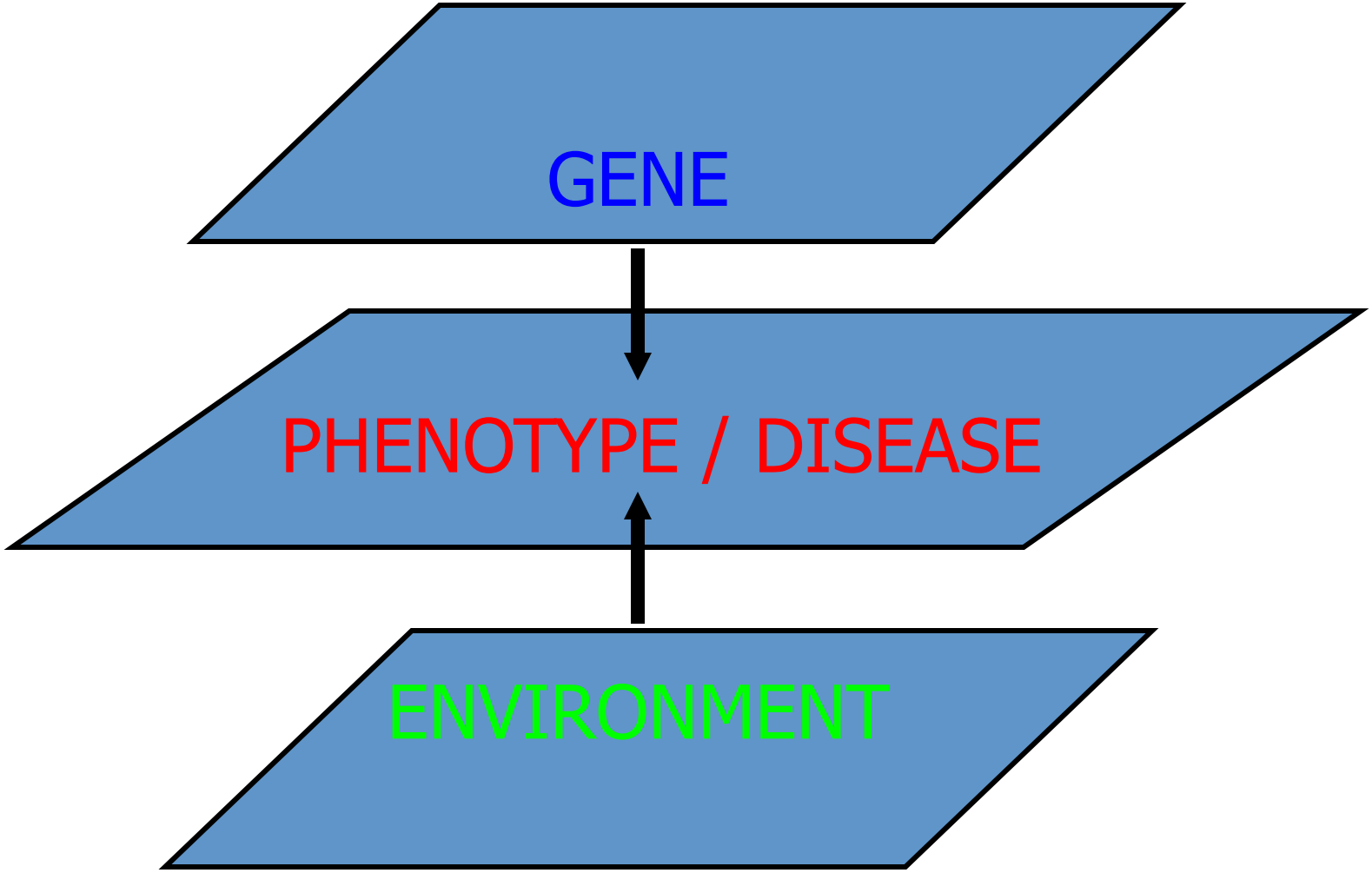
BIO306: Bioinformatics

Lecture 8

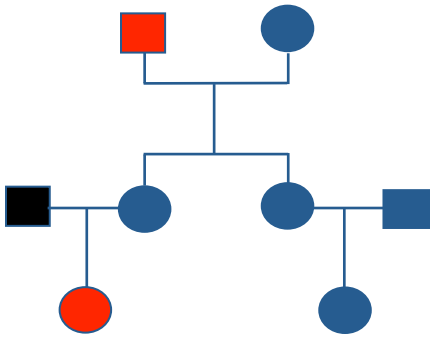
Identifying disease associated variants

Wenfei JIN PhD
jinwf@sustc.edu.cn
Department of Biology, SUSTech


GENE
PHENOTYPE / DISEASE
ENVIRONMENT

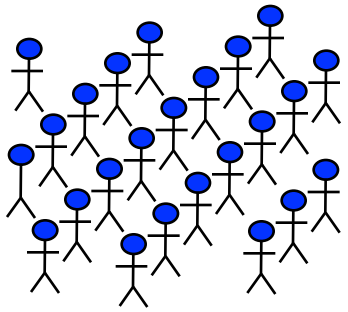
Gene Mapping

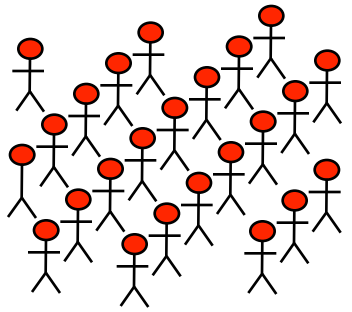
- Linkage analysis



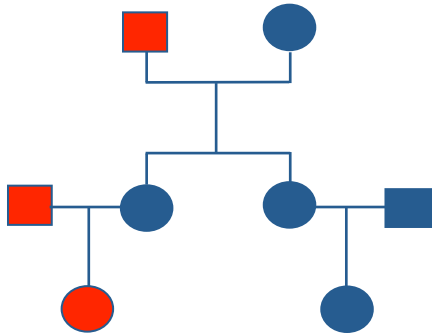
Two strategies

- Association analysis





Gene Mapping

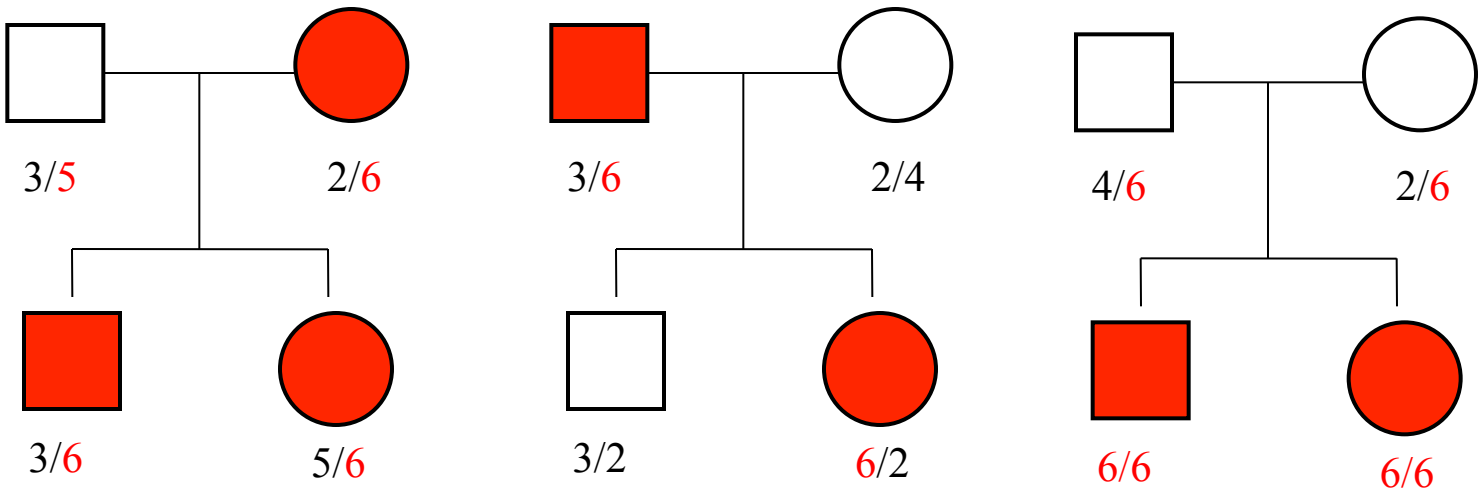


- Linkage analysis
  - Pedigree data
  - Localize chromosomal regions where disease gene might be found.
  - Low resolution (10s cM ≈ 10^7-10^8 bp in Human).
- Association analysis
  - Population data
  - Further localize the region where the disease gene is located.
  - High resolution (10s - 100s kb).

Principle always same

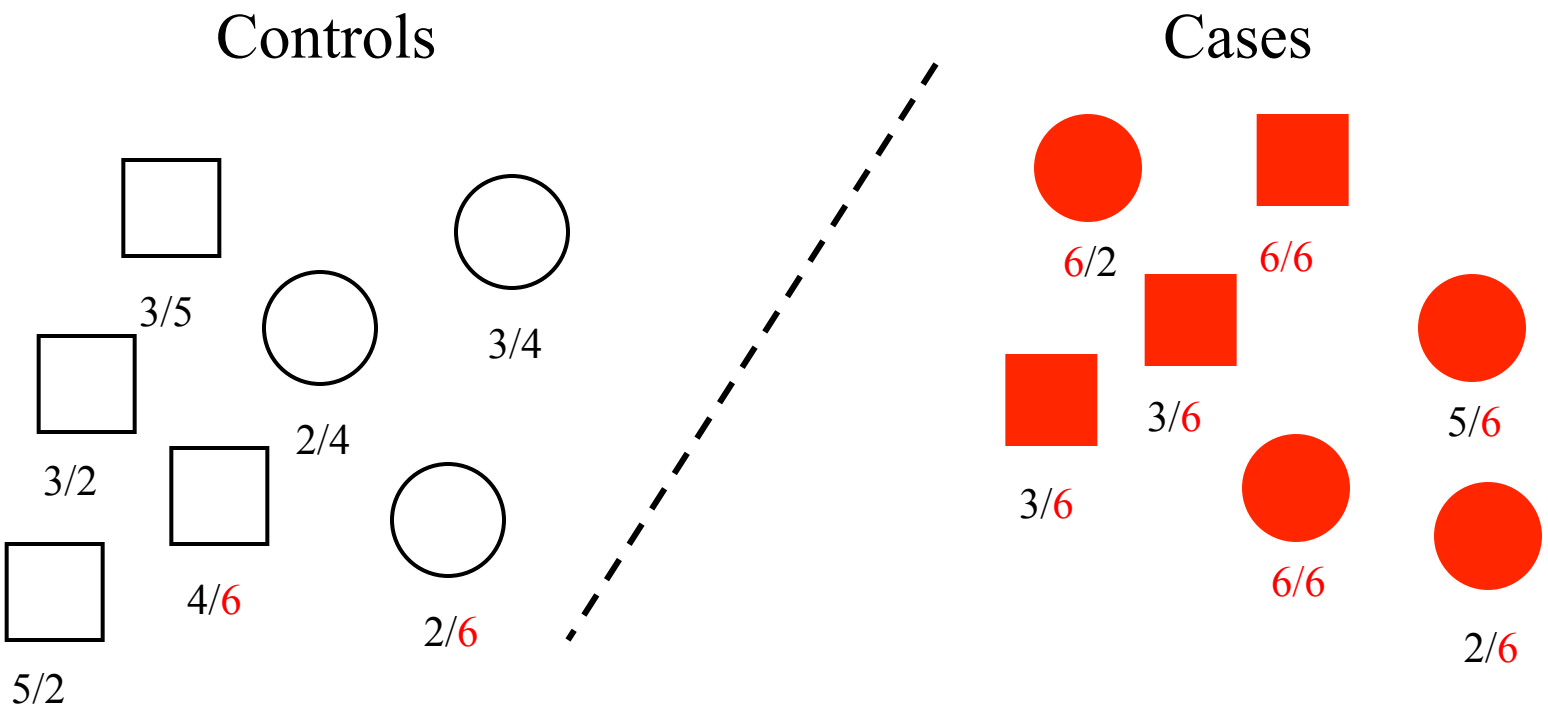
Correlate phenotypic and genotypic variability

Association AND Linkage



All families are ‘linked’ with the marker
Allele 6 is ‘associated’ with disease

Allelic Association

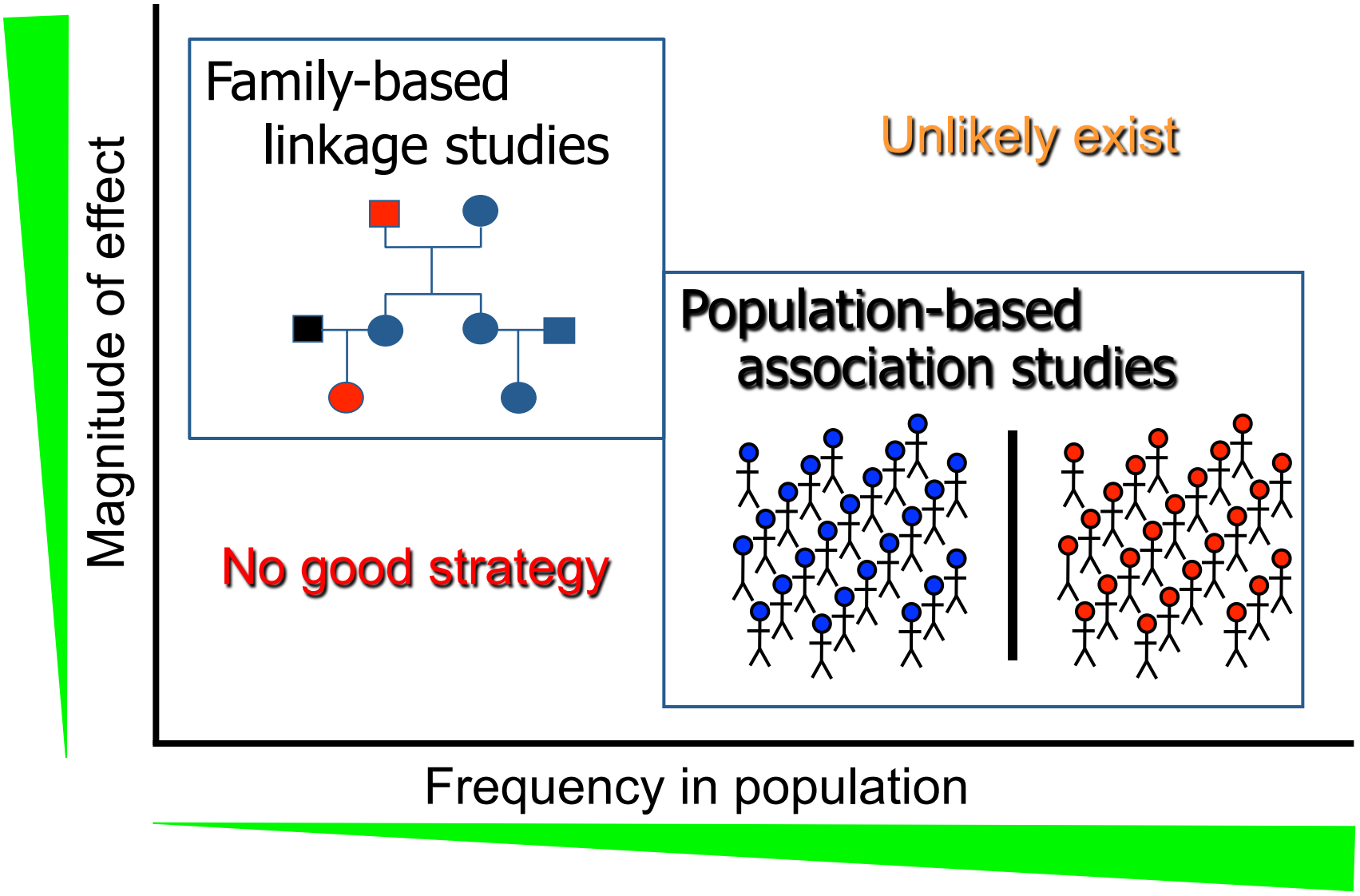


Allele 6 is ‘associated’ with disease

Linkage vs Association

Linkage	Association
1. Requires families	1. Families or unrelateds
2. Matching/ethnicity generally unimportant	2. Matching/ethnicity important
3. Few markers for genome coverage (300-400 STRs)	3. Many markers for genome coverage (10^5-10^6 SNPs)
4. Yields coarse location	4. Yields fine-scale location
5. Good for initial detection; poor for fine-mapping	5. Good for fine-mapping; poor for initial detection
6. Powerful for rare variants	6. Powerful for common variants; rare variants generally impossible

Optimal mapping strategies



Association Study Designs

- Designs
  - Family-based
    - Trio (TDT), twins/sib-pairs/extended families (QTDT)
  - Case-control
    - Collections of individuals with disease, matched with sample w/o disease
    - Some ‘case only’ designs

Family-based Designs for Association Studies

Advantages:

- Not susceptible to confounding due to population substructure
- Tests for linkage and association
- Can test for parent-of-origin effects

Disadvantages:

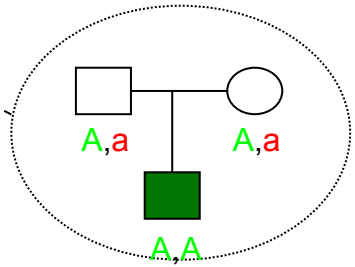
- Inefficient recruitment, only heterozygous parents informative
- Often cannot test for environmental main-effects
- Family members often not available (eg, late-onset diseases)

Basic idea underlying TDT (transmission-disequilibrium test)

- Disease alleles are transmitted from parents to offspring
- Marker alleles in LD with these alleles will also be transmitted preferentially to affected offspring
- Test if heterozygous parents transmit a particular marker allele to affected offspring more frequently than expected
  - Looks for excess transmission of particular alleles from parents to affected children
  - Controls are ‘non-transmitted alleles’

Example of TDT

For each individual, have 2x2 table of 0s, 1s, or 2s



	A-Not transmitted	a – Not transmitted
A - Transmitted	0	2
a - Transmitted	0	0

TDT (transmission-disequilibrium test)

	Non-transmitted allele		
Transmitted allele	**M1**	**M2**	**Total**
M1	a	d	a+b
M2	c	d	c+d
Total	a+c	b+d	2n

The derivation of the TDT shows that one should only use the heterozygous parents (total number b+c). The TDT tests whether the proportions b/(b+c) and c/(b+c) are compatible with probabilities (0.5, 0.5). This hypothesis can be tested using a binomial (asymptotically chi-square) test with one degree of freedom:

$$\chi^2 = \frac{[b-(b+c)/2]^2}{(b+c)/2} + \frac{[c-(b+c)/2]^2}{(b+c)/2} = \frac{(b-c)^2}{b+c}$$

Why Case/Control?

Advantages

- Methodology is well-known
- Convenient to collect
  - **Common**
  - **Very large samples**
- More efficient recruitment than family-based sampling
- Simultaneous assessment of disease allele frequency, penetrance, and AR
- Unrelated controls can provide increased power

Limitations

1. Possible Population Stratification
2. Need for highly dense marker sets (capture LD)
3. Lack of phase information
4. Lack of consistency of results

These can be overcome!

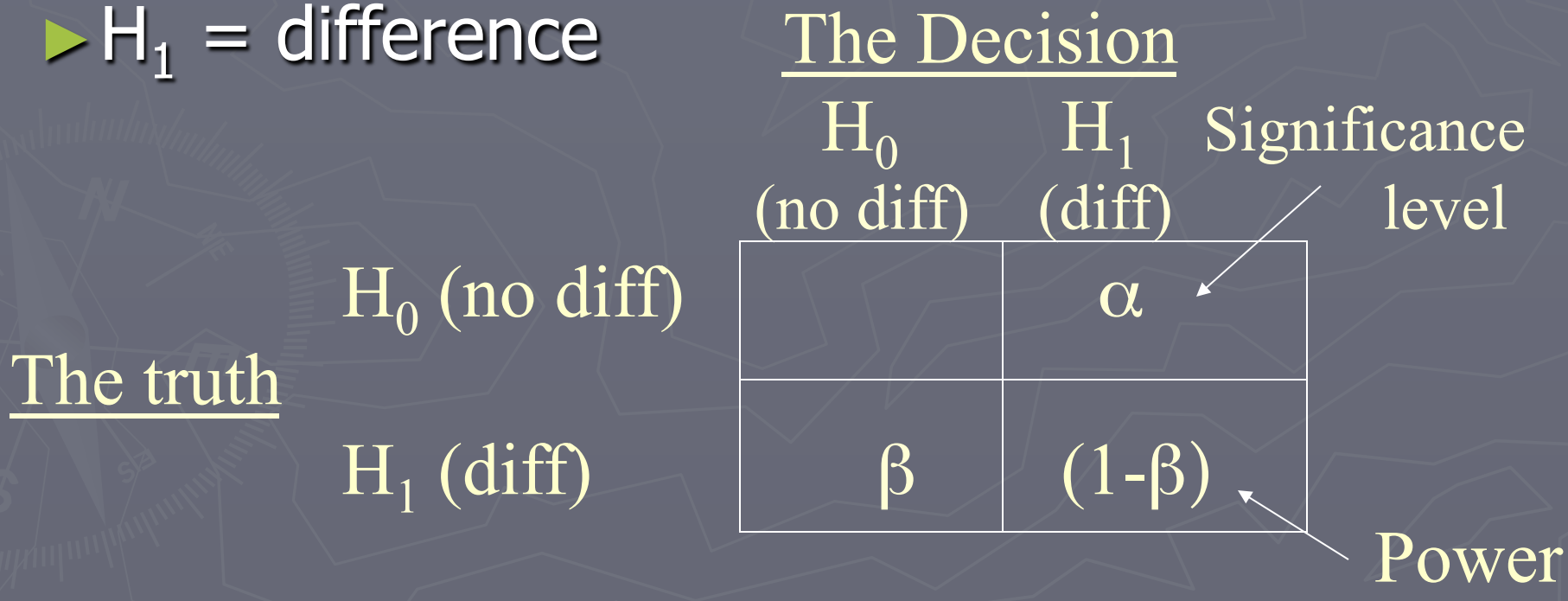
1. Assessment and 'genomic control' of stratification
2. SNP maps
3. Imputed haplotypes

Statistical basis

- The p-value
  - Under the null hypothesis the probability that you observe your data or something more extreme
  - Distribution of the test statistic under the null hypothesis (integrates to 1)
    - F
    - t
    - Chi-Square

The Decision

- Reject the null - fail to reject the null
- Truth versus decision
- H_0 = no change
- H_1 = difference

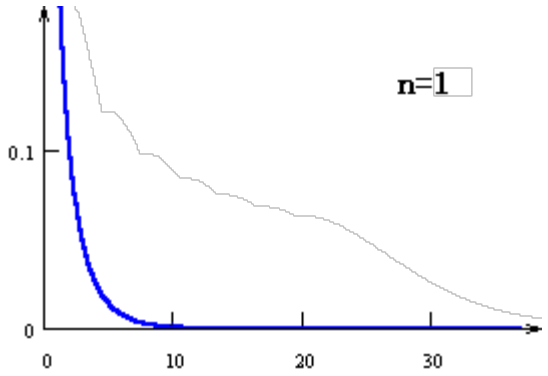


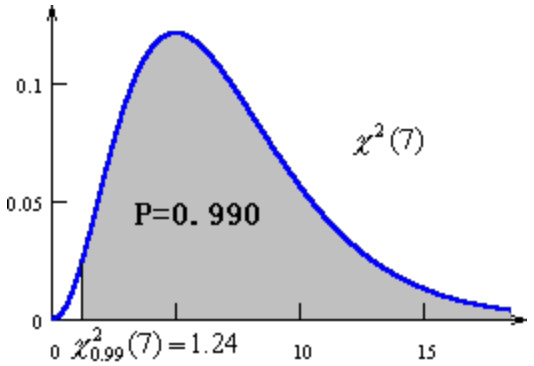
Chi-square (X^2) test

observed

expected

$$X^2 = \sum \frac{(O - E)^2}{E}$$





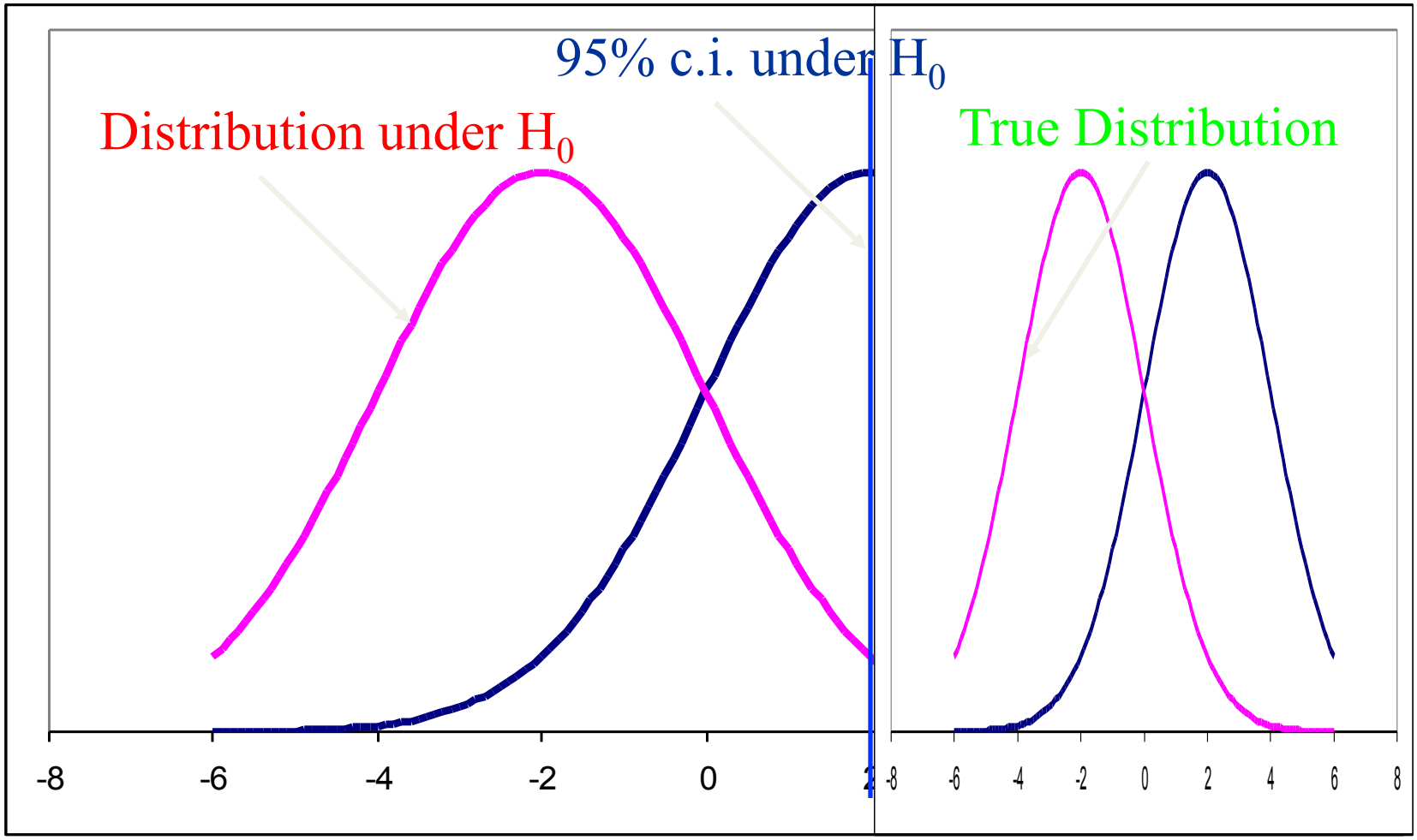
Statistical Power

- Null hypothesis: all alleles are equal risk
- Given that a risk allele exists, how likely is a study to reject the null?
- Are you ready to genotype?

Power Analysis

- Statistical significance
  - Significance = p(false positive)
  - Traditional threshold 5%
- Statistical power
  - Power = 1- p(false negative)
  - Traditional threshold 80%
- Traditional thresholds balance confidence in results against reasonable sample size

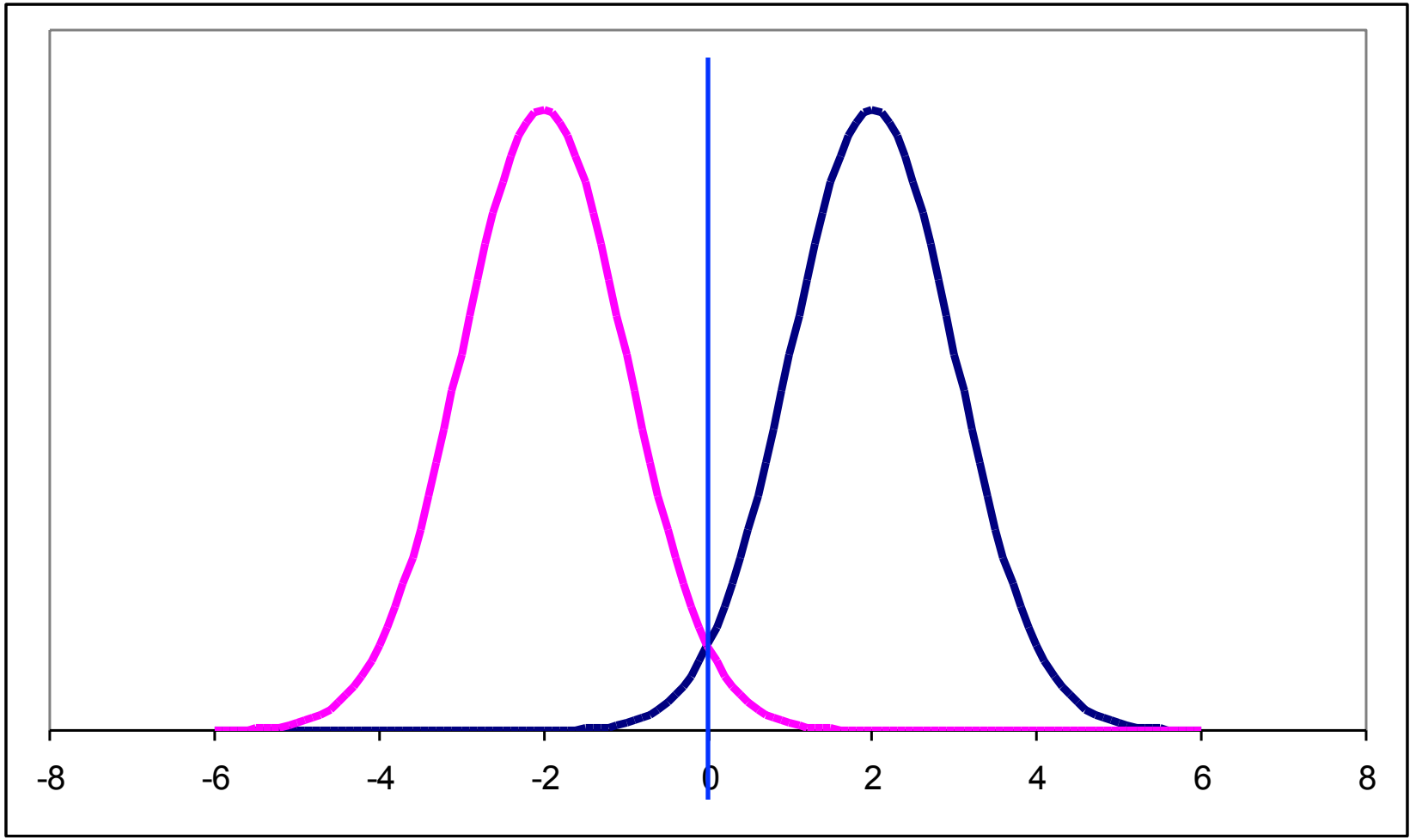
Small sample: 50% Power



Maximizing Power

- Effect size
  - Larger relative risk = greater difference between means
- Sample size
  - Larger sample = smaller SEM
- Measurement error
  - Less error = smaller SEM

Large sample: 97.5% Power

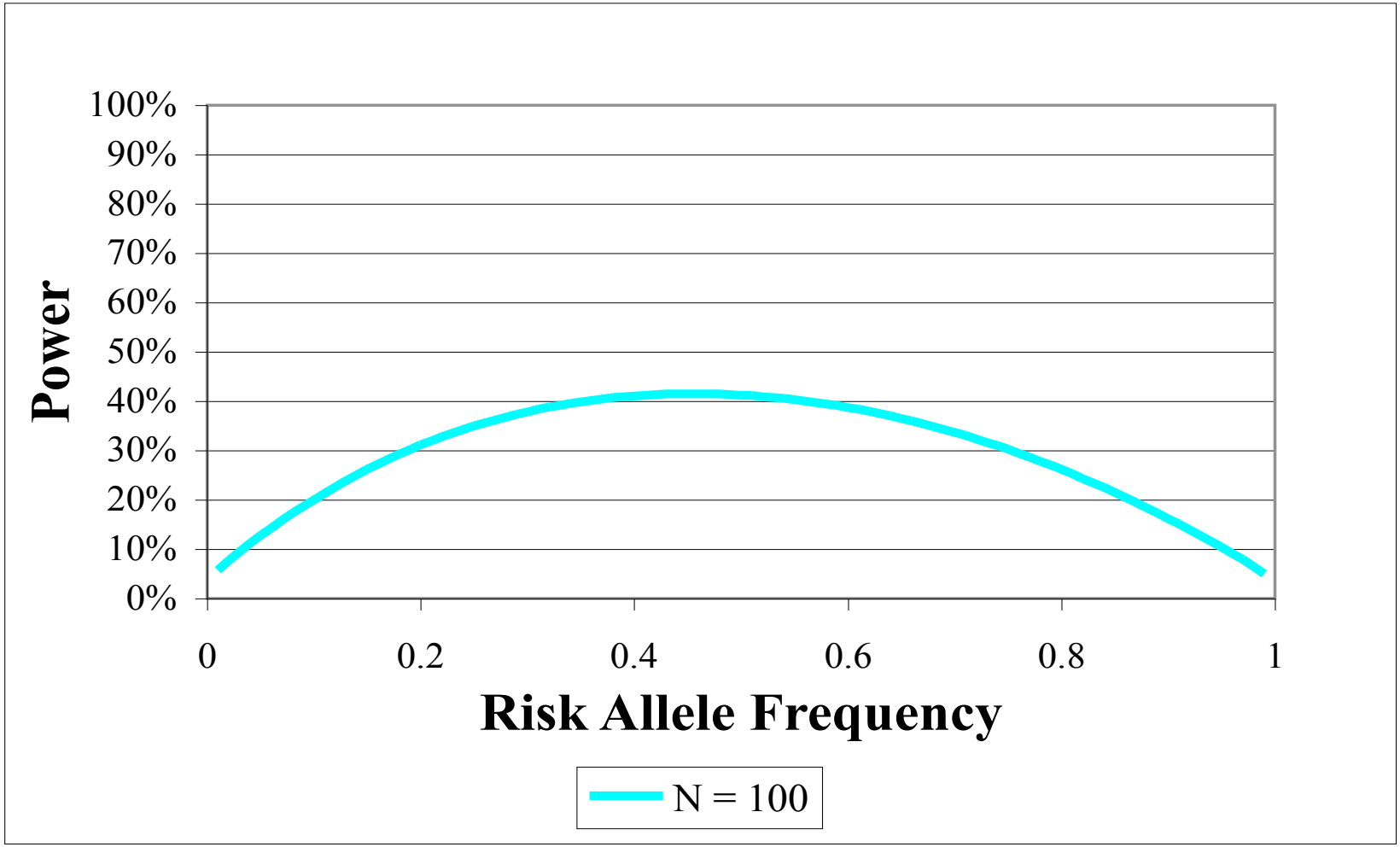


Genetic Relative Risk

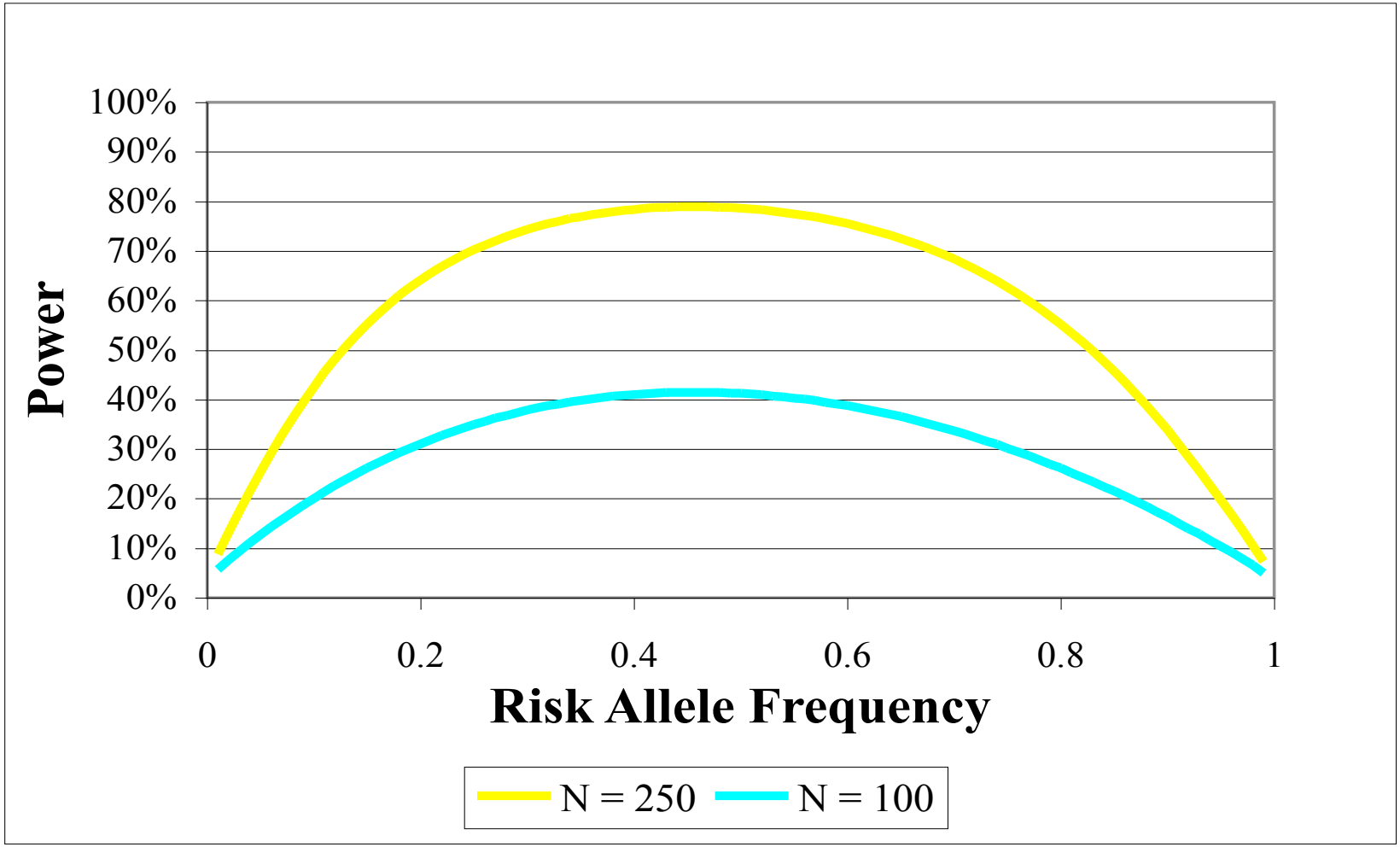
		Disease	
		Disease	Unaffected
SNP	Allele 1	p_{1D}	p_{1U}
	Allele 2	p_{2D}	p_{2U}

$$RR = \frac{p(Disease \mid Allele1)}{p(Disease \mid Allele2)} = \frac{\frac{p_{1D}}{p_{1D} + p_{1U}}}{\frac{p_{2D}}{p_{2D} + p_{2U}}}$$

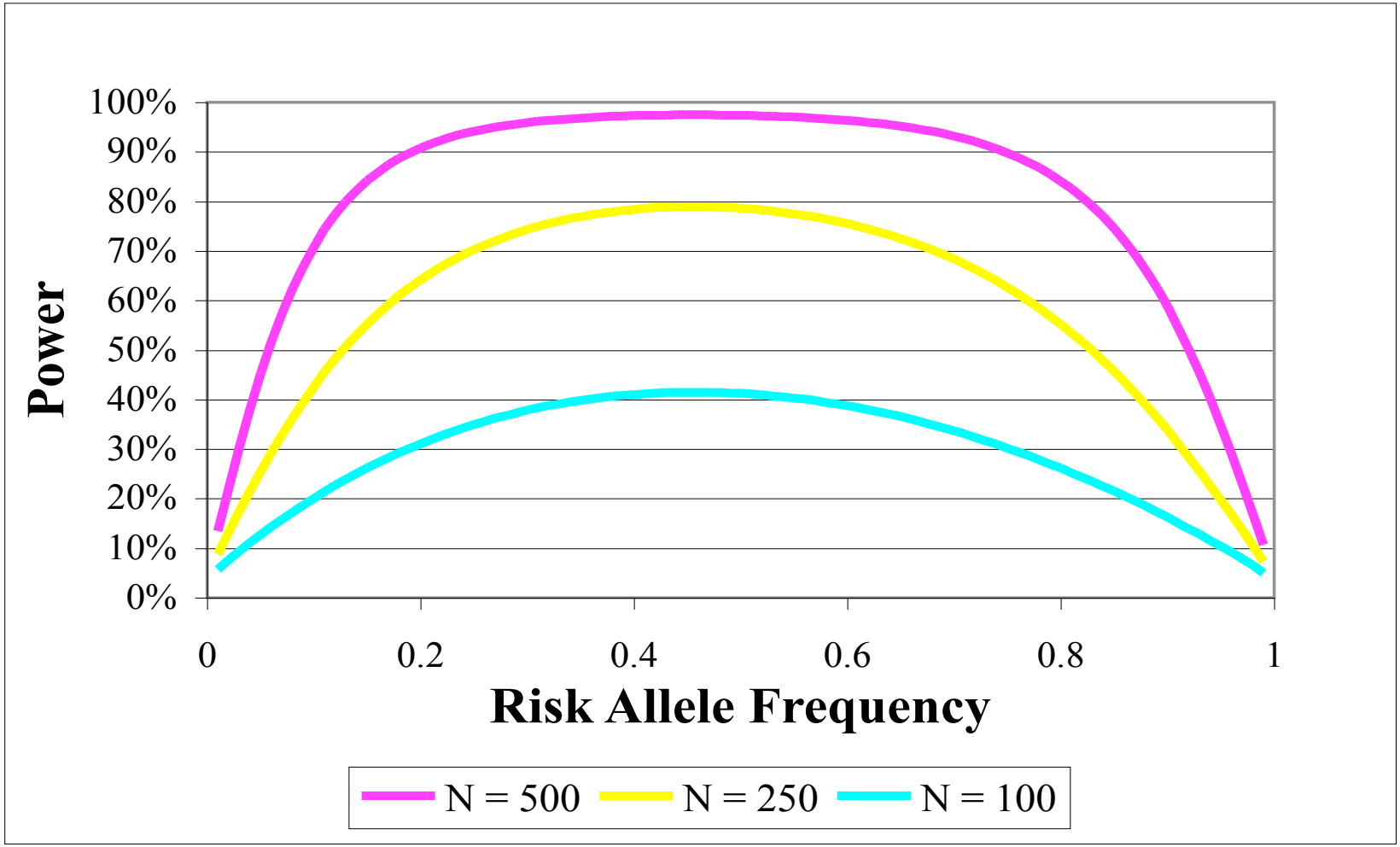
Power to Detect RR=2
N Cases, N Controls



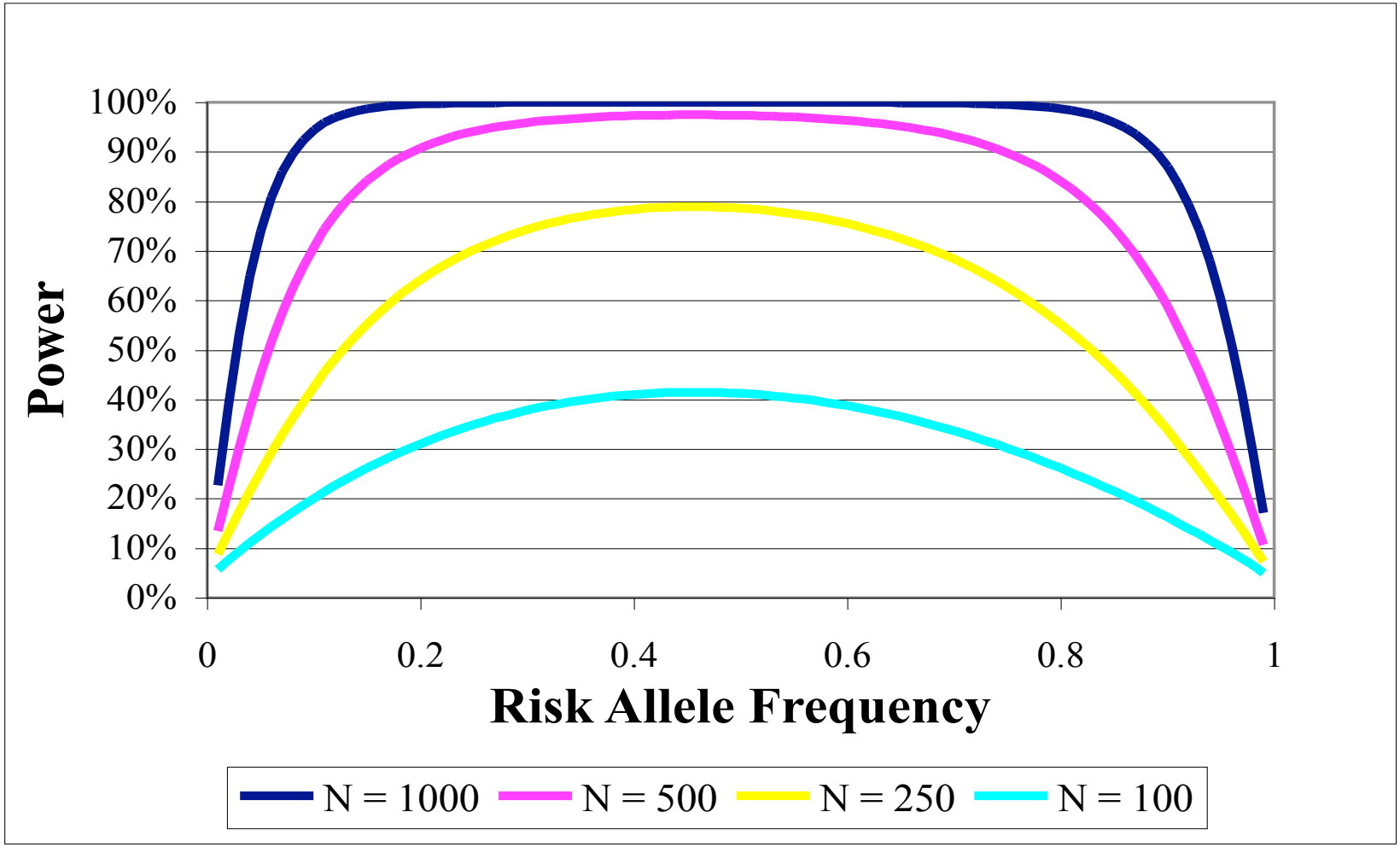
Power to Detect RR=2
N Cases, N Controls



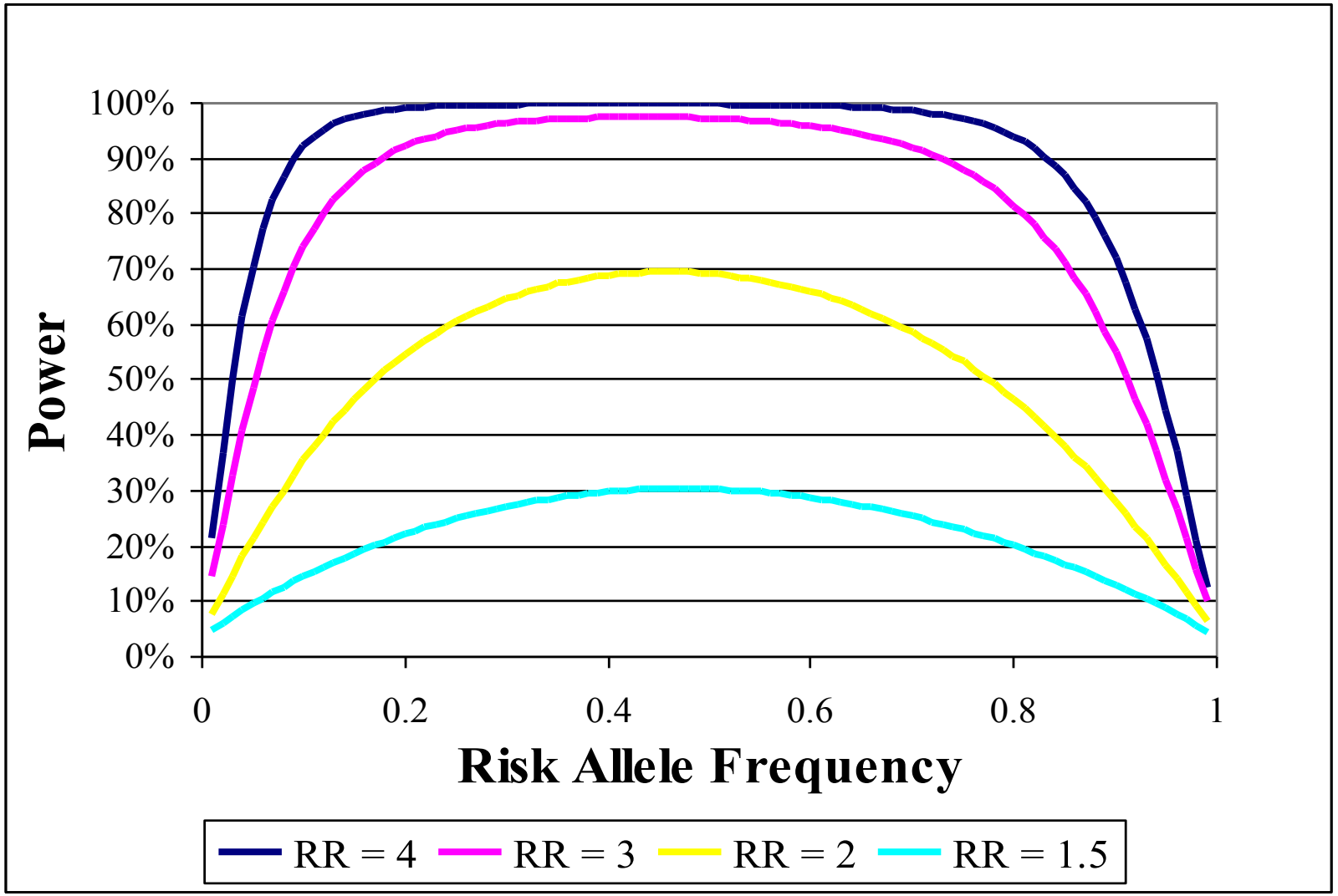
Power to Detect RR=2
N Cases, N Controls



Power to Detect RR=2
N Cases, N Controls



Power to Detect SNP Risk
200 Cases, 200 Controls



Power Analysis Summary

- For common disease, relative risk of common alleles is probably less than 4
- Maximize number of samples for maximal power
- For RR < 4, measurement error of more than 1% can significantly decrease power, even in large samples

Statistical power: an increasing concern

Sample size requirements for case-control analyses of SNPs

(2 controls per case; detectable difference of OR ≥1.5; power=80%).

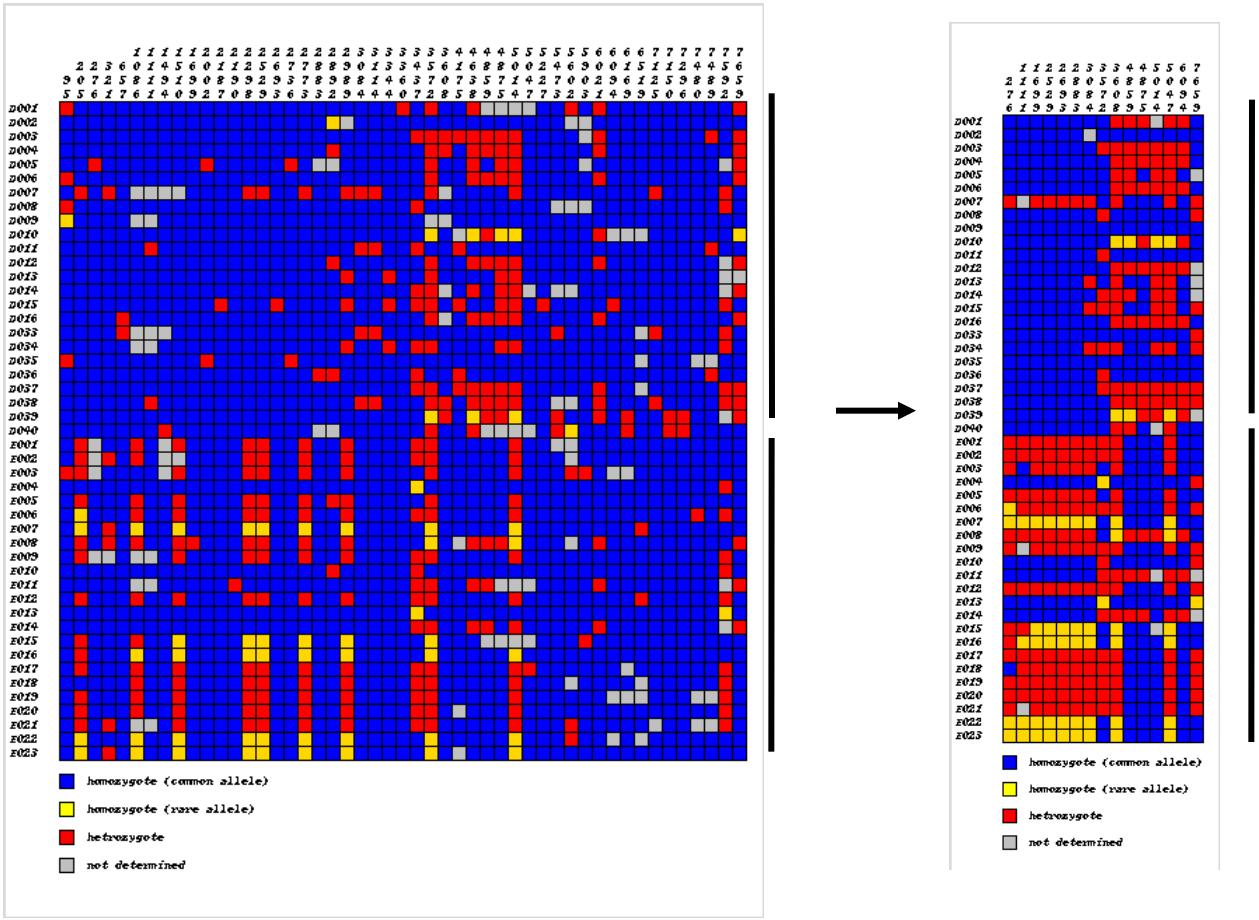
Allele frequency [a]	Dominant model [c]			Recessive model [d]		
	Exposure [b]	No. Cases required		Exposure [b]	No. Cases required	
		α=0.05	α=0.005		α=0.05	α=0.005
10%	19%	430	711	1%	6,113	10,070
20%	36%	311	516	4%	1,600	2,637
30%	51%	308	512	9%	769	1,269
40%	64%	354	590	16%	485	802
50%	75%	456	762	25%	363	602
60%	84%	661	1,107	36%	311	516

Palmer, L. J. and W. O. C. M. Cookson (2001). "Using Single Nucleotide Polymorphisms (SNPs) as a means to understanding the pathophysiology of asthma." Respiratory Research 2: 102-112.

Focus on Common Variants - Haplotype Patterns

All Gene SNPs

SNPs > 10% MAF



Why Common Variants?

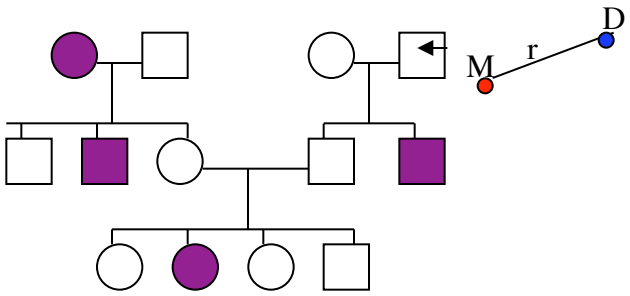
- Rare alleles with large effect (RR > 4) should already be identified from linkage studies
- Association studies have low power to detect rare alleles with small effect (RR < 4)
- Rare alleles with small effect are not important, unless there are a lot of them
- Theory suggests that it is unlikely that many rare alleles with small effect exist (Reich and Lander 2001).

CD/CV Hypothesis

- *Common Disease-Common Variant hypothesis:* Common diseases have been around for a long time. Alleles require a long time to become common (frequent) in the population. Common diseases are influenced by frequent alleles.

Pedigree Analysis & Association Mapping

Pedigree Analysis:

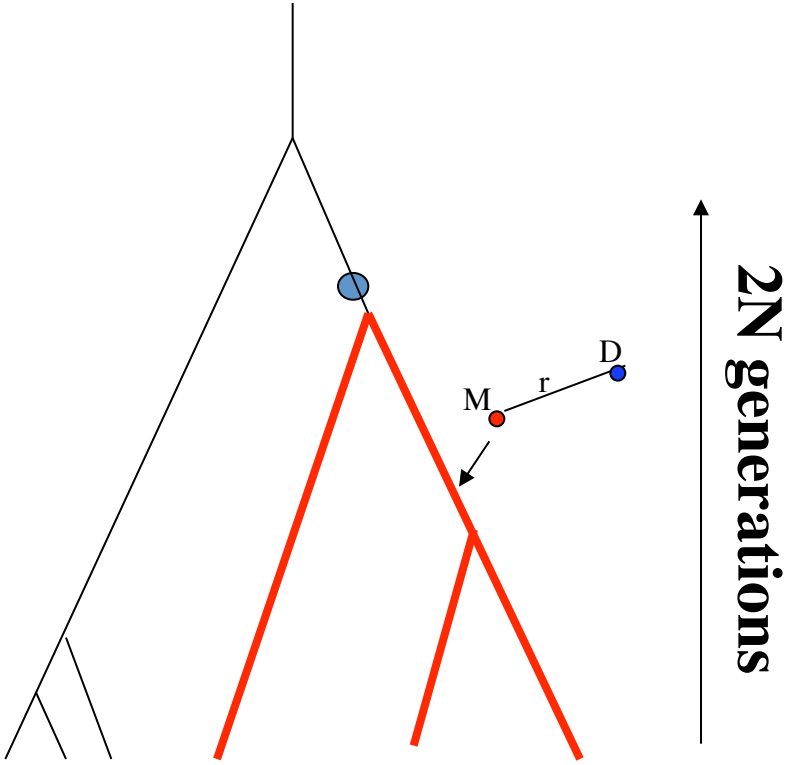


Pedigree known

Few meiosis (max 100s)

Resolution: cMorgans (Mbases)

Association Mapping:



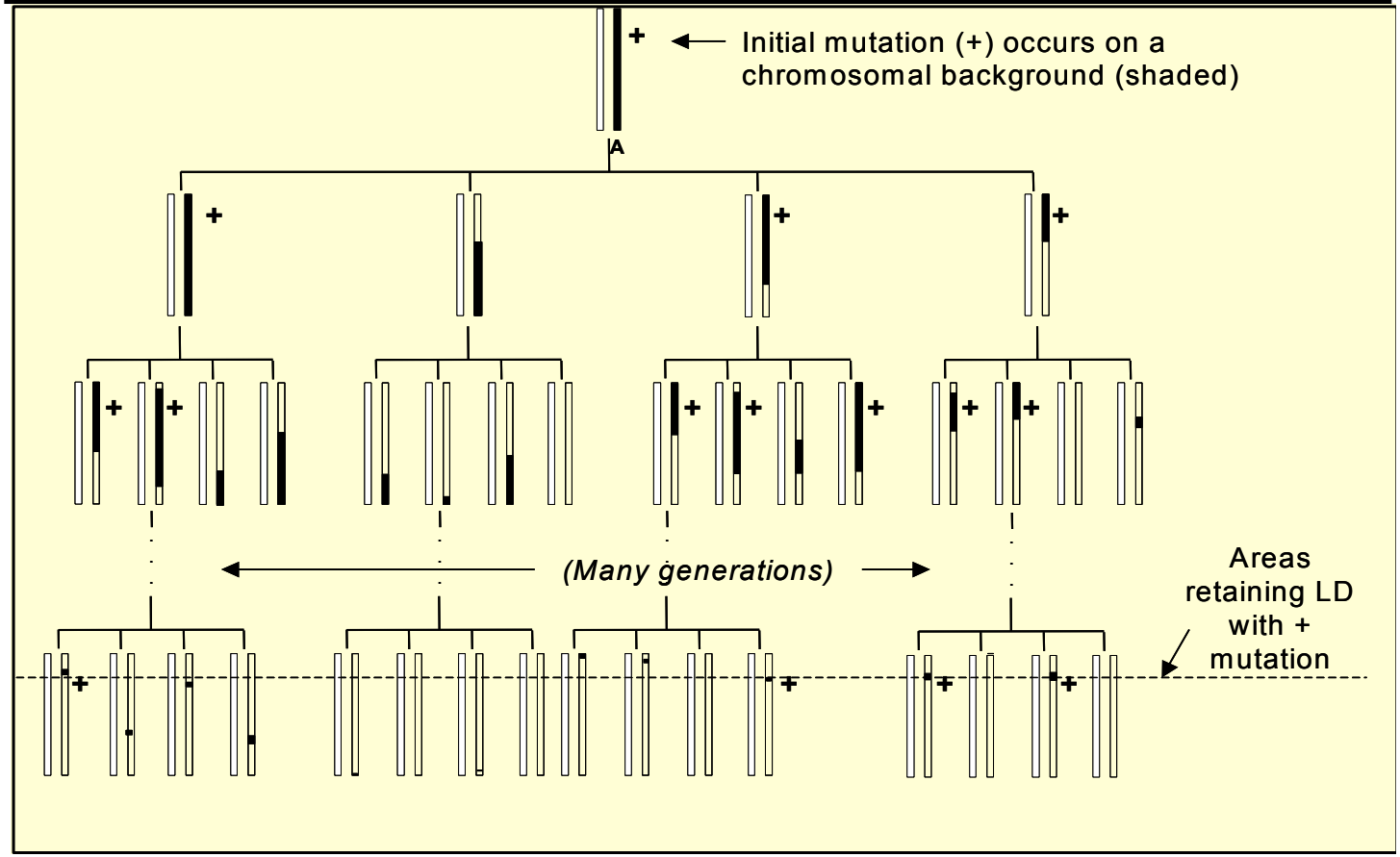
Pedigree unknown

Many meiosis (>10^4)

Resolution: 10^{-5} Morgans (Kbases)

Adapted from McVean and others

Example of Linkage Disequilibrium through generations



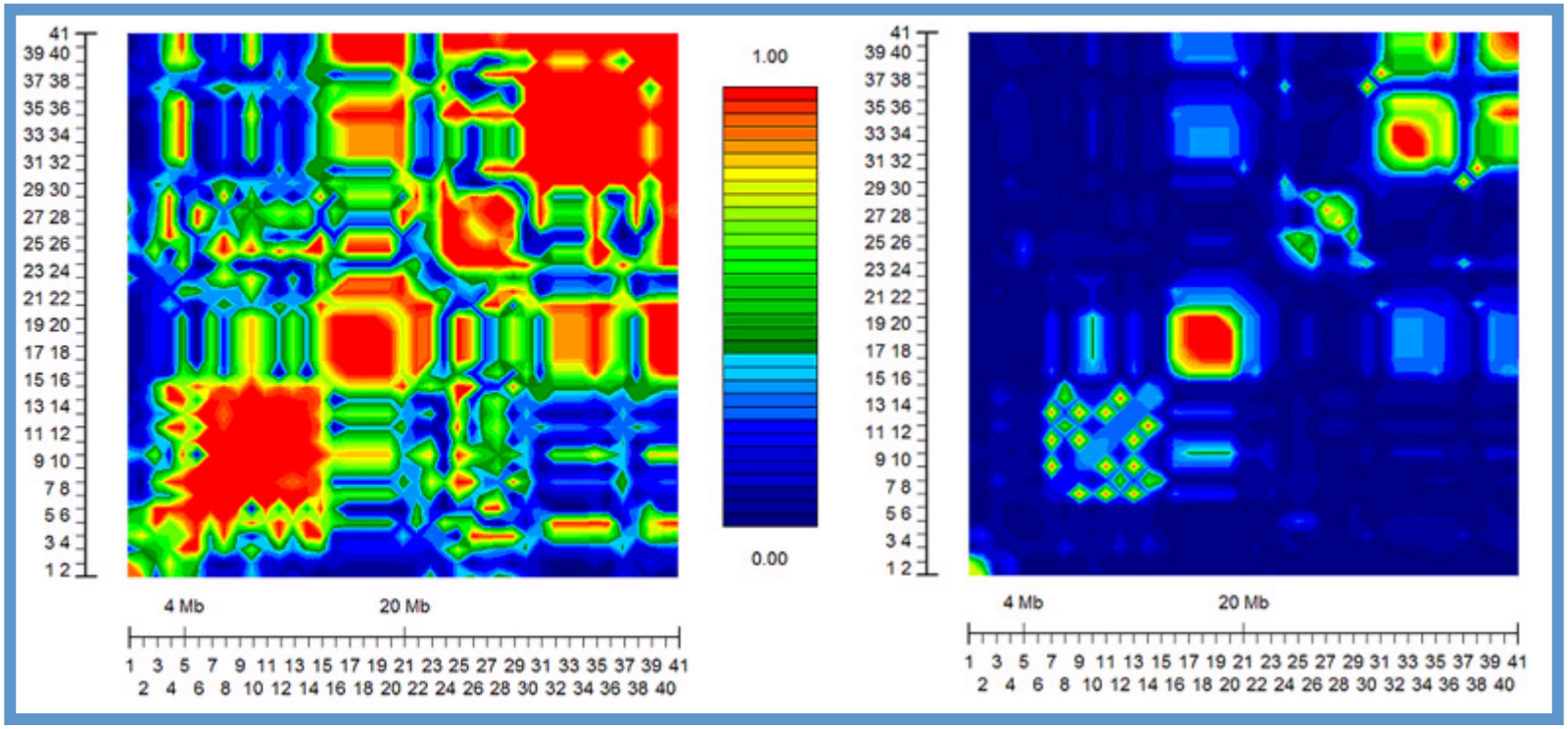
4 maps for gene localization

- Gene localization or gene mapping is based on four maps, each with additive distances.
- Two of these maps are physical:
  - the high-resolution genome map in base pairs (bp)
  - the low-resolution cytogenetic map in chromosome bands of estimated physical lengths.
- The other two maps are purely genetic:
  - the linkage map in Morgans or centimorgans (cM)
  - the map of linkage disequilibrium (LD) in LD units (LDU)

LD map

- Genetic maps in linkage disequilibrium (LD) units play the same role for association mapping as maps in centimorgans provide at much lower resolution for linkage mapping.
- Association mapping of genes determining disease susceptibility and other phenotypes is based on the theory of LD.

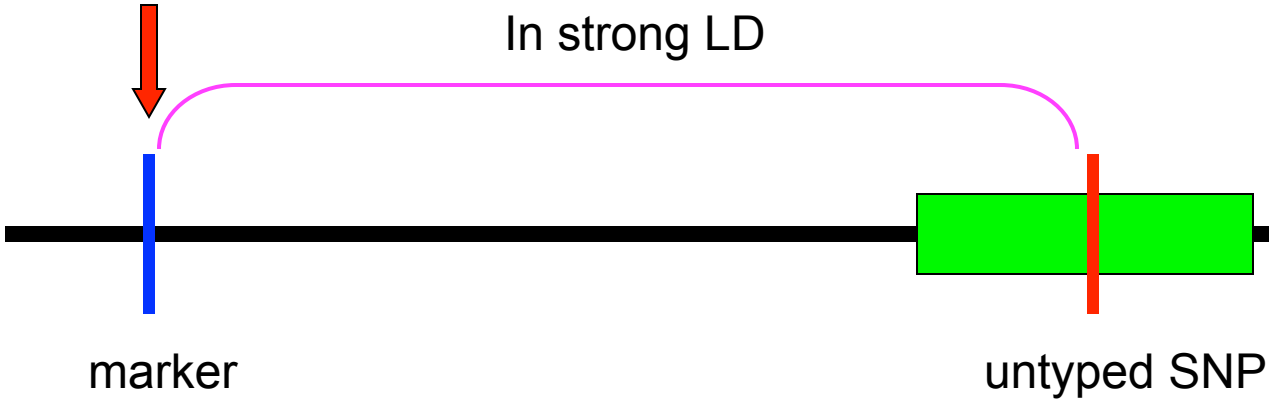
Graphic representation of LD



GOLD

LD based association studies

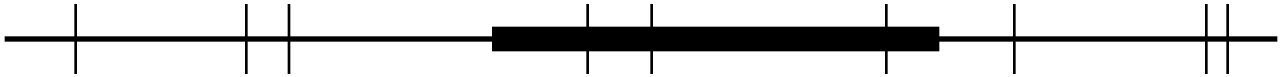
- The paradigm underlying association studies is that linkage disequilibrium can be used to capture associations between markers and nearby untyped SNPs.



Marker Selection for Association Studies

Direct:
Catalog and test all functional variants for association

Indirect:
Use dense SNP map and select based on LD



Collins, Guyer, Chakravarti (1997). Science 278:1580-81

Parameters for SNP Selection

- Allele Frequency
- Putative Function (cSNPs)
- Genomic Context (Unique vs. Repeat)
- Patterns of Linkage Disequilibrium

Association studies

- **Association between risk factor and disease: risk factor is significantly more frequent among affected than among unaffected individuals**

- **In genetic epidemiology:**

  Risk factors **=** alleles**/**genotypes**/**haplotypes

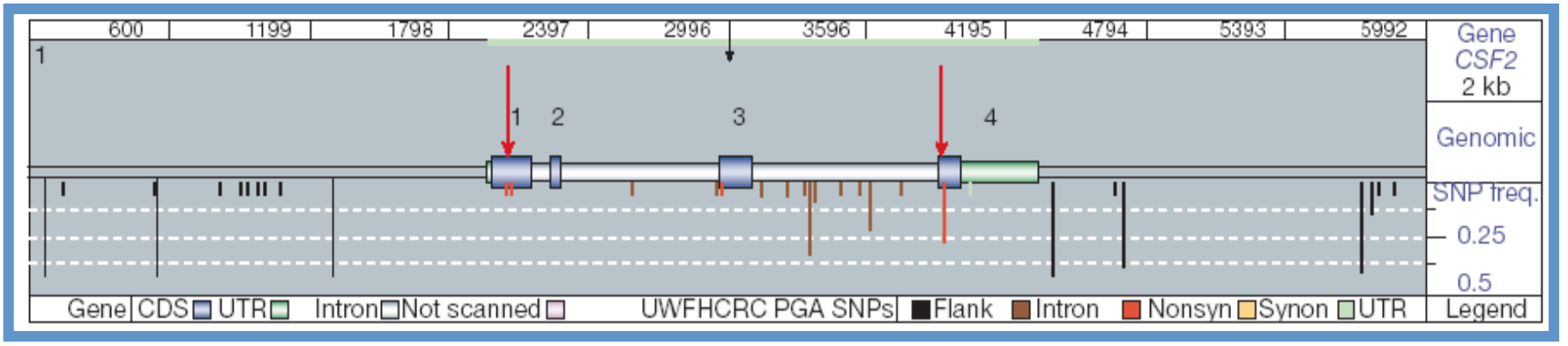
Association studies

- Candidate genes (functional or positional)
- Fine mapping in linkage regions
- Genome wide screen

Candidate gene analysis

- Direct analysis:

  Association studies between disease and functional SNPs (causative of disease) of candidate gene

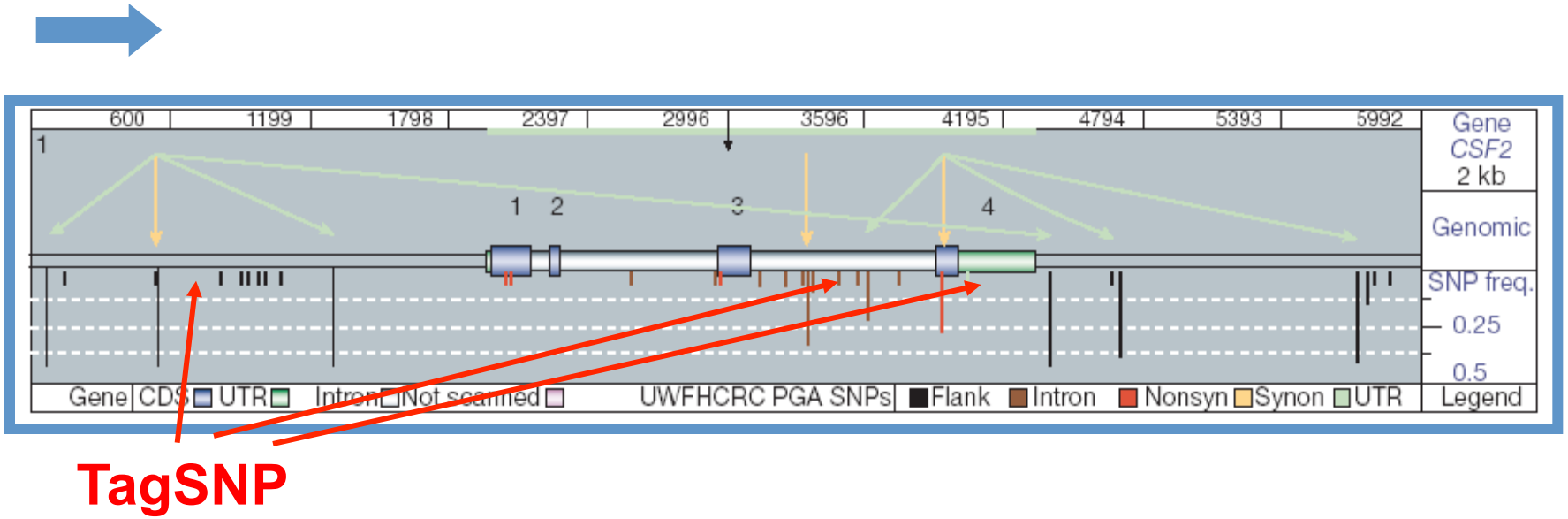


Candidate gene analysis

- Indirect analysis:

  Association studies between disease and "random" SNPs within or near candidate gene

  Linkage Disequilibrium mapping



Case-control studies: χ^2 test

Risk factor

	Yes	No	
Cases	n_{11}	n_{12}	$n_{1.}$
Controls	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

contingency table

Test of independence:

$\chi^2 = \Sigma\ (O\text{-}E)^2 / E$ with 1 df

Case-control studies: χ^2 test

2x3 contingency table

	Genotypes			
	AA	Aa	aa	
Cases	n_{AA}	n_{Aa}	n_{aa}	N
Controls	m_{AA}	m_{Aa}	m_{aa}	M
	t_{AA}	t_{Aa}	t_{aa}	N+M

Test of independence:

$\chi^2 = \Sigma\ (O\text{-}E)^2 / E$ with 2 df

Case-control studies: χ^2 test

2x2 contingency table

	Alleles A	a	
Cases	n_A	n_a	2N
Controls	m_A	m_a	2M
	t_A	t_a	2(N+M)

Test of independence:

$\chi^2 = \Sigma\ (O\text{-}E)^2 / E$ with 1 df

Odds ratio

Exposure	Disease: yes	Disease: no	total
yes	a	b	a + b
no	c	d	c + d
total	a + c	b + d	a + b + c + d

Odds for case: a/c

Odds for control: b/d

Odds ratio

Explanation of OR

- OR>1: exposure factors increase the risk of disease; positive association
- OR<1: exposure factors decrease the risk of disease; negative association
- OR=1: no association

Statistical significance of a correlation versus correlation strength

- Statistical significance is usually measured by “p-value”: the probability for observing the same amount of correlation or more if the true correlation is zero.
- Correlation strength can be measured by many many quantities: D, D’, r2…
- Correlation strength between a marker and the disease status is usually measured by odd-ratio (OR)
- The 95% confidence interval (CI) of OR contains both information on “strength” and “significance”
- When the sample size is increased, typically the p-value can become even more significant, whereas OR usually stays the same (but 95% CI of OR becomes more narrow).

Exploring Candidate Genes: Regression Analysis

- Given
  - Height as “target” or “dependent” variable
  - Sex as “explanatory” or “independent” variable
- Fit regression model

  height = β*sex + ε

Regression Analysis

- Given
  - Quantitative “target” or “dependent” variable y
  - Quantitative or binary “explanatory” or “independent” variables x_i
- Fit regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

Regression Analysis

- Works best for normal y and x
- Fit regression model

  $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$
- Estimate errors on β's
- Use t-statistic to evaluate significance of β's
- Use F-statistic to evaluate model overall

Coding Genotypes

Genotype	Dominant	Additive	Recessive
AA	1	2	1
AG	1	1	0
GG	0	0	0

- Genotype can be re-coded in any number of ways for regression analysis
- Additive ~ codominant

Fitting Models

- Given two models

  $y = \beta_1 x_1 + \varepsilon$

  $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- Which model is better?
- More parameters will always yield a better fit

- Information Criteria
  - Measure of model fit penalized for the number of parameters in model
- AIC (most common)
  - Akaike's Info Criterion
- BIC (more stringent)
  - Bayesian Info Criterion

Single-marker logistic regression

H_0: $\beta_i = 0$

Genetic model interpretations:

- Assume "11" genotype coding represents genotype with lowest absolute risk (baseline)

  $\beta_1 = \beta_2 = 0$ no association with that polymorphism

  $\beta_1 = 0, \beta_2 > 0$ (completely) recessive

  $\beta_1 = \beta_2 > 0$ (completely) dominant

  $0 < \beta_1 < \beta_2$ additive or multiplicative

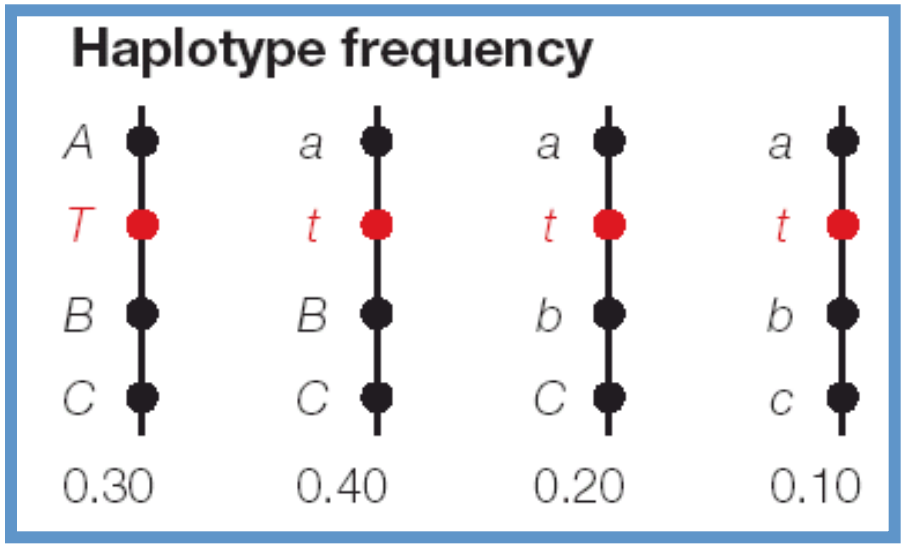
Note: This can be extended through GLM to many types of outcomes (rather than simply odds of disease/not disease, as above)

Tool References

- Haplo.stats (haplotype regression)
  - Lake et al, Hum Hered. 2003;55(1):56-65.
- PHASE (case/control haplotype)
  - Stephens et al, Am J Hum Genet. 2005 Mar;76(3):449-62
- Haplo.view (case/control SNP analysis)
  - Barrett et al, Bioinformatics. 2005 Jan 15;21(2):263-5.
- SNPHAP (haplotype regression?)
  - Sham et al Behav Genet. 2004 Mar;34(2):207-14.

Main Issues in Association Analysis

- The association is typically detected between a non-function marker and the disease, instead of the disease gene itself and the disease status. ("non-direct" role of the disease gene in association analysis)
- When the disease (case) group and the normal (control) group both are a mixture of subpopulations with a different proportion of mixing, even markers not associated with the disease will exhibit spurious association (heterogeneity)



Allele frequency	*D*	$D'_{(\text{marker}, T)}$	$r^2_{(\text{marker}, T)}$	OR_M
A = 0.30	0.21	1.0	1 .0	2.00
T = 0.30				2.00
B = 0.70	0.09	1.0	0.18	1.43
C = 0.90	0.03	1.0	0.05	1.33

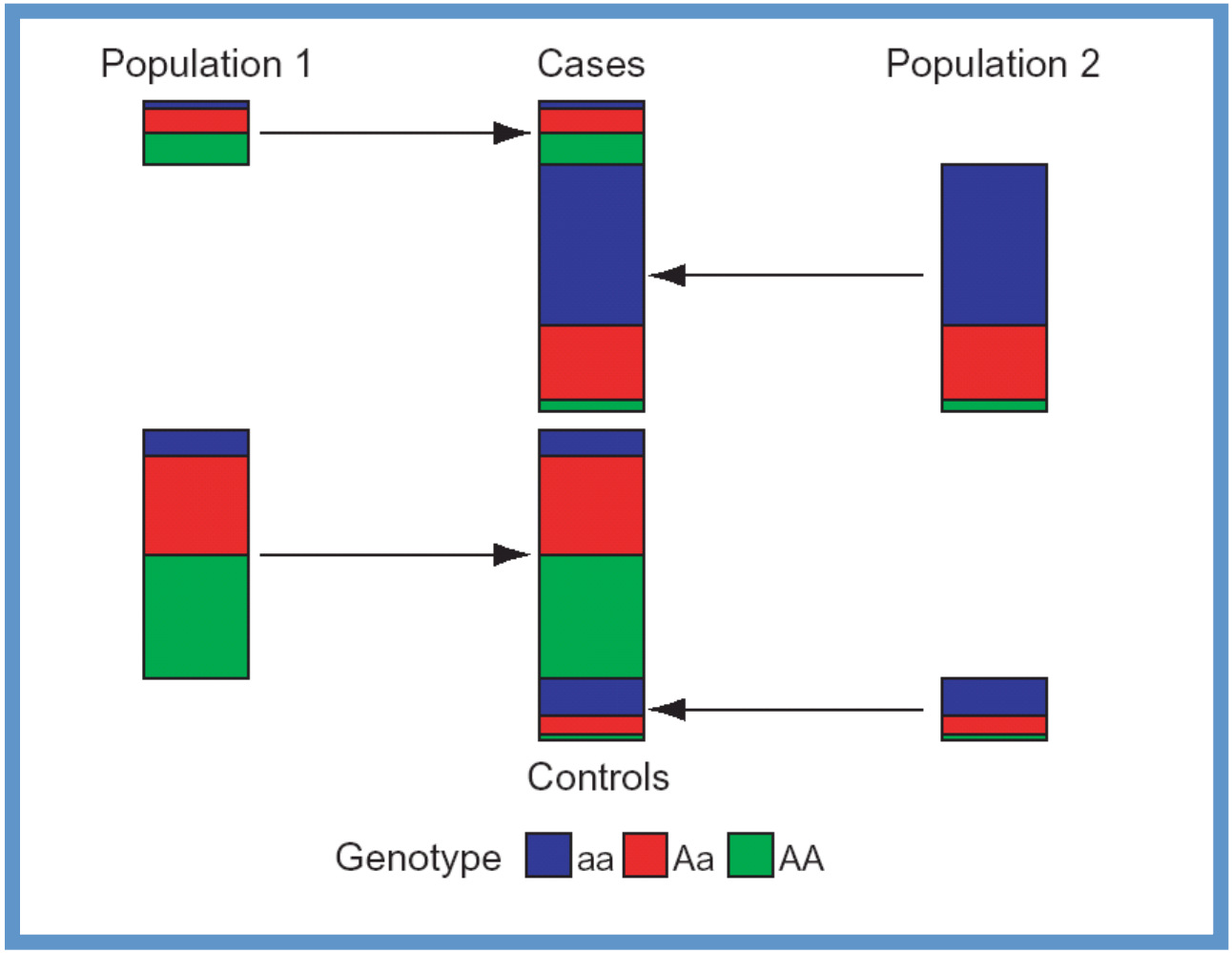
Zondervan & Cardon, 2004

Solution to the first issue

- Choose the marker, haplotype,… to have a matching (allele, haplotype,…) frequency as the disease gene.

- Whenever possible, typing a marker that is also functional (e.g. “coding SNP”, “functional SNP”, “regulatory SNP”)

Association due to population stratification



Marchini et al, 2004

Well-known problem when case/control groups consist of two different subpopulations with different mixing proportion

- Example: comparing people's height between two places: 1. prison, and 2. nurse school
- In prison, maybe 80% are men
- In nursing school, maybe 80% are women
- Men are on average taller than women
- People in prison are taller than people in nurse school

But the cause of this difference is due to the different mixing proportions, not due to "*staying in prison makes people taller*"

Solution to the second issue

- Try to use people from the same population in both case and control group.
- Use neutral marker to test whether subpopulations exist
- If possible use an isolated population (the extra benefit is to reduce the heterogeneity in the case group)
- Use family-based association design (the disadvantage is that it is more costly, and parents of late-onset patients are hard to find)

Staged Study Design

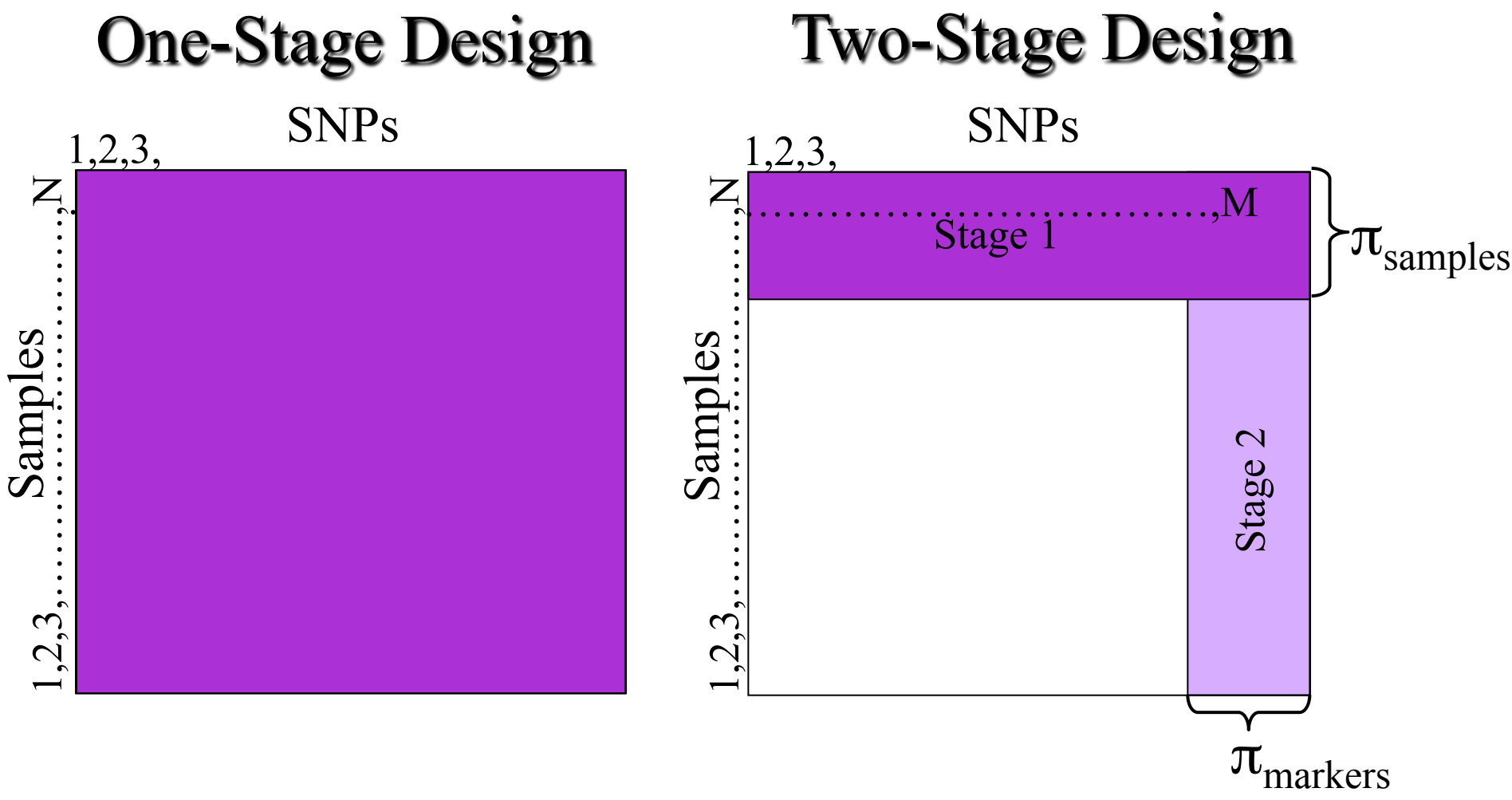
- Given 500,000 SNPs
- Bonferroni corrected significance threshold

  $p = 0.05 / 500000 = 10^{-7}$
- Significance in a single study is difficult to achieve

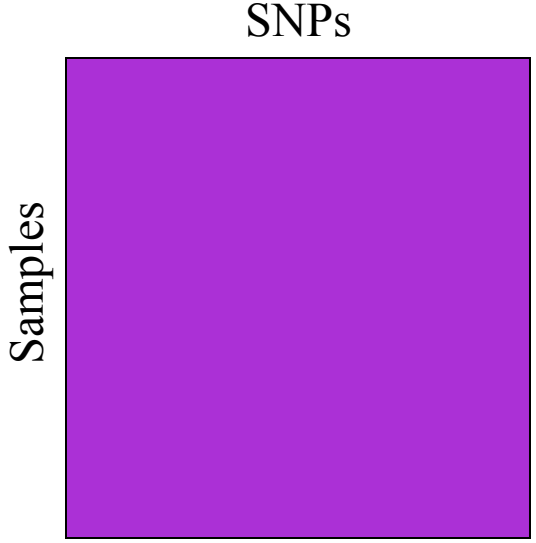
Staged Study Design

- Study I: Genotype 500k SNPs in 1000 cases/ controls
  - Expect 5,000 false positives at $p < 0.01$
- Study II: Genotype best 5000 hits from stage I in additional 1000 cases/controls
  - Expect 50 false positives at $p < 0.01$
- Study 3: Genotype best 50 hits in a third set of 1000 cases/controls
  - Expect 0.5 false positives at $p < 0.01$

One- and Two-Stage GWA Designs

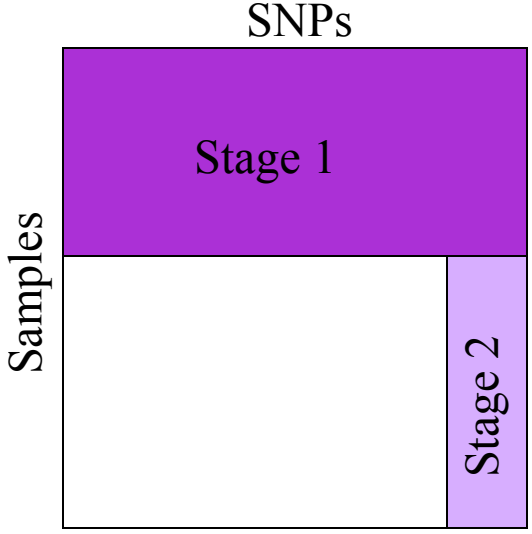


One-Stage Design

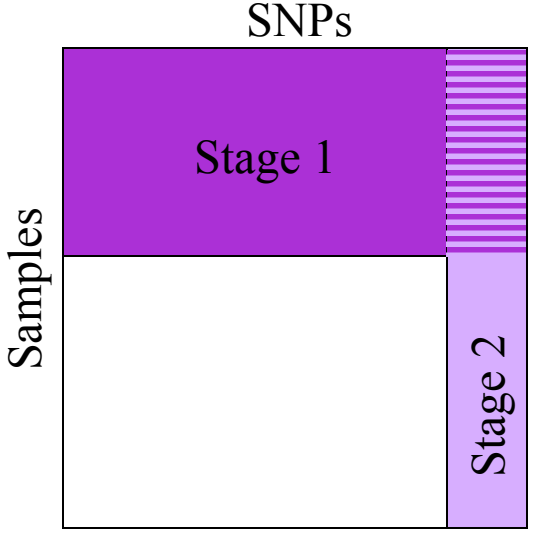


Two-Stage Design

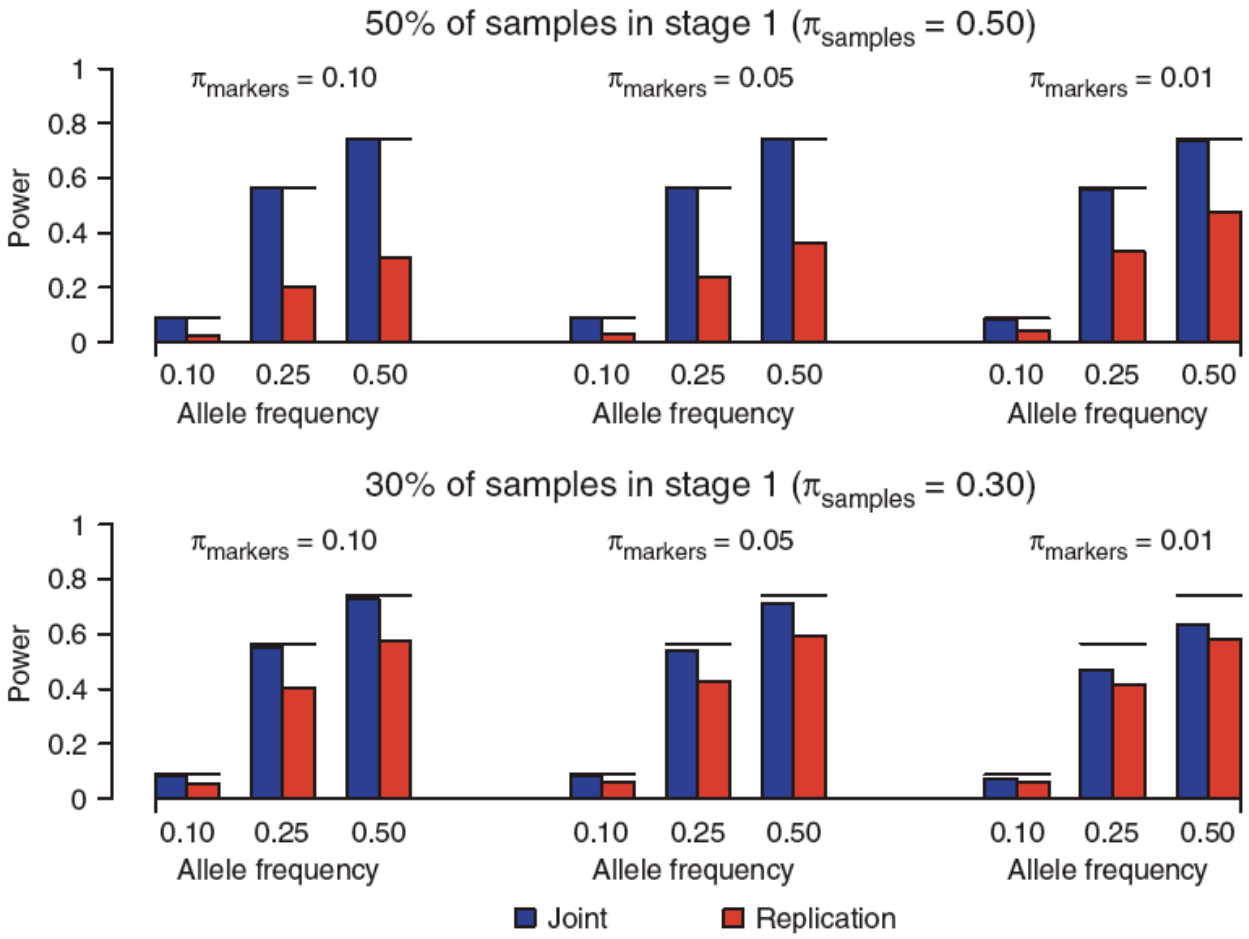
Replication-based analysis



Joint analysis



Joint Analysis



Skol et al, Nat Genet 38: 209-213, 2006

SNPs or Haplotypes

- There is no right answer: explore both
- The only thing that matters is the correlation between the assayed variable and the causal variable
- Sometimes the best assayed variable is a SNP, sometimes a haplotype

Interaction Analysis

- SNP X SNP
- Within gene: haplotype
  - Modest interaction space
  - Most haplotype splits do not matter (APOE)
- Between genes: epistasis
  - Interaction space is vast (500k X 500k)

- SNP X Environment
  - Smaller interaction space (500k X a few environmental measures)

Limiting the Interaction Space

- Not all epistatic interactions make sense
  - Physical interactions (lock and key)
  - Physical interactions (subunit stoichiometry)
  - Pathway interactions
  - Regulatory interactions

Conclusions

- Pay attention to study design
  - Sample size
  - Estimated power
  - Multiple Testing
- Analyze SNPs (and haplotypes)
- Keep population structure in mind
- Explore epistasis and environmental interactions after main effects

Genetic studies of complex diseases have not met anticipated success

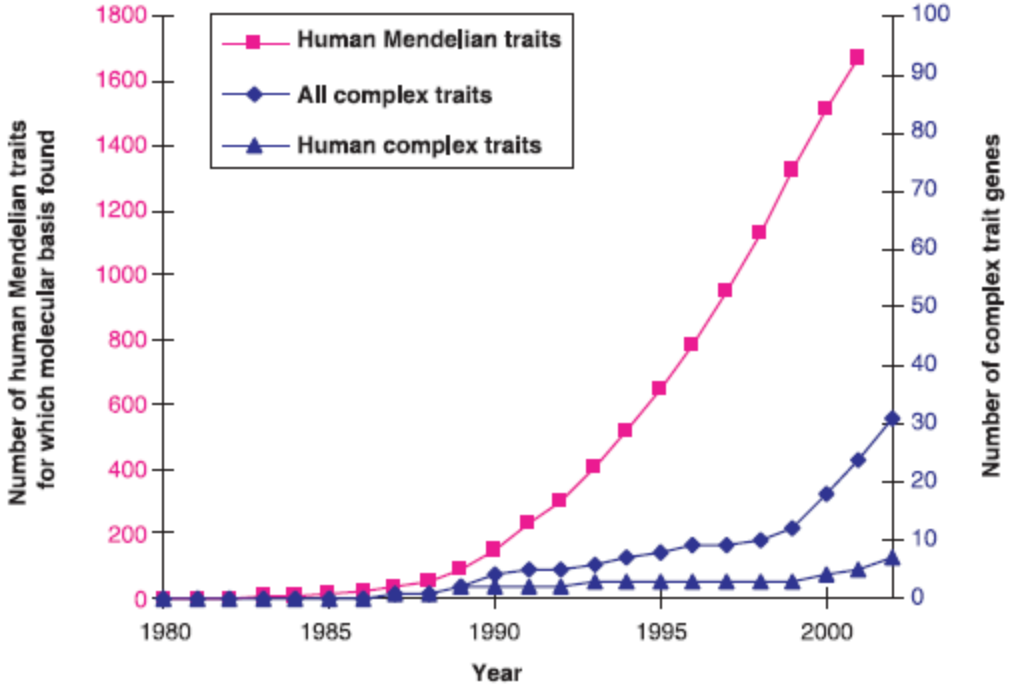


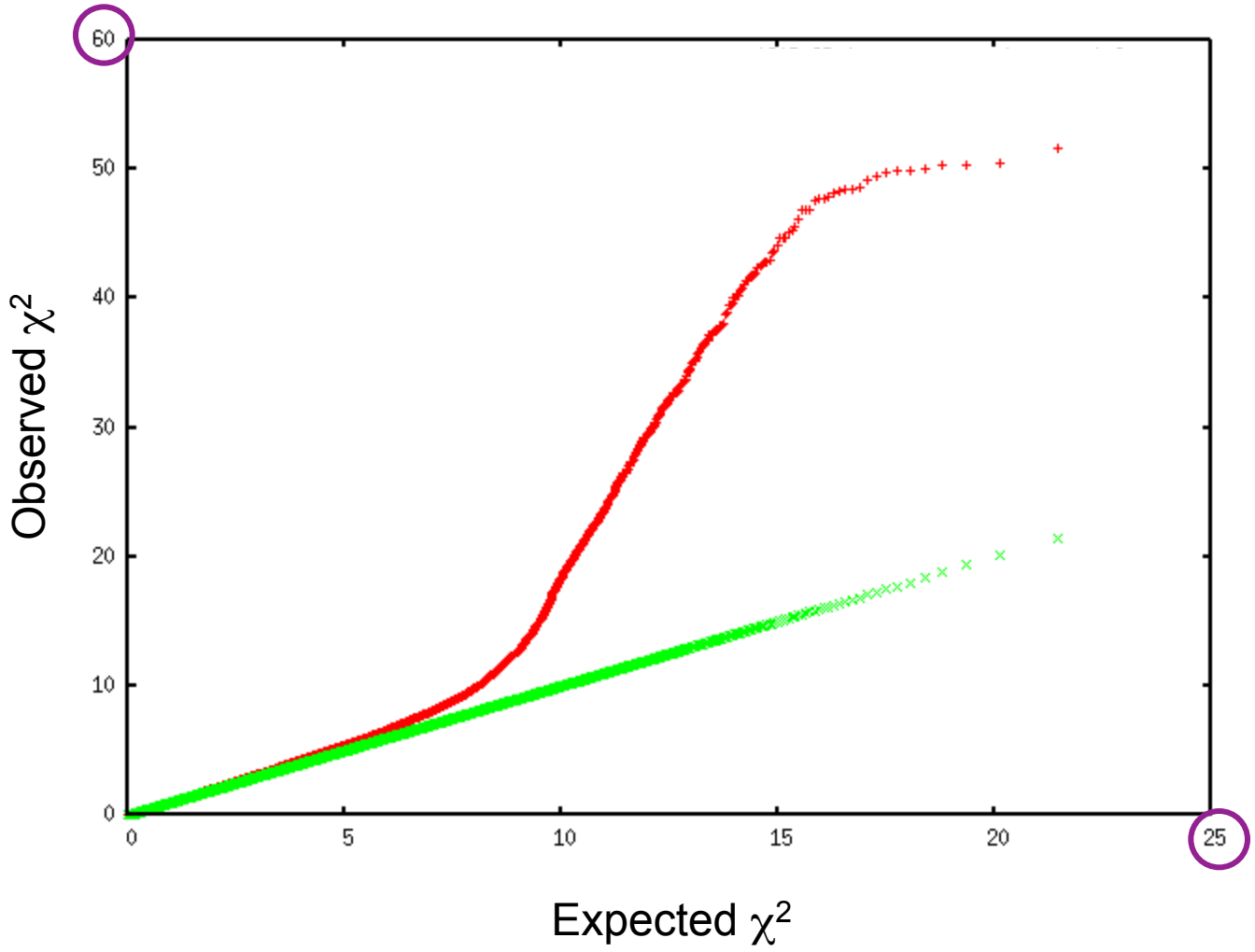
Fig. 1. Identification of genes underlying human Mendelian traits and genetically complex traits in humans and other species. Cumulative data for human Mendelian trait genes (to 2001) include all major genes causing a Mendelian disorder in which causal variants have been identified (*58*, *59*). This reflects mutations in a total of 1336 genes. Complex trait genes were identified by the whole-genome screen approach and denote cumulative year-on-year data described in this review.

Glazier et al, Science (2002) 298:2345-2349

What effect does this have on trait association?

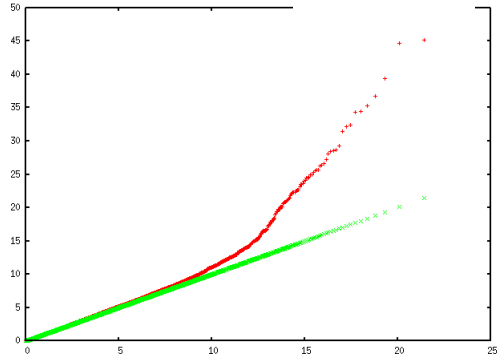
- Following data
  - Affymetrix data
  - Single locus tests
  - $\geq$ 500 cases/500 controls
  - Key issue
    - Genotype calling: batch effects, differential call rates, QC
    - e.g. Clayton et al, Nat Genet 2005

Whole Genome Association
What answer do you want?

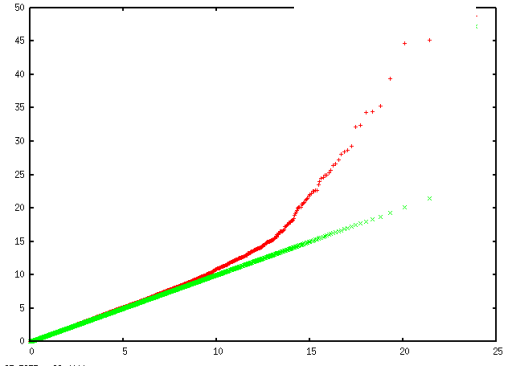


Cleaning Affymetrix Data
Batch Effects and Genotype Calling

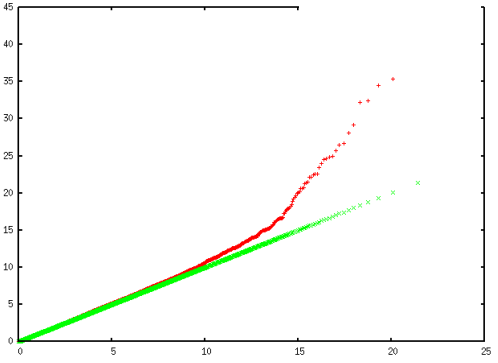
≤ 10% missing



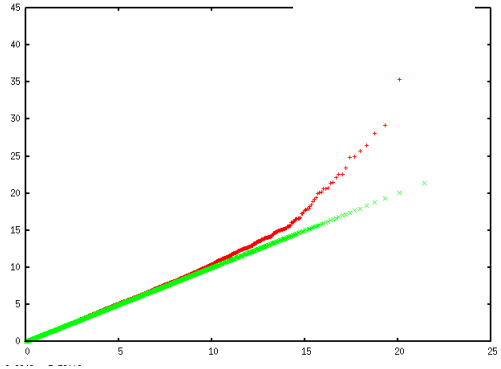
≤ 9% missing



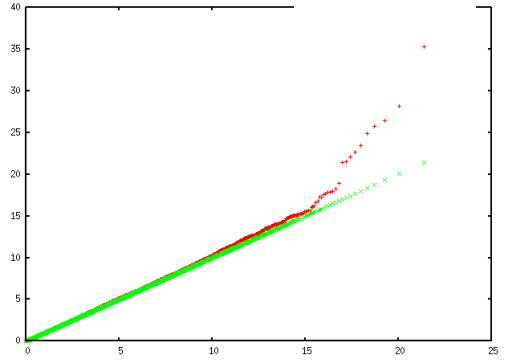
≤ 8% missing



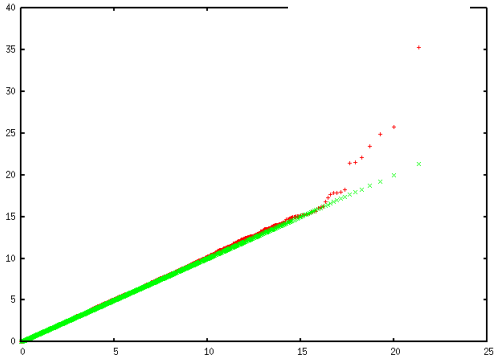
≤ 7% missing



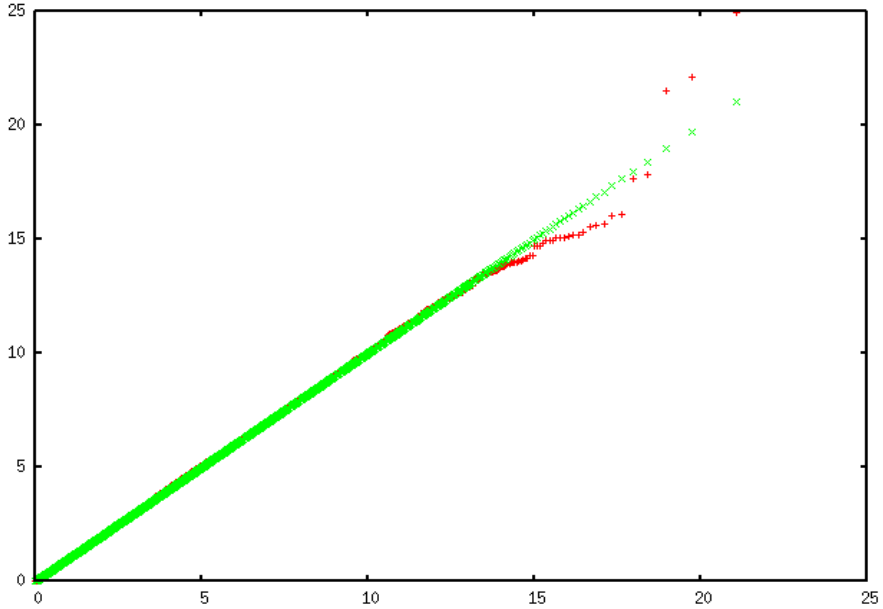
≤ 6% missing



≤ 5% missing

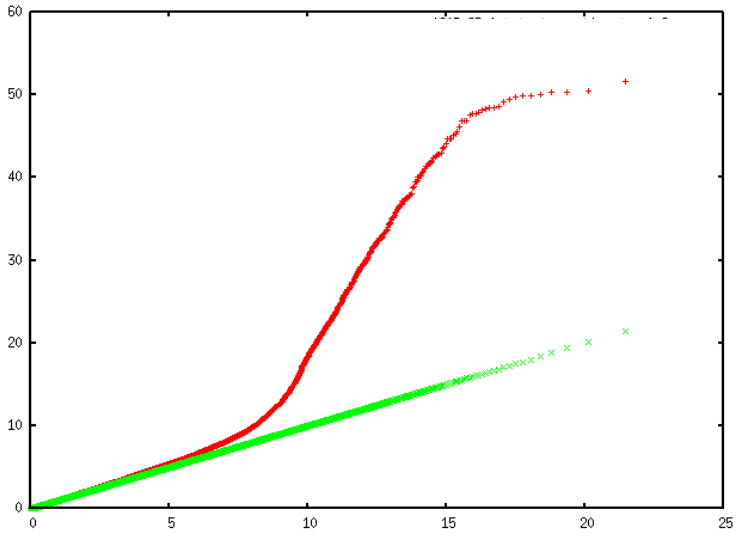


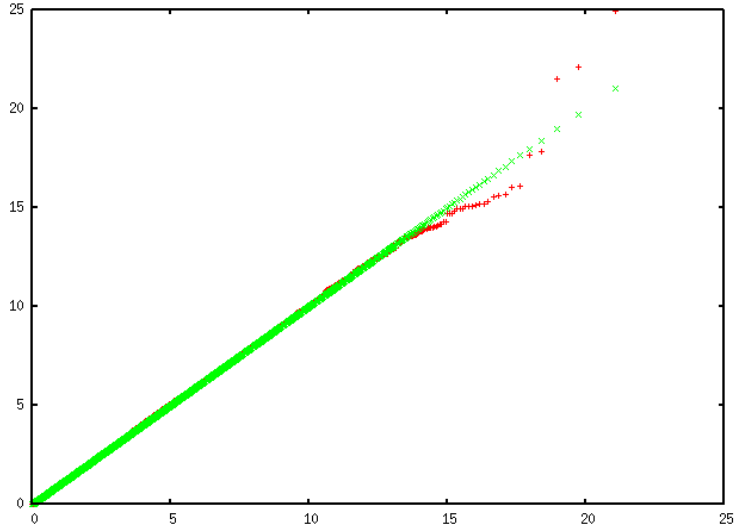
Affymetrix Data – Too Clean?



- As much as 20-30% data eliminated -- including real effects --
- Many ‘significant’ results can be data errors
  - *‘Low Hanging Fruit’ sometimes rotten*
- Real effects may not be the most highly significant (power)

Too Many or Too Few?





- Inappropriate genotype calling, study design can mask real effects or make GWA look too good
- How to address this?
    - Multiple controls (e.g., WTCCC)
    - Multiple/better calling algorithms (e.g. Affymetrix)
    - Examination of individual genotypes (manual)

Current Association Study Challenges

2) Do we have the best set of genetic markers

Table 1 | **Priorities for single-nucleotide-polymorphism selection**

Type of variant	Location	Functional effect	Frequency in genome
Nonsense	Coding sequence	Premature termination of amino-acid sequence	Very low
Missense/ non-synonymous (non-conservative)	Coding sequence	Changes an amino acid in protein to one with different properties	Low
Missense/ non-synonymous (conservative)	Coding sequence	Changes an amino acid in protein to one with similar properties	Low
Insertions/deletions (frameshift)	Coding sequence	Changes the frame of the protein-coding region, usually with very negative consequences for the protein	Low
Insertions/deletions (in frame)	Coding or non-coding	Changes amino-acid sequence	Low
Sense/synonymous	Coding sequence	Does not change the amino acid in the protein — but can alter splicing	Medium
Promoter/regulatory region	Promoter, 5′ UTR, 3′ UTR	Does not change the amino acid, but can affect the level, location or timing of gene expression	Low to medium
Splice site/intron–exon boundary	Within 10 bp of the exon	Might change the splicing pattern or efficiency of introns	Low
Intronic	Deep within introns	No known function, but might affect expression or mRNA stability	Medium
Intergenic	Non-coding regions between genes	No known function, but might affect expression through enhancer or other mechanisms	High

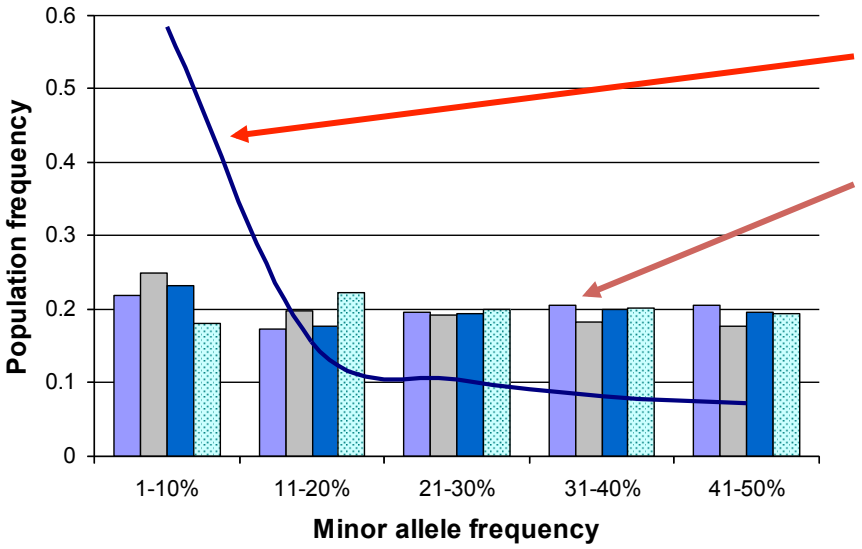
Tabor et al, Nat Rev Genet 2003

Current Association Study Challenges

2) Do we have the best set of genetic markers

There exist 6 million putative SNPs in the public domain. Are they the right markers?

Allele frequency distribution is biased toward common alleles



Expected frequency in population

Frequency of public markers

Current Association Study Challenges

3) How to analyse the data

- Allele based test?
  - 2 alleles → 1 df
    - $E(Y) = a + bX$ X = 0/1 for presence/absence
- Genotype-based test?
  - 3 genotypes → 2 df
    - $E(Y) = a + b_1A + b_2D$ A = 0/1 additive (hom); W = 0/1 dom (het)
- Haplotype-based test?
  - For M markers, 2^M possible haplotypes → 2^M -1 df
    - $E(Y) = a + \Sigma bH$ H coded for haplotype effects
- Multilocus test?
  - Epistasis, G x E interactions, many possibilities

Current Association Study Challenges

4) Multiple Testing

- Candidate genes: a few tests (probably correlated)
- Linkage regions: 100's – 1000's tests (some correlated)
- Whole genome association: 100,000s – 1,000,000s tests (many correlated)
- What to do?
  - Bonferroni (conservative)
  - False discovery rate?
  - Permutations?

  ….Area of active research

Multiple testing problem in whole-genome scan studies

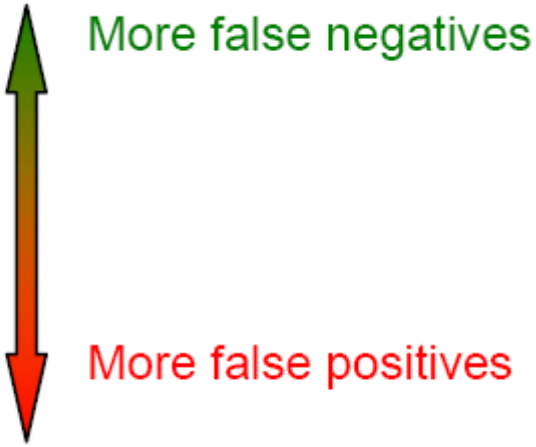
- Affymetrix 500K, Illumina 650K…
- multiple testing
  - If you have 10,000 genes in your genome, and perform a statistical analysis, a p-value cutoff of 0.05 allows a 5% chance of error. That means that 500 genes out of 10,000 could be found to be significant by chance alone.

Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

Multiple testing correct methods

- Bonferroni correction
- Bonferroni Step-down (Holm) correction
- Westfall and Young Permutation
- Benjamini and Hochberg False Discovery Rate

Bonferroni
Bonferroni Step-Down
Westfall and Young Permutation
Benjamini and Hochberg False Discovery Rate
None



Bonferroni correction

- The p-value of each gene is multiplied by the number of genes in the gene list. If the corrected p-value is still below the error rate, the gene will be significant: Corrected P-value= p-value * n (number of genes in test) <0.05
- As a consequence, if testing 1000 genes at a time, the highest accepted individual p-value is 0.00005, making the correction very stringent.
- The expected number of false positives will be 0.05.

Bonferroni Step-down (Holm) correction

- This correction is very similar to the Bonferroni, but a little less stringent:
- 1) The p-value of each gene is ranked from the smallest to the largest.
- 2) The first p-value is multiplied by the number of genes present in the gene list; if the end value is less than 0.05, the gene is significant; Corrected P-value= p-value * n < 0.05
- 3) The second p-value is multiplied by the number of genes less 1. Corrected P-value= p-value * n-1 < 0.05
- 4) The third p-value is multiplied by the number of genes less 2. Corrected P-value= p-value * n-2 < 0.05
- It follows that sequence until no gene is found to be significant.

Example:

Let n=1000, error rate=0.05

Gene name	p-value before correction	Rank	Correction	Is gene significant after correction?
A	0.00002	1	0.00002 * 1000=0.02	0.02<0.05 => Yes
B	0.00004	2	0.00004*999=0.039	0.039<0.05 => Yes
C	0.00009	3	0.00009*998=0.0898	0.0898>0.05 => No

Westfall and Young Permutation

- The Westfall and Young permutation follows a step-down procedure similar to the Holm method, combined with a bootstrapping method to compute the p-value distribution:
- 1) P-values are calculated for each gene based on the original data set and ranked.
- 2) The permutation method creates a pseudo-data set by dividing the data into artificial treatment and control groups.
- 3) P-values for all genes are computed on the pseudo-data set.
- 4) The successive minima of the new p-values are retained and compared to the original ones.
- 5) This process is repeated a large number of times, and the roportion of resampled data sets where the minimum pseudo-p-value is less than the original p-value is the adjusted p-value.
- Because of the permutations, the method is very slow.

Benjamini and Hochberg False Discovery Rate

- This correction is the least stringent of all 4 options, and therefore tolerates more false positives. There will be also less false negative genes. Here is how it works:
- 1) The p-values of each gene are ranked from the smallest to the largest.
- 2) The largest p-value remains as it is.
- 3) The second largest p-value is multiplied by the total number of genes in gene list divided by its rank. If less than 0.05, it is significant. Corrected p-value = p-value*(n/n-1) < 0.05, if so, gene is significant.
- 4) The third p-value is multiplied as in step 3: Corrected p-value = p-value*(n/n-2) < 0.05, if so, gene is significant.

Example:

Let n=1000, error rate=0.05

Gene name	p-value (from largest to smallest)	Rank	Correction	Is gene significant after correction?
A	0.1	1000	No correction	0.1 > 0.05 => No
B	0.06	999	1000/999*0.06 = 0.06006	0.06006 > 0.05 => No
C	0.04	998…	1000/998*0.04 = 0.04008	0.04008 < 0.05 => Yes

Current Association Study Challenges

5) Population Stratification

- Analysis of mixed samples having different allele frequencies is a primary concern in human genetics, as it leads to false evidence for allelic association.

- This is the main blame for past failures of association studies

Population Stratification

Sample 'A'

	M	m	Freq.
Affected	50	50	.10
Unaffected	450	450	.90
	.50	.50	

χ^2_1 is n.s.

+

Sample 'B'

	M	m	Freq.
Affected	1	9	.01
Unaffected	99	891	.99
	.10	.90	

χ^2_1 is n.s.

	M	m	Freq.
Affected	51	59	.055
Unaffected	549	1341	.945
	.30	.70	

$$\chi^2_1 = 14.84, p < 0.001$$

Spurious Association

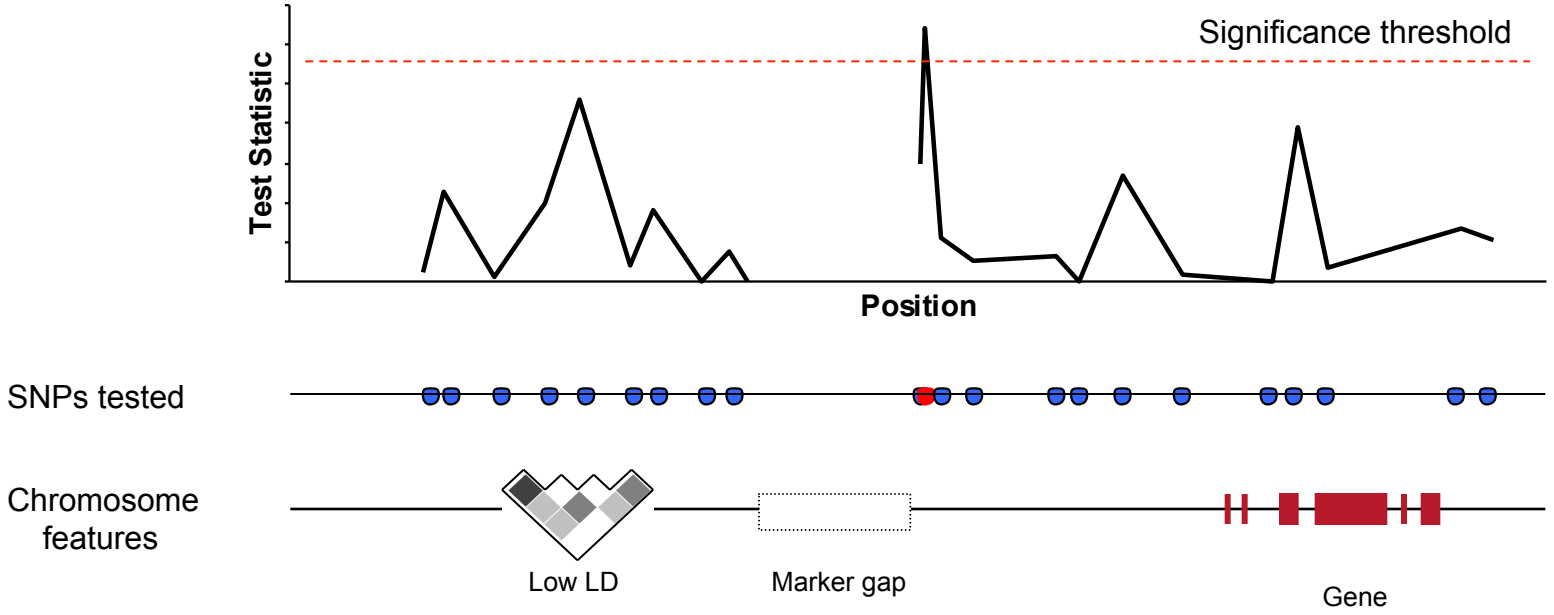
Current Association Study Challenges

6) What constitutes a replication?

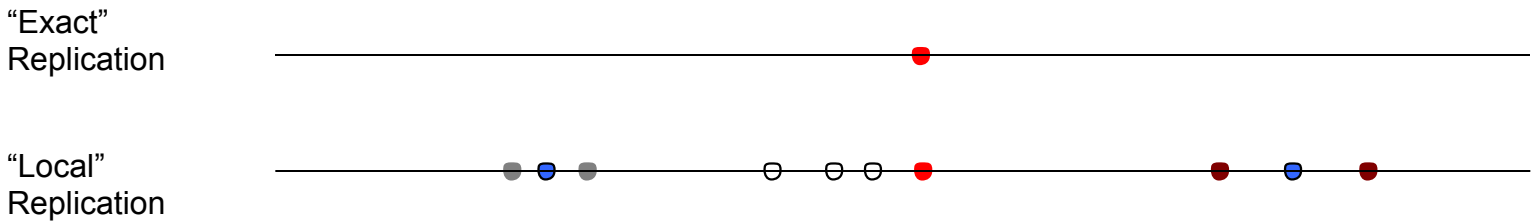
- *GOLD Standard for association studies*
- *Replicating association results in different laboratories is often seen as most compelling piece of evidence for 'true' finding*
- But…. in any sample, we measure
  - Multiple traits
  - Multiple genes
  - Multiple markers in genes
- and we analyse all this using multiple statistical tests

What is a true replication?

Initial Study



Replication Strategy



What is a true replication?

Replication Outcome	Explanation
• Association to same trait, but different gene	• Genetic heterogeneity
• Association to same trait, same gene, different SNPs (or haplotypes)	• Allelic heterogeneity
• Association to same trait, same gene, same SNP – but in opposite direction (protective ←→ disease)	• Allelic heterogeneity/ popln differences
• Association to different, but correlated phenotype(s)	• Phenotypic heterogeneity
• No association at all	• Sample size too small

Measuring Success by Replication

- Define objective criteria for what is/is not a replication *in advance*

- Design initial and replication study to have enough power
  - 'Lumper': use most samples to obtain robust results in first place
    - Great initial detection, may be weak in replication
  - 'Splitter': Take otherwise large sample, split into initial and replication groups
    - One good study → two bad studies.
    - Poor initial detection, poor replication

Despite challenges: upcoming association studies hold promise

- Large, epidemiological-sized samples emerging
- Availability of millions of genetic markers
  - Genotyping costs decreasing rapidly
- Background LD patterns characterized
  - International HapMap and other projects

2008: GWAS of lung cancer

nature genetics

Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1

Christopher I Amos[1], Xifeng Wu[1], Peter Broderick[2], Ivan P Gorlov[1], Jian Gu[1], Timothy Eisen[3], Qiong Dong[1],

Vol 452|3 April 2008|doi:10.1038/nature06885 nature

LETTERS

A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25

Ravjean J. Hung[1,2]*, James D. McKay[1]*, Valerie Gaborieau[1], Paolo Boffetta[1], Mia Hashibe[1], David Zaridze[3],

nature Vol 452|3 April 2008|doi:10.1038/nature06846

LETTERS

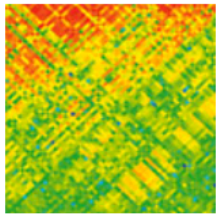
A variant associated with nicotine dependence, lung cancer and peripheral arterial disease

Thorgeir E. Thorgeirsson[1]*, Frank Geller[1]*, Patrick Sulem[1]*, Thorunn Rafnar[1]*, Anna Wiste[1,2],

- Illumina HumanHap300 Beadchip (>300k SNPs)
- European population large sample size(>1000pairs)
- multiple subsequent studies to confirm the initial findings
- identify an association between SNP variation at 15q24/25.1 and lung cancer risk (Nicotinic acetylcholine receptor subunit gene)
- **differ** on whether the link is direct or mediated through nicotine dependence (duration &dose of smoking)

http://www.nature.com/nrg/series/gwas/index.html

Genome-wide association studies



In the past 2 years genome-wide association studies in humans have revealed dozens of disease-associated loci and have provided insights into the allelic architecture of complex traits. Along the way, much has been learned about how best to carry out such studies. The articles in this series examine these design issues and the technical challenges that remain; for example, identifying association signals and interpreting the molecular mechanisms by which they exert their biological functions.

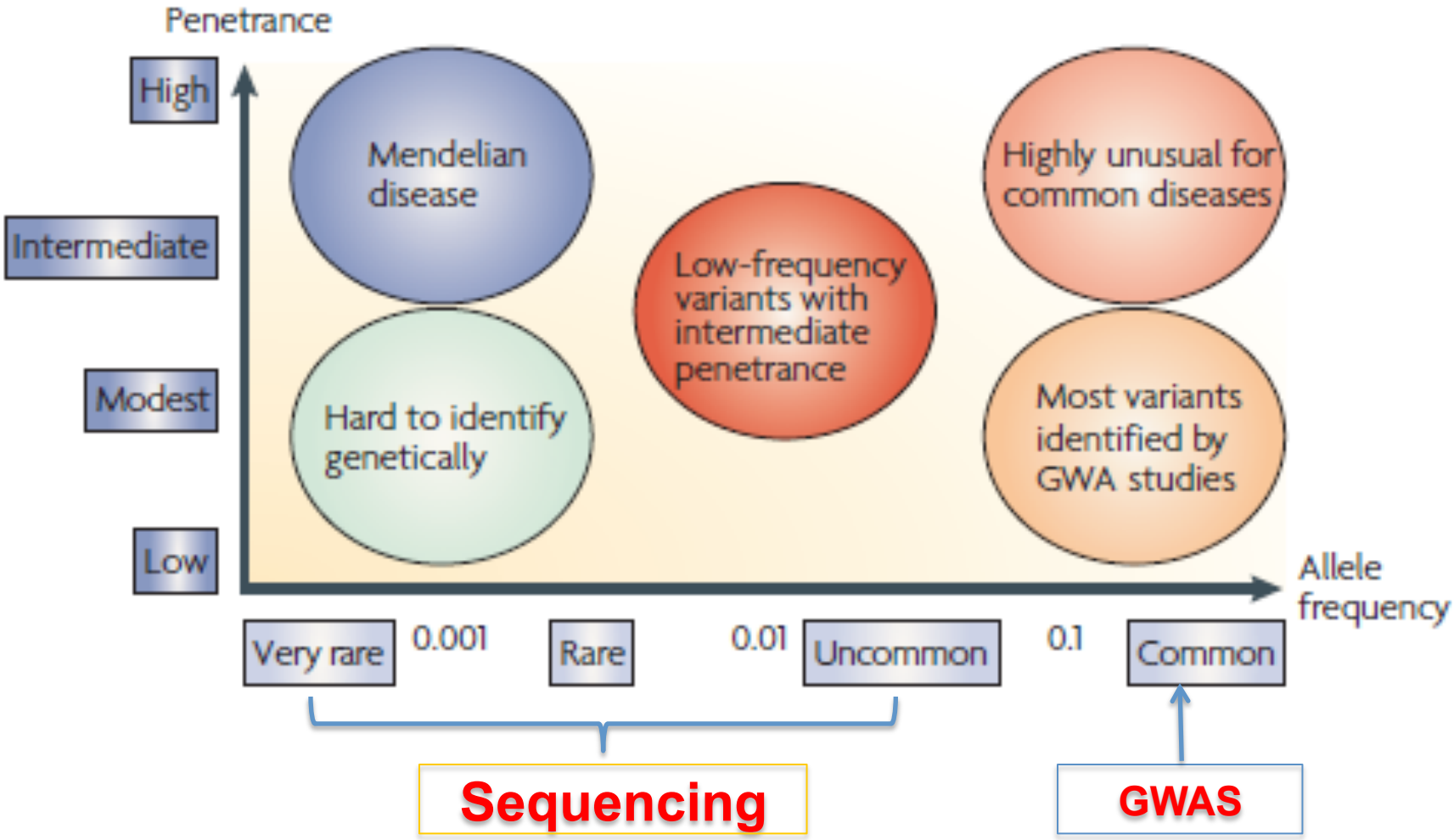
INDEX

2010
May | Apr | Feb

Issues in GWAS

- Common variants vs rare variants
- Experiment vs Imputation
- Sequencing vs genotyping
- Haplotype vs single locus analsis
- Pathway analysis vs single gene analysis
- Gene-gene interaction vs single locus analysis
- GWAS vs candidate genes AS
- Tagging markers vs functional variants
- Environmental factors (gene-environment interaction)
- Population stratification/structure
- Multiple testing problem
- Data sharing and meta-analysis

Genetic Spectrum of Complex Diseases



McCarthy MI et al. Nat Rev Genet. 2008

Association analysis based on rare variants

Several Approaches to Study Rare Variants

- **Deep whole genome sequencing**
  - Can only be applied to limited numbers of samples
  - Most complete ascertainment of variation

- **Exome capture and targeted sequencing**
  - Can be applied to moderate numbers of samples
  - SNPs and indels in the most interesting 1% of the genome

- **Low coverage whole genome sequencing**
  - Can be applied to moderate numbers of samples
  - Very complete ascertainment of shared variation

- **New Genotyping Arrays and/or Genotype Imputation**
  - Examine low frequency coding variants in 100,000s of samples
  - Current catalogs include 97-98% of sites detectable by sequencing an individual

Single SNP Test for Rare Variant

- Rare variants are hard to detect
- Power/sample size depends on both frequency and effect size
- Rare causal SNPs are hard to identify even with large effect size

Single SNP Test for Rare Variant

- Disease prevalence ~10%
- Type I error $5x10^{-6}$
- To achieve 80% power
- Equal number of cases and controls

- Minor Allele Frequency (MAF) = 0.1, 0.01, 0.001
- Required sample size = 486, 3545, 34322,

Alternatives to Single Variant Test Collapsing Method (Burden Test)

- Group rare variants in the same gene/region
- Score each individual
  - Presence or absence of rare copy
  - Weight each variant

$$X_j = \begin{cases} 1 & \text{rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

- Use individual score as a new “genotype”
- Test in a regression framework

Li and Leal (2008) *Am J Hum Genet* **83**:311-321

Challenges

- Disease is caused by multiple rare variants in an additive manner
- It is hard to separate causal and null SNPs
  - Including all rare variants will dilute the true signals
- The effect size of each rare variant varies

Power of Burden Test

	Single Variant Test	Combined Test
10 variants / all have risk 2 / All have frequency .005	.05	.86
10 variants / all have risk 2 / Unequal Frequencies	.20	.85
10 variants / average risk is 2, but varies / frequency .005	.11	.97

- Power tabulated in collections of simulated data
- Combining variants can greatly increase power
- Currently, appropriately combining variants is expected to be key feature of rare variant studies.

Impact of Null Variants

	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants + 5 null variants	.04	.70
10 disease associated variants + 10 null variants	.03	.55
10 disease associated variants + 20 null variants	.03	.33

- Including non-disease variants reduces power
- Power loss is manageable, combined test remains preferable to single marker tests

Impact of Missing Disease Alleles

	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants, 2 missed	.05	.72
10 disease associated variants , 4 missed	.05	.52
10 disease associated variants , 6 missed	.04	.28
10 disease associated variants, 8 missed	.03	.08

- Missing disease alleles loses power
- Still better than single variant test

Sequence Kernel Association Test (SKAT)

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \boldsymbol{\beta}'\mathbf{G}_i + \varepsilon_i, \qquad \text{logit}\, P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \boldsymbol{\beta}'\mathbf{G}_i,$$

1. y_i: quantitative or binary phenotypes;
2. $\boldsymbol{\alpha}'\boldsymbol{X}_i$: fixed effects of covariates;
3. $\boldsymbol{\beta}'\boldsymbol{G}_i$: genetic effects from one gene consisted of SNPs;
4. ε_i: random error.

$$\text{Assume } \beta \sim N(0, \tau W), H_0 : \beta = 0 \rightarrow \boldsymbol{H_0 : \tau = 0}$$

$$\varepsilon \sim \left(0, \sigma_E^2 I\right)$$

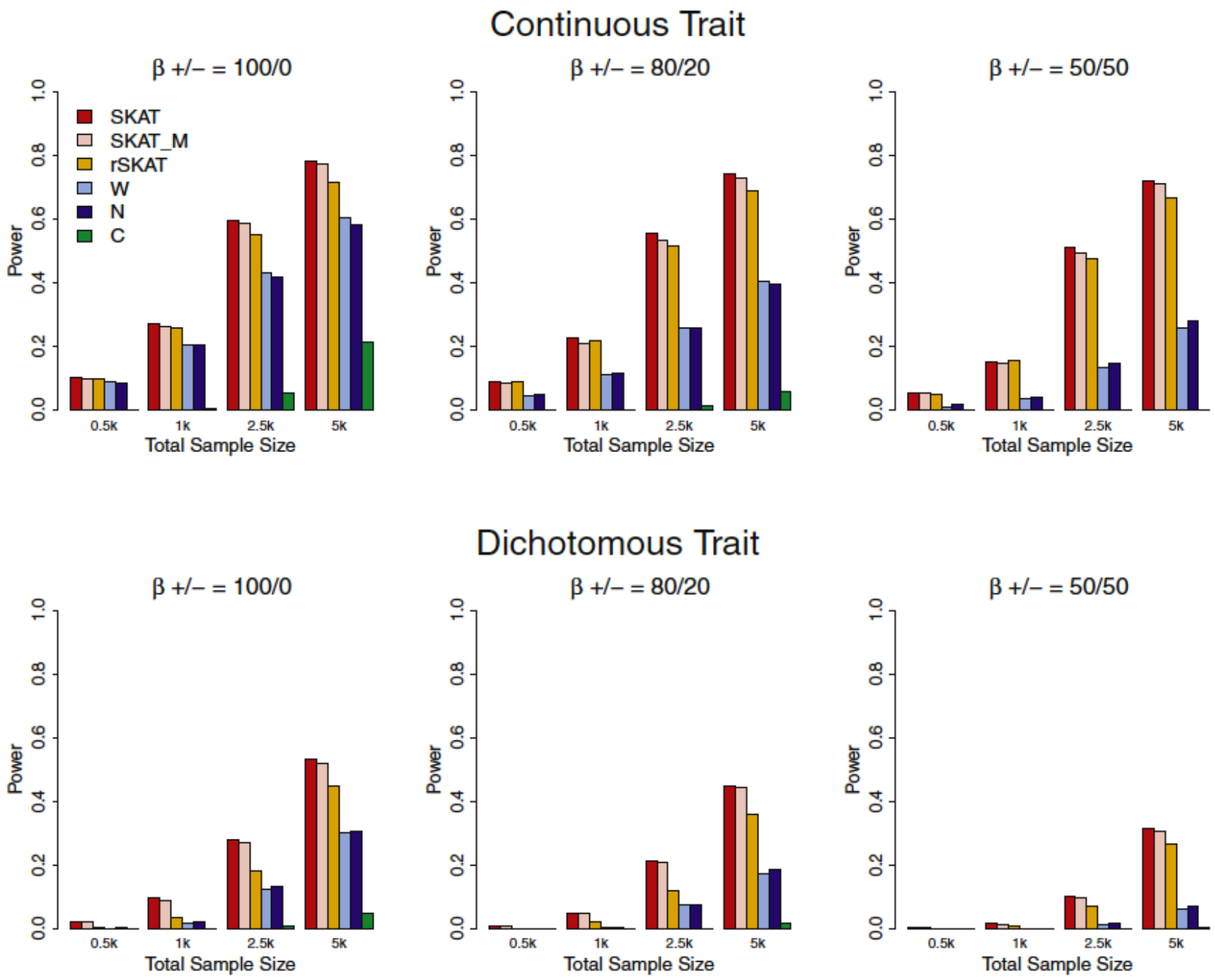
Sequence Kernel Association Test (SKAT)

- Regression based method

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \boldsymbol{\beta}'\mathbf{G}_i + \varepsilon_i, \qquad \text{logit}\, P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \boldsymbol{\beta}'\mathbf{G}_i,$$

- Score statistic $Q = (\mathbf{y} - \widehat{\boldsymbol{\mu}})'\mathbf{K}(\mathbf{y} - \widehat{\boldsymbol{\mu}}),\ \mathrm{K} = GWG'$

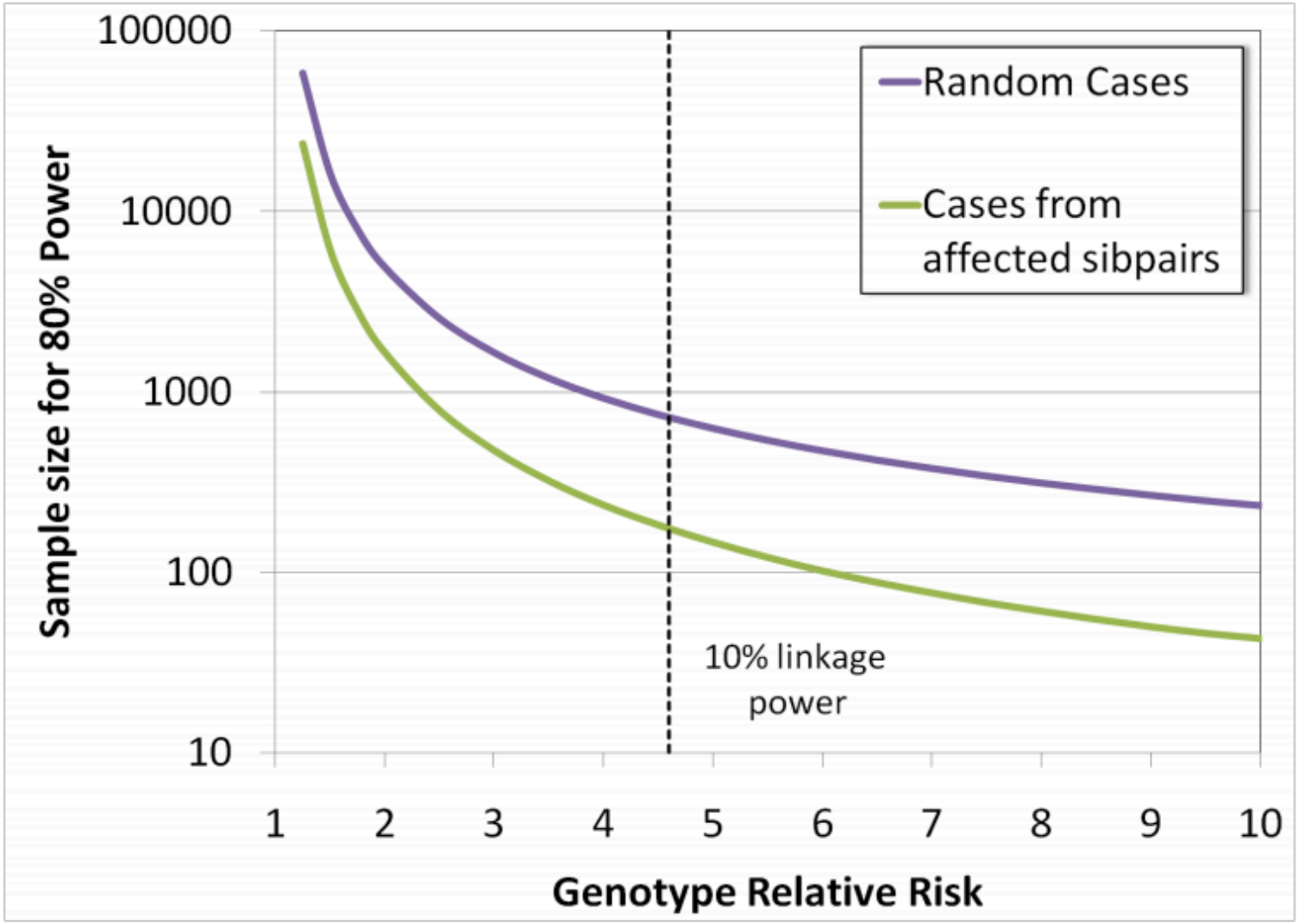
- Kernel $K(\mathbf{G}_i, \mathbf{G}_{i'}) = \sum_{j=1}^{p} w_j G_{ij} G_{i'j}.\ K(\bullet, \bullet)$ $\mathbf{W} = \text{diag}(w_1, \ldots, w_p)$


Continuous Trait
β +/− = 100/0
β +/− = 80/20
β +/− = 50/50
SKAT
SKAT_M
rSKAT
W
N
C
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.5k
1k
2.5k
5k
Total Sample Size
Dichotomous Trait
β +/− = 100/0
β +/− = 80/20
β +/− = 50/50
Power
Total Sample Size

Maximizing the Power

- Power depends summed frequency
  - Choose threshold for defining rare carefully
- Enriched functional variants in cases increase power
  - Focus on loss of function variants only
- Use more efficient design
  - For quantitative traits, focus on individuals with extreme trait values
  - For binary traits, focus on individuals with family history of disease

Benefits of Favoring Family History of Disease



Discussion

- Analysis of rare variants is an active research area
- Weight for each SNP is the key
- What to do if the samples are related
- Most tests reply on permutation
  - Computationally intensive

software

- SKAT
- PLINK (Whole genome association analysis toolset)
  - http://pngu.mgh.harvard.edu/~purcell/plink/

Thank you for your attention