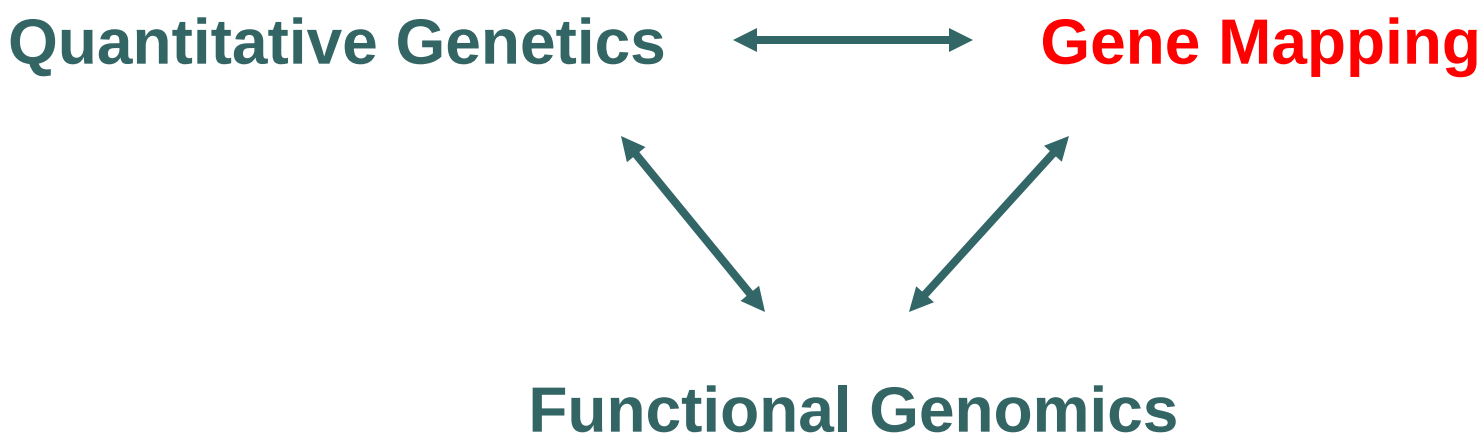
Future Directions

Pak Sham, HKU

Boulder 2007

Genetics of Complex Traits



Gene Mapping – GWA Era

The promise

- Detect all the “big” genetic players
- Understand how they interact with each other and with the environment
- Generation of detailed hypotheses regarding etiology

The challenges

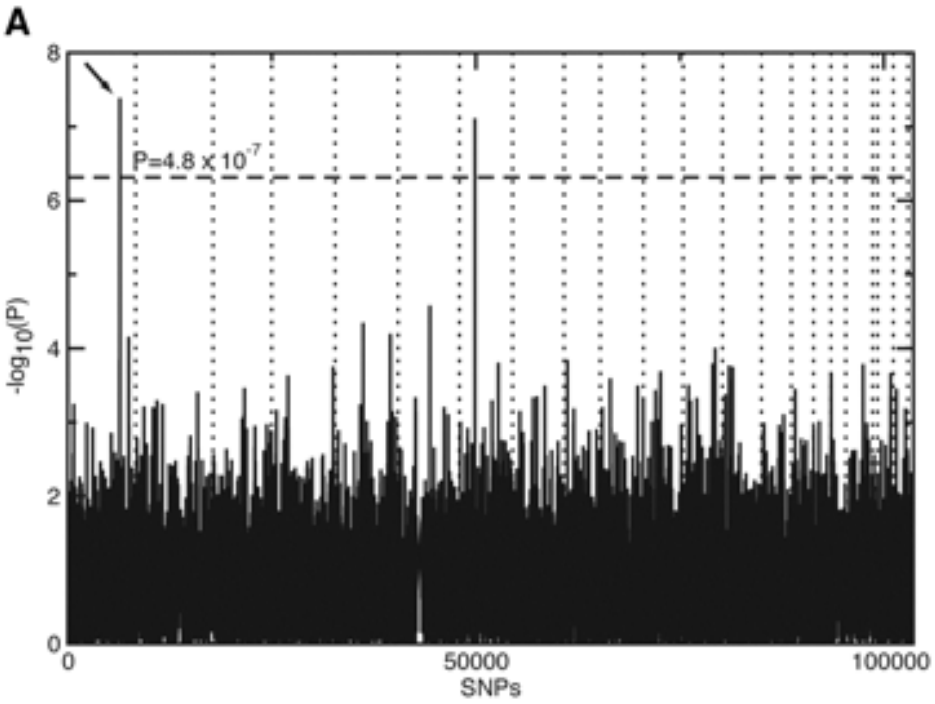
- How to get enough funding?
- How to get the most out of them?
  - Study design
  - Data analysis

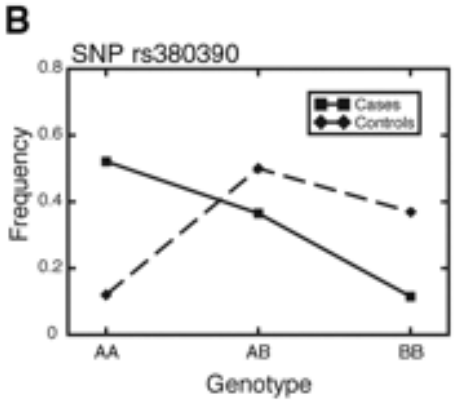
GWA: Taking the Plunge

Age-related
Macular Degeneration

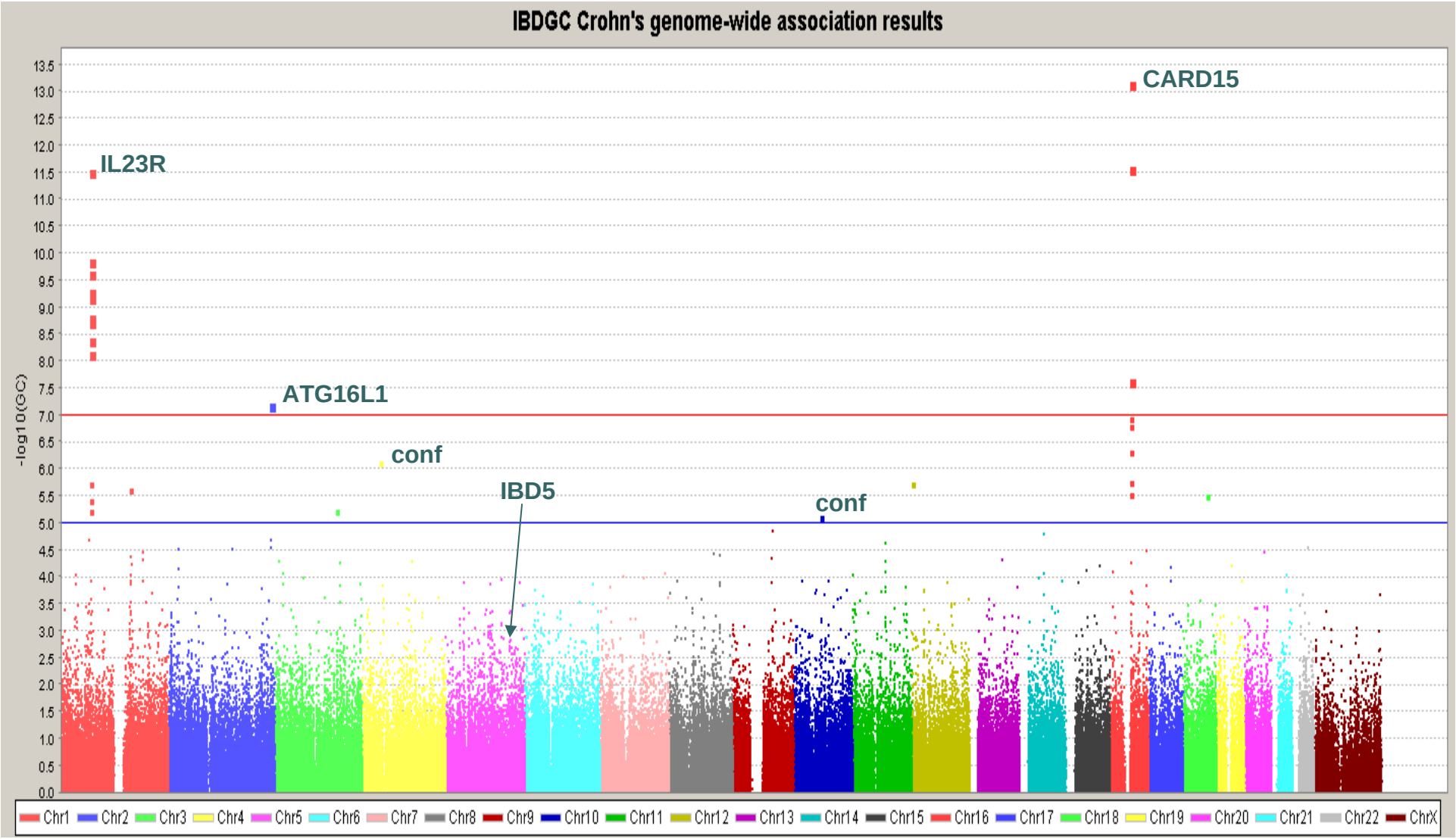
96 cases / 50 controls

100,000 SNPs





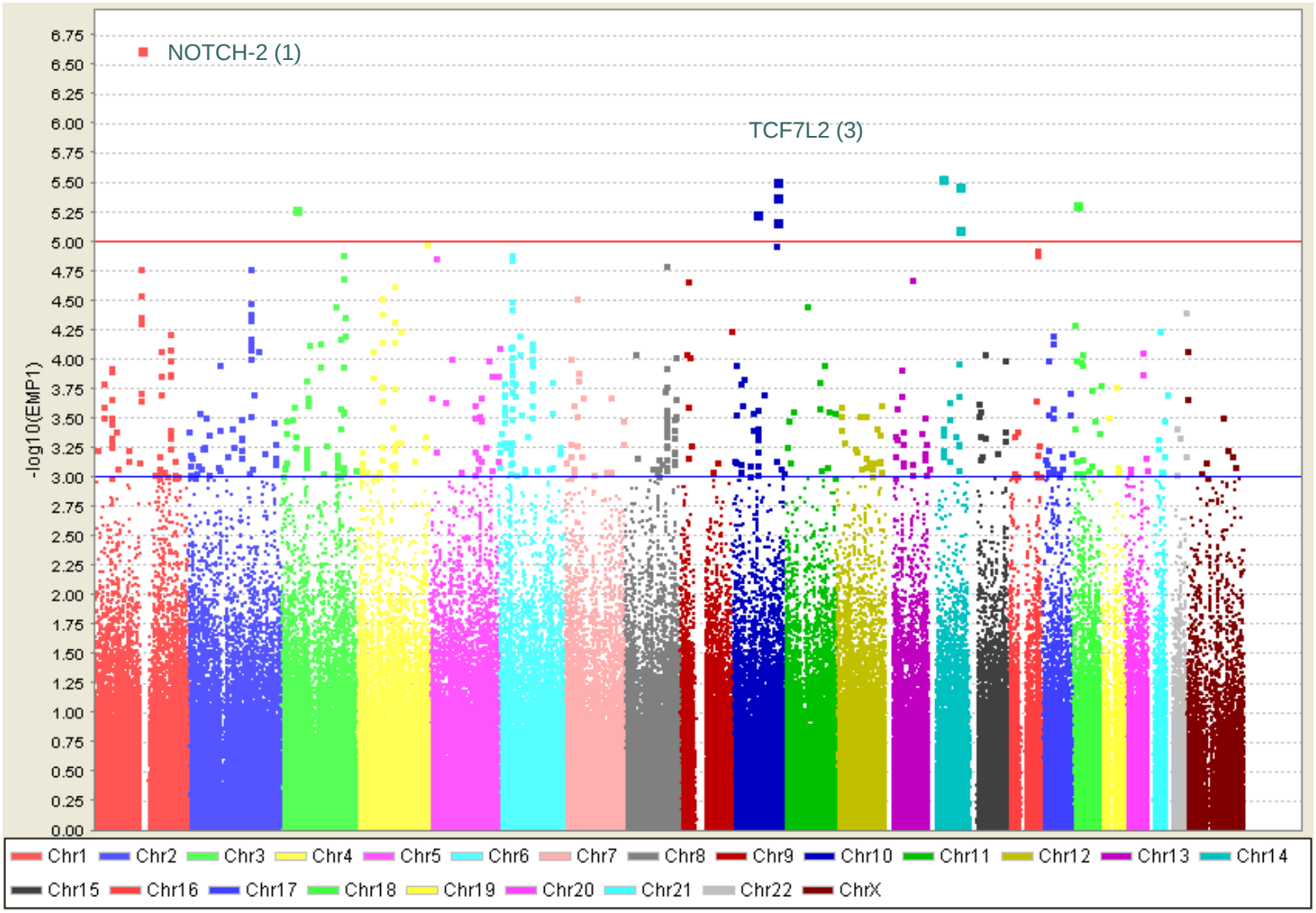
Klein *et al.* (2005)



946 cases, 977 controls

From Dr Mark Daly

Type 2 Diabetes Mellitus Genome-wide Association Results



From Dr Mark Daly

Design of GWA Studies

Reducing cost

- Shared pool of control subjects
- Split sample designs
  - Two stage: GWA → replication
  - Split-half: e.g. 250K (Sty / Nsp) in each half

Maximizing information

- Choosing most extreme (genetically loaded) subjects
- Choosing most accurately and comprehensively phenotyped subjects

Mining GWA Data

Aim

To squeeze out all the information from the vast amount of genotype data

Strategy

Optimal genotype calls

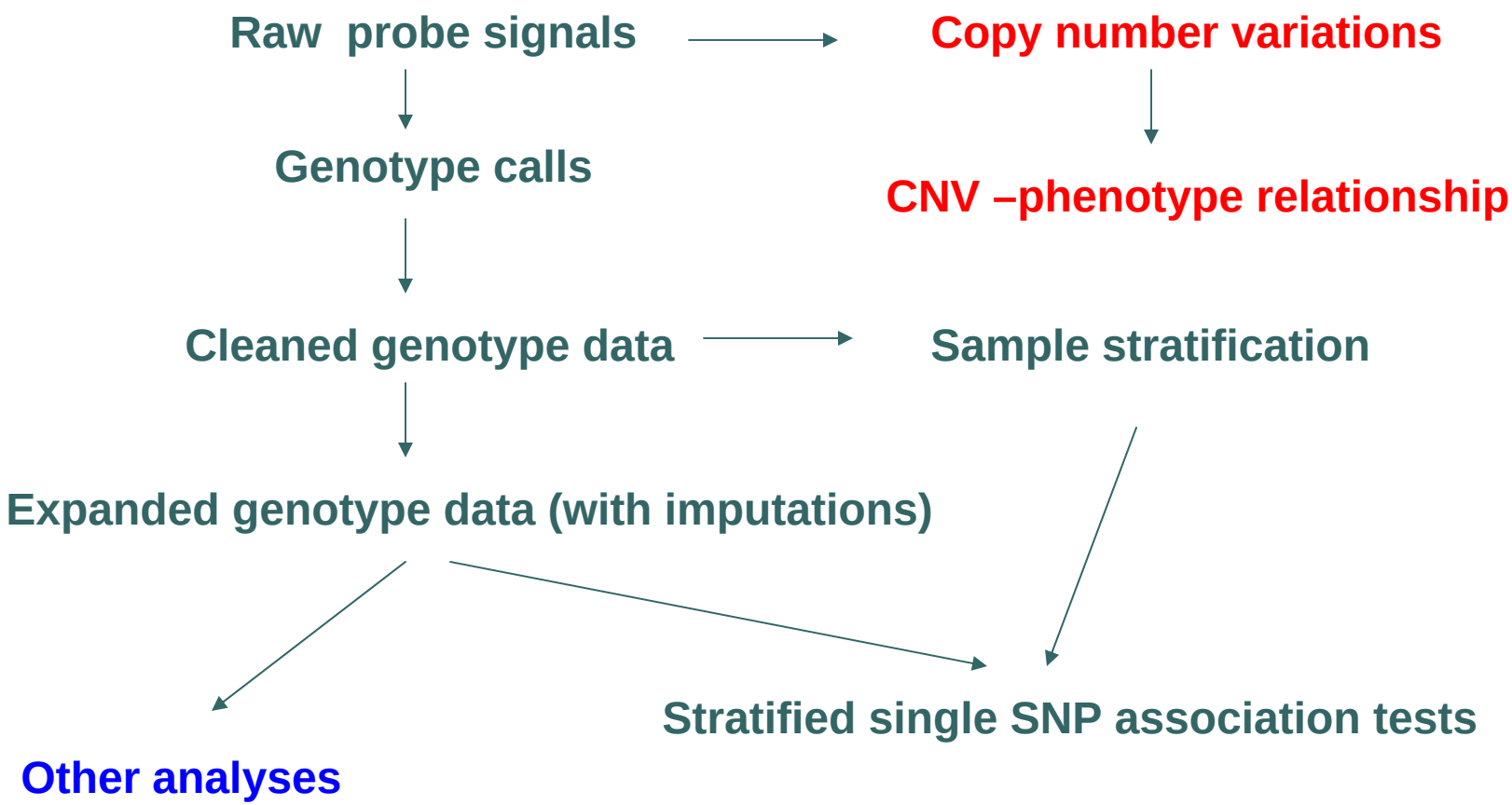
Thorough data cleaning

Sample characterization (stratification)

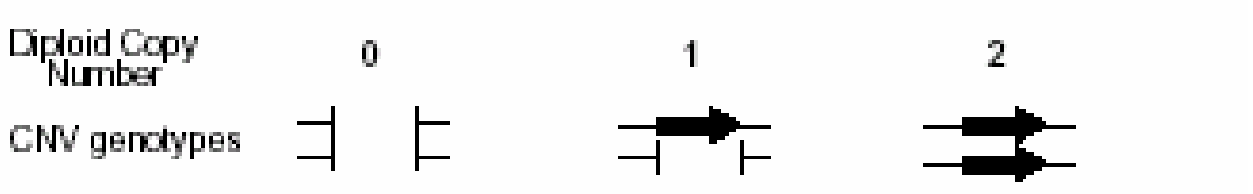
Apply multiple statistical methods

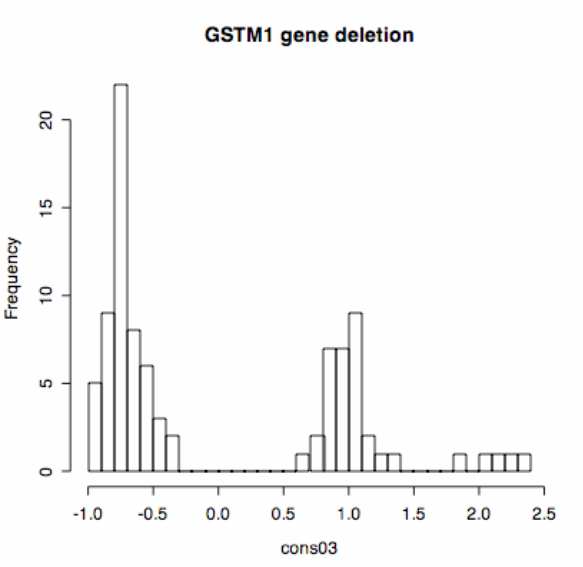
Replication

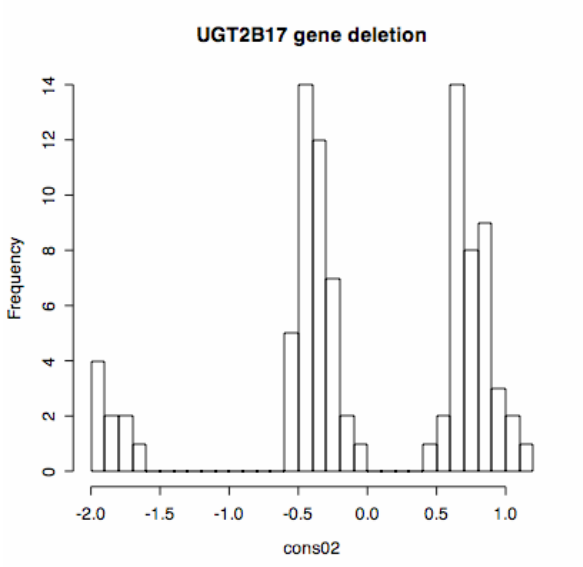
WGA Analysis Flow

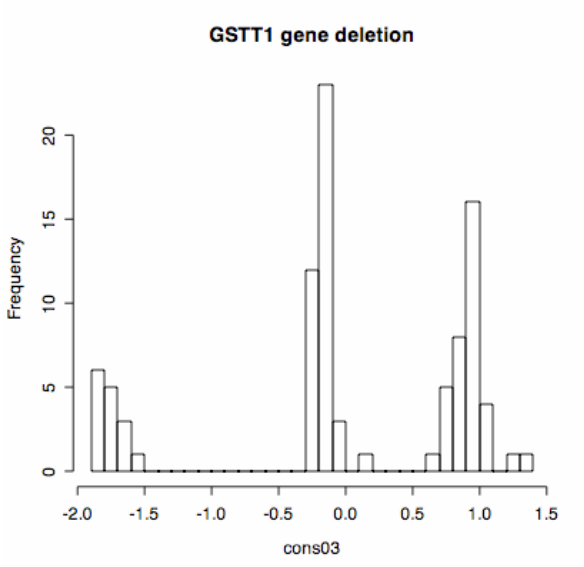


Deletion CNVs









CNV and Disease

CNVs

- are common throughout the genome
- can influence gene function (increased / decreased levels)
- can influence disease susceptibility
  - Charcot Marie-Tooth disease type A (PMP22)
  - Early-onset Parkinson's disease (alpha-synuclein)
  - Susceptibility to HIV infection (CCL3L1)

Family-based Tests

“DFAM” Test (implemented in PLINK)

- Break pedigree into nuclear families
- Consider count of minor allele (X) among affected offspring
- Calculate expectation and variance of (X) conditional on parental genotypes (if available) or sibship genotypes
- Calculate single test statistics

Combining Studies

- Imputation to establish common SNP set and then
  - Combine data and use stratified association analysis
  - Meta-analysis – combine odds ratios (inverse variance weighting)

Looking for Epistasis

- Epistatic components not detectable by single-locus association analyses
- Simple methods of epistasis analyses
  - Test of homogeneity of odds ratios or means differences across genotypic strata
  - Test of interaction in logistic or linear regression models
  - Test for correlation between unlinked loci
  - Test for difference in correlation between loci, in cases and controls
- Increases multiple testing: e.g. 500,000 SNPs leads to 124,999,750,000 possible pairs of SNPs

Epistasis: Is it worth doing?

Marchini et al (2005) compared 4 analytic strategies using simulated data with epistasis

(1) One single-locus analysis

(2) Two single-locus analyses

(3) All possible pairs of loci

(4) Two-stage: pairs of loci with low p-values from single-locus analysis

Results: Strategies (3) and (4) are often more powerful than (1) and (2) even after Bonferroni adjustment

Will GWA catch all?

Certainly not!

- Alleles of small effects
- Rare variants

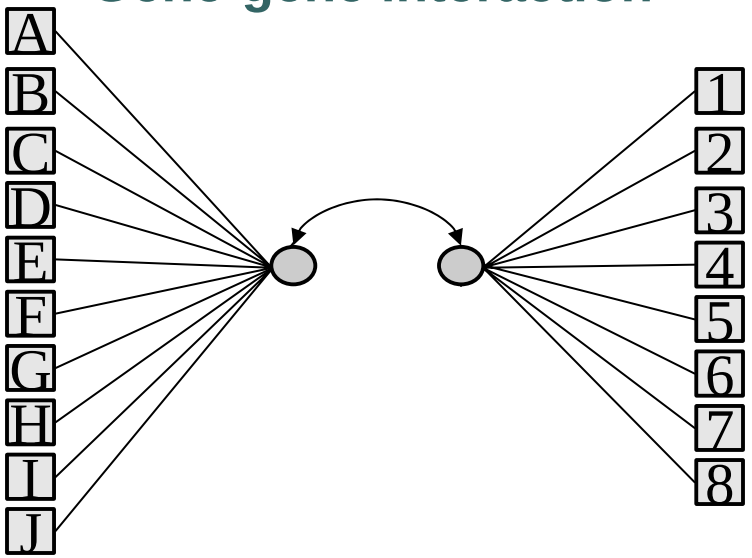
“Residual” genetic variation

Using Functional Information

Gene-based tests: Haplotypic / Allelic

Pathways –based tests

Gene-gene interaction



Canonical correlation analysis

Population-based Linkage

Linkage is possible only in families, BUT

- Every pair of individuals are related if traced back far enough
- Genetic relationship (overall genomic sharing) can be estimated from GWA data
- Local IBD sharing can be also estimated from GWA data
- Therefore IBD sharing can be correlated with phenotypic similarity in GWA data
- Likely to be useful for rare phenotypes with rare variants of moderately strong effects

Estimating Genome-wide IBD

Expected number of SNPs with IBS =

IBD	0	1	2
0	$2 \Sigma (p_i q_i)^2$	$4 \Sigma p_i q_i (1-2p_i q_i)$	$N-2 \Sigma p_i q_i (2-3p_i q_i)$
1	0	$2 \Sigma p_i q_i$	$N - 2 \Sigma p_i q_i$
2	0	0	N

N : Number of SNPs; p, q : allele frequencies

Estimating Genome-wide IBD

$$E(N_0) = 2 f_0 \sum p_i q_i$$

$$E(N_1) = 4 f_0 \sum p_i q_i (1 - 2 p_i q_i) + 2 f_1 \sum p_i q_i$$

$$E(N_2) = N - 2 f_0 \sum p_i q_i (2 - 3 p_i q_i) - 2 f_1 \sum p_i q_i$$

N_0 = Number of IBS0 SNPs
N_1 = Number of IBS1 SNPs
N_2 = Number of IBS2 SNPs

f_0 = Proportion of genome IBD0
f_1 = Proportion of genome IBD1
f_2 = Proportion of genome IBD2

Estimating Genome-wide IBD

- The estimated genome-wide IBD proportions, obtained by solving the linear equations, are:

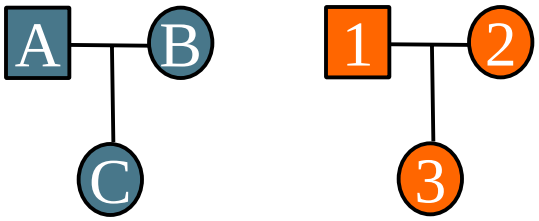
$$f_0 = N_0 / \left(2 \sum p_i q_i\right)$$

$$f_1 = \left(N_1 - 4 f_0 \sum p_i q_i (1 - 2 p_i q_i)\right) / \left(2 \sum p_i q_i\right)$$

$$f_2 = 1 - f_1 - f_2$$

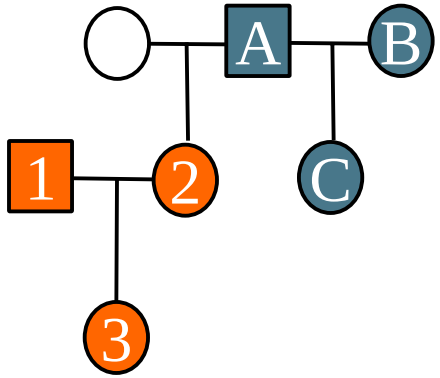
- Boundary conditions
- Small sample (rare allele) adjustment
- Inbreeding adjustment

HapMap Relationships

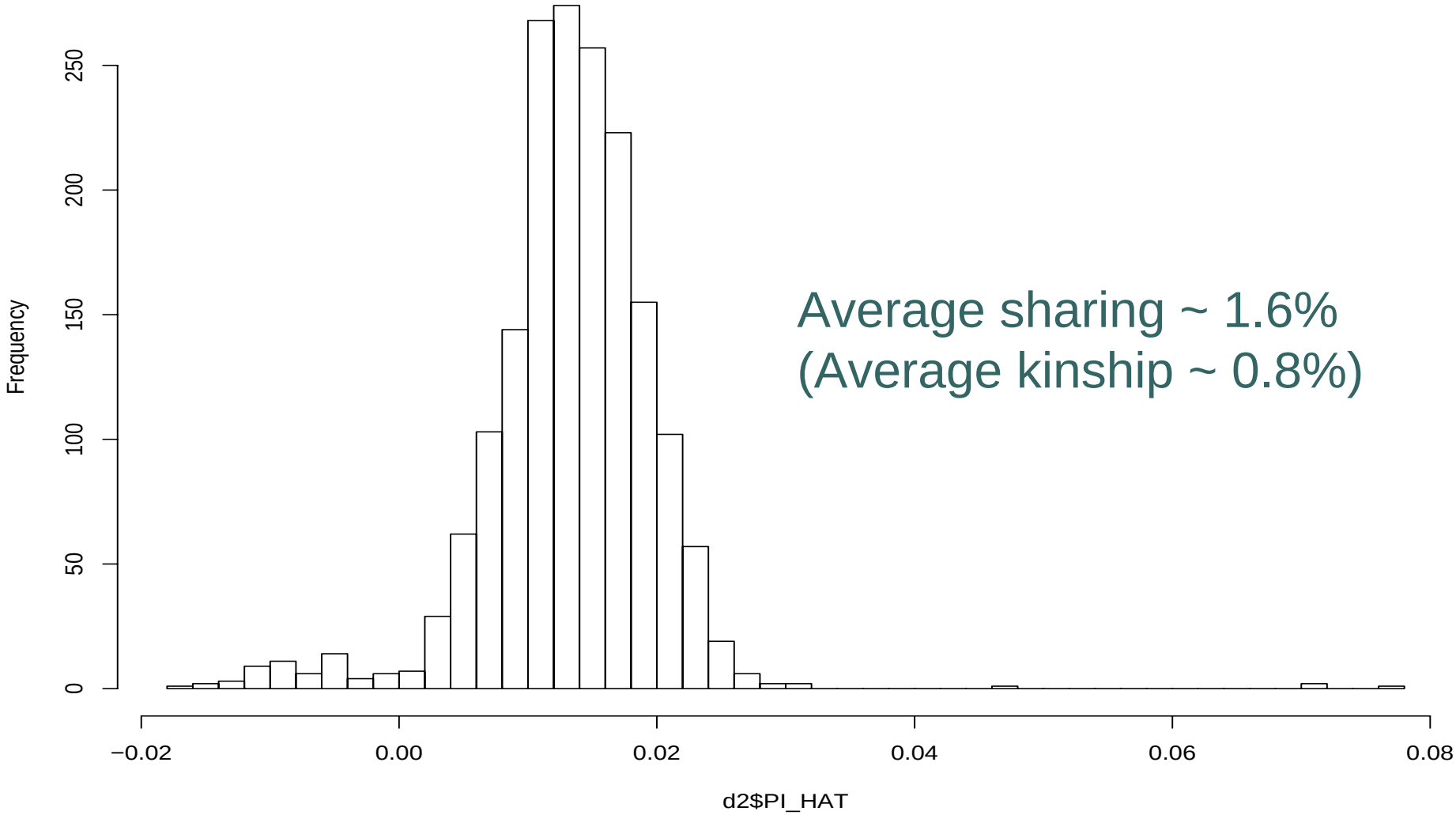


Two Yoruba Trios

		P(IBD=0)	P(IBD=1)	P(IBD=2)
A	2	0.005	0.995	0.000
A	3	0.436	0.564	0.000
C	2	0.388	0.612	0.000



Distribution of Genomic Sharing: Among CEPH Founders



Average sharing ~ 1.6%
(Average kinship ~ 0.8%)

Estimating Segmental Sharing

- Prune SNPs to reduce LD relationships
- Use allele frequencies to calculate likelihood of IBD given single SNP genotypes of the pair of individuals
- Use genome-wide IBD to estimate the least number of meioses that separate the two genomes
- Use number of meioses to calculate transition matrix of IBD states
- Use Hidden Markov Model to calculate “multipoint” IBD probabilities

Transition Matrix

IBD(i+1)

IBD(i)

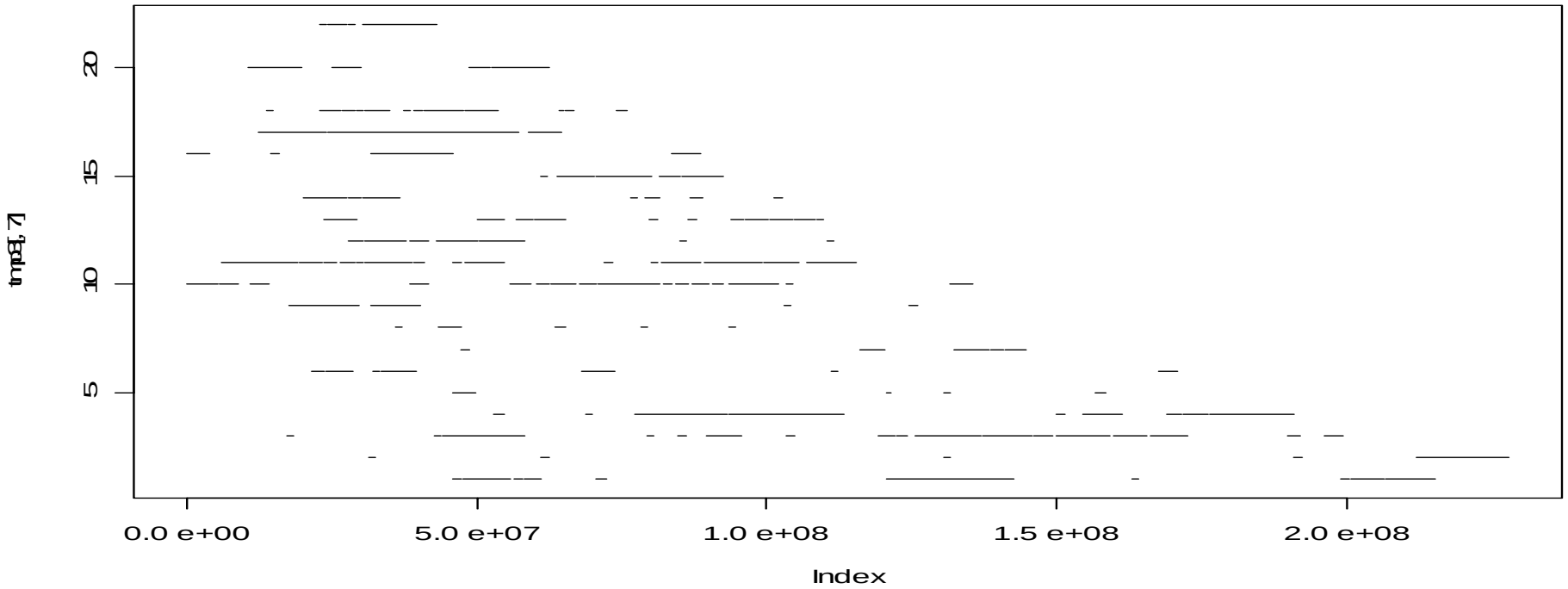
	0	1
0	$1-\frac{1-(1-\theta)^{m-2}\xi}{2^{m-1}-1}$	$\frac{1-(1-\theta)^{m-2}\xi}{2^{m-1}-1}$
1	$1-(1-\theta)^{m-2}\xi$	$(1-\theta)^{m-2}\xi$

θ : Recombination fraction $\quad \xi = \theta^2 + (1-\theta)^2$

m : Number of meioses (≥ 2)

Estimated Segmental Sharing

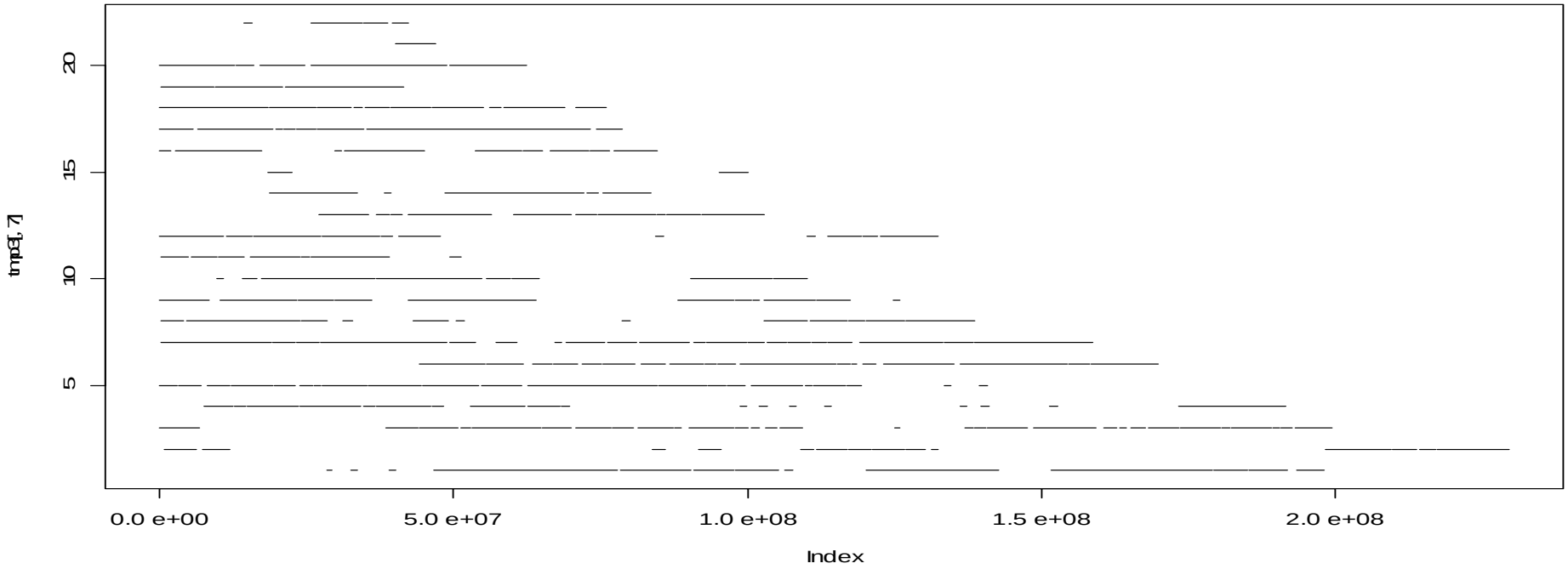
YRI: NA19130, NA191940 (Half aunt / Half niece)



Overall: IBD0 = 0.76, IBD1 = 0.24, IBD2 = 0

Estimated Segmental Sharing

YRI: NA18913, NA19240 (Grandparent / Grandchild)



Overall: IBD0 = 0.44, IBD1 = 0.56, IBD2 = 0

Other Data Mining Methods

The possibilities are endless!

Neural networks

CART

MARS etc …….

Beyond GWA

- Incorporating measured genotypes into quantitative genetic-epidemiological analysis
- Functional genomic studies – gene expression profiles, cell biology, etc.

Summary

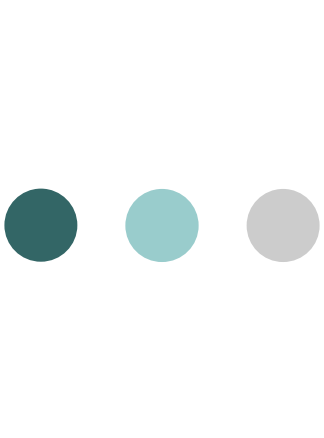
The technology for GWA is reaching maturity

GWA is already yielding novel susceptibility loci for complex diseases

GWA are increasing in number and in size

GWA data offer interesting analytical and computational challenges

The results from GWA studies will revolutionize quantitative genetics and functional genomics



The End