Mapping



- Strong LD haplotype blocks tend to be 20 kb and 100 kb long (Ch.3) and each of the blocks are represented by 5-10 SNPs
- This allows a strong probability of polymorphisms being located in haplotype blocks by known SNPs
- when multiple SNPs capture the polymorphisms they form a **tight cluster** on **fine-scale association maps**

HapMap Project

- Aim to generate a fine-scale map of haplotype blocks that can be viewed at the hapmap.ncbi.nlm.nih.gov website
- describes specific variations in our DNA, where they occur and how they are distributed among populations around the world
- Haplotype blocks in humans are very similar between human populations which makes the HapMap project beneficial as it can be used as a reference when completing associative studies

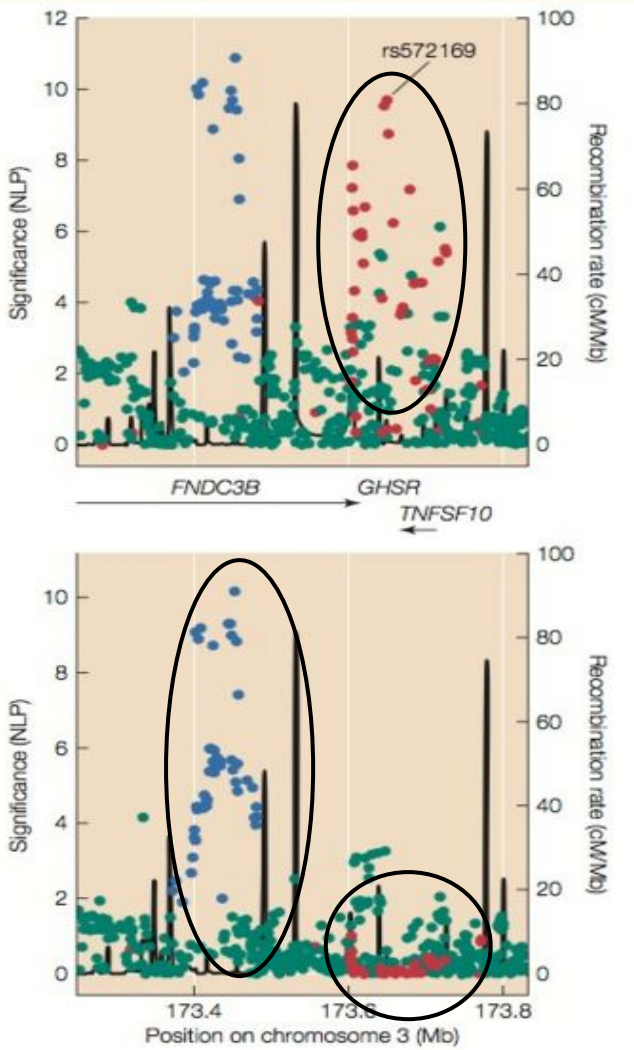
- Any disease-associated SNP's prior to the spread of humans outside of Africa should be tagged by common polymorphisms in any population
- Limits of the blocks are established by hotspots of recombination (black peaks in recombination rate profiles)

Major Commercial Genotyping Platforms

- Estimated that 90% of all common polymorphisms within Europeans and Asians are tagged by the SNPs on Illumina and Affymetrix platforms (discussed later in this lecture)
- Since LD is significantly less common in Africans the GWAS require greater depth genotyping
- Although since the LD intervals are reduced cross-population studies in Africans can provide higher-resolution maps of the causal sites

- Lead SNP is in linkage disequilibrium (red dots) with 25 other variants within a 200kb interval and is surrounded by two hot spots (show recombination rates) on the GHSR gene
- However, after fitting it into the regression the significance was highly reduced but the blue SNP LDs remained highly significant, demonstrating that the blue SNP has caught an independent association with the gene FNDC3B

Figure 6.6 Complex Association. These fine-scale association maps show an example of two independent clusters of associations with height in the vicinity of the growth hormone receptor gene *GHSR* at chromosomal location 3q26.31. The lead SNP, rs572169, is in linkage disequilibrium with 25 other variants in a 200 kb window bounded by two hotspots of recombination (the tallest black vertical spikes, which show recombination rates). After fitting the rs572169 SNP in the regression, the significance of the red SNPs was reduced to NLP < 1, but another set of 30 or so blue SNPs in the adjacent interval covering the *FNDC3B* gene remain highly significant, indicating that they capture an an independent association. (After Lango Allen et al. 2010.)



Once an Association Peak has Been Detected...

Imputation is used for further analysis of potential disease causing variants

Imputation: is achieved by using known haplotypes in a population, (HapMap or the 1000 Genomes Project) by estimating the most likely genotype at the untyped SNP, which is based on the extent of LDs, in order to determine if there is an association with a specific trait

- Association tests performed (just as indirectly typed SNPs) and association signals are detected (using open-source software such as Beagle)
- Once the novel variants are identified they are imputed into a larger sample
- Then type the new SNPs with a targeted assay to confirm the accuracy of the imputation

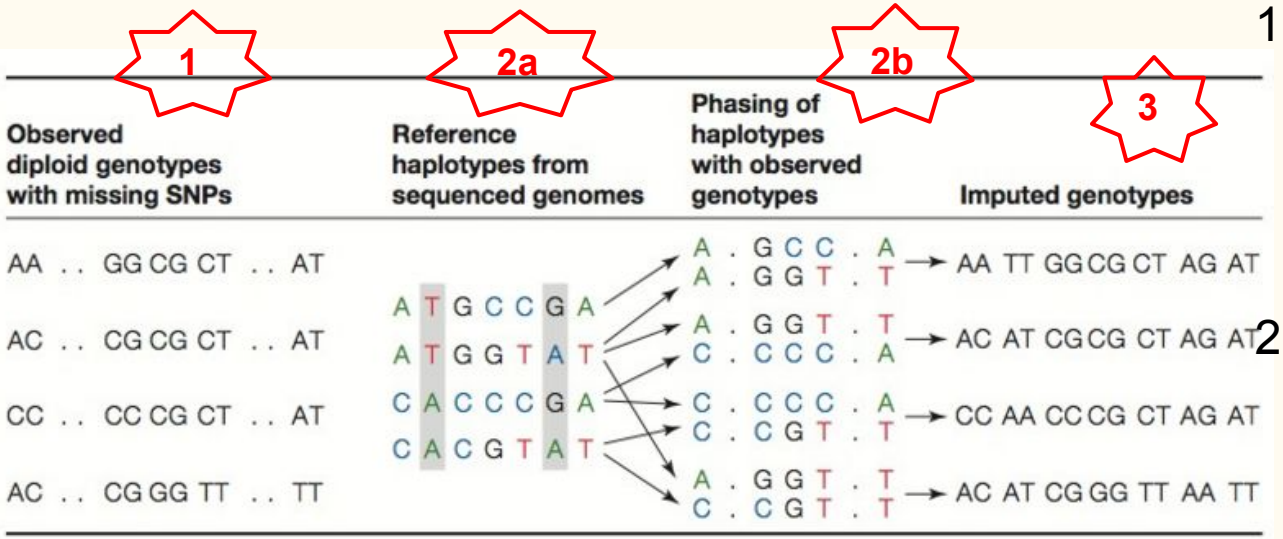


Figure 6.7 Phasing and Imputation. Genotype data is generally ascertained as a series of independent genotypes (AA, AC, CC, etc.) rather than as haplotypes. The first step in imputation is to phase the haplotypes; that is, to estimate the series of haploid genotypes on each of the two contributing chromosomes observed in a diploid individual. This step is straightforward for pairs of homozygotes or for a homozygote and heterozygote pair (AA GC could only be a combination of AG and AC haplotypes), but double heterozygotes cannot be phased directly (CG CT could either be CC and GT or CT and GC). Maximum likelihood methods are used to optimize the fit of the observed genotypes to the frequency and identity of haplotypes in a reference population. Once this is done, the identity of missing genotypes (. .) in the observed sequence is "imputed" by reference to the sequences in the reference panel (the genotypes under the gray bars). These imputed genotypes are then used for association studies.

1. Phase the the genotype data into haplotypes (estimate series of haploid genotypes on each chromosome observed in diploid individuals)
2. Use maximum likelihood methods to optimize the fit of the observed genotype to the frequency and identity of the haplotypes in the referenced population
3. The identity of missing genotypes are **imputed** by reference to the sequences in the reference panel
4. Those imputed genotypes are then used for future association studies

In addition to Imputation Methods...

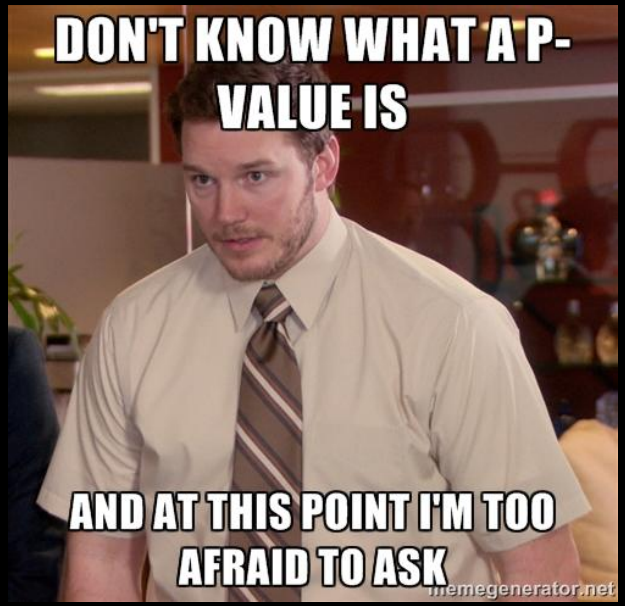
Custom genotyping chips have been manufactured such as the "immunochip" and the "metabochip" that include all SNPs that are likely to be associated with common immune and metabolic diseases respectively.

These chips function at a fraction of the cost of the whole-genome genotyping and can give detailed interrogations of hundreds of thousands of human candidates

"Goldilocks" variants (those with modest effect sizes) can also be recognized by these chips which are not normally included in standard arrays

More on this later…

Causal Inference



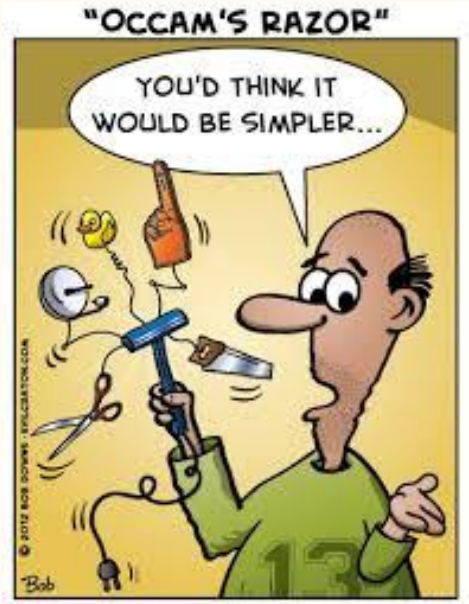
Causal Inference

- When 1 SNP @ Locus is associated with a disease or trait, it is reasonable to assume additional SNP in same gene is involved

- 2 Interpretations

- Occam's Razor

- Another Possibility?



Causal Inference

- Conditional Association used to distinguish effects of variants at one locus
- Fit SNP with smallest *p* value (linear regression model)
- By accounting for known first effect of SNP, it can be questioned whether other SNPs can explain more of the trait

Causal Inference

Haplotype	BMI
C C G A A C	21.4
C C G A A C	30.3
C C G A A C	28.2
C C G A A C	23.8
C C G A A C	33.7
C C G A A C	27.9
C C G A A C	22.0
C C G A A C	24.4
C C G A A C	30.8
C C G A A C	29.2
C C G A A C	27.7
C C G A A C	25.1
G T A A T C	21.4
G T A A T C	30.3
G T A G T C	**40.6**
G T A A T C	23.8
G T A A T C	33.7
G T A A T C	27.9
G T A A T C	22.0
G T A A T C	24.4
G T A A T C	30.8
G T A A T G	**36.2**
G T A A T C	27.7
G T A A T C	25.1

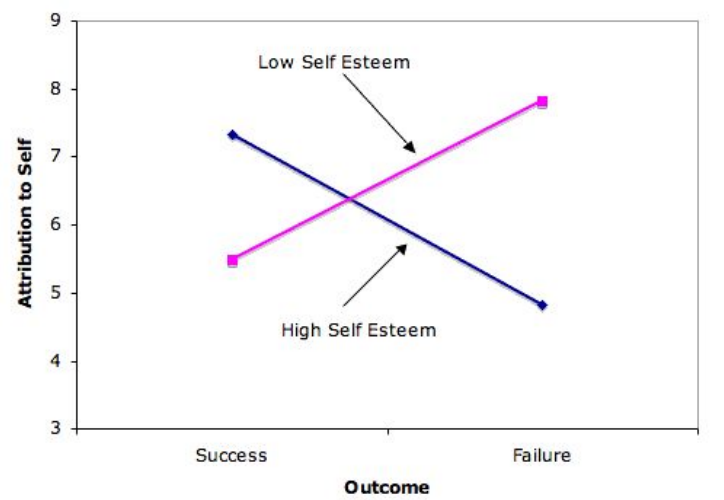
Blue haplotype mean BMI = 27.0
Orange haplotype mean BMI = 28.7
Haplotype with green allele mean BMI = 38.4
Orange without green mean BMI = 26.7

Figure 6.8 Synthetic Association. Synthetic association occurs where the apparent association of a disease or trait with a common variant or haplotype is actually driven by one or a small number of rare variants that happen to be enriched on that haplotype. In this hypothetical example, the actual distribution of body mass index (BMI) phenotypes is the same for the blue and orange haplotypes, yet the mean is higher for the orange haplotype due to the strong effects of two rare (green) variants. Without those variants, the orange haplotype actually shows a slightly reduced effect on BMI.

Interaction Effects

Interaction Effects

- What are they?
- 2 IV interact if the effect of one variable differs depending on the level of the other variable
-



Interaction Effects

- GWAS designed to capture common variant contributions to narrow sense heritability
- Should be affected by interaction effects between 2 or more genotypes (GxG) or between Genotype and Environment (GxE)
- However GWAS studies have not provided strong support for departures from additivity

Interaction Effects

- Allele effects may be independant of one another and from environmental variables
- Effects may be small when compared to main effects (low statistical power)
- Interaction effects may be heterogenous
- May be difficult to identify proper environmental variable or alternative genotype to model

Interaction Effects

- Many authors argue interaction effects are highly ubiquitous
- Only a handful of robust environmental modifications of genotypes effects on disease have been reported
- GxE interaction effects were found to make only a modest contribution to genetic regulation of transcript abundance in peripheral blood

Polygenic Risk and SNP-Based Heritability

Polygenic Risk and SNP-Based Heritability

Three Major Reasons for Conducting GWAS:

1. **Find** specific genes and variants that contribute to disease risk or phenotypic variance

2. **Understand** more about the genetic architecture of a trait

3. **Develop** genetic predictors

Genetic Architecture

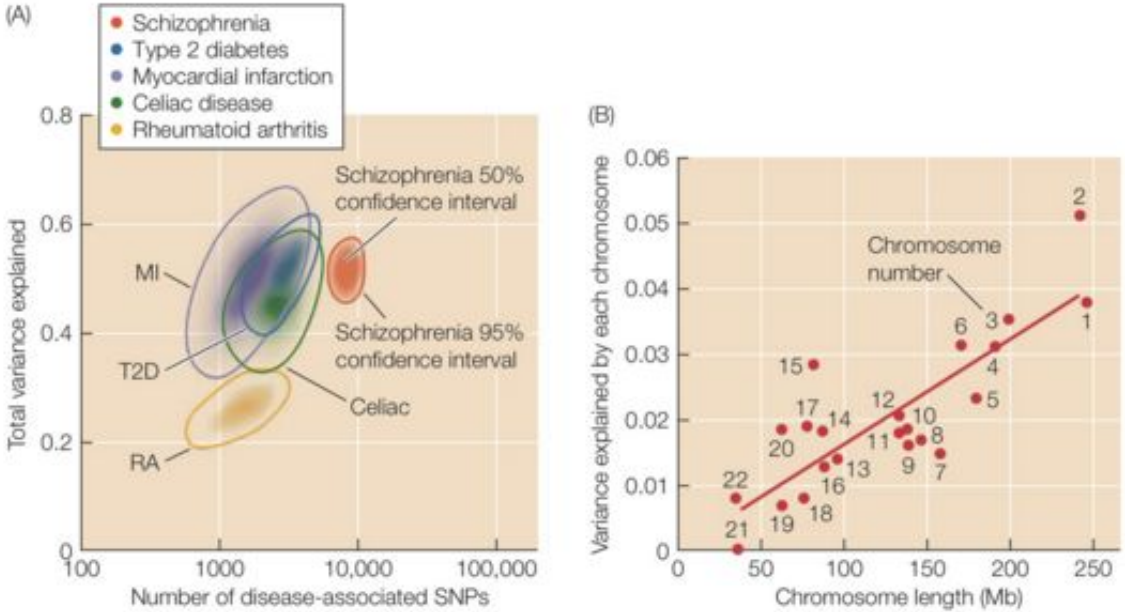
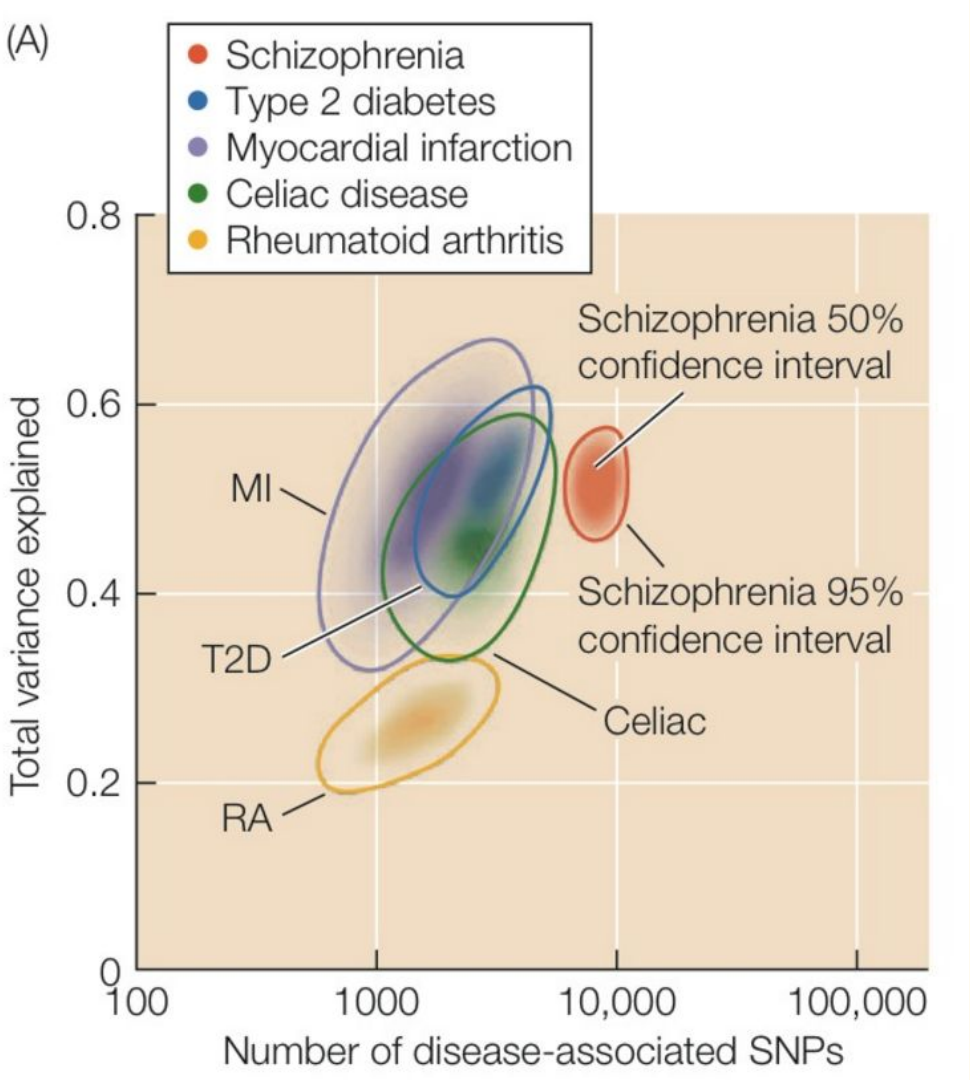
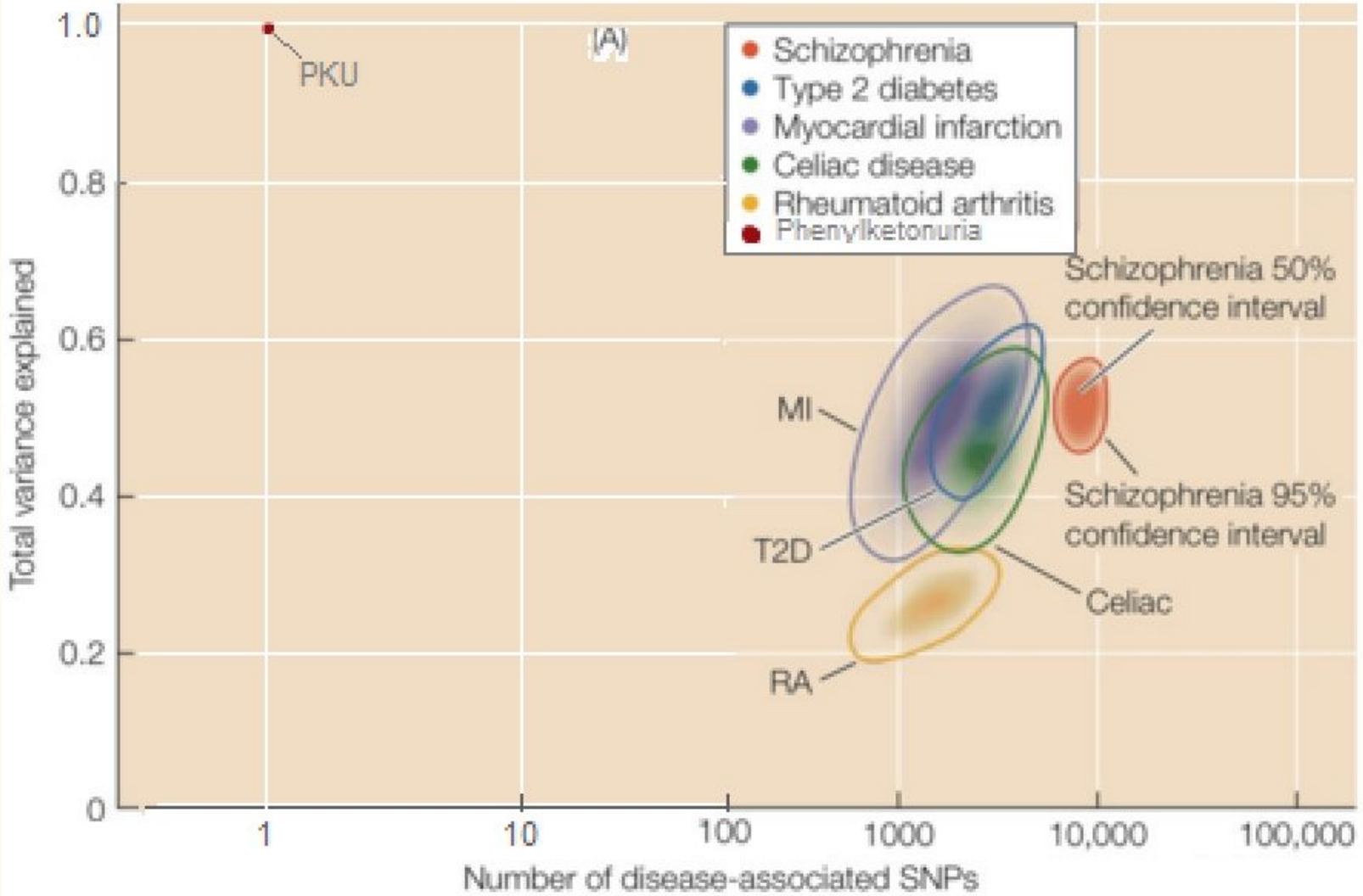
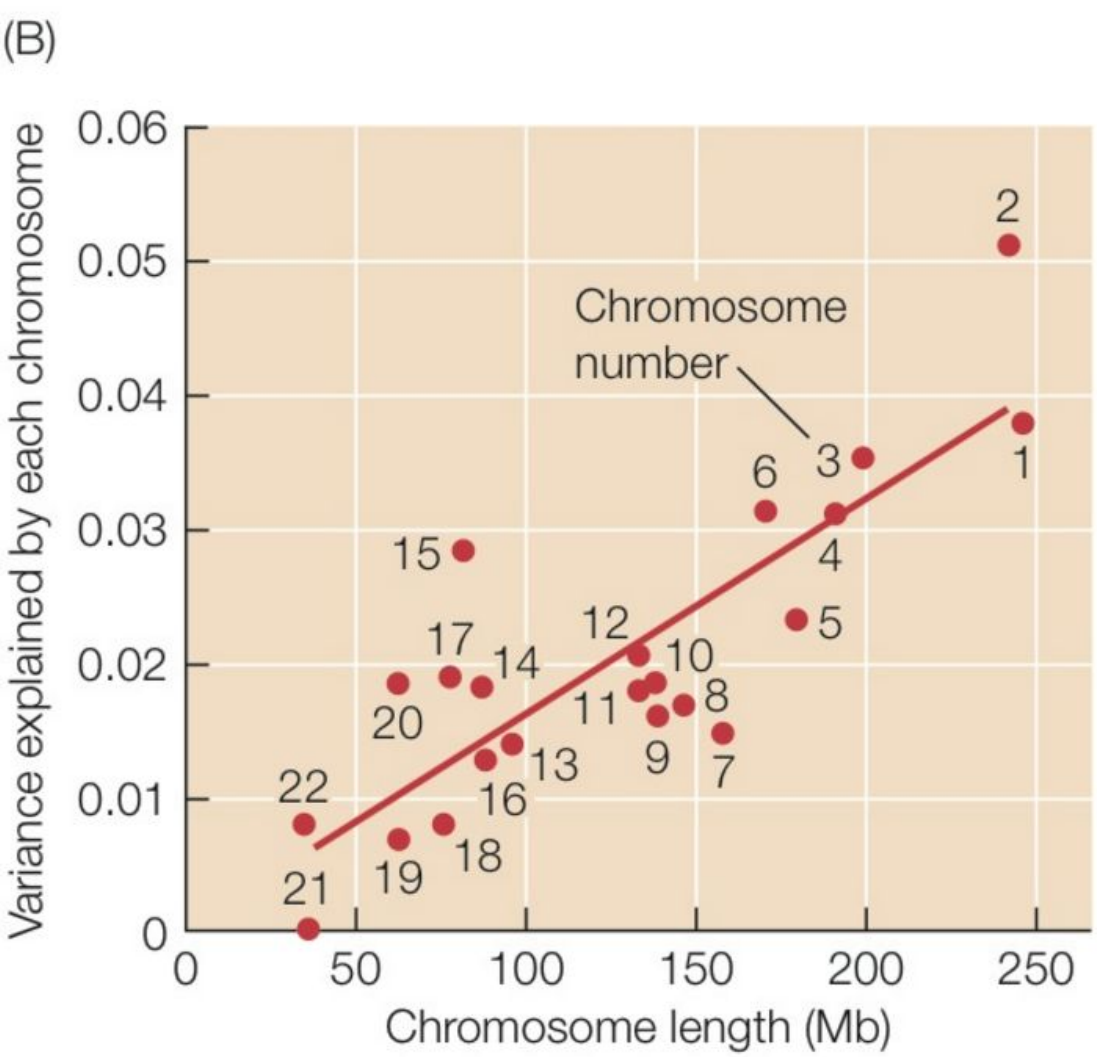
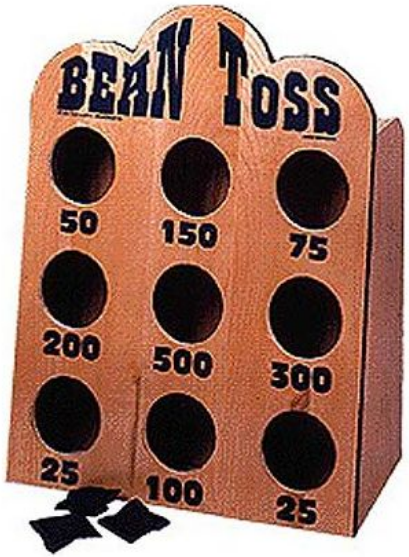


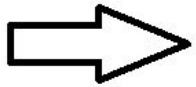
Figure 6.9 Multivariate Association. (A) Extrapolation of observed allele effect sizes led Ripke et al. (2013) to estimate the total number of common variants that explain the observed heritability for five diseases: rheumatoid arthritis (RA), celiac disease, type 2 diabetes (T2D), myocardial infarction (MI), and schizophrenia. The core colored area represents the 50% confidence interval of each estimate, and the outer circles show the 95% confidence intervals. For example, over 8300 SNPs are predicted to contribute to half the liability for schizophrenia. (B) Genomic partitioning allows estimation of the relative contributions of subsets of loci—in this case, of each chromosome—to the total variance for height. There is a highly significant linear regression of variance explained on chromosome length. (B, after Yang et al. 2011.)

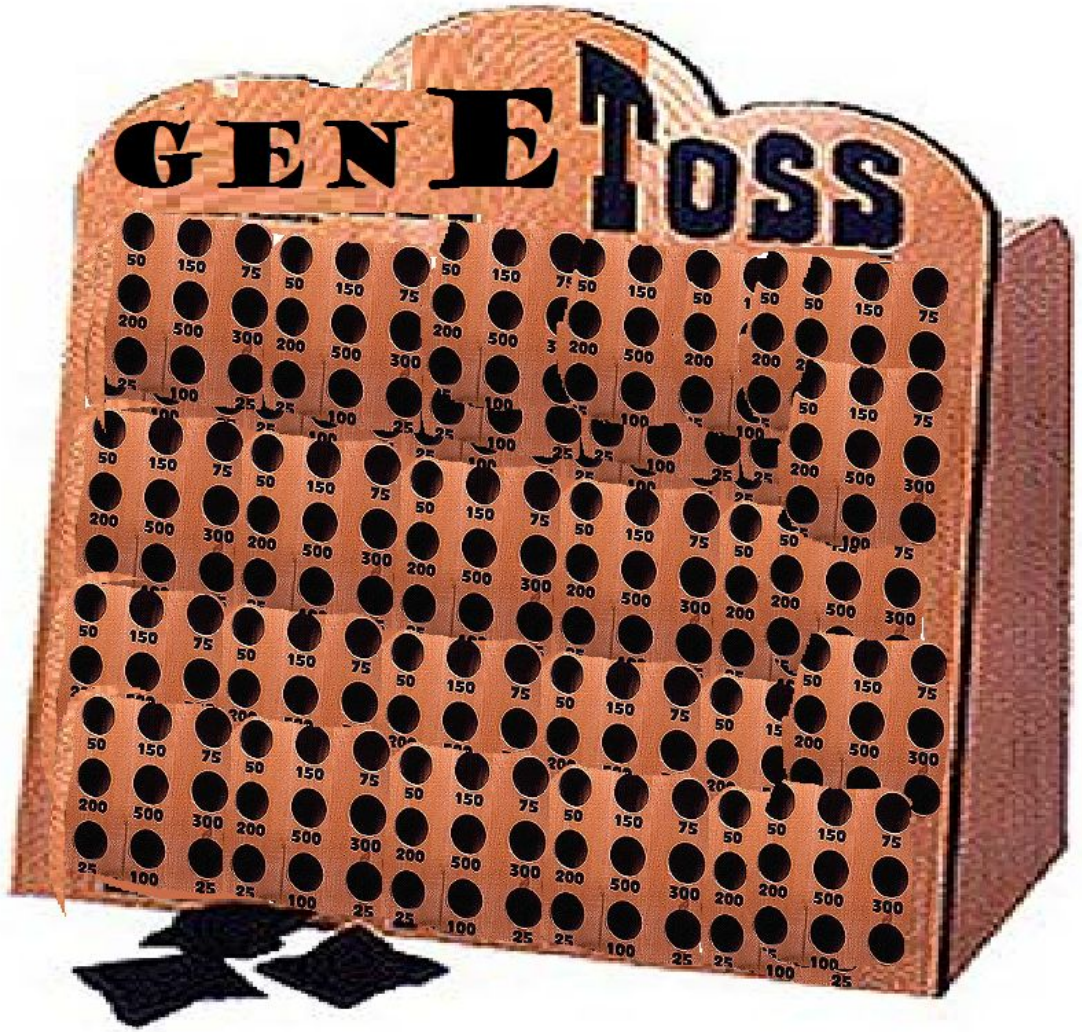

(A)
Schizophrenia
Type 2 diabetes
Myocardial infarction
Celiac disease
Rheumatoid arthritis
Schizophrenia 50% confidence interval
Schizophrenia 95% confidence interval
MI
T2D
Celiac
RA
Total variance explained
0
0.2
0.4
0.6
0.8
100
1000
10,000
100,000
Number of disease-associated SNPs


(A)
Schizophrenia
Type 2 diabetes
Myocardial infarction
Celiac disease
Rheumatoid arthritis
Phenylketonuria
PKU
Schizophrenia 50% confidence interval
Schizophrenia 95% confidence interval
MI
T2D
Celiac
RA
Total variance explained
1.0
0.8
0.6
0.4
0.2
0
1
10
100
1000
10,000
100,000
Number of disease-associated SNPs


(B)
Variance explained by each chromosome
0.06
0.05
0.04
0.03
0.02
0.01
0
0
50
100
150
200
250
Chromosome length (Mb)
Chromosome number
2
1
3
6
4
15
5
12
17
14
10
11
8
20
9
7
13
22
16
18
19
21


BEAN TOSS
50
150
75
200
500
300
25
100
25




GENE TOSS

Predictors

Predictors

- Explain **fraction** of variance
  - False positives
  - Effect size estimates imprecise
  - Incomplete LD between causal variants and tagging SNPs

- The goal is not **prediction**, but **classification**

- Allelic sum scores estimate risk by weighing effect size of each allele

Genotyping Technologies

Genotyping Technologies

Illumina

- Arrays built on Infinium II Assay
- Produces between 730 000 to 4.3 million markers
- Single-base extension reaction

Affymetrix

- Human SNP and Axiom Arrays
- Gene chips

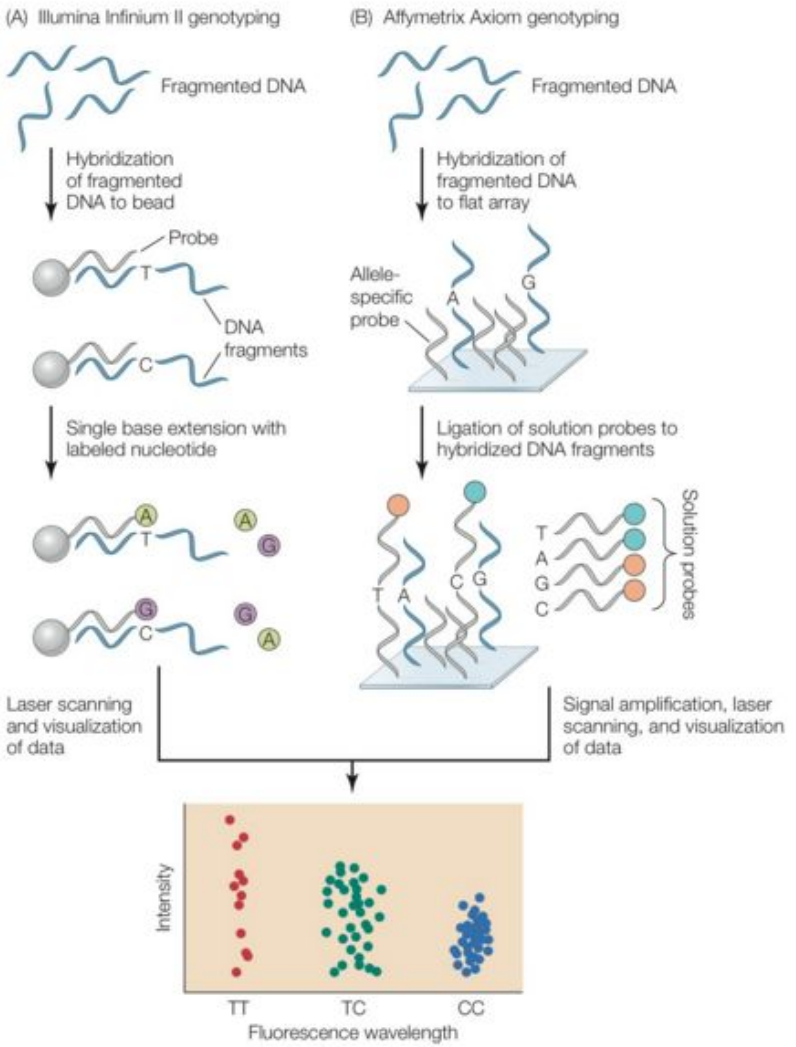


Figure 6.10 Genotyping Technologies. (A) Illumina's Infinium II assay is a single-base extension protocol. Genomic DNA is fragmented and hybridized to beads coated in a probe oligonucleotide that terminates at the base adjacent to the SNP. The assay is designed so that one of two colored nucleotides is incorporated at each polymorphism. The ratios of intensities across all the samples in a batch are deconvoluted to generate clouds of points corresponding to each genotype (in the graph shown here, minor homozygotes are red, heterozygotes green, and major homozygotes blue). (B) The Affymetrix Axiom assay also relies on capture of fragmented genomic DNA, but visualization of the polymorphism follows from amplification of the signal associated with a short oligonucleotide designed to ligate to the template according to the identity of the SNP. This platform, too, calls genotypes called from the clouds of points from multiple samples.

Reference

Gibson, G. (2015). A Primer of Human Genetics. Sunderland, MA: Sinauer Associates, Inc.

National human genome research institute. Retrieved from: https://www.genome.gov

Wellcome Trust Case Control Consortium. Retrieved from: http://www.wtccc.org.uk/info/070606.html

The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research, 2014, Vol. 42. Retrieved from: http://www.ebi.ac.uk/gwas/