Linkage and Association

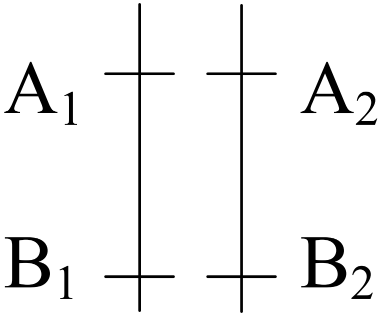
John P. Rice, Ph.D.

Washington University School of Medicine

Outline

- Linkage
- Linkage Disequilibrium
- Haplotypes
- History of GWAS
- dbGaP
- Methods
  - Genomic Inflation Factor
  - False Discovery Rate
  - Ethnic Stratification
  - QQ-Plots

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 ⇔ 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)

LOD Score

- LOD score is $\log_{10}$ (odds for linkage/odds for no linkage) Traditional (1955) cut-off is LOD=3 (linkage 1000 times more likely)
- A LOD of 3 corresponds to $\alpha = 0.0001$
- Lander and Kruglyak (1995) A LOD score cut-off of 3.6 for a genome screen using an infinitely dense map corresponds to a “genome-wide significance of 0.05”
- This is the criteria often cited today

Effective Number of Tests For genome-wide p=.05

Marker Spacing	LOD	P-value	$N_{effective}$
10 cM	2.88	.000135	370
5 cM	3.06	.000088	568
2 cM	3.24	.000057	877
1 cM	3.35	.000044	1,136
0.1 cM	3.63	.000022	2,273

Am. J. Hum. Genet. 77:582–595, 20

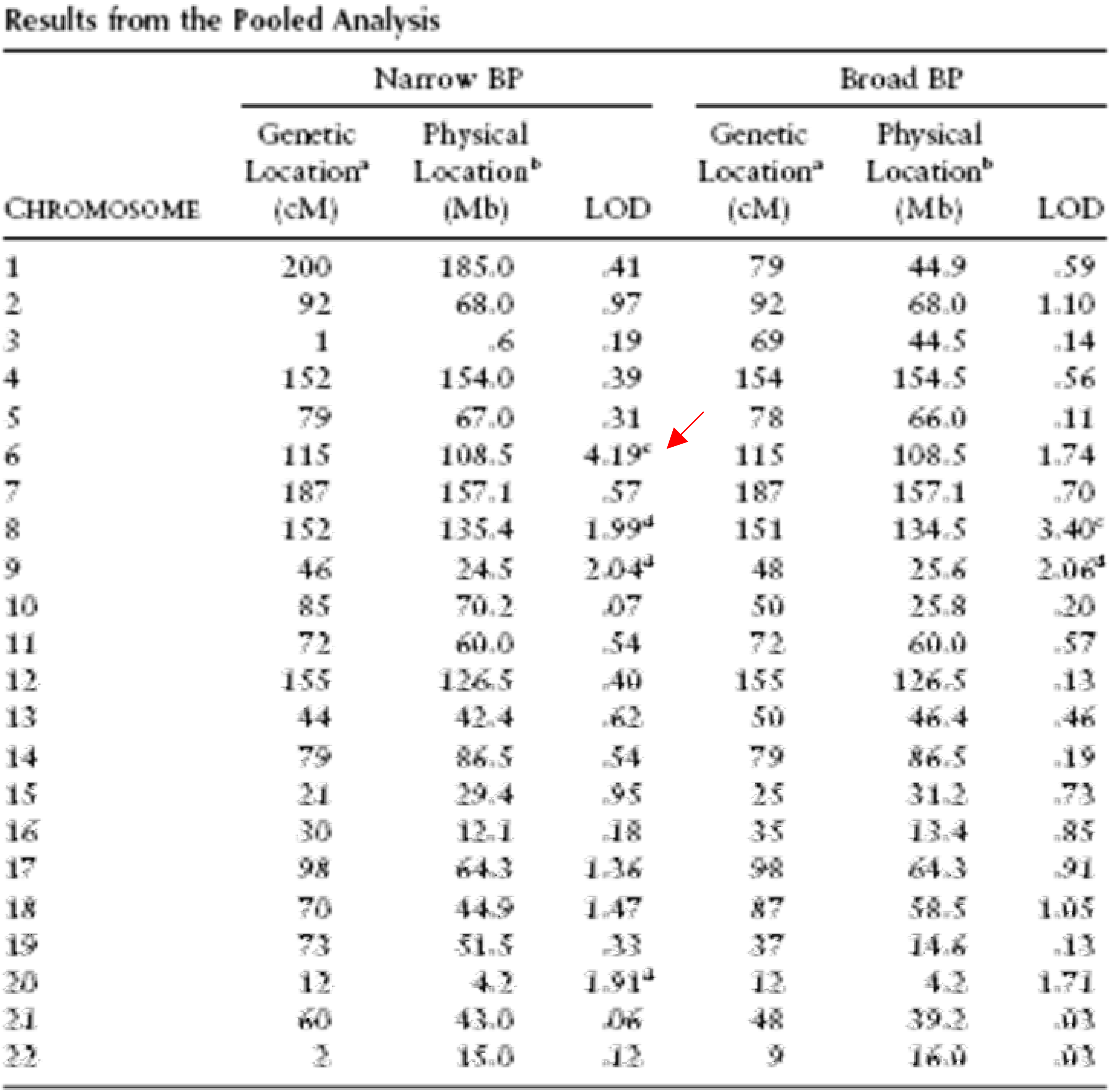
Combined Analysis from Eleven Linkage Studies of Bipolar Disorder Provides Strong Evidence of Susceptibility Loci on Chromosomes 6q and 8q

Matthew B. McQueen,[*] B. Devlin,[*] Stephen V. Faraone,[*] Vishwajit L. Nimgaonkar,[*] Pamela Sklar,[*] Jordan W. Smoller,[*] Rami Abou Jamra, Margot Albus, Silviu-Alin Bacanu, Miron Baron, Thomas B. Barrett, Wade Berrettini, Deborah Blacker, William Byerley, Sven Cichon, Willam Coryell, Nick Craddock, Mark J. Daly, J. Raymond DePaulo, Howard J. Edenberg, Tatiana Foroud, Michael Gill, T. Conrad Gilliam, Marian Hamshere, Ian Jones, Lisa Jones, Suh-Hang Juo, John R. Kelsoe, David Lambert, Christoph Lange, Bernard Lerer, Jianjun Liu, Wolfgang Maier, James D. MacKinnon, Melvin G. McInnis, Francis J. McMahon, Dennis L. Murphy, Markus M. Nöthen, John I. Nurnberger Jr., Carlos N. Pato, Michele T. Pato, James B. Potash, Peter Propping, Ann E. Pulver, John P. Rice, Marcella Rietschel, William Scheftner, Johannes Schumacher, Ricardo Segurado, Kristel Van Steen, Weiting Xie, Peter P. Zandi, and Nan M. Laird[*,†]

Bipolar Disorder

- Lifetime prevalence of BP1 ≈ 1%, BPII ≈ 0.5%
- Risk of suicide 10 – 15%
- Treatment not curative, treatments not completely effective in mitigating symptoms
- Heritability estimates ≈ 80%
- Linkage reports for ½ the chromosomes, with a lack of replication
- Lack of power in original reports?

Data Set	No. of Edigrees	No. of Genotyped Individuals	No. of Genetic Markers[a]	
			Genotyped	Mapped
Bonn	75	387	389	386
Columbia	40	358	334	333
Johns Hopkins 1	63	562	823	802
Johns Hopkins 2	40	175	381	380
NIMH Wave 1	95	525	357	351
NIMH Wave 2	55	348	465	458
NIMH Wave 3	220	982	372	372
NIMH Wave 4	274	1,053	384	384
Portuguese	16	102	346	342
UCSD	20	163	331	324
Wellcome Trust	151	509	380	378
Total	1,067	5,179	4,562	4,510



Results from the Pooled Analysis

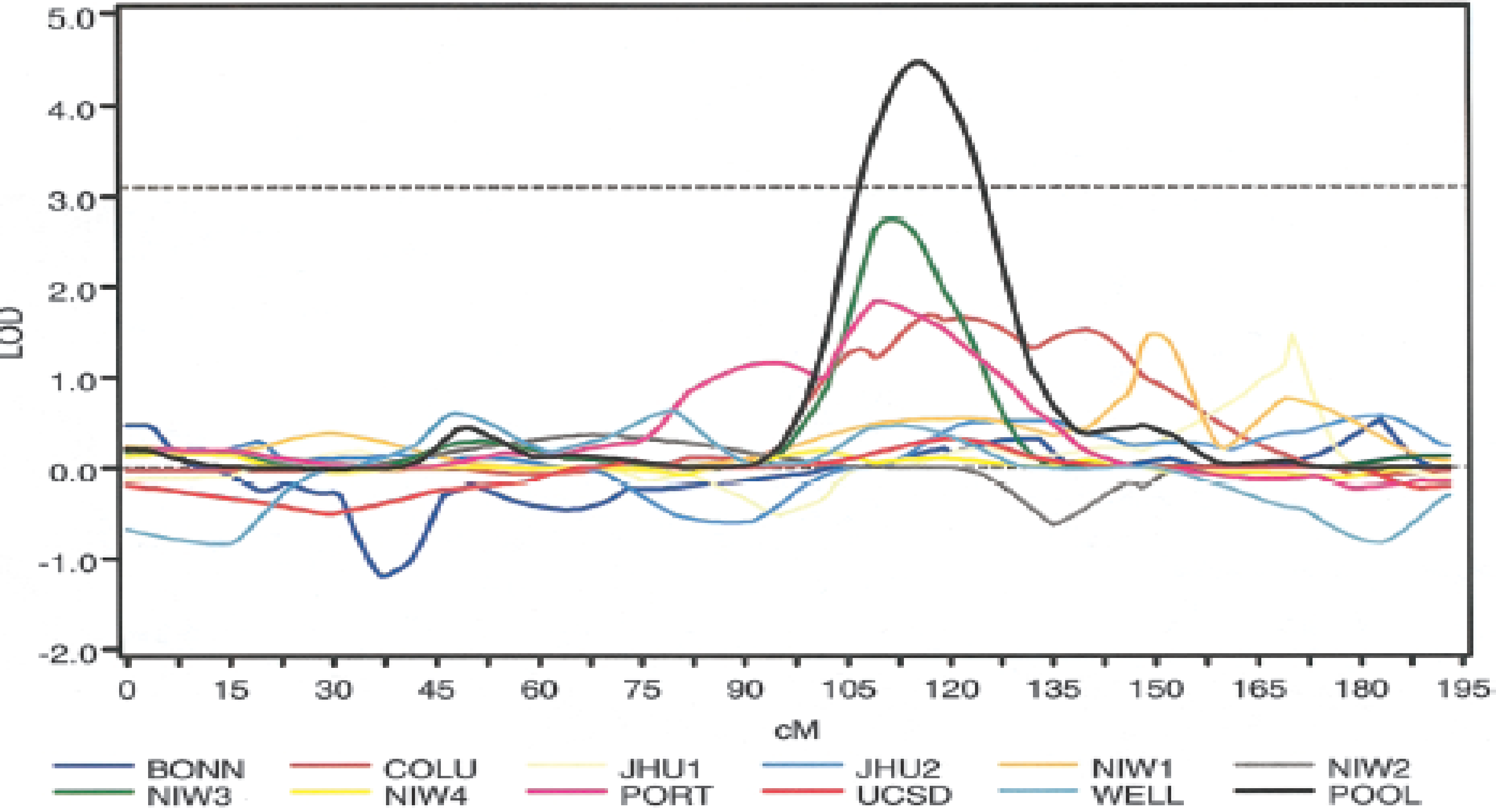
	Narrow BP			Broad BP		
CHROMOSOME	Genetic Location[a] (cM)	Physical Location[b] (Mb)	LOD	Genetic Location[a] (cM)	Physical Location[b] (Mb)	LOD
1	200	185.0	.41	79	44.9	.59
2	92	68.0	.97	92	68.0	1.10
3	1	.6	.19	69	44.5	.14
4	152	154.0	.39	154	154.5	.56
5	79	67.0	.31	78	66.0	.11
6	115	108.5	4.19[c]	115	108.5	1.74
7	187	157.1	.57	187	157.1	.70
8	152	135.4	1.99[d]	151	134.5	3.40[c]
9	46	24.5	2.04[d]	48	25.6	2.06[d]
10	85	70.2	.07	50	25.8	.20
11	72	60.0	.54	72	60.0	.57
12	155	126.5	.40	155	126.5	.13
13	44	42.4	.62	50	46.4	.46
14	79	86.5	.54	79	86.5	.19
15	21	29.4	.95	25	31.2	.73
16	30	12.1	.18	35	13.4	.85
17	98	64.3	1.36	98	64.3	.91
18	70	44.9	1.47	87	58.5	1.05
19	73	51.5	.33	37	14.6	.13
20	12	4.2	1.91[d]	12	4.2	1.71
21	60	43.0	.06	48	39.2	.03
22	2	15.0	.12	9	16.0	.03

Significant and Suggestive Linkage

- Given density of markers, significant linkage is LOD > 3.03
- Suggestive linkage is LOD > 1.75
- These take into account that 2 genome screens were analyzed (narrow and broad)
- **Significant** – Occurs once in twenty genome screens

  Suggestive – Occurs once in a genome screen

Chromosome 6



Linkage Analysis (Summary)

- Approximately 2,000 “independent “ tests with an infinitely dense genetic map (Multiple testing a much bigger problem in GWAS)
- Linkage studies have been unsuccessful for complex diseases
- May be useful as input into GWAS analysis?
- Today – GWAS (using SNP chips) have taken over
- My opinion – pursue chromosomes 6 and 8, even if not genome-wide significant in GWAS

Genome-Wide Association Studies (GWAS)

- Chips by Illumina and Affymetrix genotype 1 million SNPs (Single Nucleotide Polymorphisms) as well as CNVs (Copy Number Variations)
- Affordable on a large scale
- Capitalize on Linkage Disequilibrium between the markers and variation at a susceptibility gene

Disequilibrium

Let $P(A_1)=p_1$

Let $P(B_1)=q_1$

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

No association if $h_{11}=p_1q_1$

$D = h_{11}-p_1q_1$

Linkage Disequilibirum:

- Linkage
- Random Genetic Drift
- Founder Effect
- Mutation
- Selection
- Population admixture/stratification

Population Stratification

Population 1

1	9
9	81

Odds ratio = 1

Population 2

25	25
25	25

Odds ratio = 1

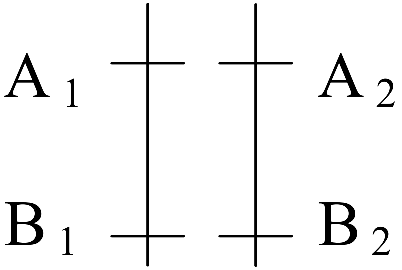
Combined Population

26	34
34	106

Odds ratio = 2.38

Linkage Disequilibrium

———————————————————



A_1 A_2

B_1 B_2

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

θ = P (recombinant)

————————————————————

Consider haplotype $A_i B_j$, frequency h_{ijo} in generation 0, what is the frequency in the next generation?

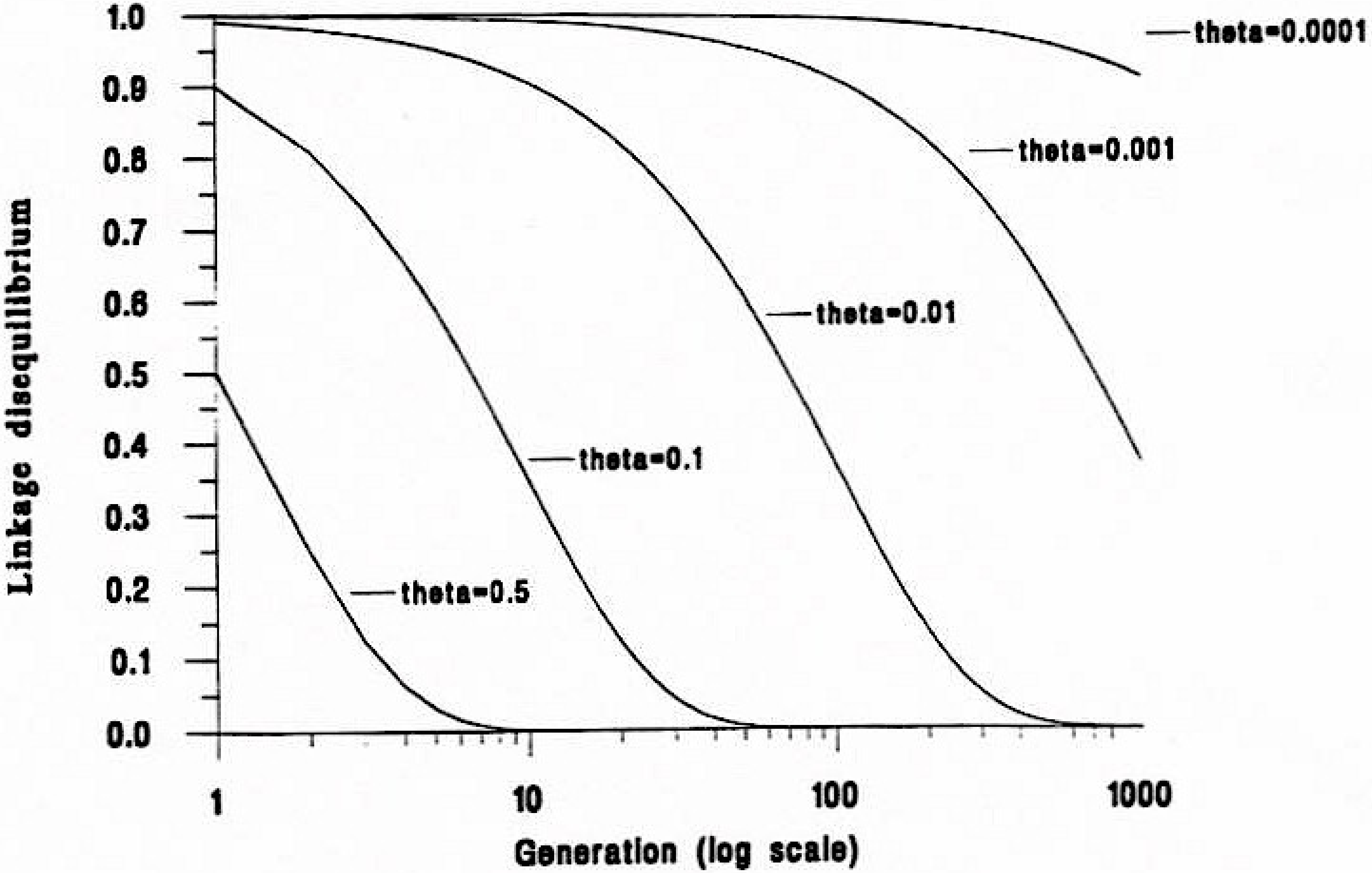


Figure 4.1 Decay of linkage disequilibrium by generation.

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 (eg, no recombination)

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

Prediction of one SNP by another depends on r^2

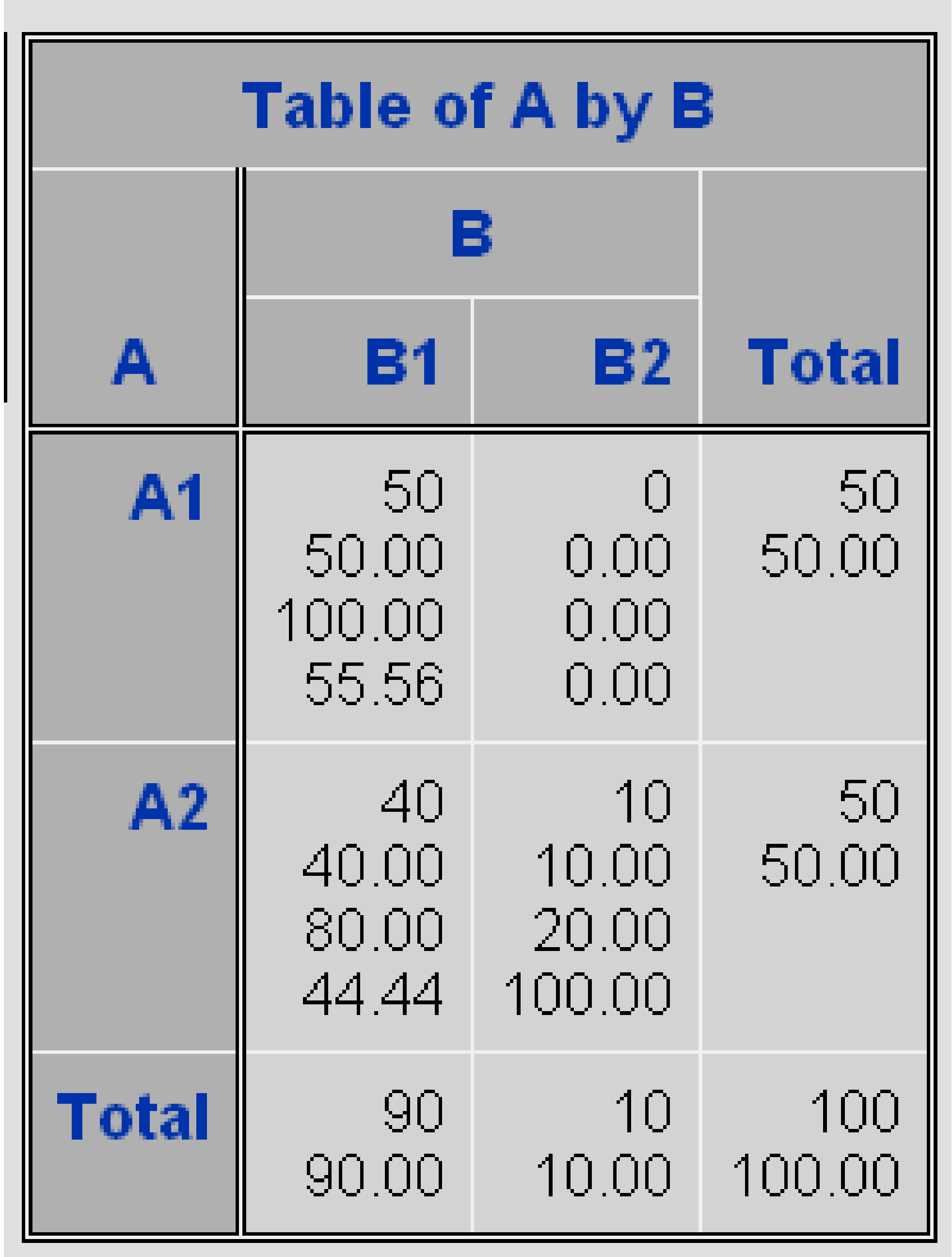


Table of A by B

	B		
A	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

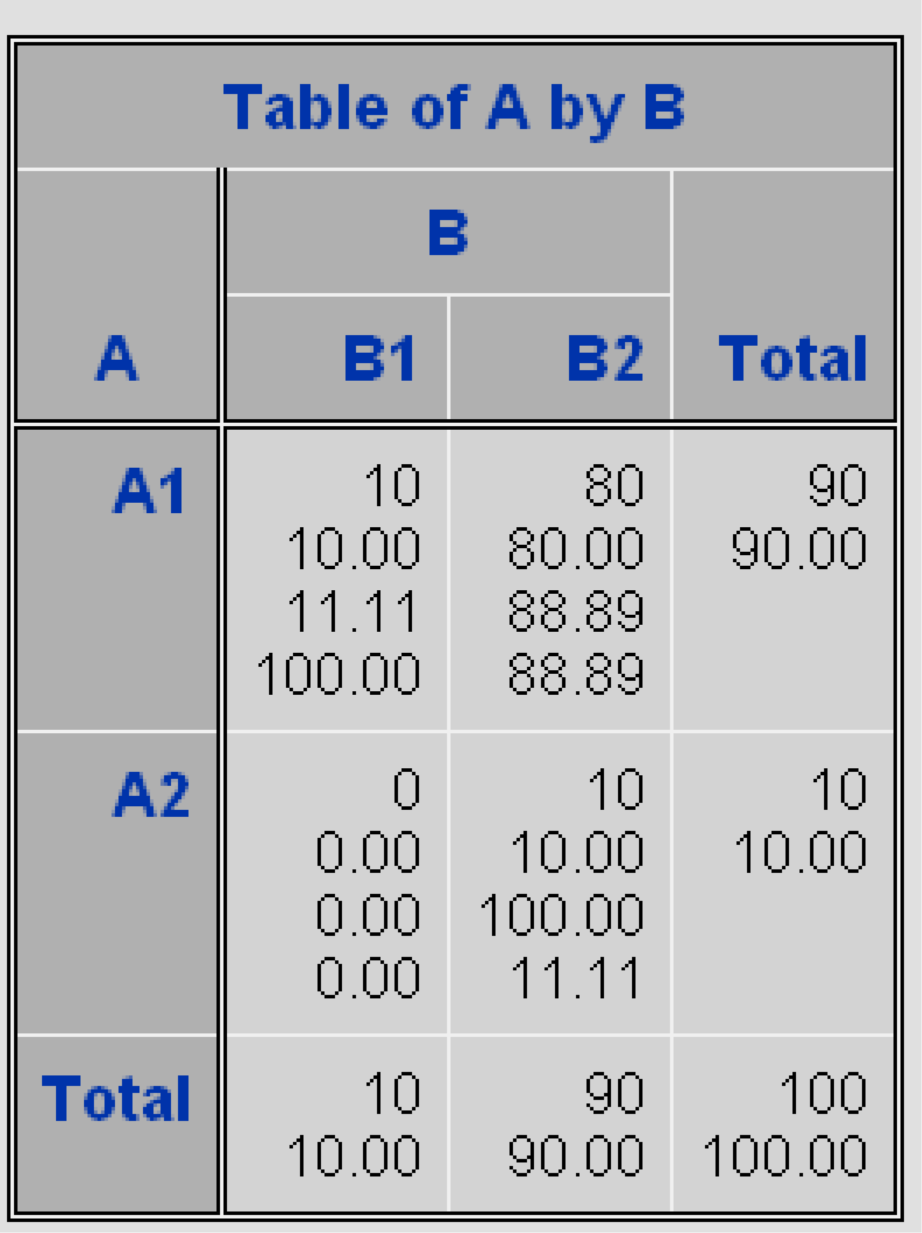


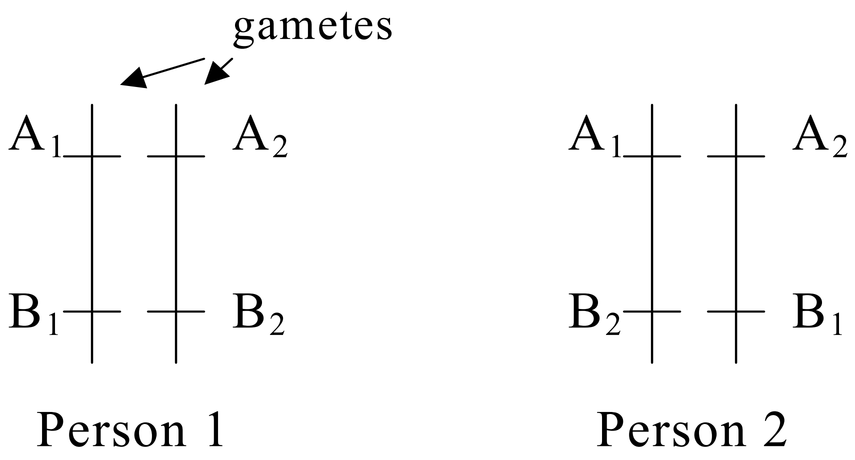
Table of A by B			
A	B		Total
	B1	B2	
A1	10 10.00 11.11 100.00	80 80.00 88.89 88.89	90 90.00
A2	0 0.00 0.00 0.00	10 10.00 100.00 11.11	10 10.00
Total	10 10.00	90 90.00	100 100.00

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

Haplotypes

- We measure genotypes
- A double heterozygote is ambiguous
- Must estimate haplotype frequencies from genotype frequencies – usually assume random mating and use EM algorithm
- The program haploview is commonly used to estimate and depict LD

Note:



Different Haplotypes; same genotypes $A_1\ A_2\ B_1\ B_2$

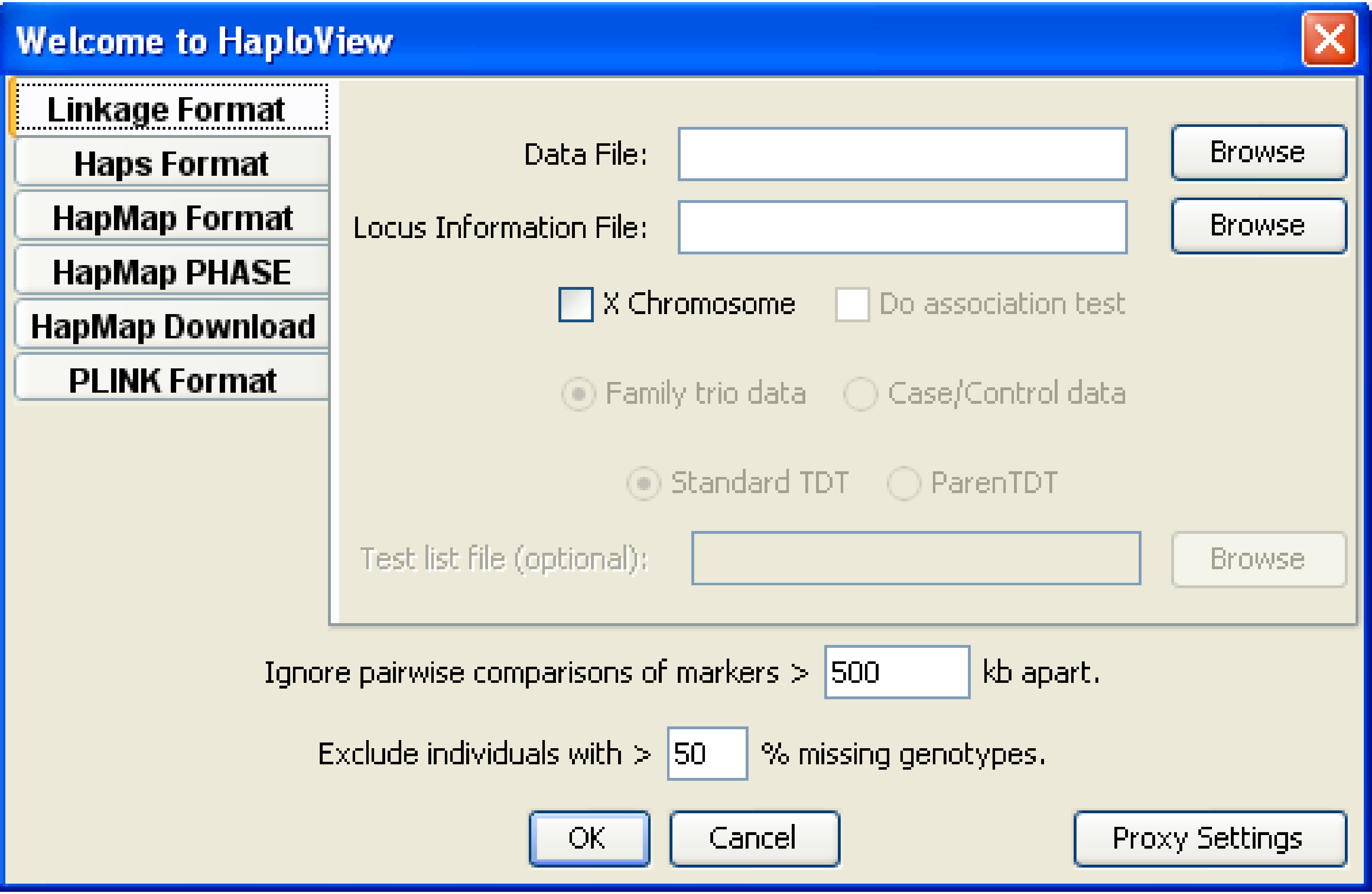
Haplotypes $A_1\ B_1$, $A_2\ B_2$; $A_1\ B_2$, $A_2\ B_1$

Independence	$h_{ij} = p_i\ q_j$
Positive Association	$h_{ij} > p_i\ q_j$
Negative Association	$h_{ij} < p_i\ q_j$

Assume random mating but allow for disequilibrium

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	h_{11}^2	$2h_{11}h_{12}$	h_{12}^2
A_1A_2	$2h_{11}h_{21}$	$\boxed{\alpha}$	$2h_{12}h_{22}$
A_2A_2	h_{21}^2	$2h_{21}h_{22}$	h_{22}^2

A_1B_1	A_1B_2	A_2B_1	A_2B_2
h_{11}	h_{12}	h_{21}	h_{22}



Welcome to HaploView

Data File:

Locus Information File:

□ X Chromosome □ Do association test

● Family trio data ○ Case/Control data

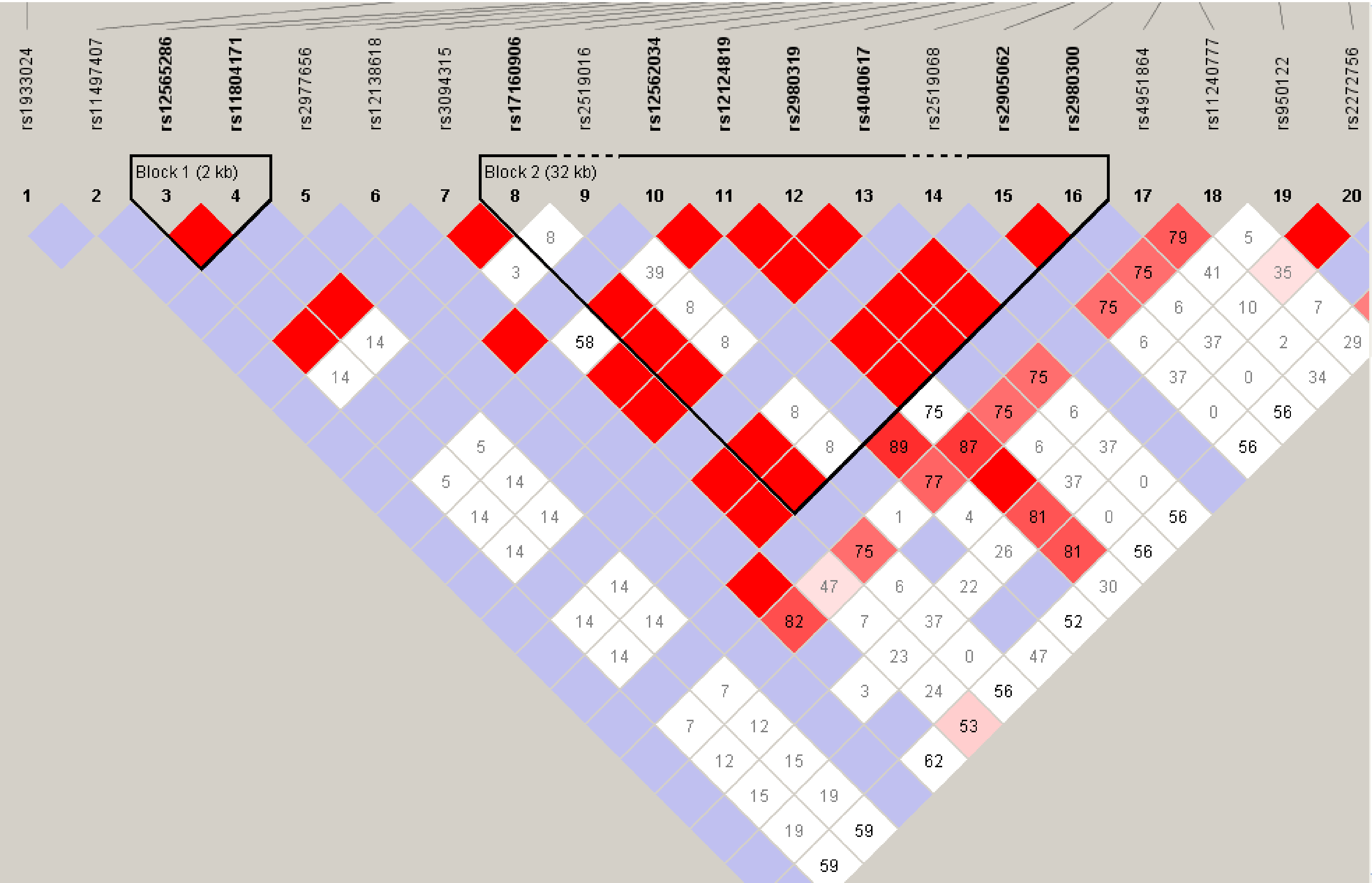
● Standard TDT ○ ParenTDT

Test list file (optional):

Ignore pairwise comparisons of markers > 500 kb apart.

Exclude individuals with > 50 % missing genotypes.

D′ plot from Haploview



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 (usually defined in terms of D')
- 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.

Summary (Blocks and Bins)

- Blocks using **D ′** may have a “biological” interpretation (long stretches with **|D ′|** =1 and indicates no recombination)
- Selection of Tag SNPs is a statistical issue, want to predict untyped SNPS from those that are typed – r^2 is natural measure
- Most current WGA studies use bins based on r^2 (typically $r^2 > 0.8$)
- Sample size needed is N/ r^2 with reduced r^2

Analysis

- Case/ control studies are common. Use logistic regression with case/control status as the dependent variable. Use SNP genotype as an independent variable with other covariates and test one SNP at a time
- PLINK is my program of choice to do this
- Family based studies are also used. TDT (case and both parents) designs are used in GWAS but less efficient

SNP Marker Coding:

Genotype	X_1
1/1	0
1/2	1
2/2	2

Testing Marker Effects

$\log(\text{odds}) = \alpha + \beta_1 X_1$

$\text{odds} = e^{\alpha} e^{\beta 1\ X1}$

Genotype	Odds
11	e^{α}
12	$e^{\alpha}e^{\beta 1}$
22	$e^{\alpha}e^{2\beta 1}$

Test $\beta_1 = 0$, all odds $= e^{\alpha}$

Note: No dominance effect

SNP Marker Coding:

Genotype	X1	X2
1 1	0	0
1 2	1	1
2 2	2	0

Testing Marker Effects

$$\log(\text{odds}) = \alpha + \beta_1 X_1 + \beta_2 X_2$$

$$\text{odds} = e^{\alpha} e^{\beta 1\, X1}\, e^{\beta 2\, X2}$$

Genotype	Odds
1 1	e^{α}
1 2	$e^{\alpha}\, e^{\beta 1} e^{\beta 2}$
2 2	$e^{\alpha} e^{2\beta 1}$

Test $\beta_1 = \beta_2 = 0$, all odds $= e^{\alpha}$

If $\beta_2 = 0$, then have additive model

Haplotypes?

- We may wish to consider more than one SNP at a time in the linear regression.
  - More information in a set of close SNPs
  - May wish to study a set of SNPs to see if one explains the case/control difference, i.e., does the evidence for one SNP disappear when controlling for other SNPs.

Haplotype Trend Analysis

- Zaykin et al (2002) Hum Hered 53:79-91
- Use haplotypes in logistic regression
- For a pair of SNPs, there are 4 haplotypes, so there will be 3 “dummy” variables
- Assume pair of haplotypes in an individual are “additive”, so only need 3 regression coefficients
- If haplotypes are known with certainty, then:

Haplotype	X1	X2	X3
h_1/ h_1	2	0	0
h_1/ h_2	1	1	0
h_1/ h_3	1	0	1
h_1/ h_4	1	0	0
h_2/ h_2	0	2	0
h_2/ h_3	0	1	1
h_2/ h_4	0	1	0
h_3/ h_3	0	0	2
h_3/ h_4	0	0	1
h_4/ h_4	0	0	0

Estimated Haplotypes

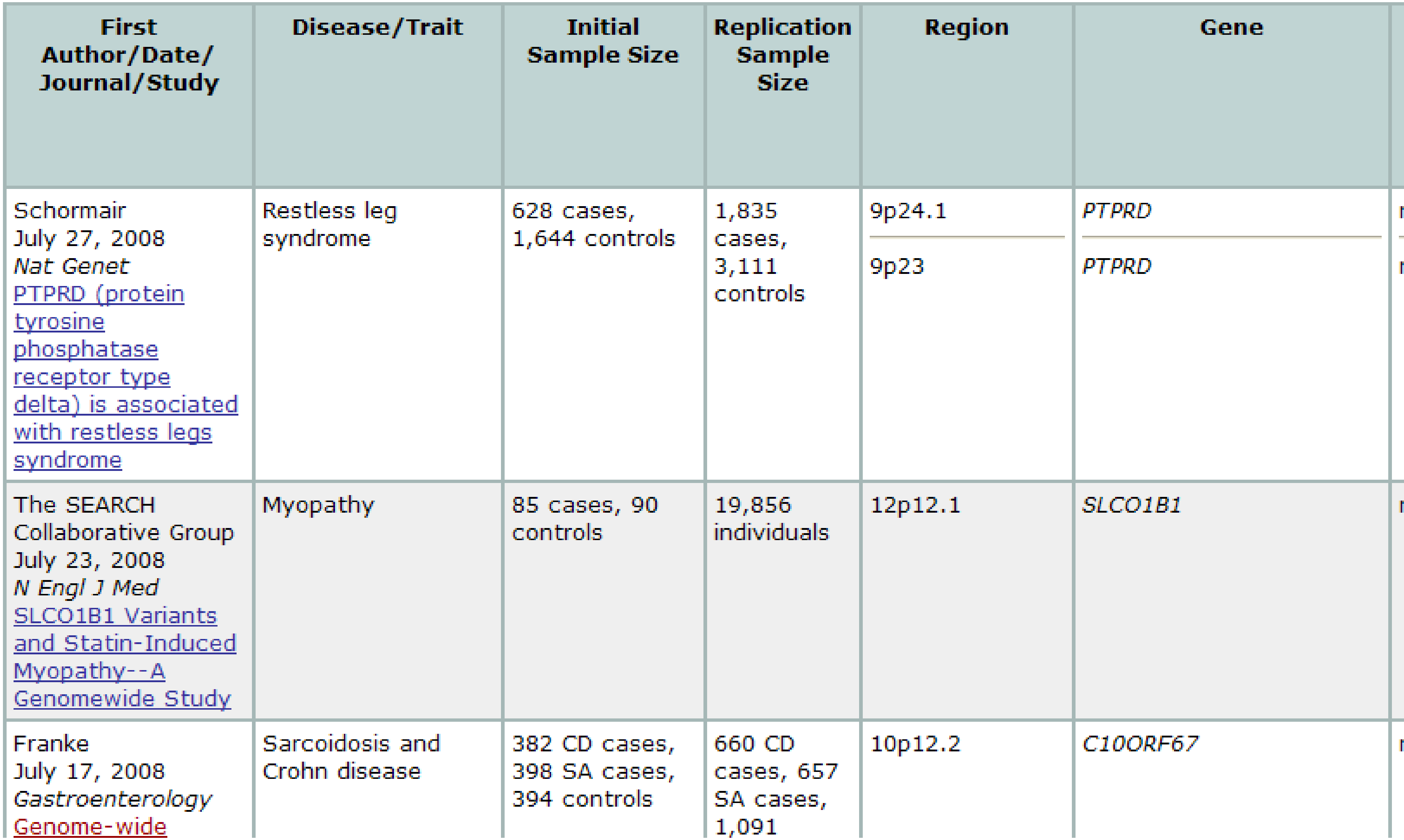
- One can get estimates of the haplotype probabilities for each individual (LD between SNPs OK)
- Put the estimated probabilities into the logistic regression

GWAS Studies

How do we keep up?

A Catalog of Published GWAS

- www.genome.gov/26525384
- Number of Studies:
  - 2005 2 – Includes Age-related Macular Degeneration
  - 2006 8
  - 2007 87
  - 2008 70 (through July 27)
- Bipolar Disorder:
  - 3 studies (1 used pooled genotypes)
  - No convincing signals



First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Gene	
Schormair July 27, 2008 *Nat Genet* PTPRD (protein tyrosine phosphatase receptor type delta) is associated with restless legs syndrome	Restless leg syndrome	628 cases, 1,644 controls	1,835 cases, 3,111 controls	9p24.1 9p23	*PTPRD* *PTPRD*	r r
The SEARCH Collaborative Group July 23, 2008 *N Engl J Med* SLCO1B1 Variants and Statin-Induced Myopathy--A Genomewide Study	Myopathy	85 cases, 90 controls	19,856 individuals	12p12.1	*SLCO1B1*	r
Franke July 17, 2008 *Gastroenterology* Genome-wide	Sarcoidosis and Crohn disease	382 CD cases, 398 SA cases, 394 controls	660 CD cases, 657 SA cases, 1,091	10p12.2	*C10ORF67*	r

“History” of GWAS

- Early studies used pooled designs – too expensive to do individual genotypes
- Affymetrix and Illumina come out with affordable SNP chips
- First study to generate enthusiasm – Age-related macular degeneration (Klein, 2007) found a “real” signal
- Type II diabetes studies found “real” signals – linkage studies were problematic

Welcome Trust (WTCCC) Initiative

- Common set of 3,000 controls
- Several disorders (including Bipolar) with 2,000 cases each
- Results in the public domain
- Published in Nature in 2007

Major U.S. GWAS Initiatives

- New NIH Policy – All NIH Funded GWAS studies must deposit individual genotypes and phenotypic data in dbGaP at NCBI
- GAIN and GEI RFAs funded studies with existing DNA, subjects consented to allow data to go to dbGaP, and genotyping done at associated genotyping centers
- New RFA from NIMH to collect very large (~10,000) samples

GAIN Proposals

Genetic Association Information Network

- 6 WGA projects were selected across NIH
- Projects:
  - Schizophrenia
  - Bipolar Disorder
  - Depression
  - ADHD
  - Psoriasis
  - Type 1 Diabetes (nephropathy)
- Data at dbGap (1 year embargo on publication)
- Note: 4/6 Mental Health related!!

Gene Environment Initiative (GEI)

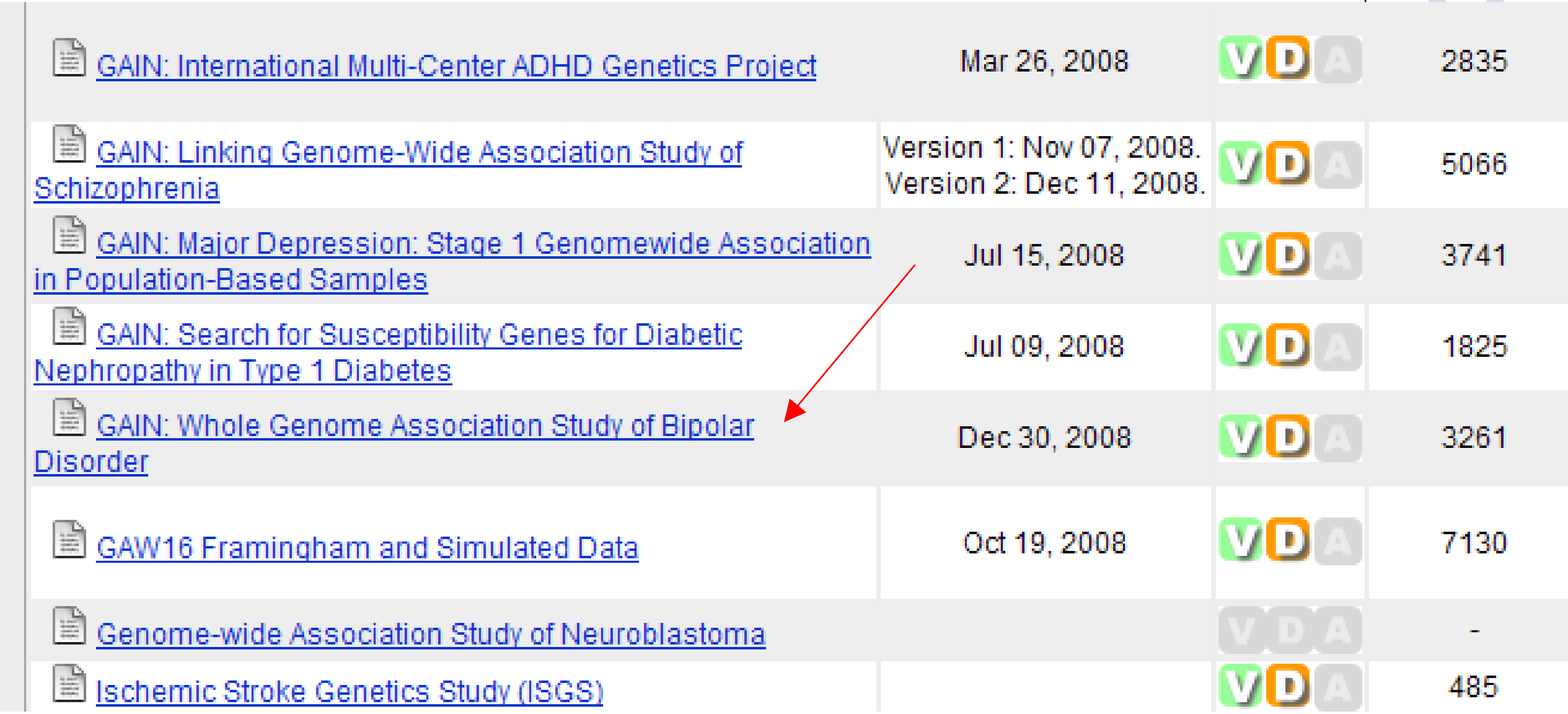
- 8 GWAS funded – oral cleft, addiction, coronary heart disease, lung cancer, type 2 diabetes, birth weight, dental caries, premature birth
- Required existing DNA and subjects consented to share
- Issued Supplement for replication samples
- Addiction (Bierut) samples genotyped first – we got genotypes from CIDR in May; once cleaned, they go to dbGaP

Good News for Analysts

- Cleaned data available goes to investigators who collected data at the same time as everyone else
- It takes years to collect subjects
- Cleaning GWAS data is hard and time consuming
- Opportunity for combining data from multiple studies
- Is this fair?

dbGaP

- Genotype and Phenotype Database
- Data made available to investigators and others at the same time – 1 year publication embargo
- Request access using eRA Commons sign on – requires Institutional sign-off
- Request must be approved by a DAC (data access committee)



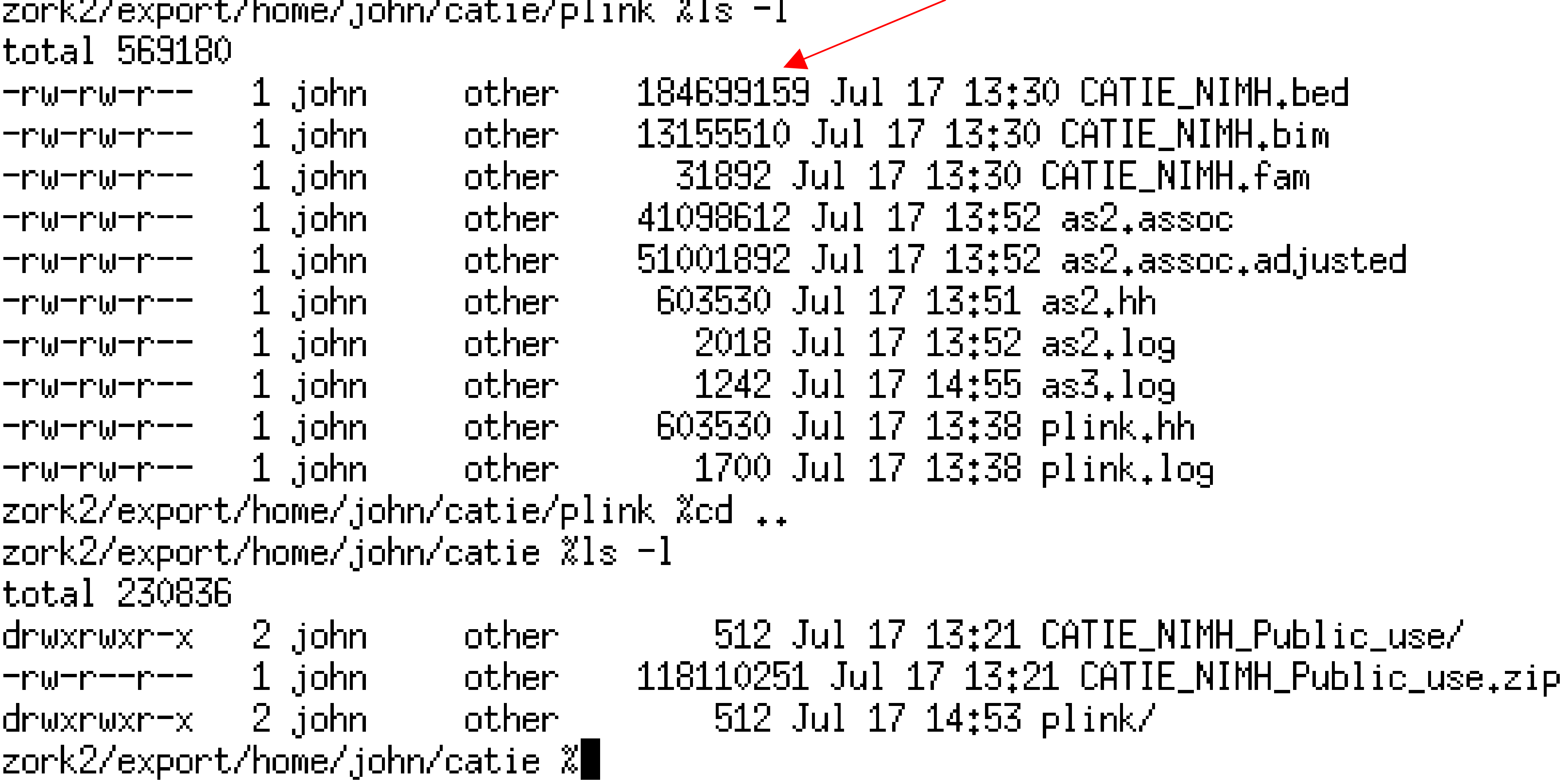
GAIN: International Multi-Center ADHD Genetics Project	Mar 26, 2008	V D A	2835
GAIN: Linking Genome-Wide Association Study of Schizophrenia	Version 1: Nov 07, 2008. Version 2: Dec 11, 2008.	V D A	5066
GAIN: Major Depression: Stage 1 Genomewide Association in Population-Based Samples	Jul 15, 2008	V D A	3741
GAIN: Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes	Jul 09, 2008	V D A	1825
GAIN: Whole Genome Association Study of Bipolar Disorder	Dec 30, 2008	V D A	3261
GAW16 Framingham and Simulated Data	Oct 19, 2008	V D A	7130
Genome-wide Association Study of Neuroblastoma		V D A	-
Ischemic Stroke Genetics Study (ISGS)		V D A	485

Some statistical and data management issues

- Genomic Inflation Factor
- We illustrate with admixed schizophrenia data (CATIE) where we don’t control for ethnicity

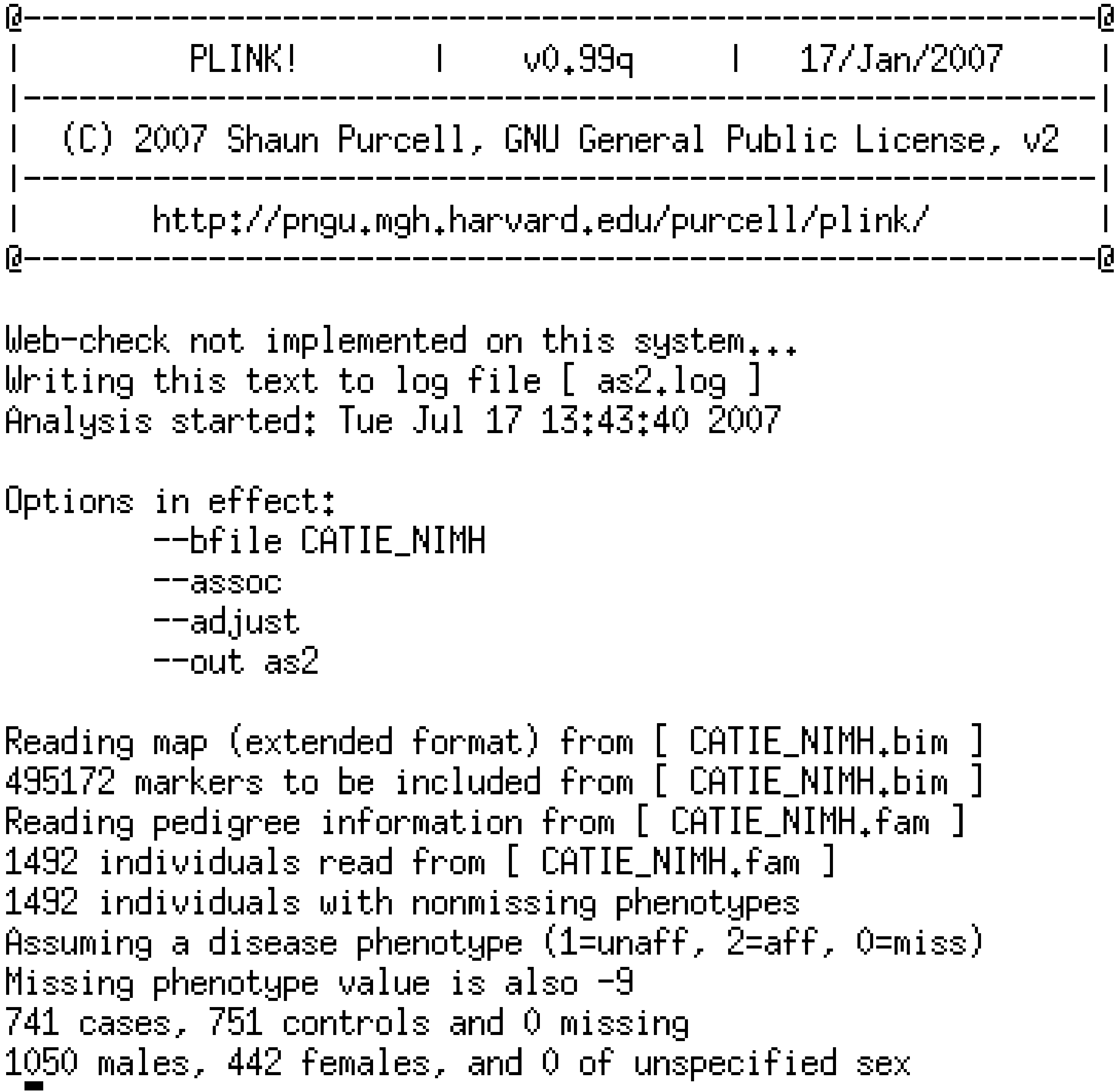
Genomic inflation factor -- lambda

- When testing 300K to 1M SNPs, most tests are under the null
- Median chi-square should be .445
- Lambda = median chi-sq/.445
- Can use lambda to correct chi-sqs for this inflation
- Better – look for source (eg, ethnic admixture), and correct for that



```
zork2/export/home/john/catie/plink %ls -l
total 569180
-rw-rw-r--   1 john     other    184699159 Jul 17 13:30 CATIE_NIMH.bed
-rw-rw-r--   1 john     other    13155510 Jul 17 13:30 CATIE_NIMH.bim
-rw-rw-r--   1 john     other      31892 Jul 17 13:30 CATIE_NIMH.fam
-rw-rw-r--   1 john     other    41098612 Jul 17 13:52 as2.assoc
-rw-rw-r--   1 john     other    51001892 Jul 17 13:52 as2.assoc.adjusted
-rw-rw-r--   1 john     other     603530 Jul 17 13:51 as2.hh
-rw-rw-r--   1 john     other       2018 Jul 17 13:52 as2.log
-rw-rw-r--   1 john     other       1242 Jul 17 14:55 as3.log
-rw-rw-r--   1 john     other     603530 Jul 17 13:38 plink.hh
-rw-rw-r--   1 john     other       1700 Jul 17 13:38 plink.log
zork2/export/home/john/catie/plink %cd ..
zork2/export/home/john/catie %ls -l
total 230836
drwxrwxr-x   2 john     other        512 Jul 17 13:21 CATIE_NIMH_Public_use/
-rw-r--r--   1 john     other    118110251 Jul 17 13:21 CATIE_NIMH_Public_use.zip
drwxrwxr-x   2 john     other        512 Jul 17 14:53 plink/
zork2/export/home/john/catie %
```

Unzipped (binary) file is 185MB

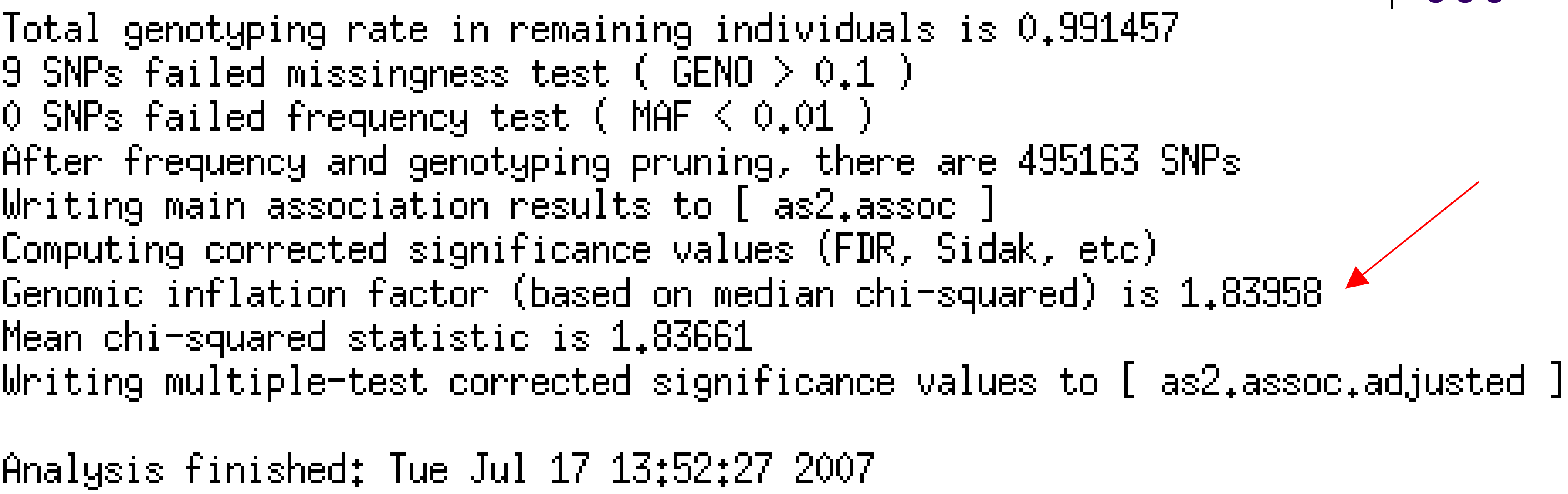


```
@----------------------------------------------------------@
|        PLINK!       |     v0.99q      |   17/Jan/2007      |
|----------------------------------------------------------|
|  (C) 2007 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|       http://pngu.mgh.harvard.edu/purcell/plink/         |
@----------------------------------------------------------@

Web-check not implemented on this system...
Writing this text to log file [ as2.log ]
Analysis started: Tue Jul 17 13:43:40 2007

Options in effect:
        --bfile CATIE_NIMH
        --assoc
        --adjust
        --out as2

Reading map (extended format) from [ CATIE_NIMH.bim ]
495172 markers to be included from [ CATIE_NIMH.bim ]
Reading pedigree information from [ CATIE_NIMH.fam ]
1492 individuals read from [ CATIE_NIMH.fam ]
1492 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
741 cases, 751 controls and 0 missing
1050 males, 442 females, and 0 of unspecified sex
```



```
Total genotyping rate in remaining individuals is 0.991457
9 SNPs failed missingness test ( GENO > 0.1 )
0 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 495163 SNPs
Writing main association results to [ as2.assoc ]
Computing corrected significance values (FDR, Sidak, etc)
Genomic inflation factor (based on median chi-squared) is 1.83958
Mean chi-squared statistic is 1.83661
Writing multiple-test corrected significance values to [ as2.assoc.adjusted ]

Analysis finished: Tue Jul 17 13:52:27 2007
```

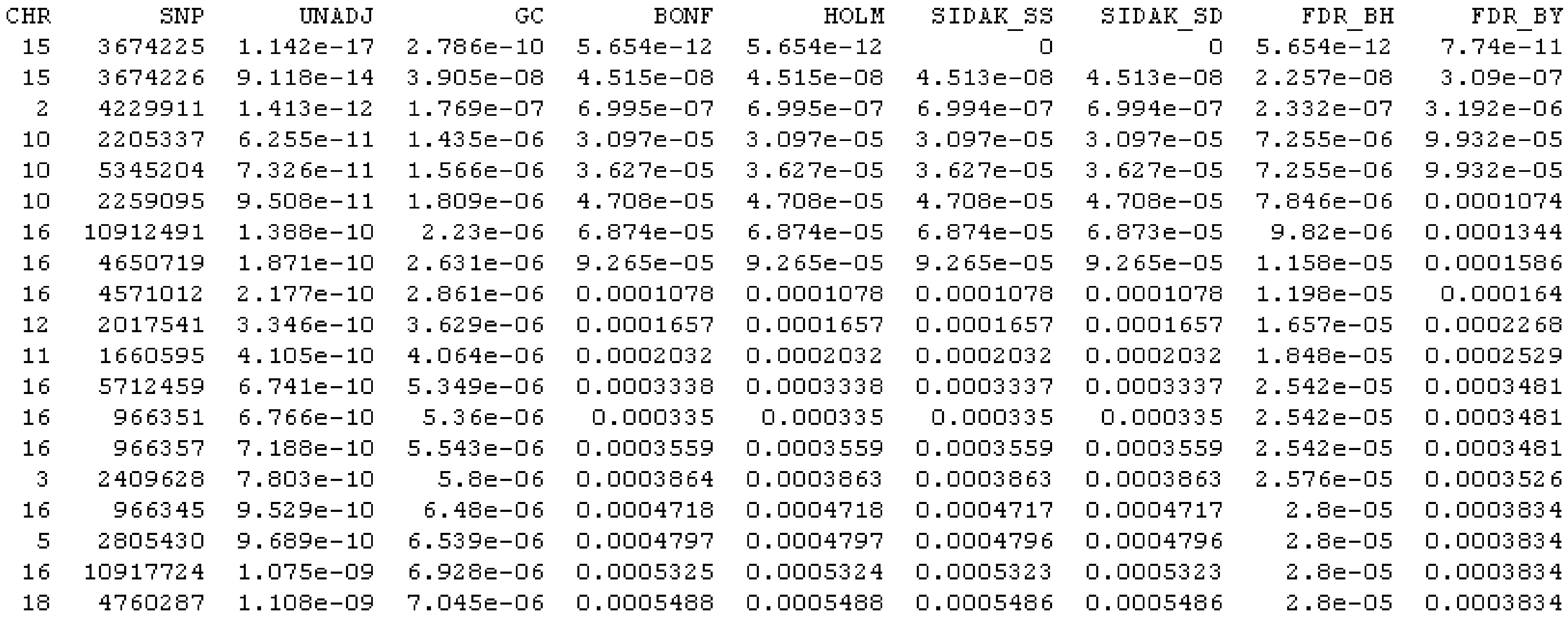
495,163 SNPs Analyzed

Total Time: 9 min!

Terrible lambda

Note: Mixture of EU and AAs

Plink Output



```
CHR         SNP       UNADJ          GC        BONF        HOLM    SIDAK_SS    SIDAK_SD      FDR_BH      FDR_BY
 15     3674225   1.142e-17   2.786e-10   5.654e-12   5.654e-12           0           0   5.654e-12    7.74e-11
 15     3674226   9.118e-14   3.905e-08   4.515e-08   4.515e-08   4.513e-08   4.513e-08   2.257e-08    3.09e-07
  2     4229911   1.413e-12   1.769e-07   6.995e-07   6.995e-07   6.994e-07   6.994e-07   2.332e-07   3.192e-06
 10     2205337   6.255e-11   1.435e-06   3.097e-05   3.097e-05   3.097e-05   3.097e-05   7.255e-06   9.932e-05
 10     5345204   7.326e-11   1.566e-06   3.627e-05   3.627e-05   3.627e-05   3.627e-05   7.255e-06   9.932e-05
 10     2259095   9.508e-11   1.809e-06   4.708e-05   4.708e-05   4.708e-05   4.708e-05   7.846e-06   0.0001074
 16    10912491   1.388e-10    2.23e-06   6.874e-05   6.874e-05   6.874e-05   6.873e-05    9.82e-06   0.0001344
 16     4650719   1.871e-10   2.631e-06   9.265e-05   9.265e-05   9.265e-05   9.265e-05   1.158e-05   0.0001586
 16     4571012   2.177e-10   2.861e-06   0.0001078   0.0001078   0.0001078   0.0001078   1.198e-05    0.000164
 12     2017541   3.346e-10   3.629e-06   0.0001657   0.0001657   0.0001657   0.0001657   1.657e-05   0.0002268
 11     1660595   4.105e-10   4.064e-06   0.0002032   0.0002032   0.0002032   0.0002032   1.848e-05   0.0002529
 16     5712459   6.741e-10   5.349e-06   0.0003338   0.0003338   0.0003337   0.0003337   2.542e-05   0.0003481
 16      966351   6.766e-10    5.36e-06    0.000335    0.000335    0.000335    0.000335   2.542e-05   0.0003481
 16      966357   7.188e-10   5.543e-06   0.0003559   0.0003559   0.0003559   0.0003559   2.542e-05   0.0003481
  3     2409628   7.803e-10     5.8e-06   0.0003864   0.0003863   0.0003863   0.0003863   2.576e-05   0.0003526
 16      966345   9.529e-10    6.48e-06   0.0004718   0.0004718   0.0004717   0.0004717      2.8e-05   0.0003834
  5     2805430   9.689e-10   6.539e-06   0.0004797   0.0004797   0.0004796   0.0004796      2.8e-05   0.0003834
 16    10917724   1.075e-09   6.928e-06   0.0005325   0.0005324   0.0005323   0.0005323      2.8e-05   0.0003834
 18     4760287   1.108e-09   7.045e-06   0.0005488   0.0005488   0.0005486   0.0005486      2.8e-05   0.0003834
```

P-values

- Uncleaned, admixed data – small p-values are an artifact.
- Welcome Trust used significance level of 5×10^{-7} based an Bayesian arguments
- Bonferroni correction assumes independent tests
- PLINK also computes q-values based on FDR (false discovery rate)

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 $p_i \le (i/m)\, q^*$
  Then reject $H_1, \ldots H_m$

Theorem: Above controls FDR at q*

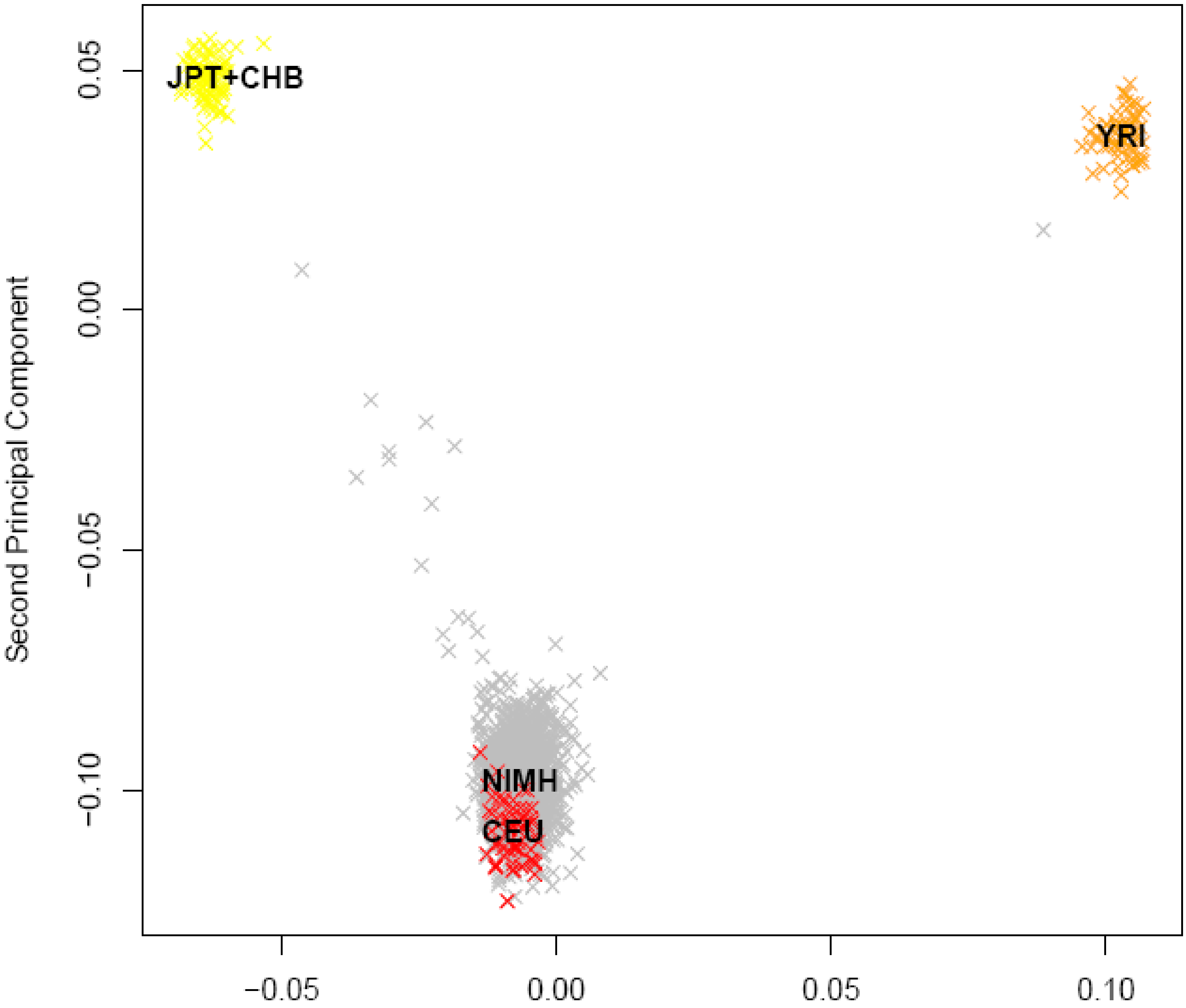
Computer program: QVALUE; computed by PLINK

Interpretation of FDR

- If q-value is 0.1, 1/10 is false positive.
- If we identify 10 SNPs and 9 are real and 1 is false positive – major success.
- Usual experiment-wise error (Bonferroni correction) only one false positive at the chosen p-value.

Some statistical and data management issues

- Population stratification
- Perform principal components analysis (10,000 markers probably enough), and plot your samples along with hapmap samples
- Eigenstrat is commonly used
- We illustrate with NIMH repository control data who self report as “white”


JPT+CHB
YRI
NIMH
CEU
Second Principal Component
0.05
0.00
−0.05
−0.10
−0.05
0.00
0.05
0.10

Problem Samples (to be removed)

- One subject clusters with Yoruba sample
- A handful of subjects trail off to Asian sample. Some reported American Indian ancestry
- In addition, several samples had phenotypic sex differ from genetic sex – probably sample swaps

Cleaning of GENEVA addiction GWAS data (SAGE)

- 1 million Illumina chips were done at CIDR
- Data should be at dbGaP in a few weeks
- We just completed cleaning, but haven’t received the final data

Study Design

- Case/ Control (4,400 individuals)
- Samples come from 3 studies
  - Alcohol Dependence (COGA)
  - Nicotine Dependence (COGEND)
  - Cocaine Dependence (FSCD)
- Cases have a diagnosis of alcohol dependence
- Controls do not have a dx of alc, nic, or cocaine dependence; must have drunk alcohol
- Mixture of EUs, AAs and Hispanics

Primary Model

- Dependent variable (s)
  - Case control status (diagnosis of alcohol dependence)—simple logistic model
- Independent variables
  - Genotype --(1 df trend test)
  - EU vs AA vs Hispanic (Asians, Mixed, etc excluded)
  - Study (alc, cocaine, nicotine)
  - Gender
- Test each SNP with 1 df

Relatedness

- Identify unexpected relatedness, correct pedigree and identify one representative from each family
- Use IBD – Identity by Descent
- Two individuals can share 0, 1 or 2 alleles from a common ancestor
- MZ twins (or duplicates) always share 2 alleles IBD; Parent-offspring pairs always share 1 allele IBD, etc.
- PLINK can estimate these probabilities from the SNP data (which is IBS data since parents are not genotyped)

Prob of IBD by Relationship

Z2	Z1	Z0	kinship	Relationship		
1	0	0	0.5	MZ twin (or duplicate)		
0	1	0	0.25	parent-offspring		
0.25	0.5	0.25	0.25	full siblings		
0	0.5	0.5	0.125	half siblings		
0	0.5	0.5	0.125	avuncular (uncle/aunt - niece/nephe		
0	0.5	0.5	0.125	grandparent-grandchild		
0	0.25	0.75	0.0625	great grandparent - great grandchild		

We found "unexpected" relatedness

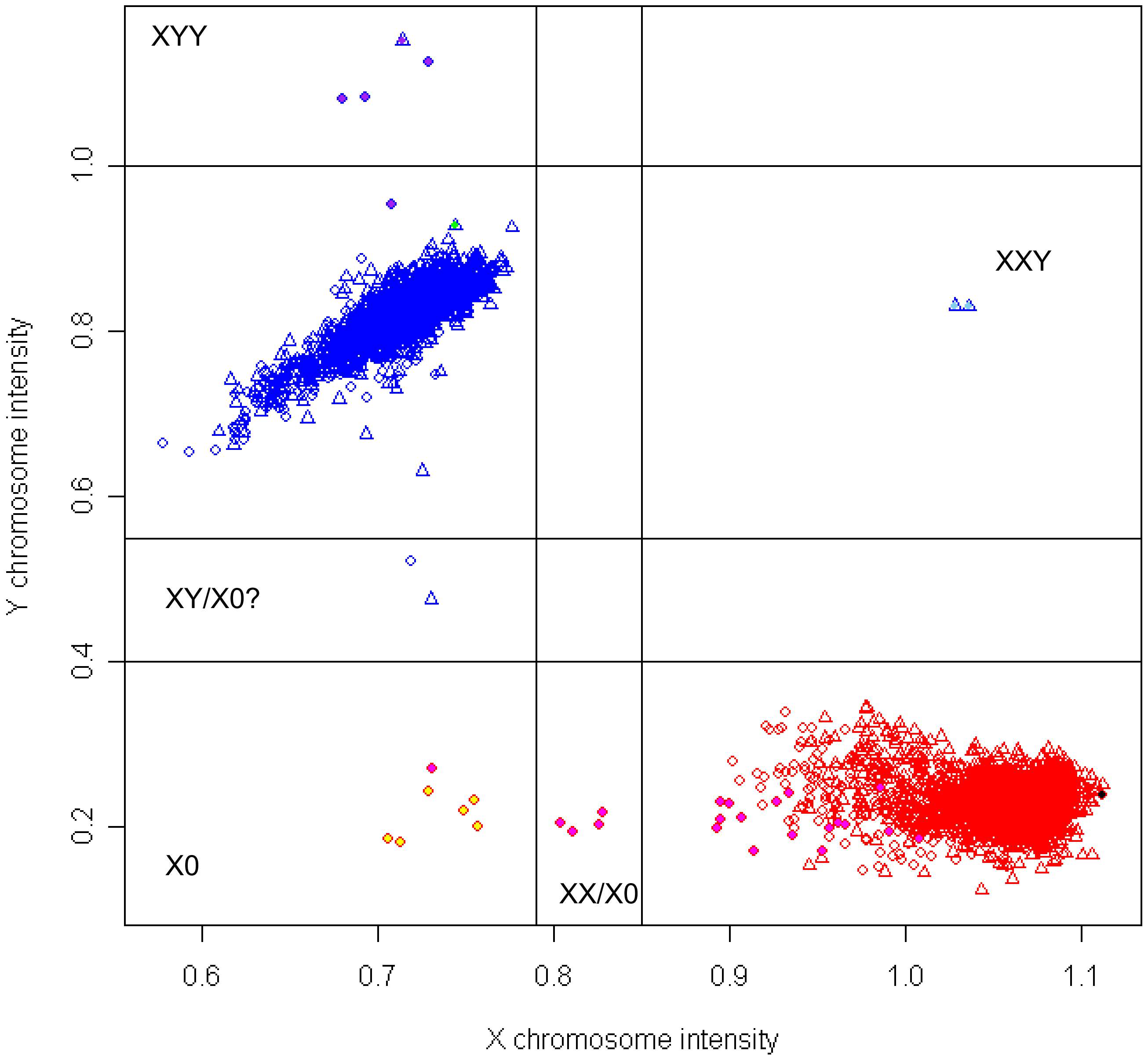
- Duplicates:
  - 8 subjects were both in FSCD and COGA
  - This will be documented by dbGaP
- Some full sibs were selected for SAGE and were known – Others were identified in cleaning
- Other unexpected relatedness found
- Data from "extra" samples will be distributed by dbGaP

Aneuploidy

- Normal male – XY; Normal Female – XX
- Phenotypically male if at least one Y chromosome
- Found XXY (male who genotypes like a female), XYY, XO individuals, mosaics
- Most of this is due to DNA from cell lines
- Some detected by looking at intensity plots

CIDR X0/XX=magenta, XYY=purple, XXY=skyblue, X0=yellow, XXX=black, XY/XXY/XYY=green

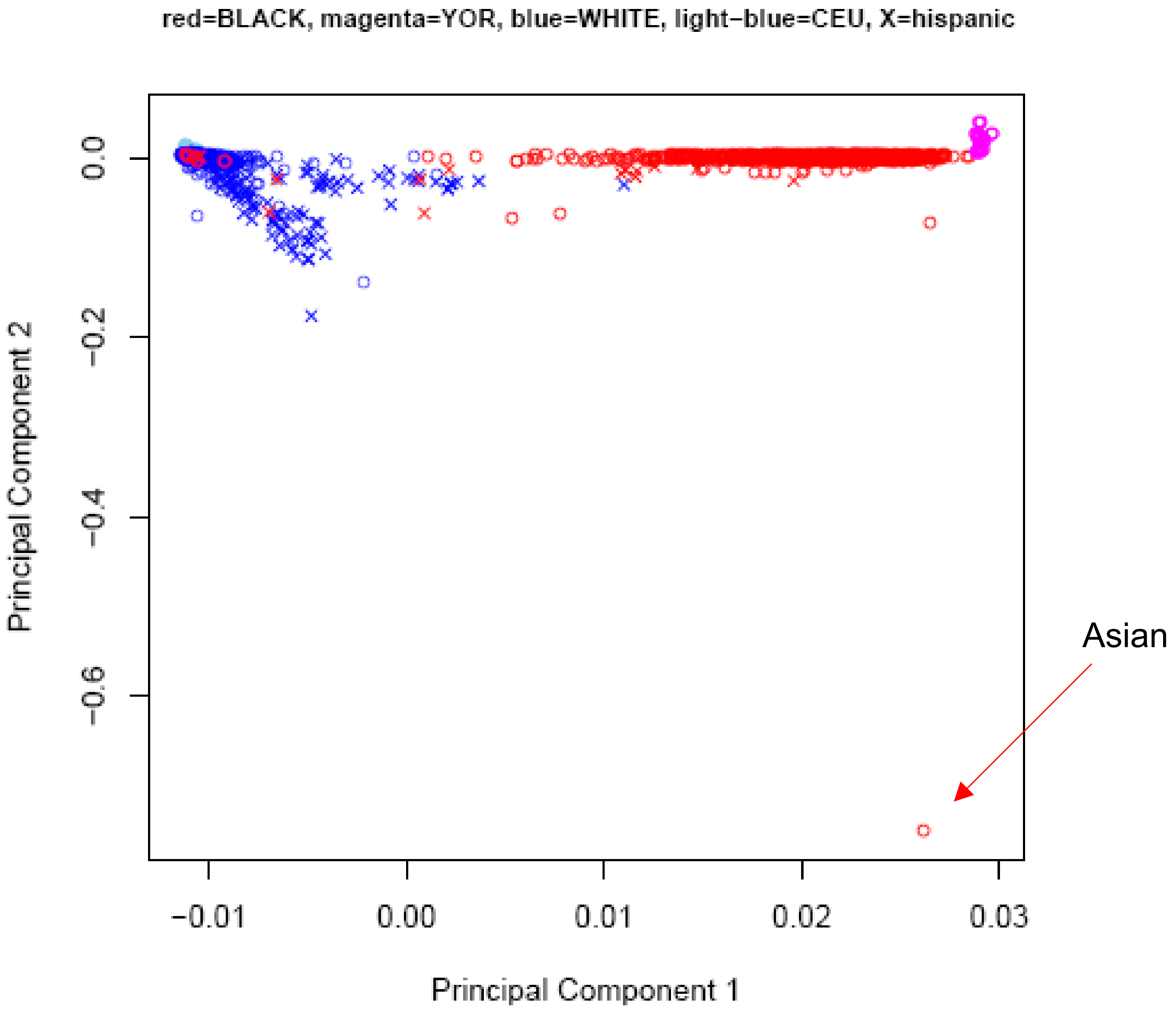
red=F, blue=M, circle=cell line, triangle=blood



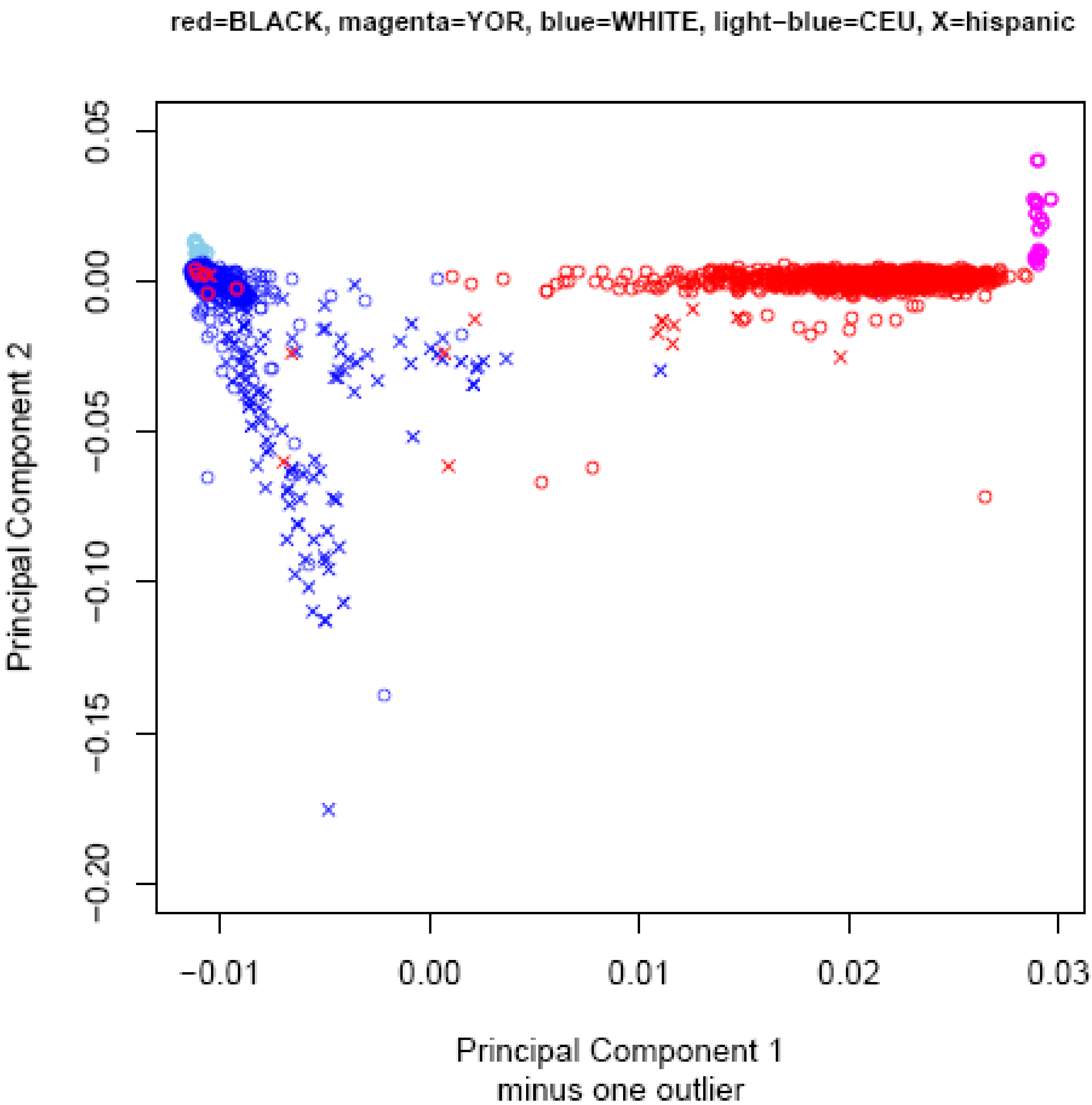
CIDR X0/XX=magenta, XYY=purple, XXY=skyblue, XO=yellow, XXX=black, XY/XXY/XYY=green

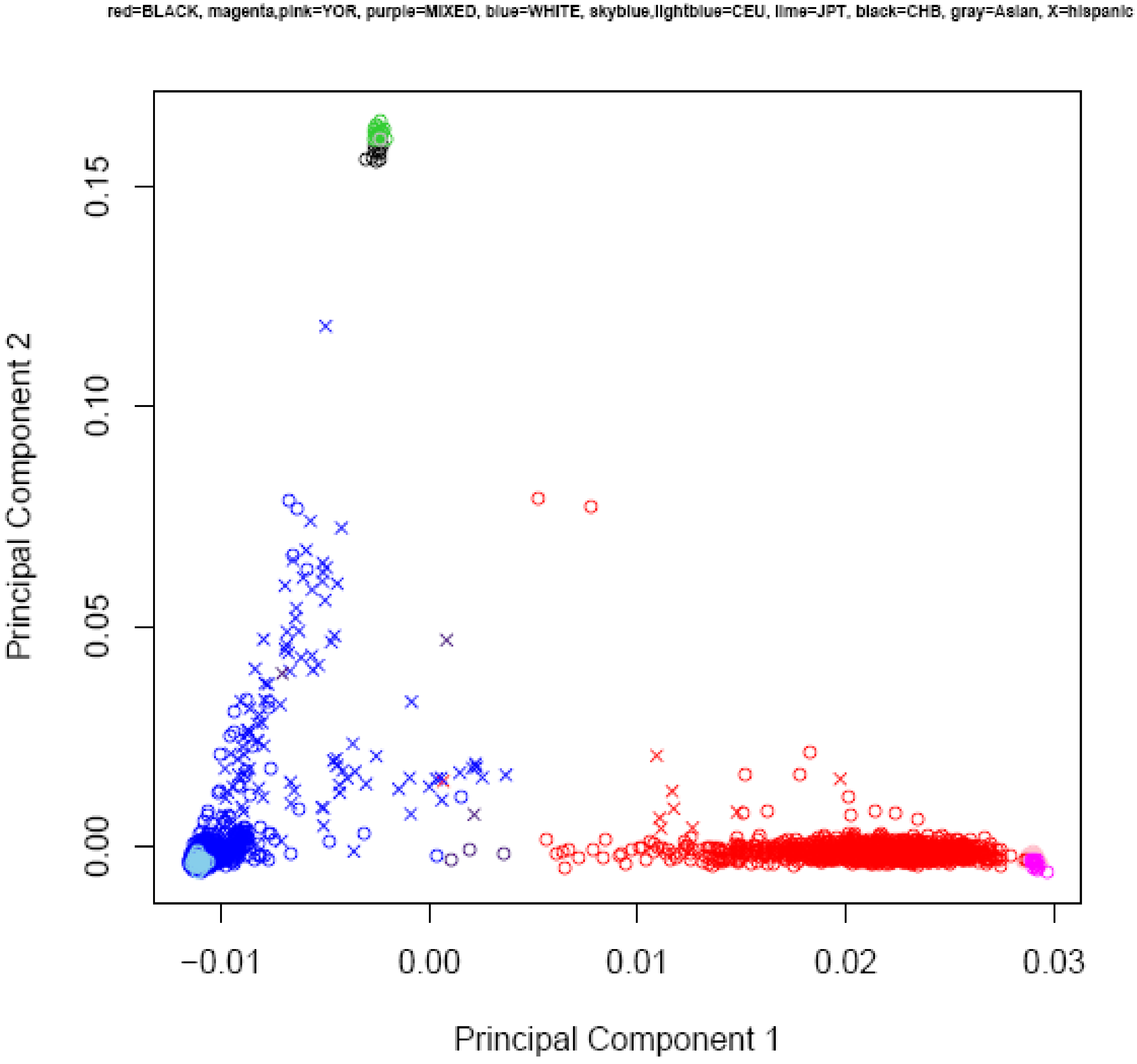
Population structure

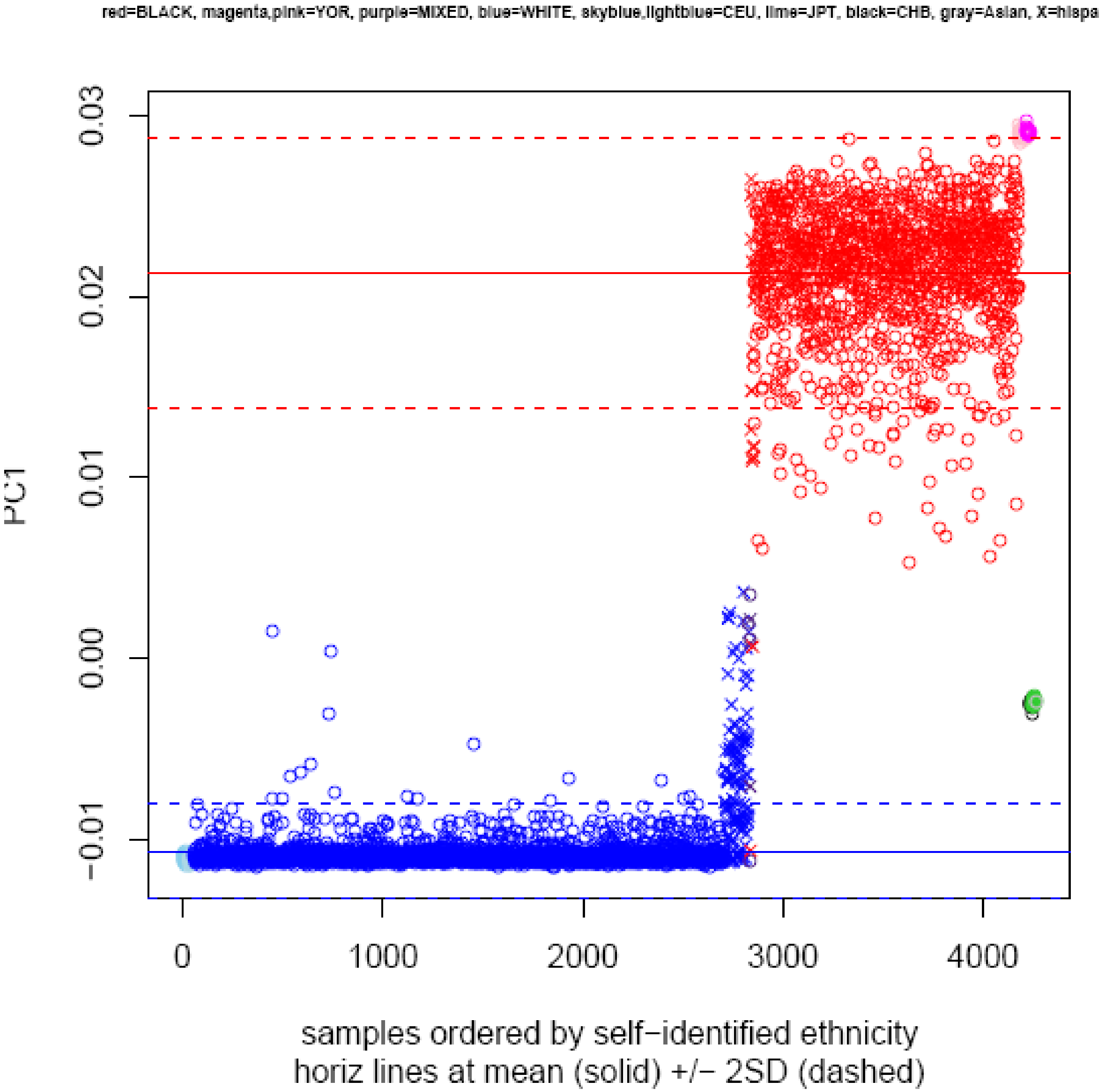
- Assign samples to population groups for allele frequency estimation, HW testing, etc.
- Alternatively, produce quantitative covariates to control for population admixture
- Use the program Eigenstrat to perform Principal Component Analysis



red=BLACK, magenta=YOR, blue=WHITE, light−blue=CEU, X=hispanic
Principal Component 2
0.0
−0.2
−0.4
−0.6
−0.01
0.00
0.01
0.02
0.03
Principal Component 1
Asian


red=BLACK, magenta=YOR, blue=WHITE, light−blue=CEU, X=hispanic
Principal Component 2
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.01
0.00
0.01
0.02
0.03
Principal Component 1
minus one outlier


red=BLACK, magenta,pink=YOR, purple=MIXED, blue=WHITE, skyblue,lightblue=CEU, lime=JPT, black=CHB, gray=Asian, X=hispanic
Principal Component 2
0.00
0.05
0.10
0.15
-0.01
0.00
0.01
0.02
0.03
Principal Component 1


red=BLACK, magenta,pink=YOR, purple=MIXED, blue=WHITE, skyblue,lightblue=CEU, lime=JPT, black=CHB, gray=Asian, X=hispa
PC1
0.03
0.02
0.01
0.00
-0.01
0
1000
2000
3000
4000
samples ordered by self-identified ethnicity
horiz lines at mean (solid) +/- 2SD (dashed)

Admixture

- First PC separates EUs and AAs
- Second PC separates Hispanics
- Some self reported ethnicities were in error and turned out to be data entry mistakes
- One “unexpected” Asian was found

Hardy-Weinberg Equilibrium

Hardy, Godfrey Harold
(1877-1947)

Four greatest wishes: (1) to prove the Riemann Hypothesis ∑, (2) to make a brilliant play in a crucial cricket match, (3) to prove the non-existence of God, (4) to murder Mussolini.

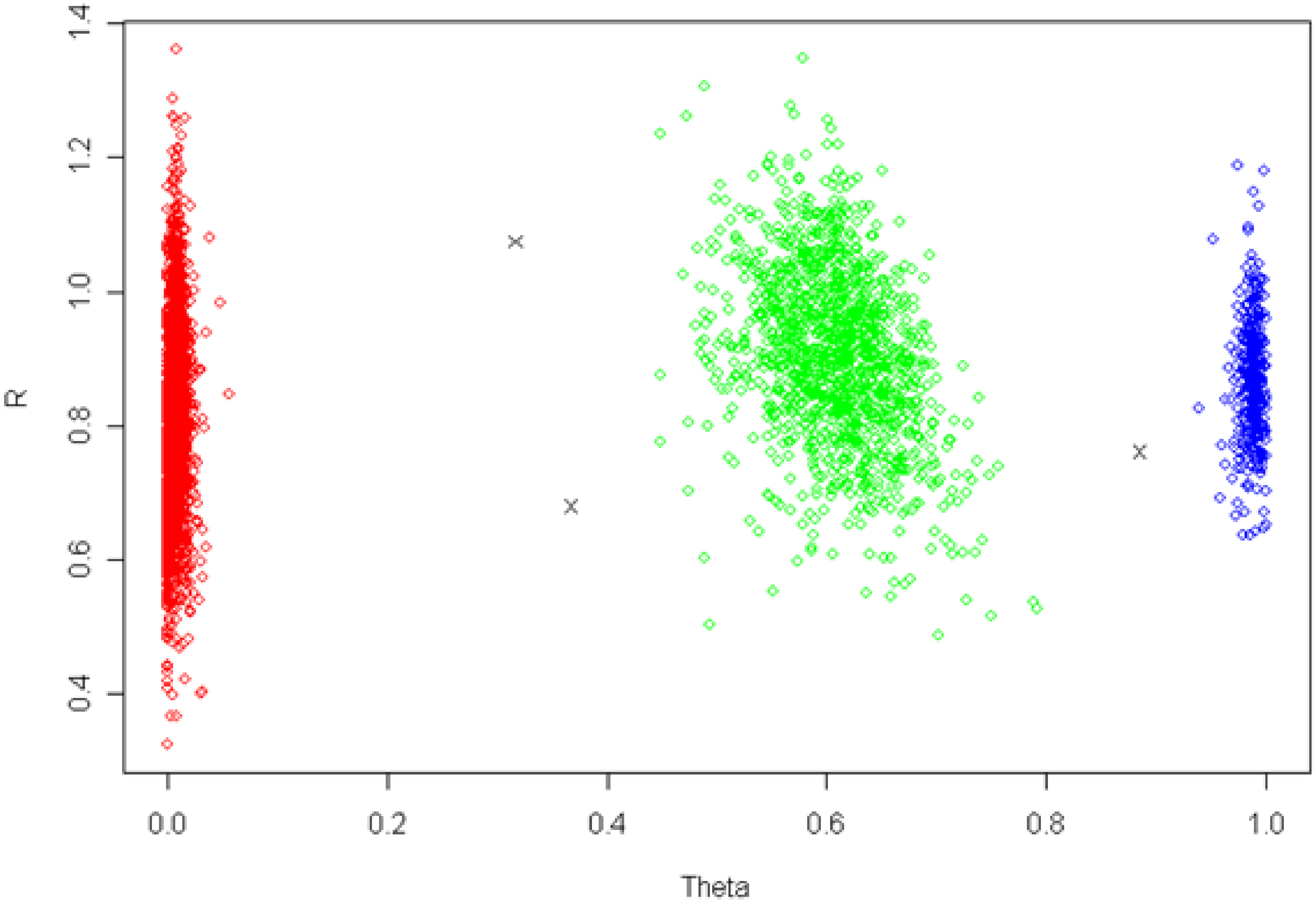
HWE

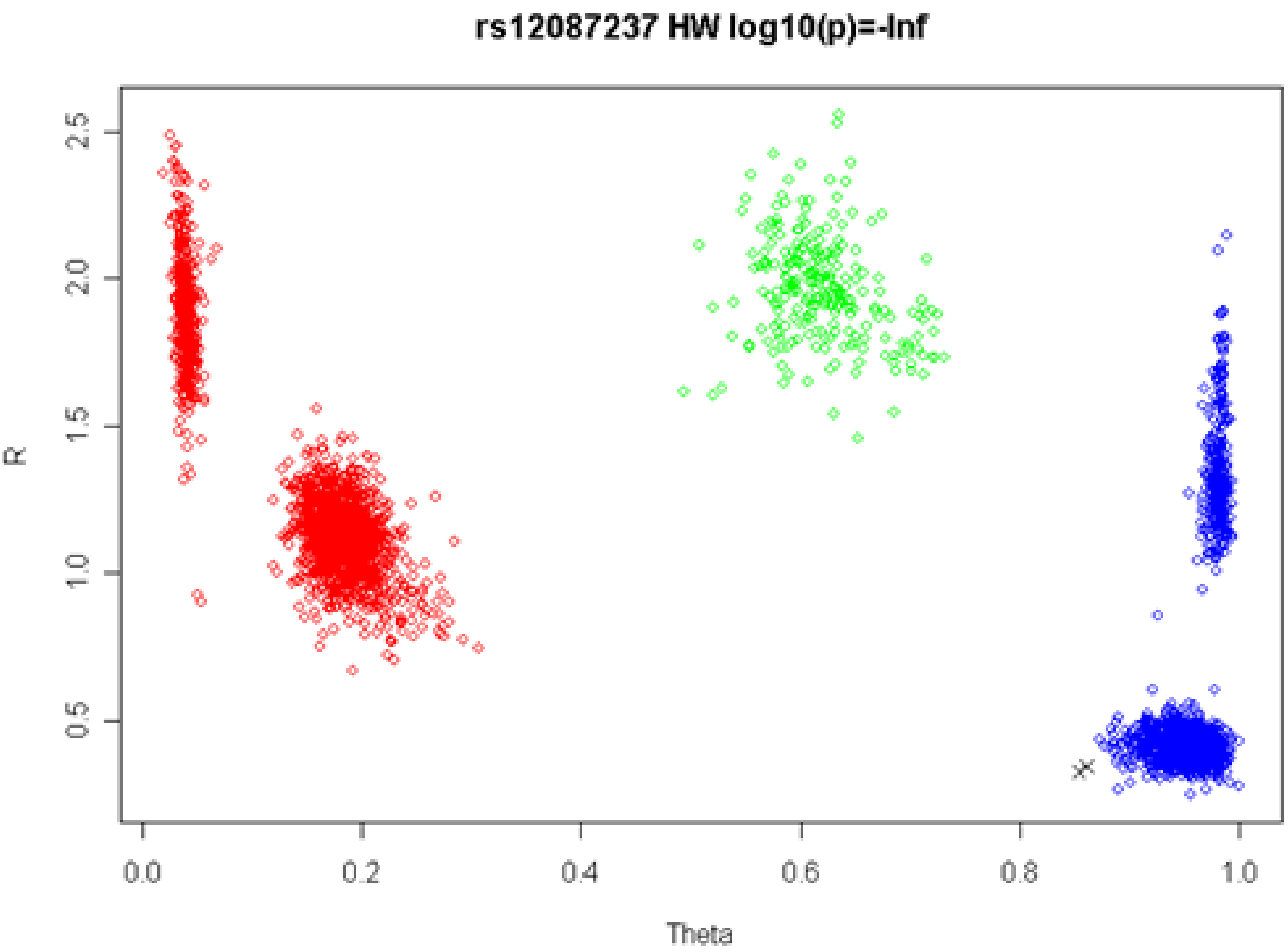
- Let a SNP have two alleles 1,2 with frequencies p and q =1 – p, respectively.
- The SNP is in HWE if the genotypic frequencies are p^2, 2pq, and q^2 for genotypes 11, 12, 22.
- Hardy and Weinberg showed a population reaches HWE in a single generation of random mating.
- Usually see HWE for markers.

HWE

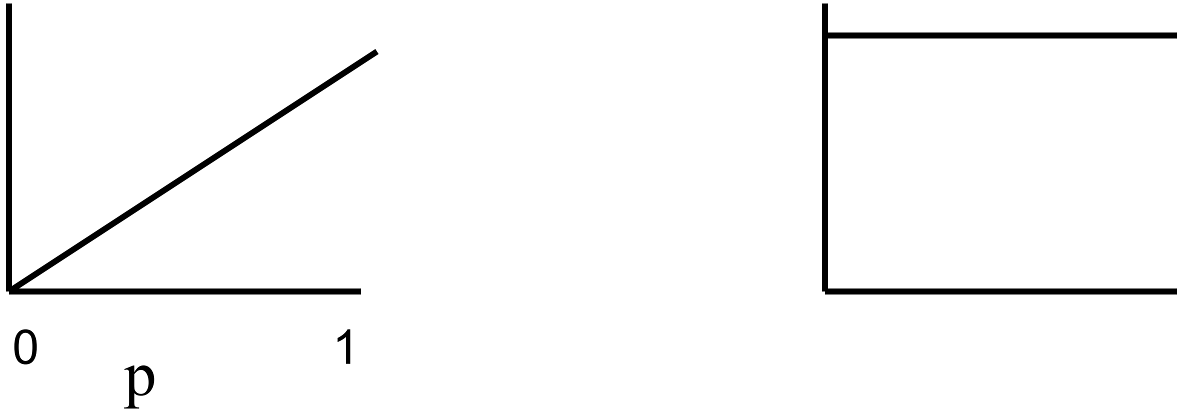
- Filter out SNPs with $p < 10^{-06}$ when testing for HWE
- Note: test done separately within ethnic groups – mixing populations with different allele frequencies leads to non-HWE
- CNVs (copy number variations) can cause non-HWE
- Bottom line – always inspect intensity plots for signals of interest.

Intensity Plot – good SNP




rs12087237 HW log10(p)=-Inf
R
2.5
2.0
1.5
1.0
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Theta

Uniform Distribution

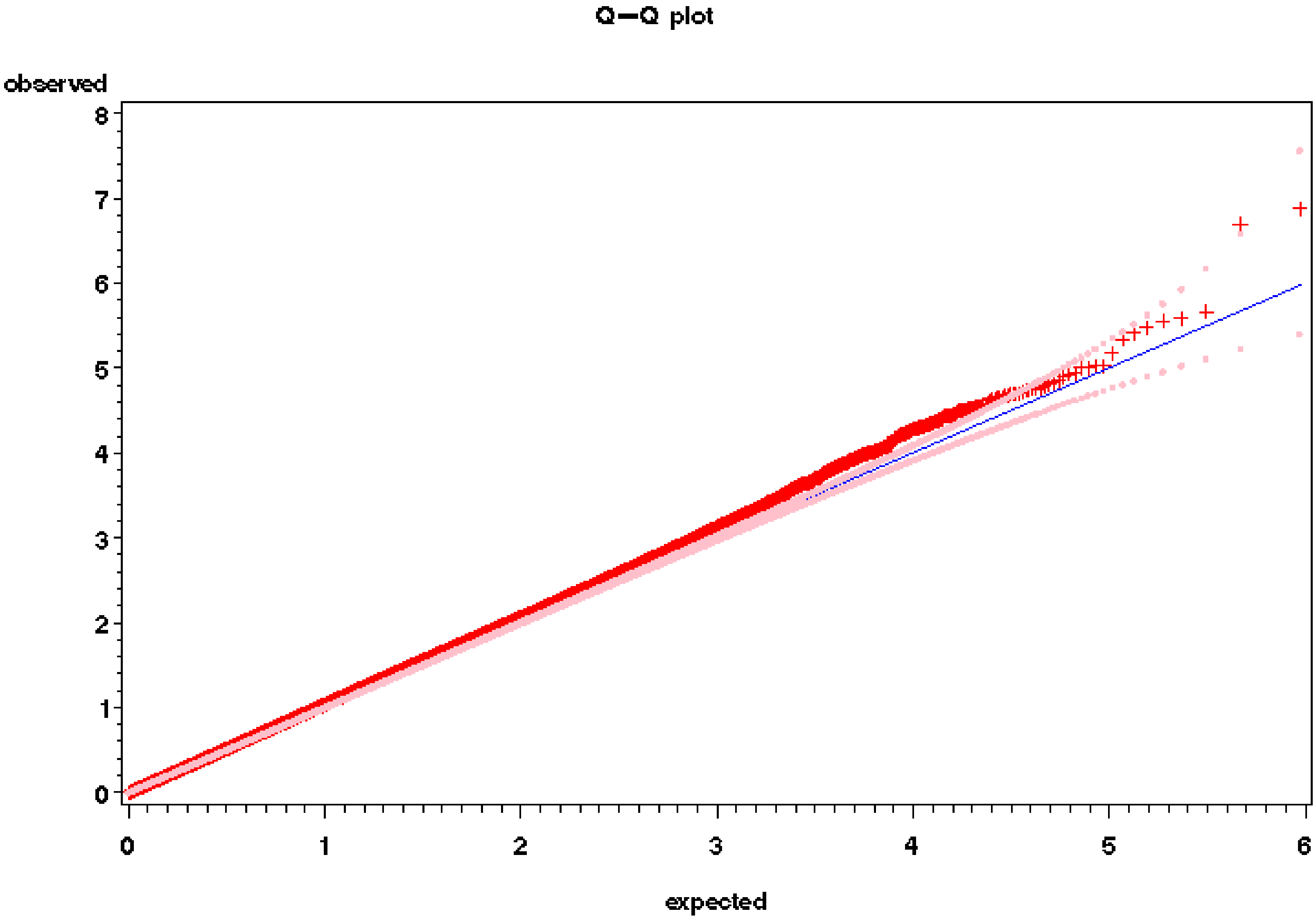


If we perform N independent statistical tests for which all null Hypotheses are true, we expect a uniform distribution.

QQ-plot of association test

- When we test 1 million SNPs, most are not truly associated. Plot - log(p) for observed tests against a uniform distribution as a final check
- Genomic inflation factor – If using a chi-square test with 1 df, median value should be 0.445. λ=observed median / .445. Usually correct chi-sq by dividing by λ
- Always best to control for pop admixture, eliminate CNVs, etc first

GEI/GENEVA: Bierut Addiction



$\lambda = 1.045$

GEVEVA Acknowledgement

- U. Washington
  - Bruce Weir, Thomas Lumley, Ken Rice, Tushar Bhangale, Xiuwen Zheng, Ian Painter, Fred Boehm, CathyLaurie
- CIDR
  - Kim Doheny, Elizabeth Pugh, Kurt Hetrick.
- NCBI
  - Justin Pashall, Mike Feolo, Stephanie Pretel
- Washington U.
  - Laura Bierut, John Rice, Nancy Saccone, Sherri Fisher
- NHGRI
  - Emily Harris, Teri Manolio

Conclusions

- GWAS has already been successful for many complex traits – linkage has not been
- Many GWAS are in progress
- We use plink and SAS for data management, data cleaning and analysis
- The only way to learn this is to really be involved in one
- Availability at dbGaP is a major event –

“can’t herd cats, but you can move their food”

Final Words

- Current GWAS – Chi-Square on steroids
- Only pick low fruit – genome-wide significant; test one SNP at a time
- How to identify true signals mixed in with noise due to chance?
- How to identify gene-gene interactions and G x E interactions?
- Where is the heritability of 50-80%?