



Biostatistics

Collaboration Center

Basic Biostatistics in Medical Research

Lecture 6: Statistics in Genetics Research

Kwang-Youn Kim, PhD
Biostatistics Collaboration Center (BCC)
Department of Preventive Medicine

Outline

- Terminology and Concepts
- Linkage Analysis
- Genome-wide Association studies
- Quantitative Trait Analysis
- Next Generation Sequencing (e.g. RNA-seq)
- Conclusions

Statistics in Genetics?

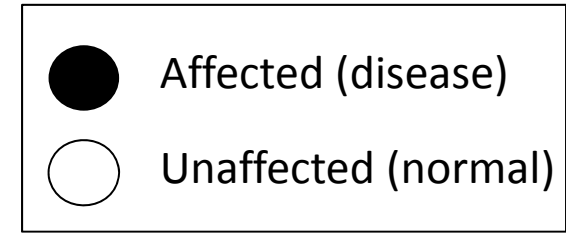
- Mendel
- R.A. Fisher

Linkage Analysis

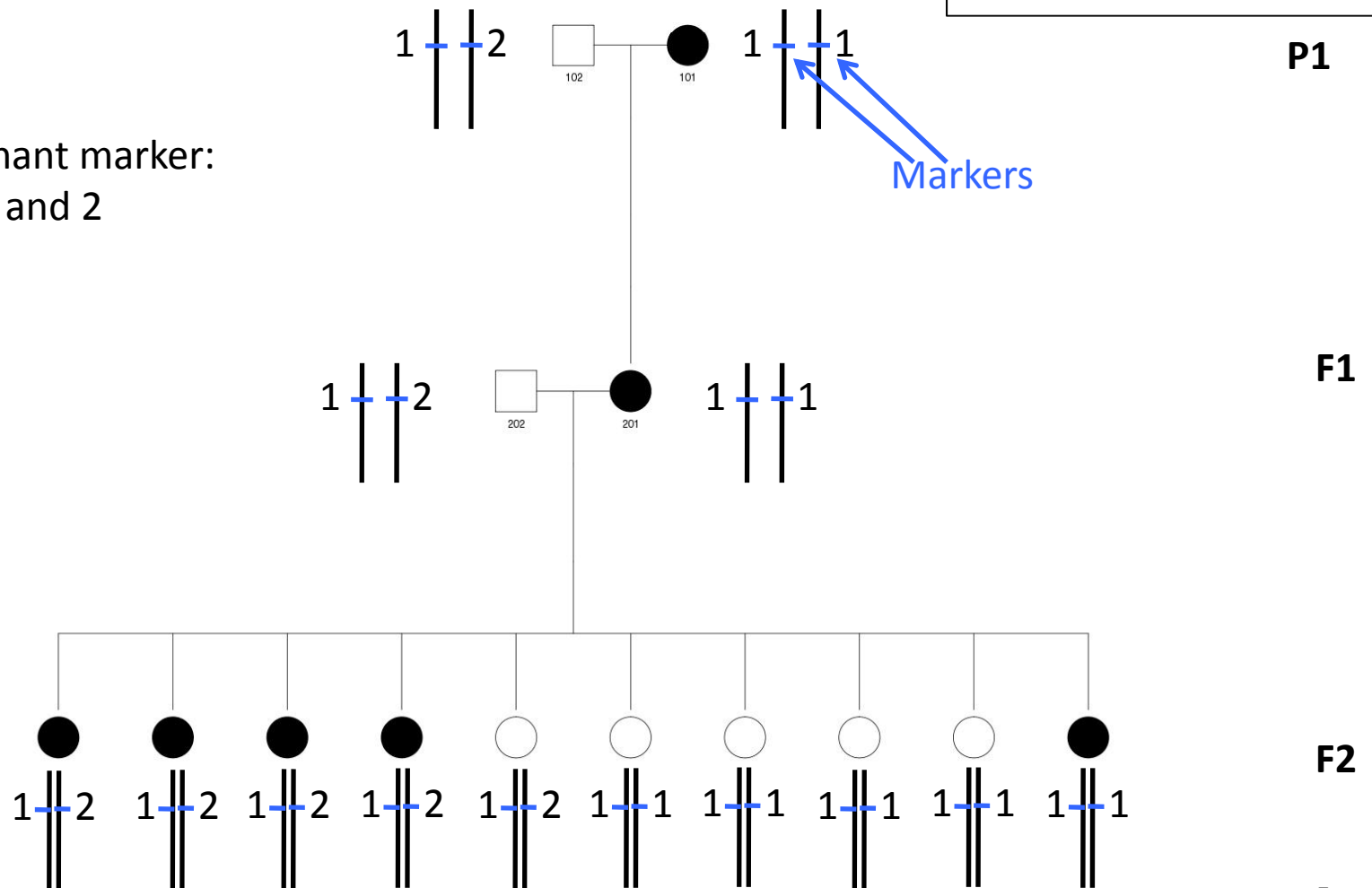
- Linkage: joint inheritance of genetic loci due to physical proximity (violation of Mendel II: law of independent assortment)
- Linkage analysis: identifying (disease) locus by tracking deviations from Mendel II
- Requires writing down likelihoods, estimating and testing for recombination fraction
- Requires family data with markers and phenotypes

Linkage Analysis

Phenotypes



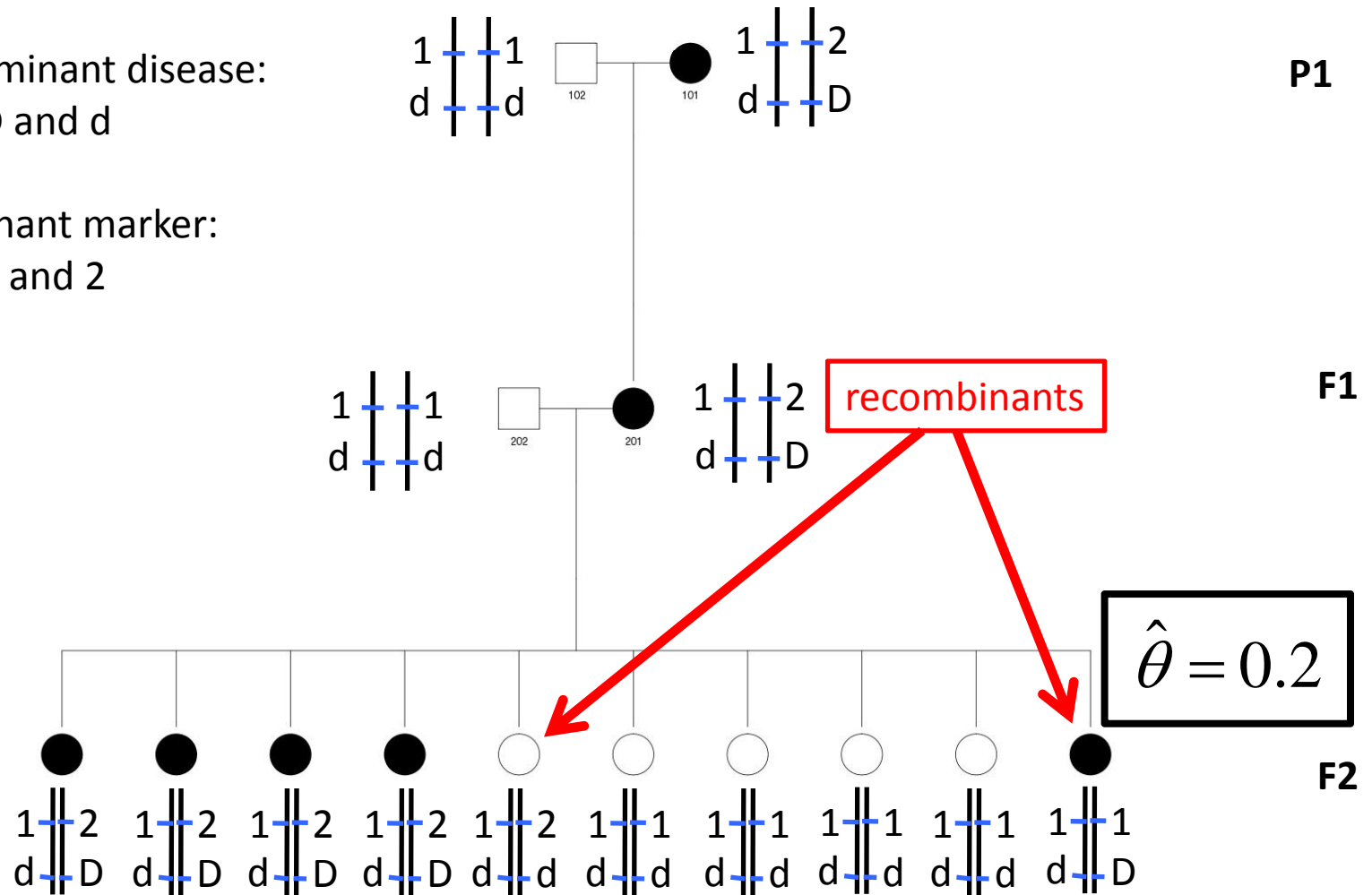
Codominant marker:
alleles 1 and 2



Linkage Analysis

Rare dominant disease:
alleles D and d

Codominant marker:
alleles 1 and 2



Linkage Analysis

A polymorphic DNA marker genetically linked to Huntington's disease

**James F. Gusella^{*}, Nancy S. Wexler^{†||}, P. Michael Conneally[†], Susan L. Naylor[§],
Mary Anne Anderson^{*}, Rudolph E. Tanzi^{*}, Paul C. Watkins^{*||}, Kathleen Ottina^{*},
Margaret R. Wallace[‡], Alan Y. Sakaguchi[§], Anne B. Young^{||}, Ira Shoulson^{||},
Ernesto Bonilla^{||} & Joseph B. Martin^{*}**

^{*} Neurology Department and Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA

[†] Hereditary Disease Foundation, 9701 Wilshire Blvd, Beverley Hills, California 90212, USA

[‡] Department of Medical Genetics, Indiana University Medical Center, Indianapolis, Indiana 46223, USA

[§] Department of Human Genetics, Roswell Park Memorial Institute, Buffalo, New York 14263, USA

^{||} Venezuela Collaborative Huntington's Disease Project^{*}

Family studies show that the Huntington's disease gene is linked to a polymorphic DNA marker that maps to human chromosome 4. The chromosomal localization of the Huntington's disease gene is the first step in using recombinant DNA technology to identify the primary genetic defect in this disorder.

Linkage Analysis

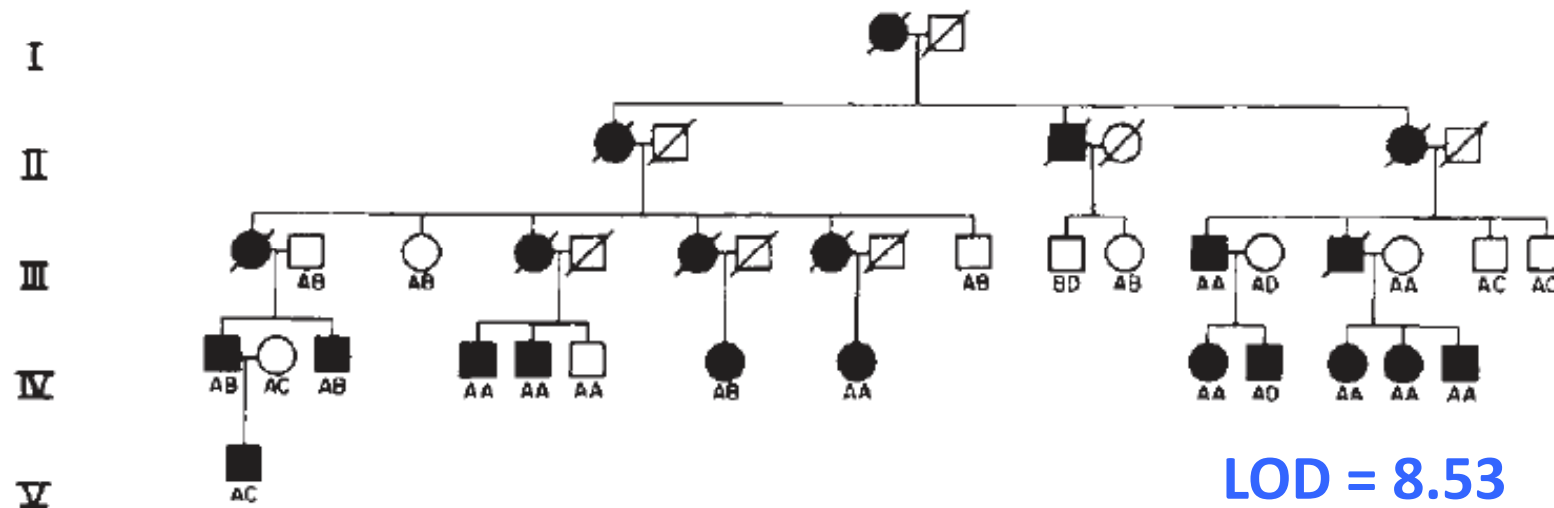


Fig. 1 Pedigree of an American Huntington's disease family. Symbols: circles, females; squares, males; a black symbol indicates that an individual is affected with Huntington's disease; a slashed symbol indicates that an individual is deceased. This pedigree was identified through the National Research Roster for Huntington's Disease Patients and Families at Indiana University. Relevant

Gusella et al. (1983)

Linkage Analysis

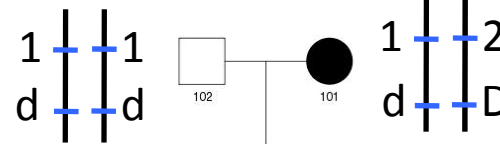
- LOD score: test if $\theta = 1/2$
- More formally,
$$\text{LOD} = \log_{10}[L(\hat{\theta})/L(\theta = 1/2)]$$
- $L(\theta)$: likelihood of the pedigree data
- LOD score to likelihood ratio statistic (LRS) conversion
$$\text{LRS} = 4.6 \times \text{LOD}$$

Linkage Analysis – Dominant Model

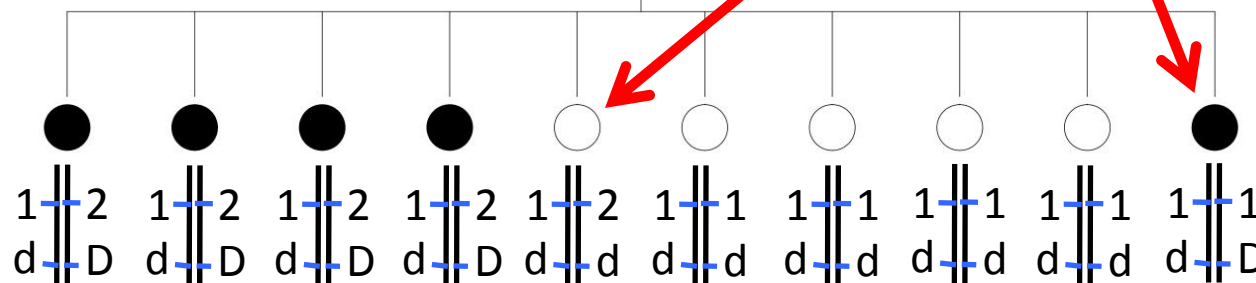
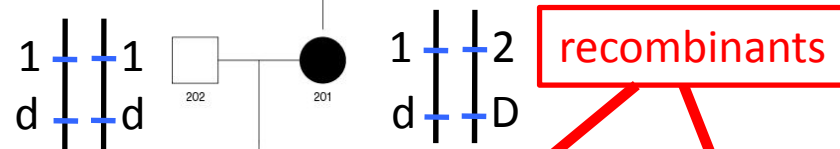
Rare dominant disease:
alleles D and d

Codominant marker:
alleles 1 and 2

Grandparents provide phase info



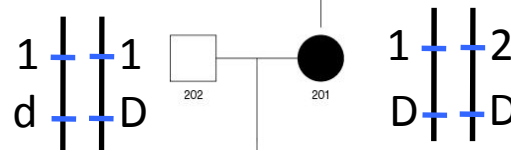
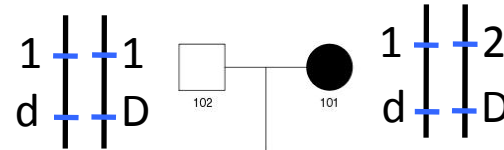
Dominant Model: LOD = 0.84
Recessive Model: non-informative



Linkage Analysis – Recessive Model

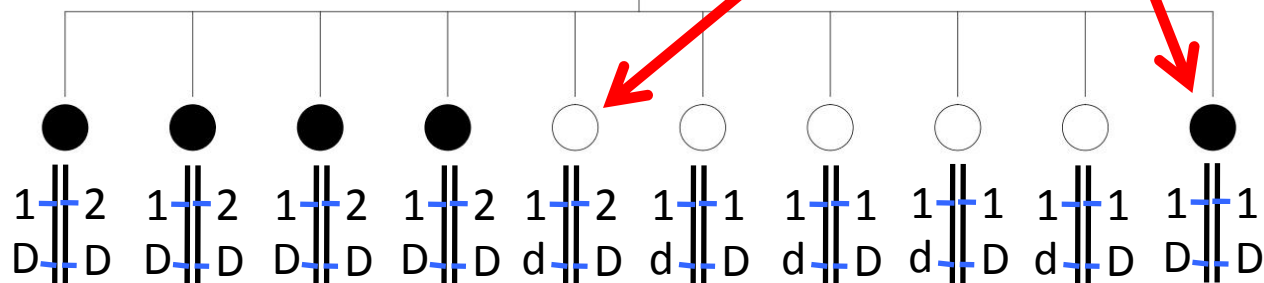
Rare recessive disease:
alleles D and d

Codominant marker:
alleles 1 and 2



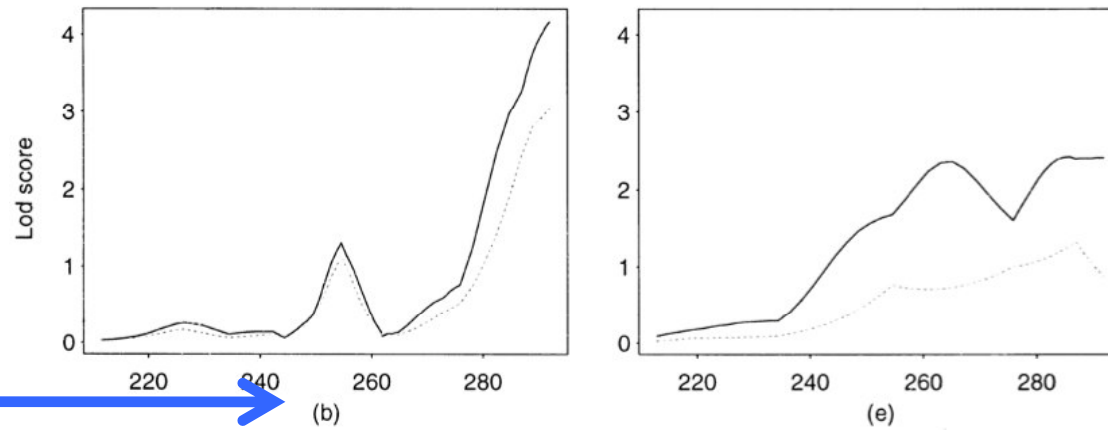
recombinants

Dominant Model: LOD = 0.84
Recessive Model: non-informative



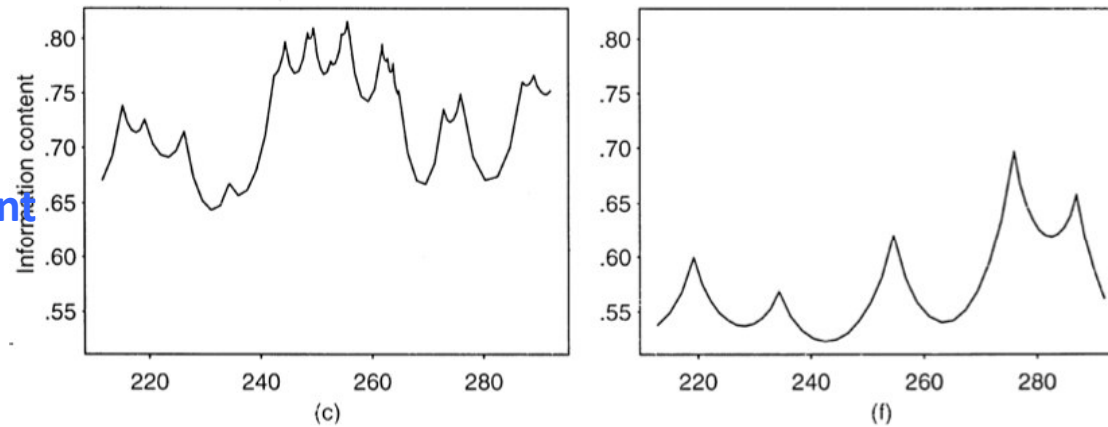
Linkage Analysis of NIDDM

Lod Score



Chromosomal position

Information Content



Source: Kong and Cox (1997)

Linkage Analysis

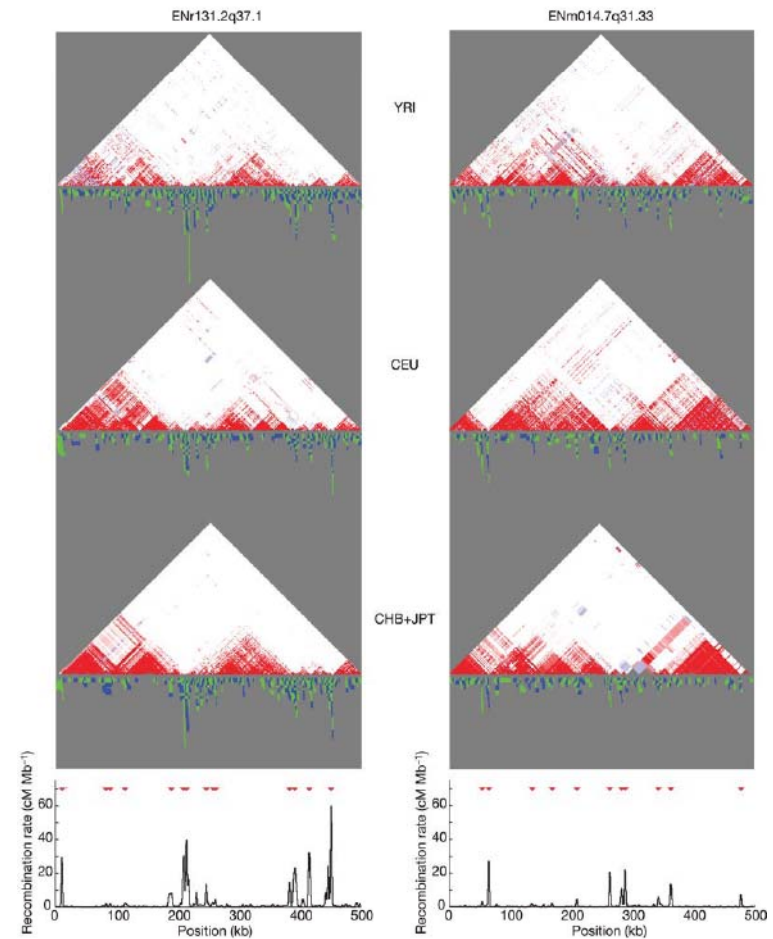
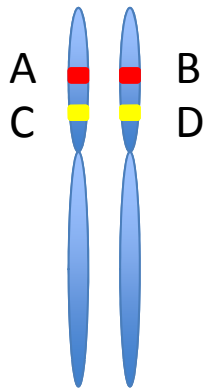
- LOD score of 3 is considered a strong signal (Newton Morton)
- Challenges
 - Incomplete penetrance
 - Heterogeneity
 - Missing genotype
 - What is the resolution of linkage analysis?

Association Studies

- Relating genetic variation to phenotypic variation at the population level
- Rely on linkage disequilibrium
- Linkage disequilibrium is not linkage!
- Linkage disequilibrium is not linkage!

Linkage Disequilibrium

- Linkage Disequilibrium (LD): non-random association of alleles at two or more loci
- Measures of LD: D , D' , r^2



Linkage Disequilibrium – Measures of Strength

- r^2 (Hill and Robertson 1968)

$$r^2(p_A, p_B, p_{AB}) = \frac{(p_{AB} - p_A p_B)^2}{p_A(1 - p_A)p_B(1 - p_B)}$$

– $r^2=1$ implies perfect LD

– $r^2=0$ implies no LD

- D' (Lewontin 1962)

$$D' = \frac{D_{AB}}{\min(P(A)[1 - P(B)], [1 - P(A)]P(B))}, \text{ if } D_{AB} > 0$$

$$D' = \frac{D_{AB}}{\max(P(A)P(B), [1 - P(A)][1 - P(B)])}, \text{ if } D_{AB} < 0$$

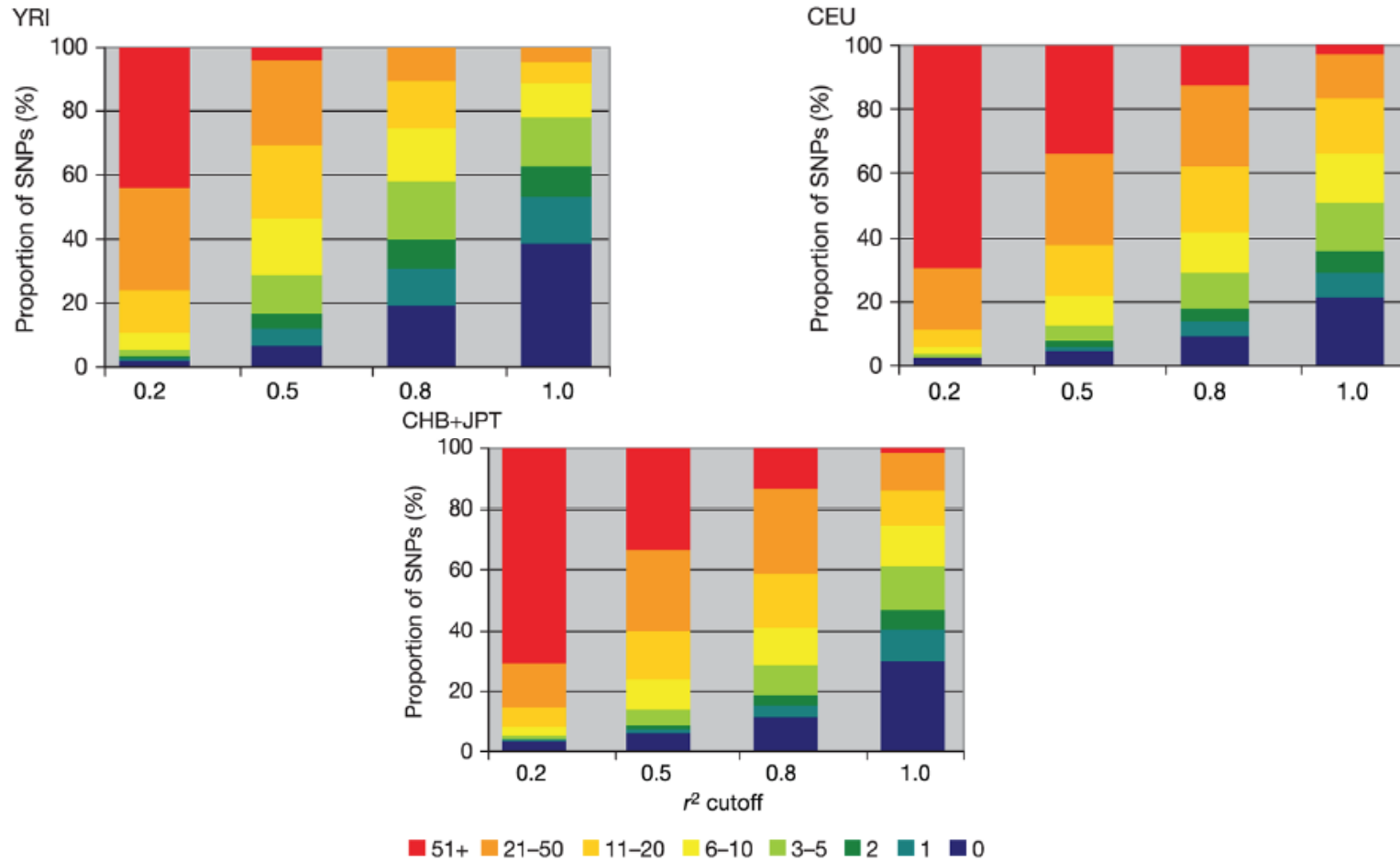
– $D'=1$ implies perfect LD

– $D'=0$ implies no LD

SNPs and tagSNPs

- Single Nucleotide Polymorphism (SNP)
- AGC**G**CCTCCC
AGC**C**CCTCCC
- Unlikely that the marker SNP is a causal variant
- Rely on SNPs that are in LD with the causal SNPs, namely tagSNPs

Coverage of tag SNPs



Source: Altshuler et al. 2005

Genome-wide Association Studies (GWAS)

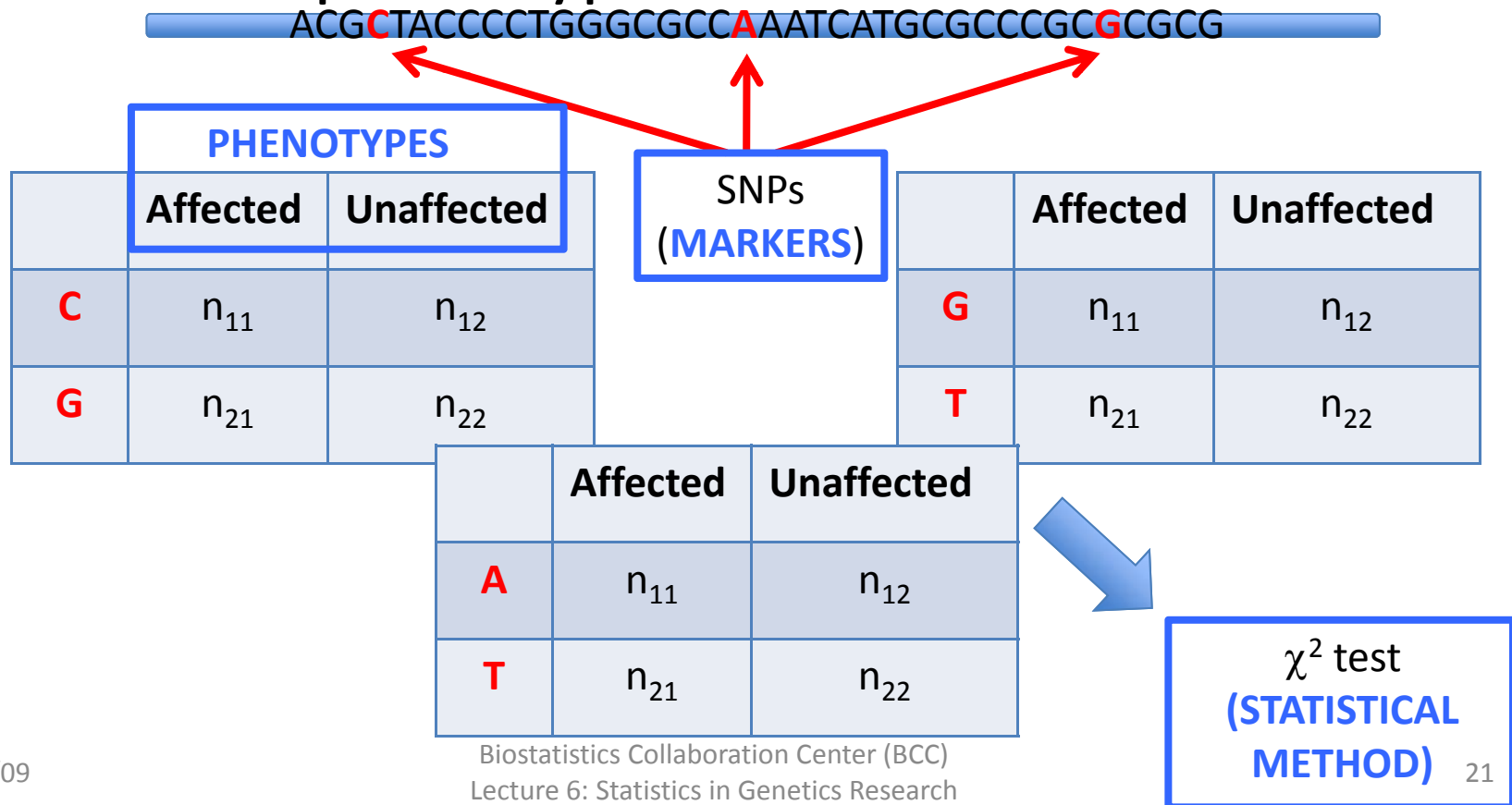
- Extension of association study to genomewide
- Study of **genetic variation** across the entire human genome that is designed to identify genetic associations with **observable traits** (such as blood pressure or weight), or the **presence or absence of a disease or condition** (NIH)
- Predominantly case-control study design
- Prospective studies also emerging

GWAS Ingredients

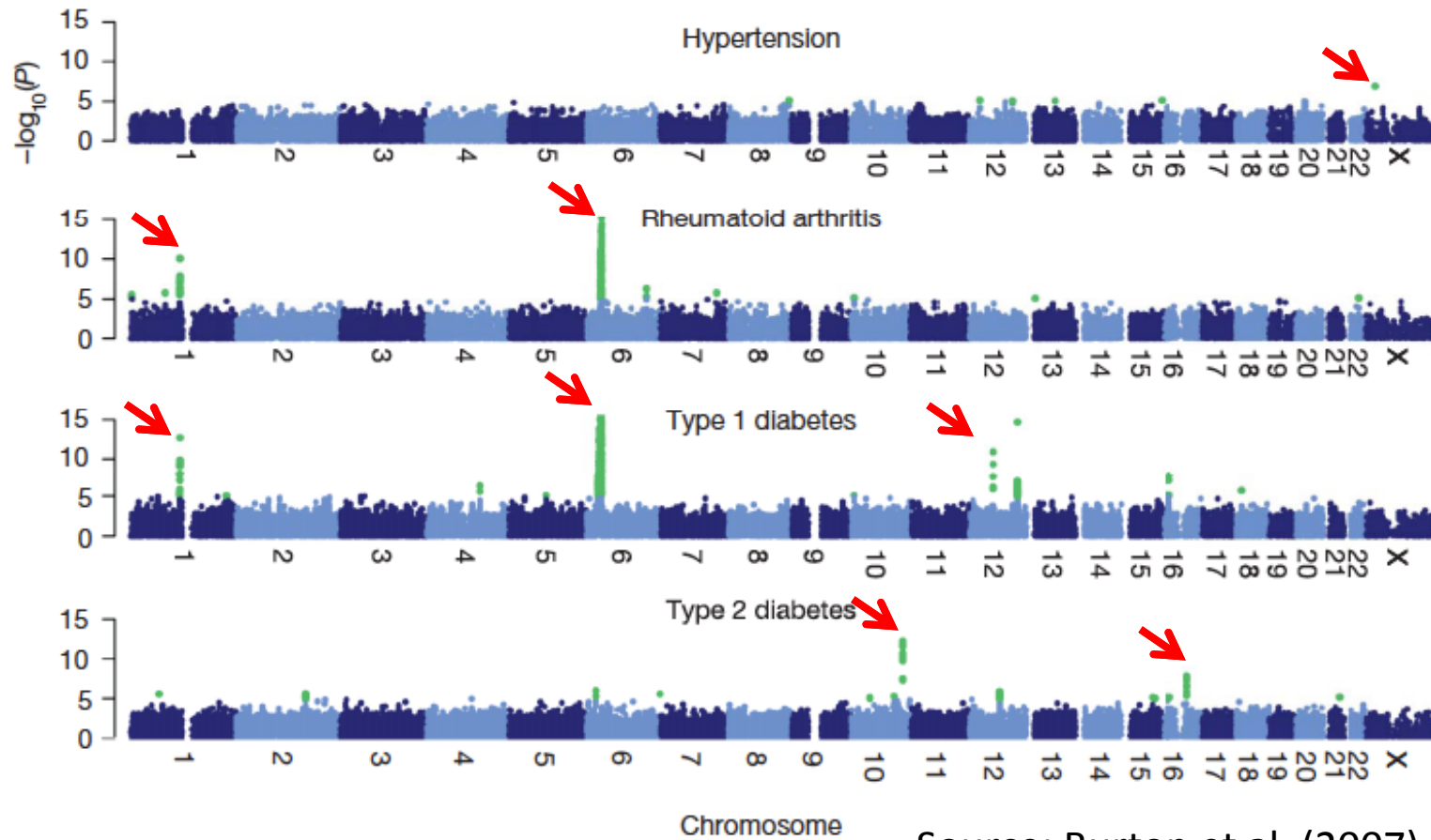
- **Markers**, e.g. SNPs
- **Phenotypes**, e.g. disease status
- **Statistical methods**, e.g. chi-square test

GWAS Ingredients

- At each SNP location, test for association b/w SNPs and phenotype

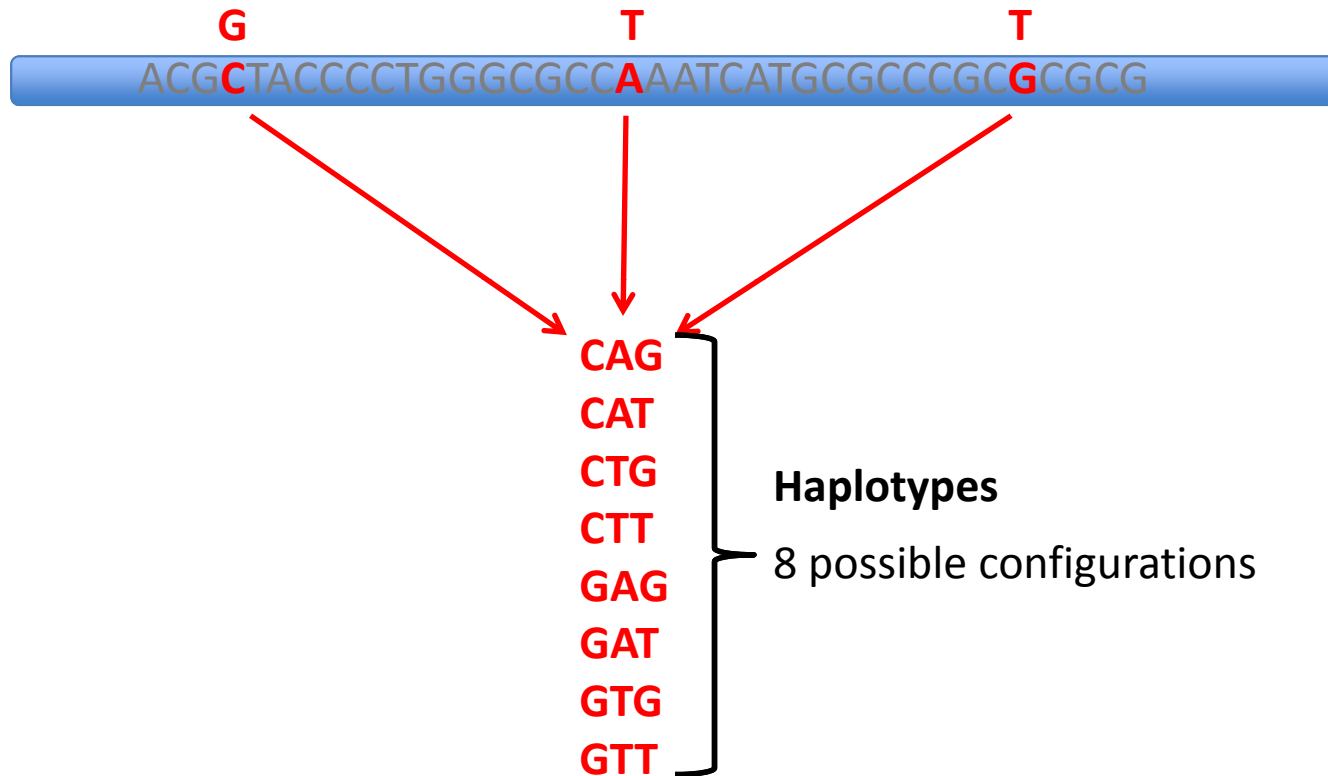


GWAS: Manhattan Plot



Source: Burton et al. (2007)

GWAS: Haplotype Analysis



GWAS: Haplotype Analysis

Person 1 ACG**C**TACCCCTGGGGCGCC**A**AATCATGCGCCCGC**G**CGCG

Person 2 ACG**G**TACCCCTGGGGCGCC**A**AATCATGCGCCCGC**G**CGCG

Person 3 ACG**C**TACCCCTGGGGCGCC**T**AATCATGCGCCCGC**G**CGCG

Person 4 ACG**G**TACCCCTGGGGCGCC**A**AATCATGCGCCCGC**G**CGCG

Person 5 ACG**C**TACCCCTGGGGCGCC**A**AATCATGCGCCCGC**G**CGCG

Person 6 ACG**G**TACCCCTGGGGCGCC**A**AATCATGCGCCCGC**G**CGCG

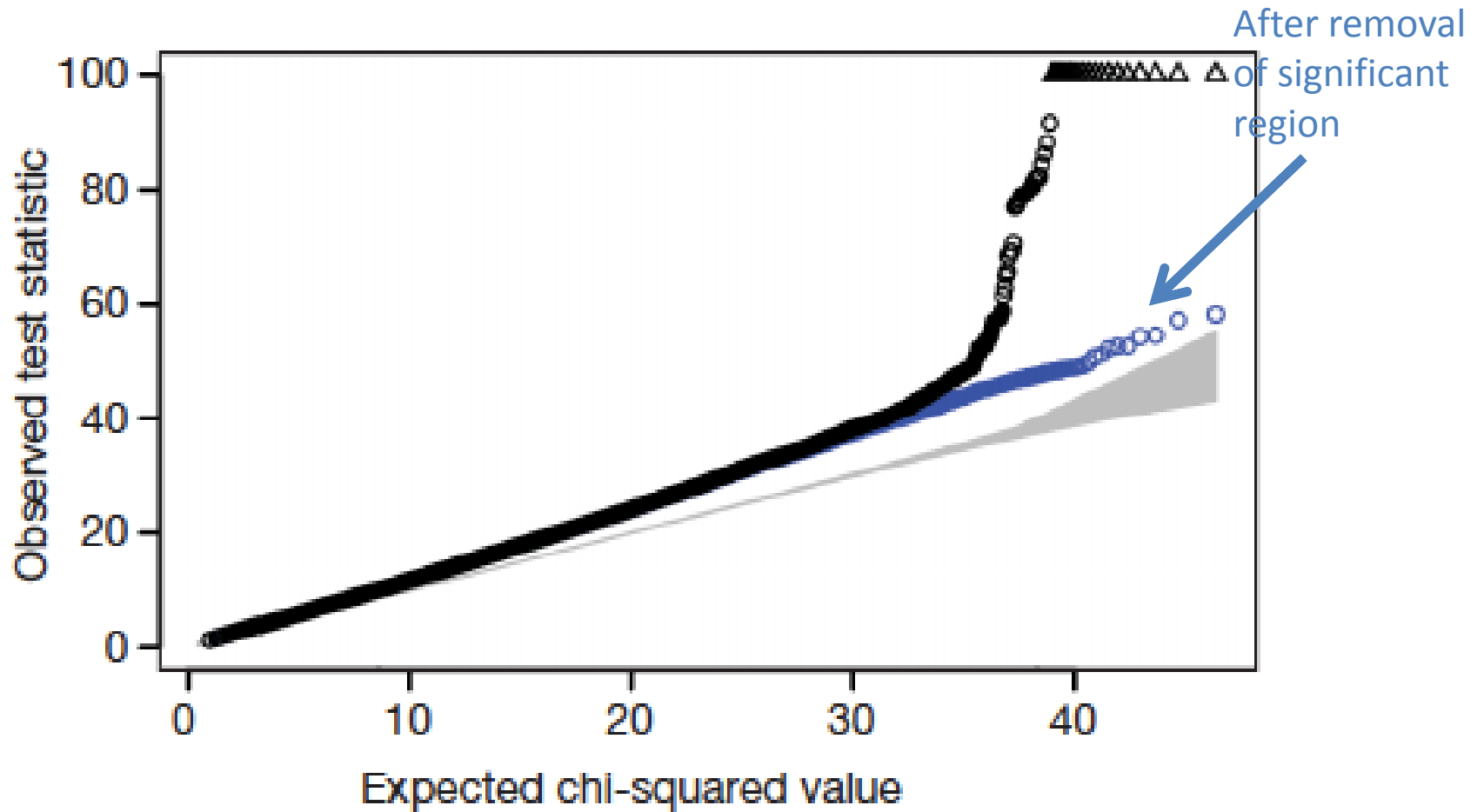
Observed Haplotypes

CAG

GAG

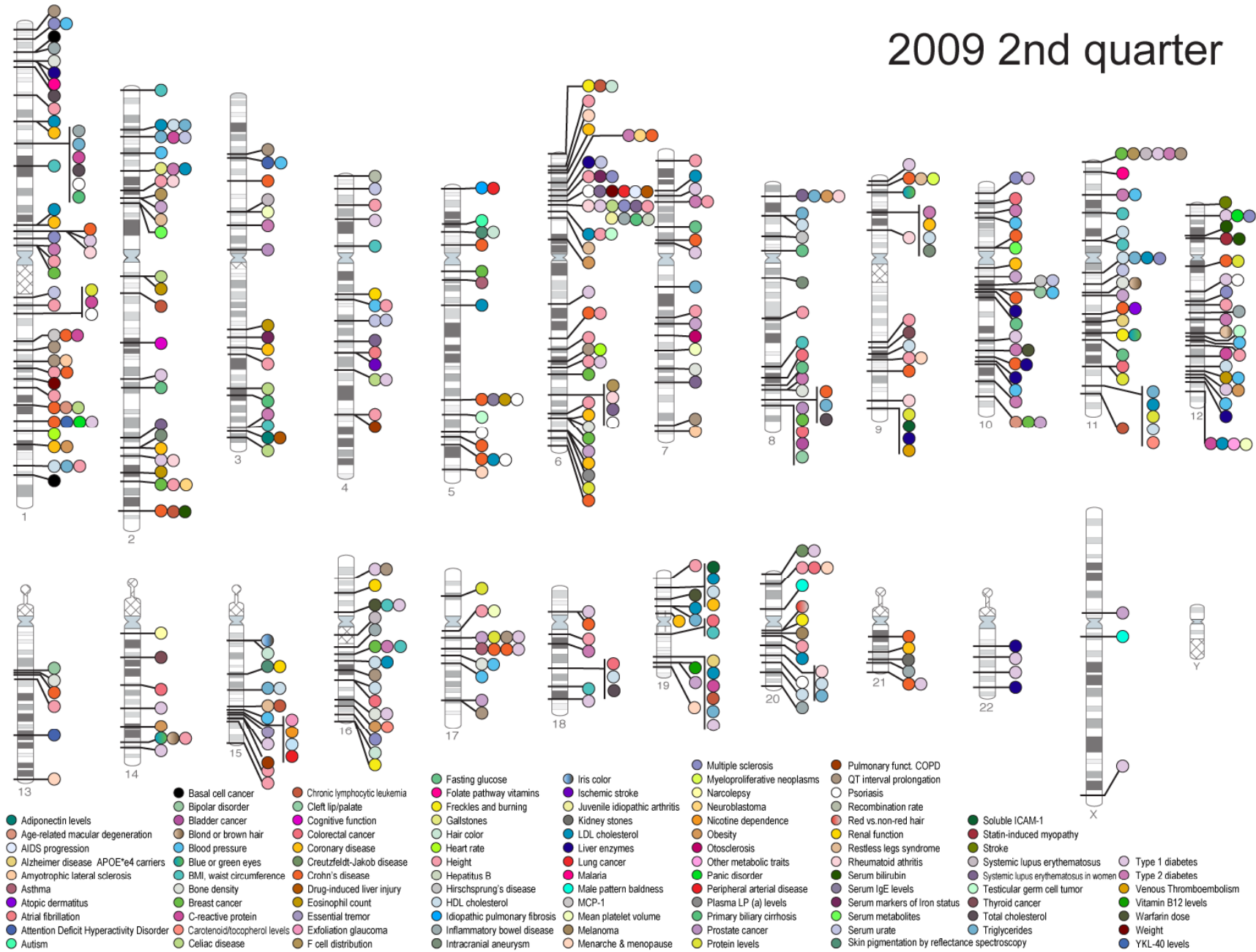
CTG

GWAS: Q-Q Plot



Source: WTCCC (2007)

2009 2nd quarter



11/5/09

Source: NHGRI catalog of published GWAS

26

Multiple Comparisons

- GWAS involves tests at multiple SNPs
- E.g. 500,000 SNPs = 500,000 tests
- At $\alpha = 0.05$, there are 25,000 false positives!
- Need a clever way of controlling for false positives

Controlling for False Positives

- Simple way to control for family-wise false positive rate is using Bonferroni method
- Revise the stringency criteria by the number tests performing
- With 500,000 SNPs
$$\alpha = 0.05 \Rightarrow \alpha = 0.05/500,000$$
$$= 10^{-7}$$
- Technicality: Bonferroni correction does not change the p -value


False Discovery Rate (FDR)

- Alternative method to control for multiple tests (Benjamini and Hochberg 1995)
- Control for $\{\# \text{ falsely rejected}\} / \{\# \text{ total rejected}\}$
- More useful for exploratory analysis
- q -values (Storey, 2001)

Population Stratification

- Systematic difference in allele frequencies b/w subpopulations
- Let's say that a SNP in gene X has allele frequencies of

	A	T
CHB	0.7	0.3
CEU	0.5	0.5



	Use chopsticks	Do not use chopsticks
CHB	350	250
CEU	150	250

$\chi^2 = 40.84$; p-value $< 10^{-9}$

Population Stratification

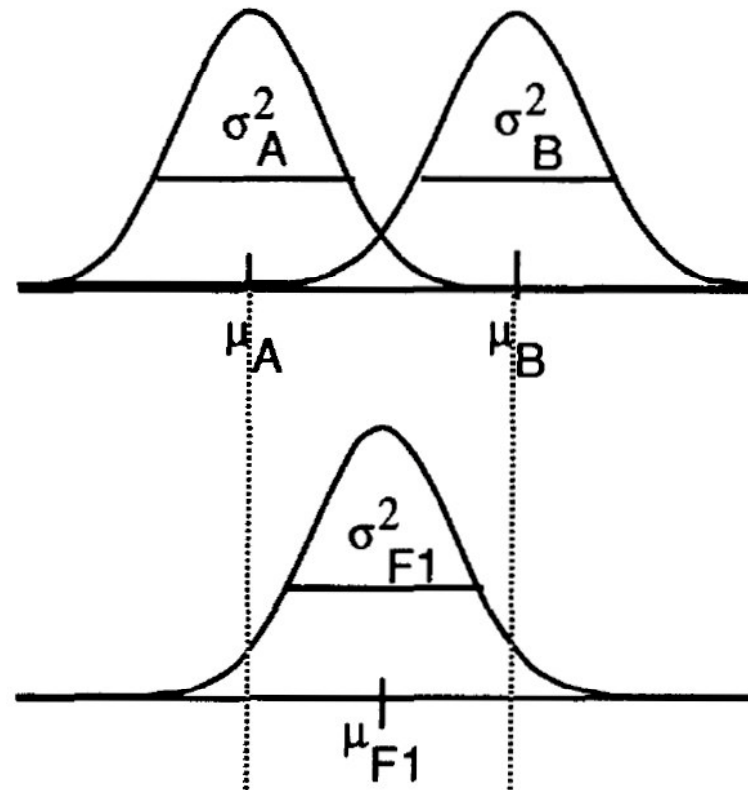
- Methods to control for population stratification
 - Family-member controls: e.g. TDT
 - Genomic adjustment

GWAS Shortcomings

- Studies are difficult to replicate
- Account for small variation
- Population stratification
- Multiple comparisons

Quantitative Trait Analysis

- Traits are continuous
- Also requires markers
- Much work has been done on animal crosses

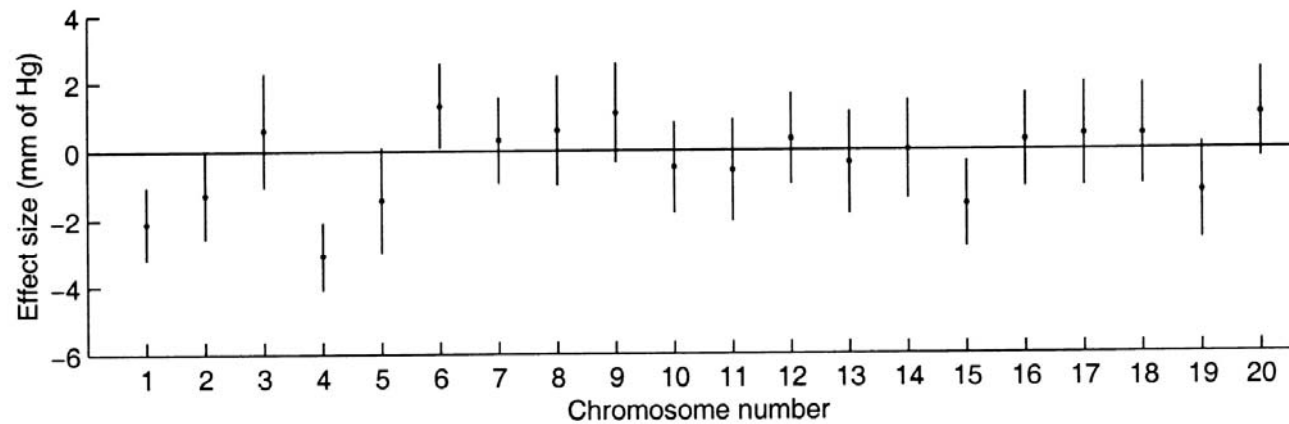
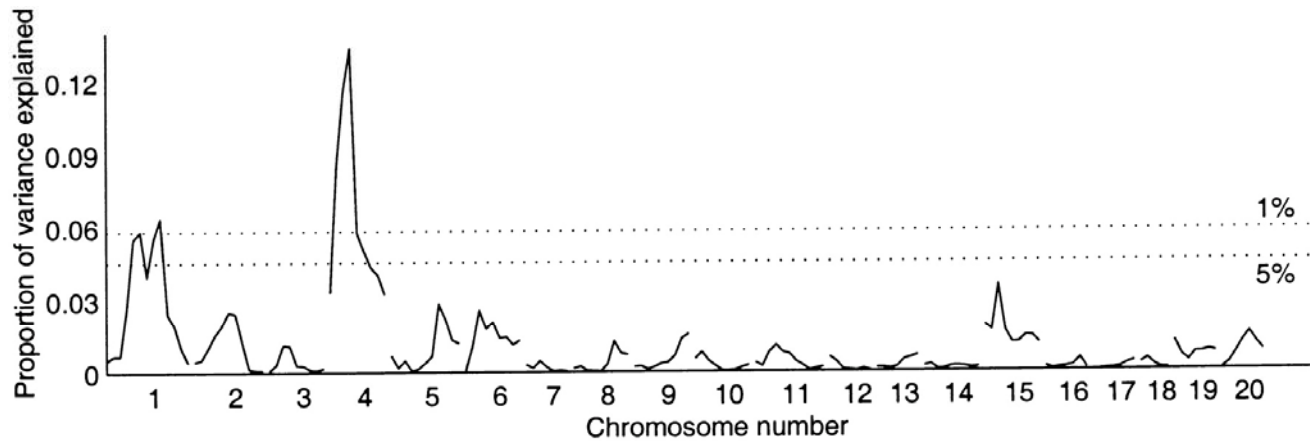


Source: Lander and Botstein (1988)

Quantitative Trait Analysis

- Continuous trait: e.g. height, weight, cornea thickness, cognitive ability, etc.
- Methods
 - ANOVA
 - Interval Mapping
 - Composite Interval Mapping

Map of QTL analysis

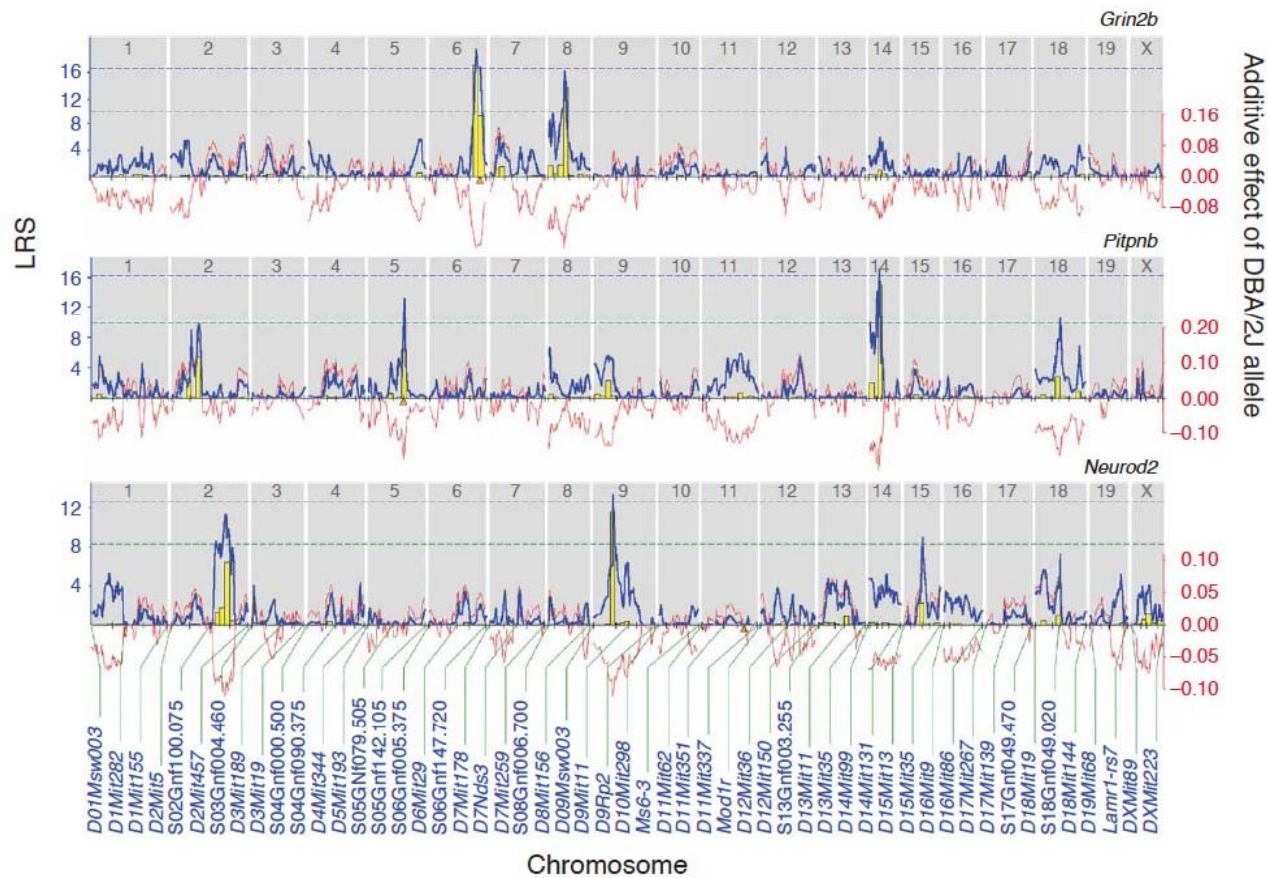


Source: Sen and Churchill (2001)

Expression QTL (eQTL) Analysis

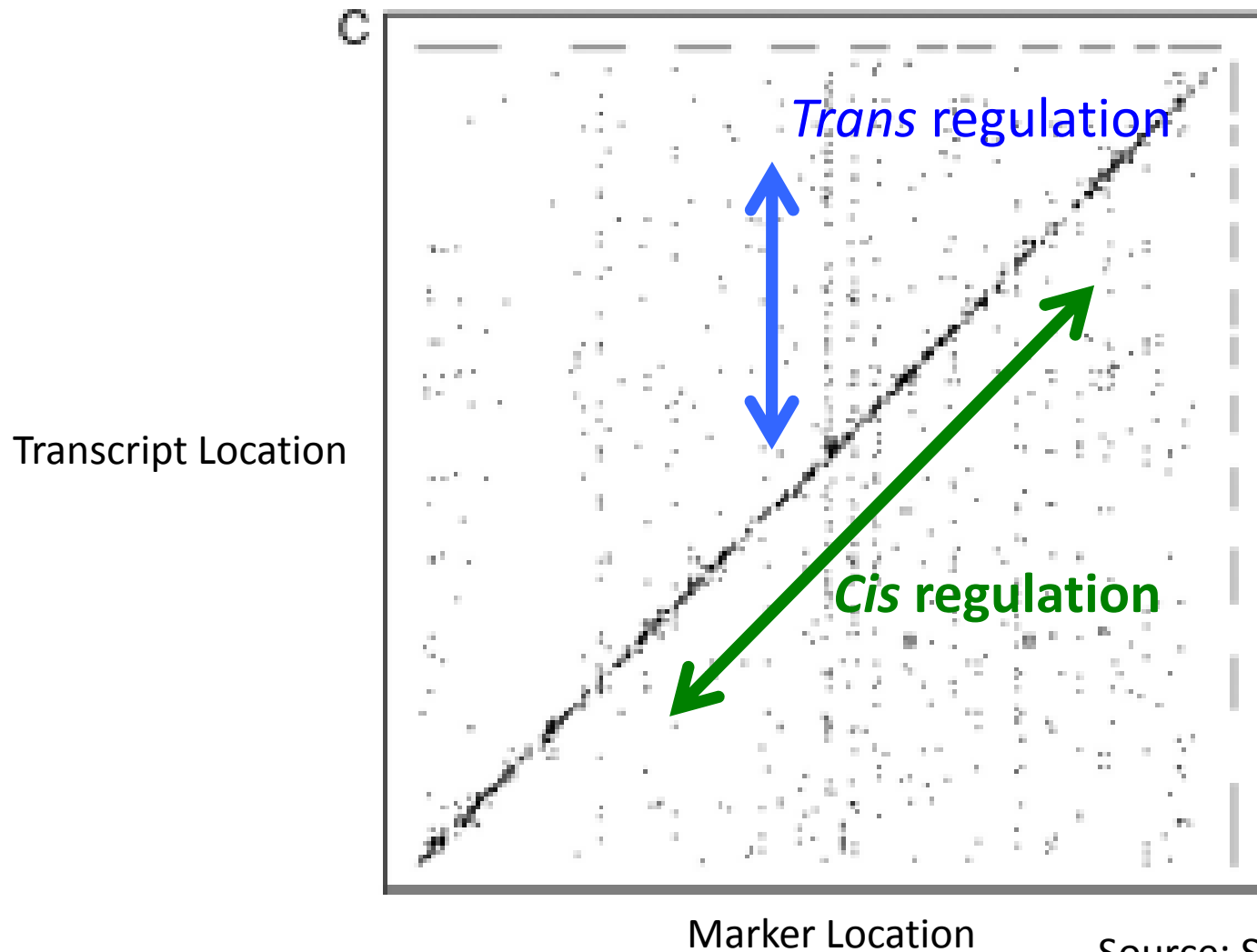
- Specific case of QTL analysis
- Gene expression is the trait of interest
- Microarray data is the driving force
- Identify *cis*- and *trans*- effects
- Acts as further evidence of GWAS findings

Example of eQTL



Source: Chesler et al. (2005)

Heatmap

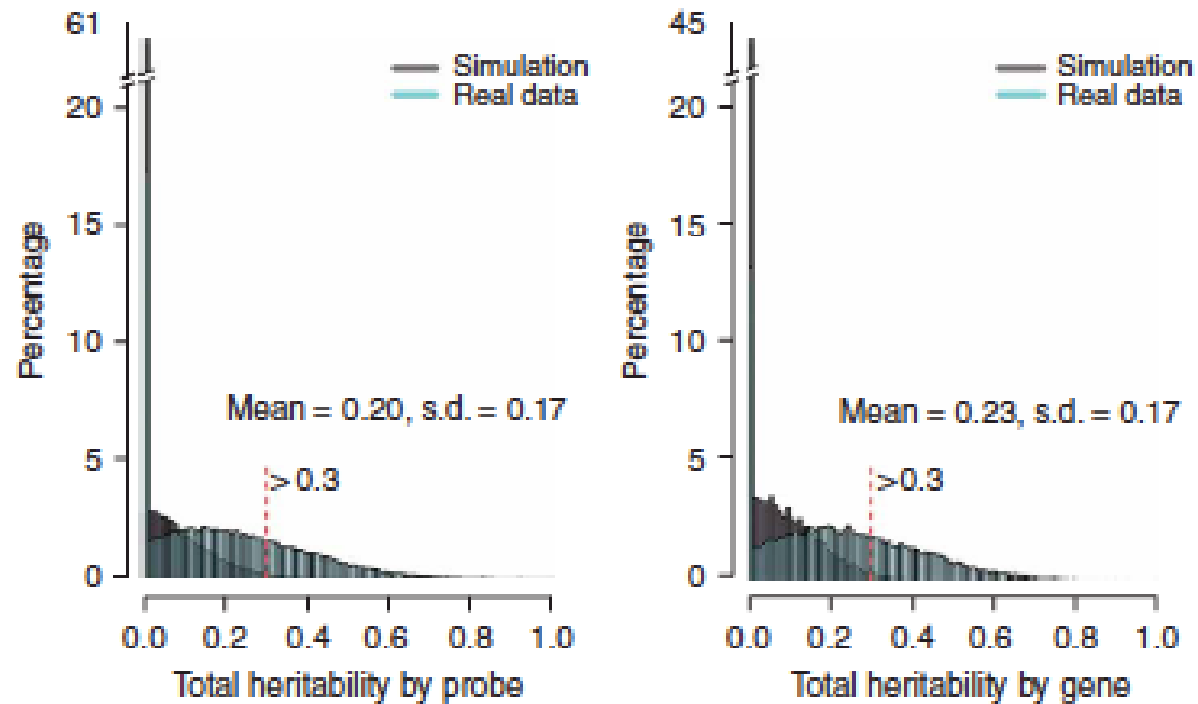


Source: Scheetz et al. (2006)

Heritability

- Phenotypic variability attributable genetic variation
- Broad sense, narrow sense heritability
- Methods of estimation
 - Parent-offspring regression
 - Full-sib comparison
 - Half-sib comparison
 - Twin studies

Heritability of Gene Expression

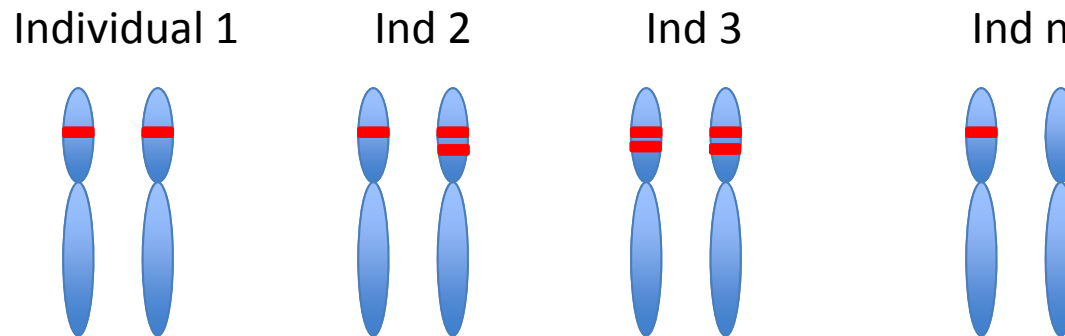


Source: Dixon et al. (2007)

Misconceptions of Heritability

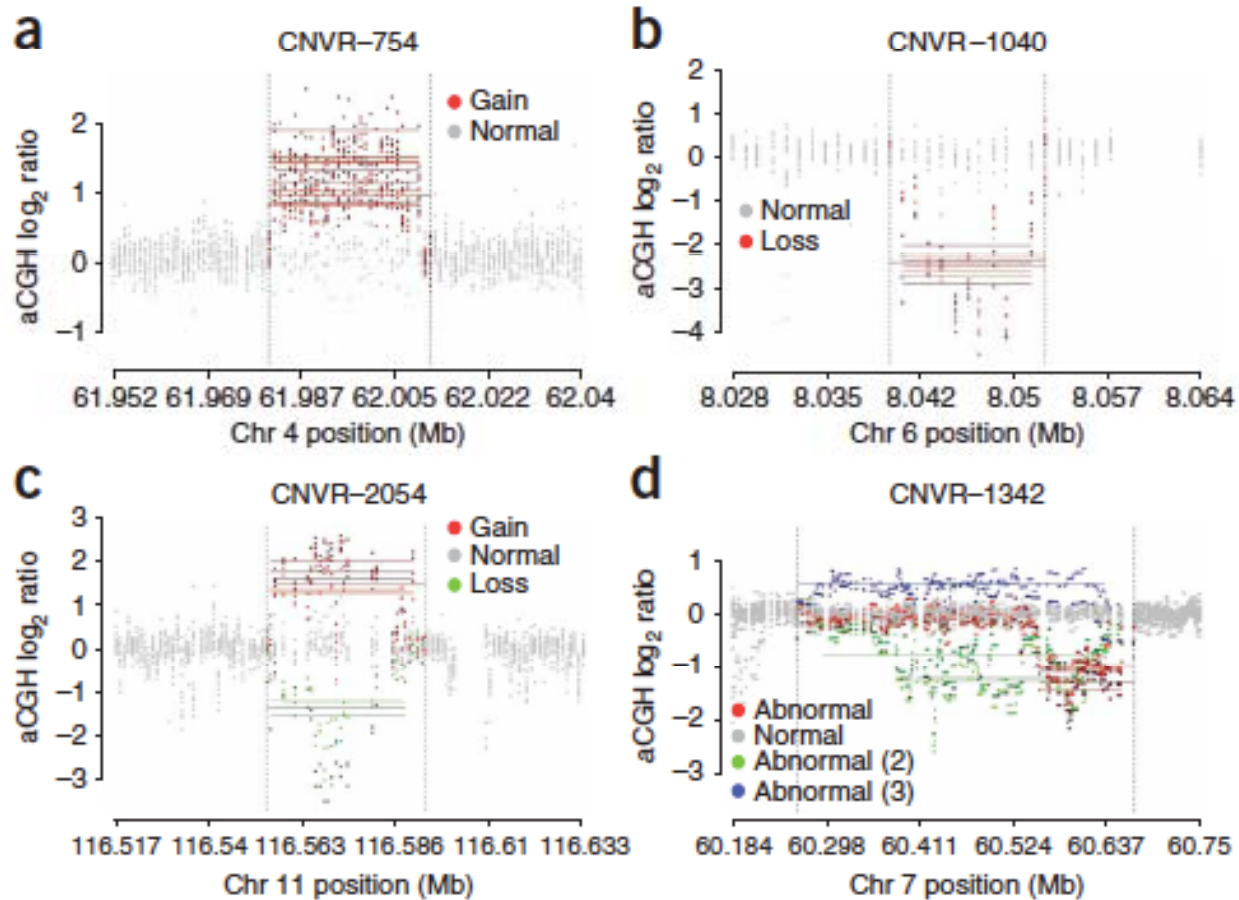
- Heritability is not the proportion of phenotype that is genetic
- Rather, it is the proportion of phenotypic variance that is due to genetic factors
- It is a population parameter that depends on allele frequencies and environmental factors

Copy Number Variation (CNV)



- Structural variant polymorphism
- Also referred to as Copy Number Polymorphism (CNP)
- Useful tool to investigate disease association
- Detected by array CGH or SNP chips

Visual Inspection of CNVs



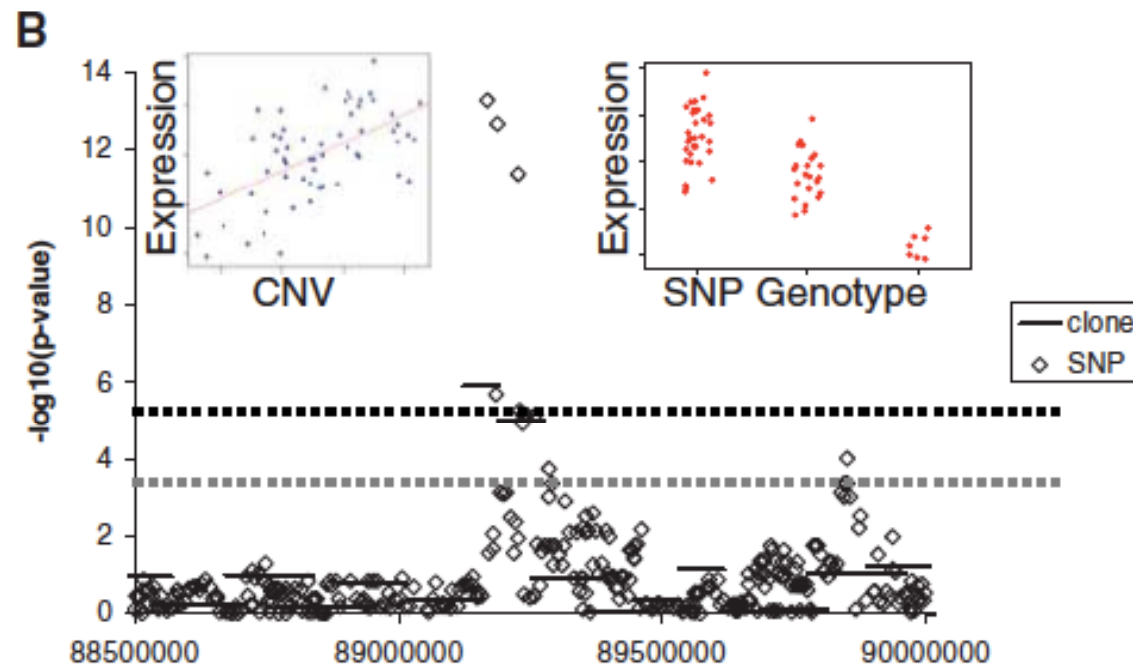
Source: Cahan et al. (2009)

Properties of CNV

- CNVs are similar to SNPs but often multiallelic
- Estimation of the boundaries are tricky
- CNV does not always occur as integers

CNV and Gene expression

- Similar to eQTL analysis, association b/w CNV and gene expression can be investigated



Source: Stranger et al. (2007)

Next Generation Sequencing: RNA-seq

- Next-generation sequencing method for transcriptomics
- “RNA-Seq is expected to revolutionize the manner in which eukaryotic transcriptomes are analysed” (Wang et al. 2009)
- Allows identification of unknown transcripts as well as alternatively spliced forms
- We need methods to account for alternatively spliced forms

Conclusion

- Genetics and statistics are intertwined
- Linkage analysis and Association Studies complement one another
- Many statistical methods are reusable in genetics research
- Further statistics research is needed for analyzing next generation sequencing



Biostatistics

Collaboration Center

- Thank you!
- Please go to our website at <http://www.feinberg.northwestern.edu/depts/bcc/>
or google “BCC Northwestern”
- Fill out our online request form for further collaboration