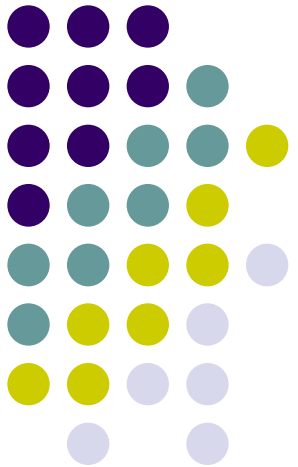
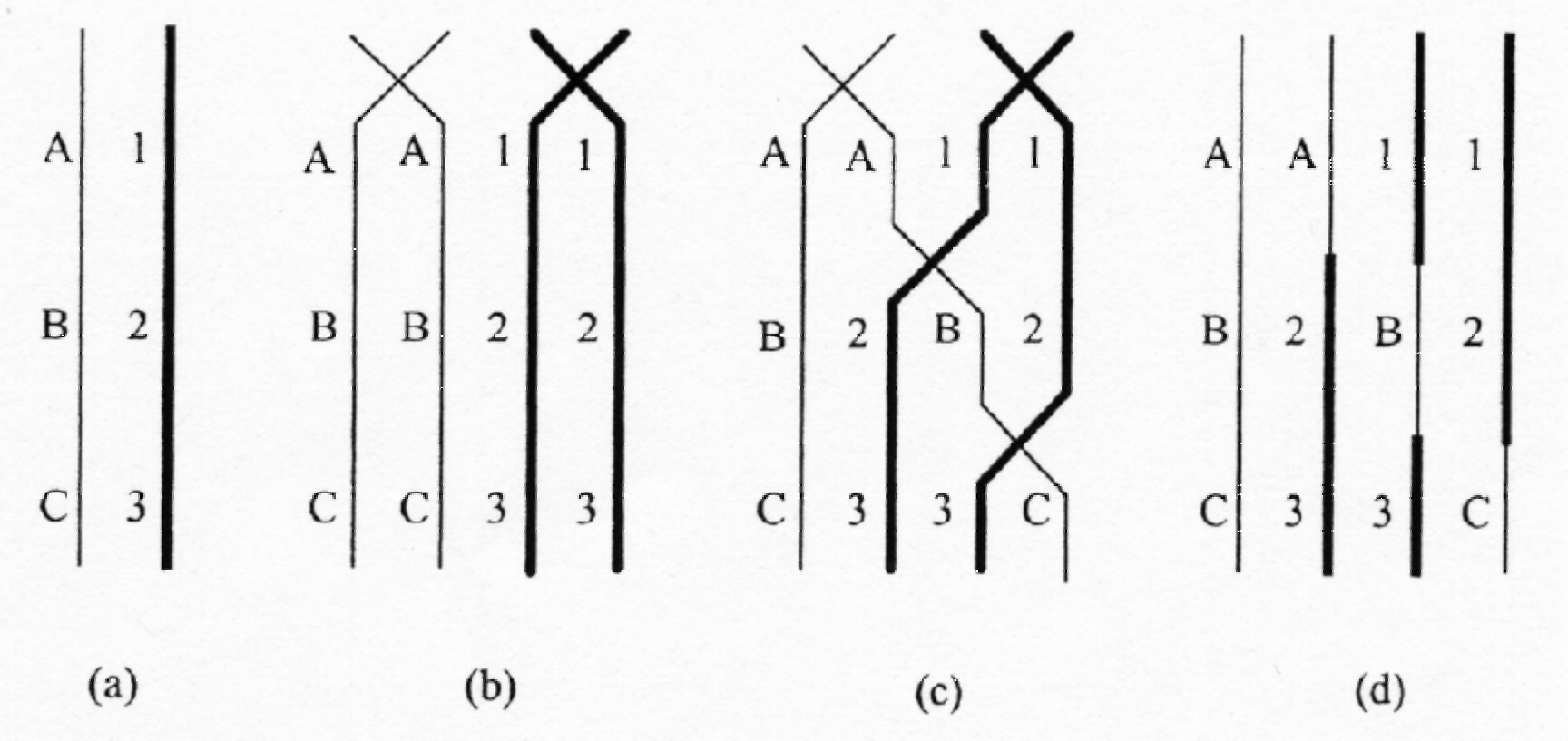
Statistical Methodologies for Analyzing Whole Genome Association Data

John P. Rice, Ph.D.

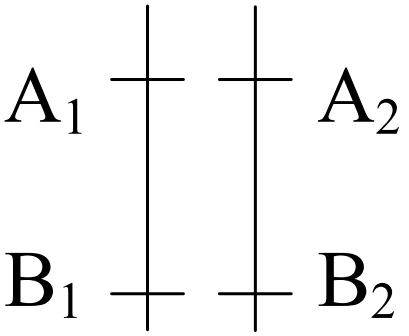
Washington University School of Medicine



Crossing Over During Meiosis



Definition of centimorgan (cM)



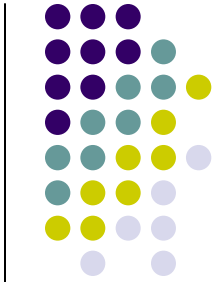
Gametes $A_1 B_2$, $A_2 B_1$ are recombinants

$A_1 B_1$, $A_2 B_2$ are non-recombinants

θ = Prob (recombinant)

θ=.01 $\Leftrightarrow$ A and B are 1cM apart

Genome Arithmetic

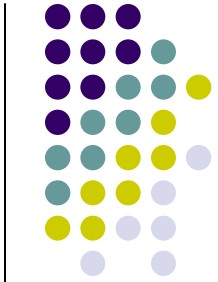


- Kb=1,000 bases; Mb=1,000Kb
- 3.3 billion base pairs; 3,300 cM in genome

  3,300,000,000/3,300 = 1 Mb/cM
- 33,000 genes

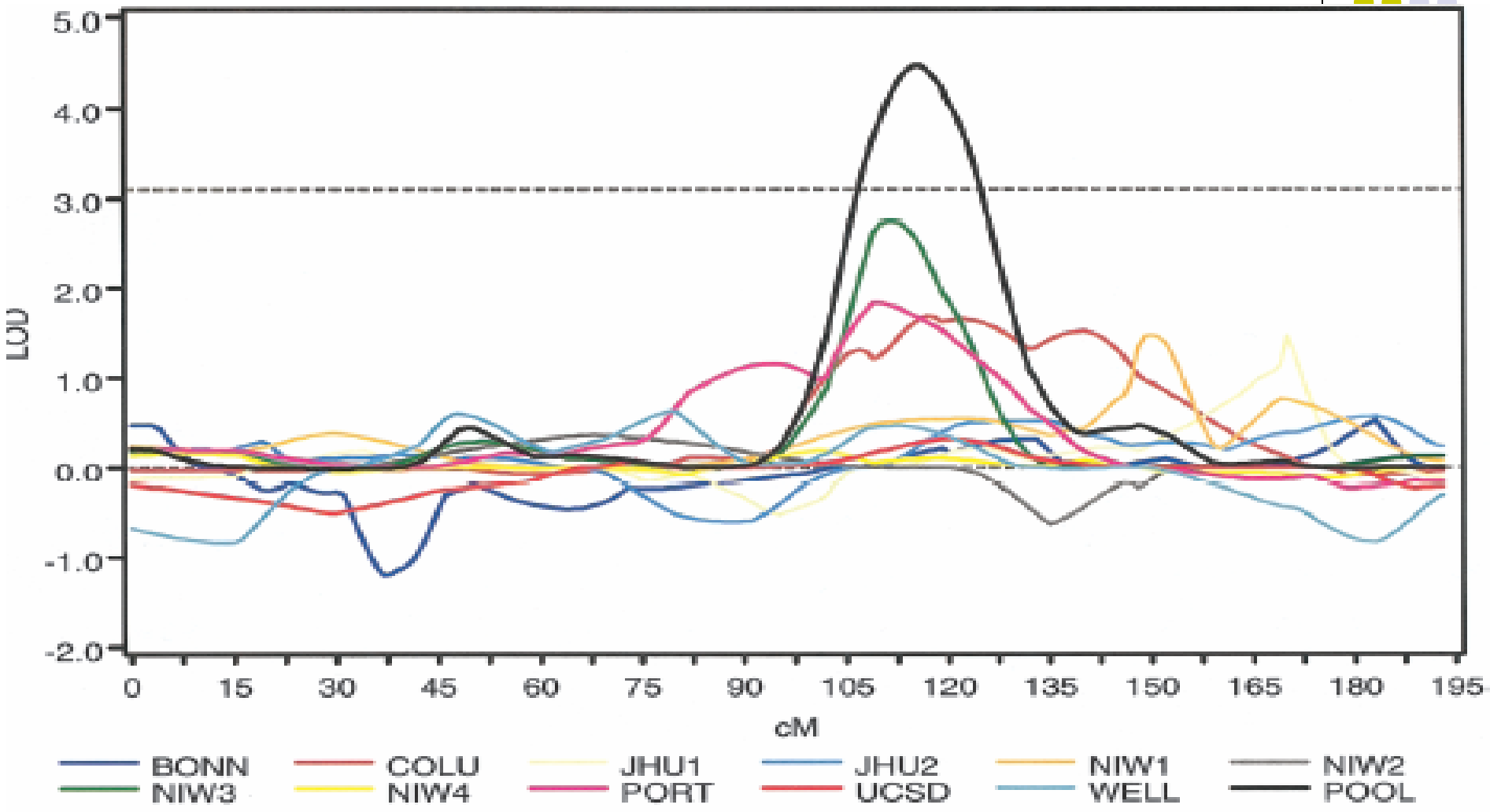
  33,000/3,300 Mb = 10 genes / Mb
- Thus, 20 cM region may have 200 genes to examine
- Erratum – closer to 20,000 genes in humans

Linkage Vs. Association

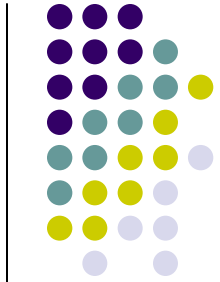


- Linkage:
  - -Disease travels with marker within families
  - -No association within individuals
  - -Signals for complex traits are wide (20MB)
- Association:
  - -Can use case/control or case/parents design
  - -Only works if association in the population
  - -Allelic heterogeneity (eg, BRAC1) a problem
- Linkage – large scale; Association fine scale (<200kb)

Exanple of a LOD Curve



Disequilibrium



	A1	A2
B1		
B2		

Let $P(A_1)=p_1$

Let $P(B_1)=q_1$

Let $P(A_1B_1)=h_{11}$

No association if $h_{11}=p_1q_1$

$D = h_{11}-p_1q_1$

D´ and r²

D tends to take on small values and depends on marginal gene frequencies

$D' = D / \max(D)$

$r^2 = D^2 / (p_1 p_2 q_1 q_2)$

= square of usual correlation coefficient (ϕ)

Note: $r^2 = 0 \Leftrightarrow D' = 0$

$D' = \pm 1$ if one cell is zero

r^2 can be small even when $D' = \pm 1$

Prediction of one SNP by another depends on r^2

Basic Idea

- If SNP A is a disease susceptibility gene, and if we genotype SNP B (for example in a whole genome association study), and if A and B are in disequilibrium, then cases and controls will have different frequencies of alleles at B
- Power to detect A is related to N/r^2

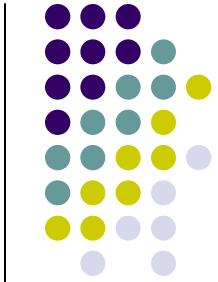
Table of A by B			
A	B		
	B1	B2	Total
A1	50 50.00 100.00 55.56	0 0.00 0.00 0.00	50 50.00
A2	40 40.00 80.00 44.44	10 10.00 20.00 100.00	50 50.00
Total	90 90.00	10 10.00	100 100.00

$D' = 1, r^2 = .1$

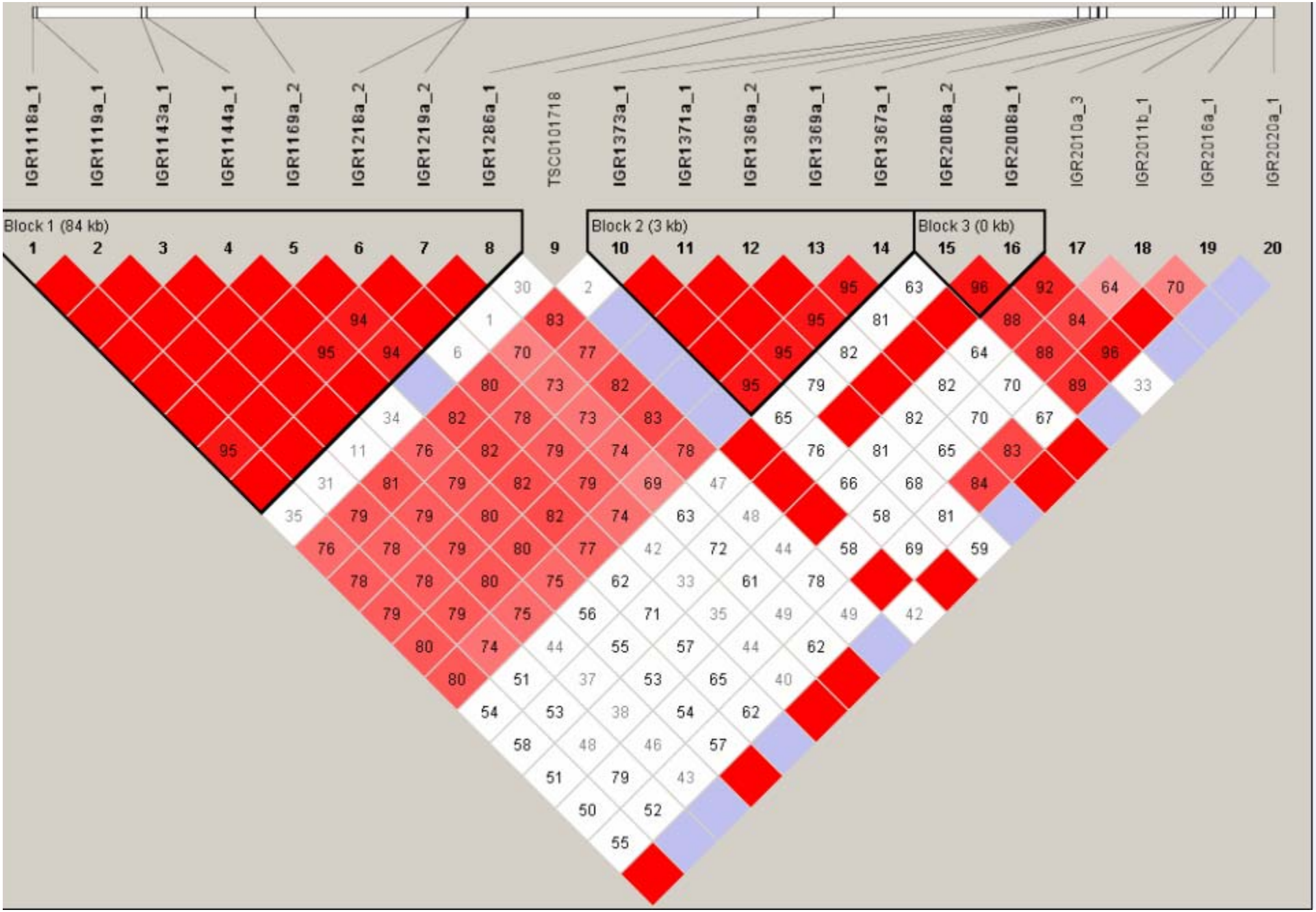
Table of A by B			
A	B		
	B1	B2	Total
A1	10 10.00 11.11 100.00	80 80.00 88.89 88.89	90 90.00
A2	0 0.00 0.00 0.00	10 10.00 100.00 11.11	10 10.00
Total	10 10.00	90 90.00	100 100.00

$D' = 1, r^2 = .01$

Blocks and Bins



- Predictability of one SNP by another best described by r^2 – basic statistics
- Block – set of SNPs with all pair-wise LD high (Please specify measure)
- If one uses r^2 – insert a SNP with low frequency in between SNPs with freqs close to 0.5, then block breaks up!
- Perlegen (Hinds et al, Science, 2005) -– use bins where a tag SNP has r^2 of 0.8 with all other SNPs. Bins may not be contiguous.

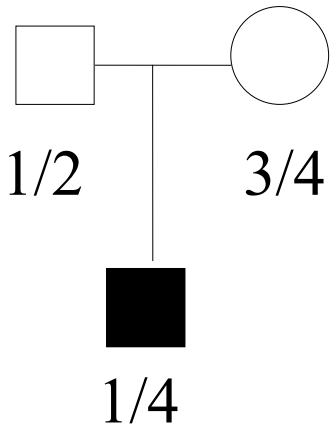

IGR1118a_1
IGR1119a_1
IGR1143a_1
IGR1144a_1
IGR1169a_2
IGR1218a_2
IGR1219a_2
IGR1286a_1
TSC0101718
IGR1373a_1
IGR1371a_1
IGR1369a_2
IGR1369a_1
IGR1367a_1
IGR2008a_2
IGR2008a_1
IGR2010a_3
IGR2011b_1
IGR2016a_1
IGR2020a_1
Block 1 (84 kb)
Block 2 (3 kb)
Block 3 (0 kb)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Summary (Blocks and Bins)

- Blocks using **D ′** may have a “biological” interpretation (long stretches with **|D ′|** =1)
- Selection of Tag SNPs is a statistical issue, want to predict untyped SNPS from those that are typed – r^2 is natural measure
- Phase of SNPs is important – usually ignored
- Most current WGA studies use bins based on r^2 (typically $r^2 > 0.8$)
- There is an art to selecting tag SNPs

Statistical Analysis

- Case/Control Design
  - Use standard statistical tests (logistic regression) to test whether the distribution of the SNP differs between cases and controls
  - Sensitive to population stratification
- Family Based Design



Alleles 1 and 4 are transmitted -- CASE
Alleles 2 and 3 are non-transmitted –CONTROL

NOTE: Genotype 3 people to get 1 case and 1 control
NOT sensitive to population stratification

Problem of Multiple Tests

Significant level = α

We perform N (independent) tests

We expect to reject $N\alpha$ tests if null hypothesis is true for each test.

Example:

N = 100, α = .05, x = # of rejections

$P(x \geq 1) = 1 - P(x = 0)$

$= 1 - (1 - \alpha)^{100}$

$= .99408$

Note: $1 - (1 - \alpha)^N \approx N\alpha$ for α small

Choose $\alpha' = \alpha/N = .0005$

The $1 - (1 - \alpha')^{100} = .0488$

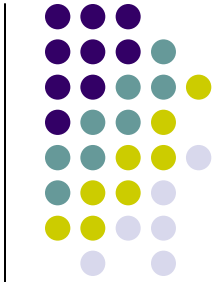
Bonferroni Correction

Problem: Power goes down as α decreases

Multiple tests for association

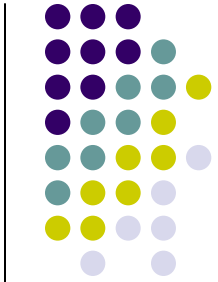
- Intuition: LD extents over smaller regions than linkage
- More “independent” tests for LD -- There must be at the equivalent of at least 200,000 independent tests in one experiment (linkage about 2,000 independent tests)
- Multiple testing for whole genome association studies will be problematic
- Practical question – How to correct for multiple tests

Multiple Testing



- Suppose we use 600,000 SNPs, and there are 10 true susceptibility loci. Test at significance level p=0.001, and power is 60%
- We expect 10 x .6 = 6 true positives, and 600,000 x .001= 600 false positives. We expect one false positive to be significant at the 0.0000002 level.
- Tests are not independent, so use of Bonferroni correction of 0.05/600,000=.000000008 is too conservative. Even with appropriate p-value, there would be little power without massive sample sizes. A gene with the effect size needed to be detected would already be known.

False Discovery Rate (FDR)



- V= # true null hypotheses called significant
  S= # non-true hypotheses called significant
  Q=V/(V + S) (false positives/all positives)
  FDR = E(Q)
- Benjamini & Hochberg (1995)
  When testing m hypotheses $H_1,\ldots,H_m$, order p-values $p_1, \ldots p_m$, let k be largest i for which $p_i \leq (i/m)\, q^*$
  Then reject $H_1, \ldots H_m$

Theorem: Above controls FDR at q*

Computer program: QVALUE

Multiple Testing

- FDR helps and is commonly used
- Question: Should all markers be tested using same p-value?
- Roeder et al (2006) Am J Hum Genet, 78:243

  Use a set of weights in the FDR computations.

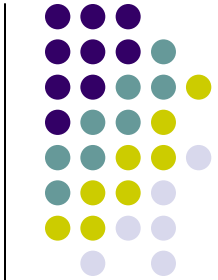
  If a small proportion are over-weighted, does not reduce the power to detect the others very much, but helps the detection of the ones to “bet” on.

  Use of prior linkage evidence may be a way to increase power.

Example: Top 10 SNPs from Analysis of 1,500 SNPs

Obs	rs ID	Primary p-value (p_V3_MI)	Storey q-value (smoother pi0 = 0.89)	LD Bin	Number of markers in bin	Min r^2 of tag with nontags	MAF	MAF (HapMap)	Chr	Pos (bp)	Gene	Function
1	12334778	1.21E-04	0.113	85	2	0.968	0.484	0.500	8	123,567,696		.
2	4506214	2.56E-04	0.113	85	2	0.968	0.477	0.492	8	123,564,218		.
3	4336618	3.47E-04	0.113	68	2	0.982	0.325	0.408	8	123,536,162		.
4	6986303	3.66E-04	0.113	.	.	.	0.294	0.275	8	134,547,711	ST3GAL1	INTRON
5	7846137	4.35E-04	0.113	68	2	0.982	0.318	0.400	8	123,532,225		.
6	6470170	5.44E-04	0.118	.	.	.	0.226	0.150	8	124,736,503		.
7	10101440	1.26E-03	0.234	.	.	.	0.127	0.125	8	127,604,613		.
8	1394051	1.71E-03	0.279	.	.	.	0.099	0.100	8	134,927,184		.
9	4870857	2.74E-03	0.397	.	.	.	0.398	0.383	8	124,661,555		.
10	4259418	3.51E-03	0.457	.	.	.	0.318	0.458	8	123,518,309		.

Conclusions



- WGA studies will be done (6 GAIN studies have just been selected) and be in the public domain
- Candidate gene studies have been problematic (the prior probability of selecting the right gene may be 1/10,000), so may be very low power.
- Multiple testing issues a major challenge for WGA studies, but these will be overcome