Estimating experiment-wise error rates by permutation

ESP 2009

Bertram Müller-Myhsok

Multipe testing problem in Genome-wide Analysis

– Many tests performed

– 100,000 SNPs
– 300,000 SNPs
– 500,000 SNPs
– 1,000,000 SNPs

Multiple testing in GWA cont‘d

- Often several phenotypes
- Often several genetic models

- Example:
  - 10 phenotypes
  - 4 models
  - 500,000 SNPs

Multiple testing in GWA cont‘d

- Number of tests then:
  - 10 * 4 * 500,000 = 20 million = 2e+07
- Necessary significance level for experiment-wise p of 0.05 using Bonferroni:
  - 0.05 / 2e+07 = 2.5e-09

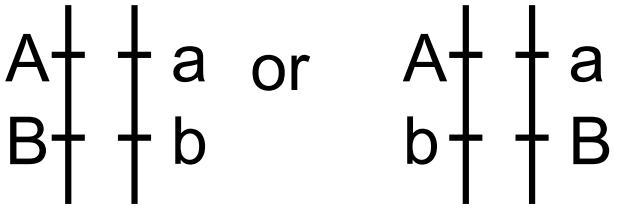
But:

- Tests are not necessarily independent
  - LD between SNPs
  - Correlation between phenotypes
  - Correlation between genotypes
- The effective number of tests is less than the number of tests performed

LD: From Genotypes to Haplotypes

	BB	Bb	bb
AA	AABB	AABb	Aabb
Aa	AaBB	AaBb	Aabb
aa	aaBB	aaBb	aabb

Joint genotype AaBb



Haplotypes

- For unrelated individuals – Haplotypes can be reconstructed with 100% accuracy if:
  - All loci are homozygous
  - No more than 1 locus is heterozygous
- If more 2 or more loci are heterozygous then
  - The haplotypes are ambiguous
  - However usually one set (pair) of haplotypes is more likely

Haplotype Reconstruction

- Haplotype construction is not robust over large genetic distances
- Beware of using “best” haplotype configuration
  - It is important to take account of uncertainty in phase assignment if haplotypes are to be used in subsequent analysis
- The ability to estimate accurate haplotype assignments is dependent on the size of data set.
  - Generally smaller data sets give less accurate results

Linkage Equilibrium

- Alleles in random association are said to be in linkage equilibrium.
  - Where the gametic frequencies are:
    - A_1B_1: $p_1 \times q_1$
    - A_1B_2: $p_1 \times q_2$
    - A_2B_1: $p_2 \times q_1$
    - A_2B_2: $p_2 \times q_2$
- Eg for marker A $p_1=0.2$ and $p_2=0.8$ and for marker B $q_1=0.6$ and $q_2=0.4$
  - Under linkage equilibrium for this example the expected frequencies are:
    - $A_1B_1 = Pe_{11} = 0.12$
    - $A_1B_2 = Pe_{12} = 0.08$
    - $A_2B_1 = Pe_{21} = 0.48$
    - $A_2B_2 = Pe_{22} = 0.32$
- Alleles not in random association are said to be in linkage disequilibrium

Linkage Disequilibrium

- Observed the following data for 100 Chromosomes

Observed	Frequency
• A_1B_1 = 18	Po_{11} = 0.18
• A_1B_2= 2	Po_{12} = 0.02
• A_2B_1= 42	Po_{21} = 0.42
• A_2B_2= 38	Po_{22} = 0.38

- $D = Po_{11}\,Po_{22} - Po_{12}\,Po_{21}$
- For this example
  - D=0.06

Linkage Disequilibrium

- D can be standardized to [-1,1] or [0,1]

- The strength of association is often indicated as the standardized disequilibrium, D´.

  - $D' = D/D_{max}$ if D is positive
  - $D' = D/D_{min}$ if D is negative

  - D_{max} = the smaller of p_1q_2 (Pe_{12}) and p_2q_1 (Pe_{21})
  - D_{min} = the larger of $-p_1q_1$ ($-Pe_{11}$) and $-p_2q_2$ ($-Pe_{22}$)

  For this example $D' = 0.06/0.08 = 0.75$

Linkage Disequilibrium

- Another measure of the strength of association is r.

$r=D/(p_1p_2q_1q_2)^{1/2}$

$X^2=r^2N$ (1 df)

For this example $r = 0.06/(0.2 \cdot 0.8 \cdot 0.4 \cdot 0.6)^{1/2} = 0.3062$

$X^2=0.093746 \cdot 100=9.376$ $p=0.0022$

LD

- If one haplotype is not observed |D'|=1
  - Complete LD
    - There is not a 100% correlation between the allele at one locus and the allele at the second locus
- If two haplotypes are not observed r^2=1.0
  - Perfect LD
    - There is 100% correlation between the allele at one locus and the allele at the second locus
    - If r^2 =1.0 for two loci genotyping one locus provides as much information as genotyping both loci

LD

- D’ provides information on historic recombination events
- r^2 provides information on the correlation of two loci
  - Better measure for association studies

r^2 and D'

- There are a variety of programs which calculate r^2 and D'
- For Example HAPLOVIEW

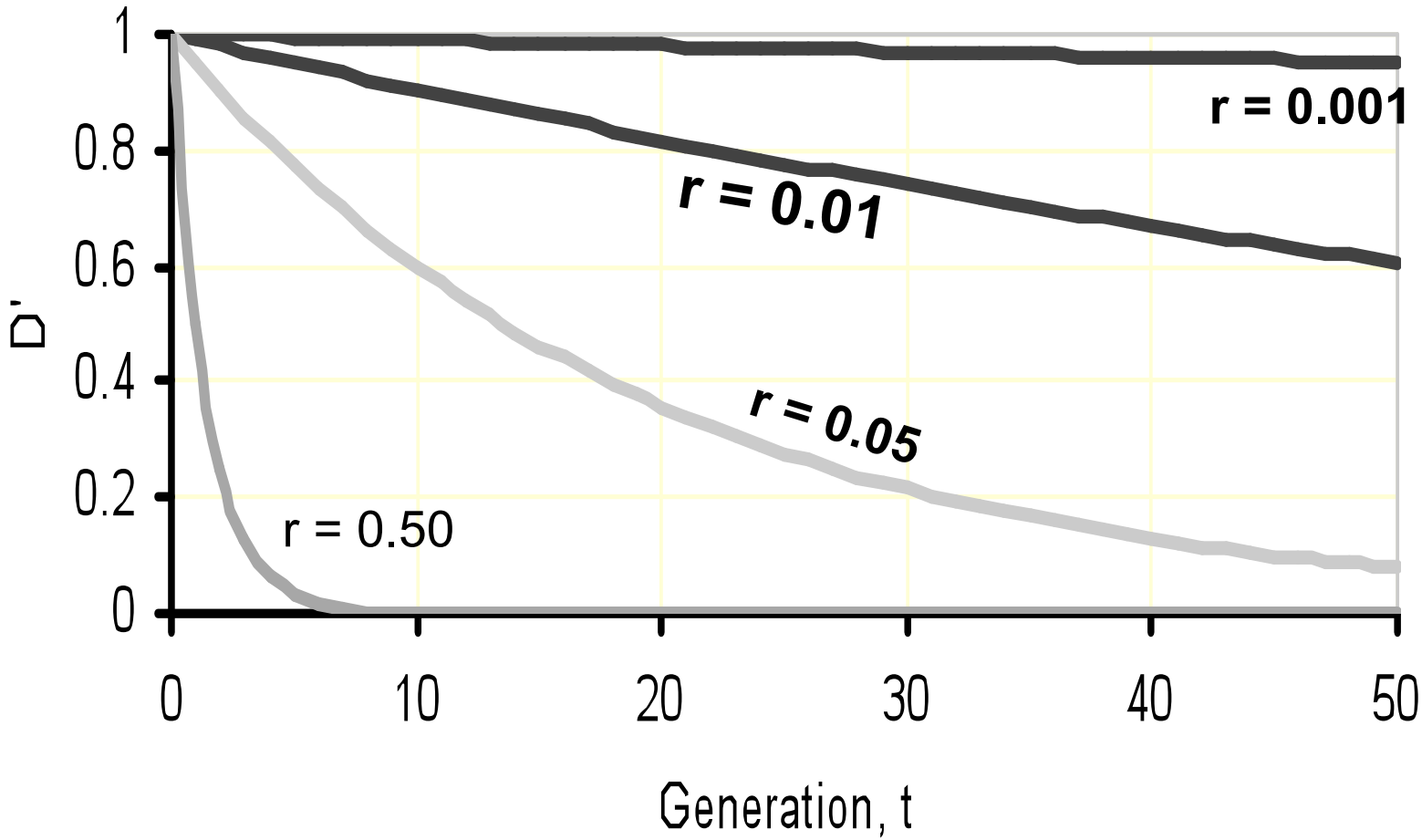
- Calculates pair-wise haplotypes using the EM algorithm

  – Then calculates D' and r^2

Linkage Disequilibrium

- With random mating (assumptions: large population with no mutation, migration or selection) linkage equilibrium is eventually obtained
- The rate of decay will depend on the recombination fraction between loci.
- The greater the rate of recombination the quicker the decay.

Decay of linkage disequilibrium over time



Three examples …

Ex1:

Two SNPs, SNP1 and SNP2, 1000 people genotyped

	A/A	A/B	B/B
A/A	199	176	0
A/B	269	0	227
B/B	68	61	0

Measures of LD: D' = 0.0638

r2 = 0.0036

Three examples

Ex2:
Two SNPs, SNP1 and SNP2, 1000 people genotyped

	A/A	A/B	B/B
A/A	375	0	0
A/B	176	214	106
B/B	0	0	129

Measures of LD: D' = 0.7357
r2 = 0.4650

Three examples

Ex3:
Two SNPs, SNP1 and SNP2, 1000 people genotyped

	A/A	A/B	B/B
A/A	551	0	0
A/B	0	214	0
B/B	0	0	235

Measures of LD: D' = 1.0000
r2 = 1.0000

Analysis

- Tested against binary phenotype
  - 449 controls
  - 551 cases

How to correct?

- Bonferroni:
  - By the number of tests
- What is the number of tests?
  - ?
- Permutation ?

Permutation in principle

- Predictor variables, e.g. SNPs
- Outcome variable, e.g. affection status

Affection status	SNP1	SNP2
1	A/B	A/B
0	A/A	A/B
0	A/A	B/B
1	B/B	A/A
1	B/B	A/B
0	A/B	A/A

Permutation in principle

- Permute order of outcomes
- Keep order of independent variables

Affection status	SNP1	SNP2
1	A/B	A/B
0	A/A	A/B
0	A/A	B/B
1	B/B	A/A
1	B/B	A/B
0	A/B	A/A

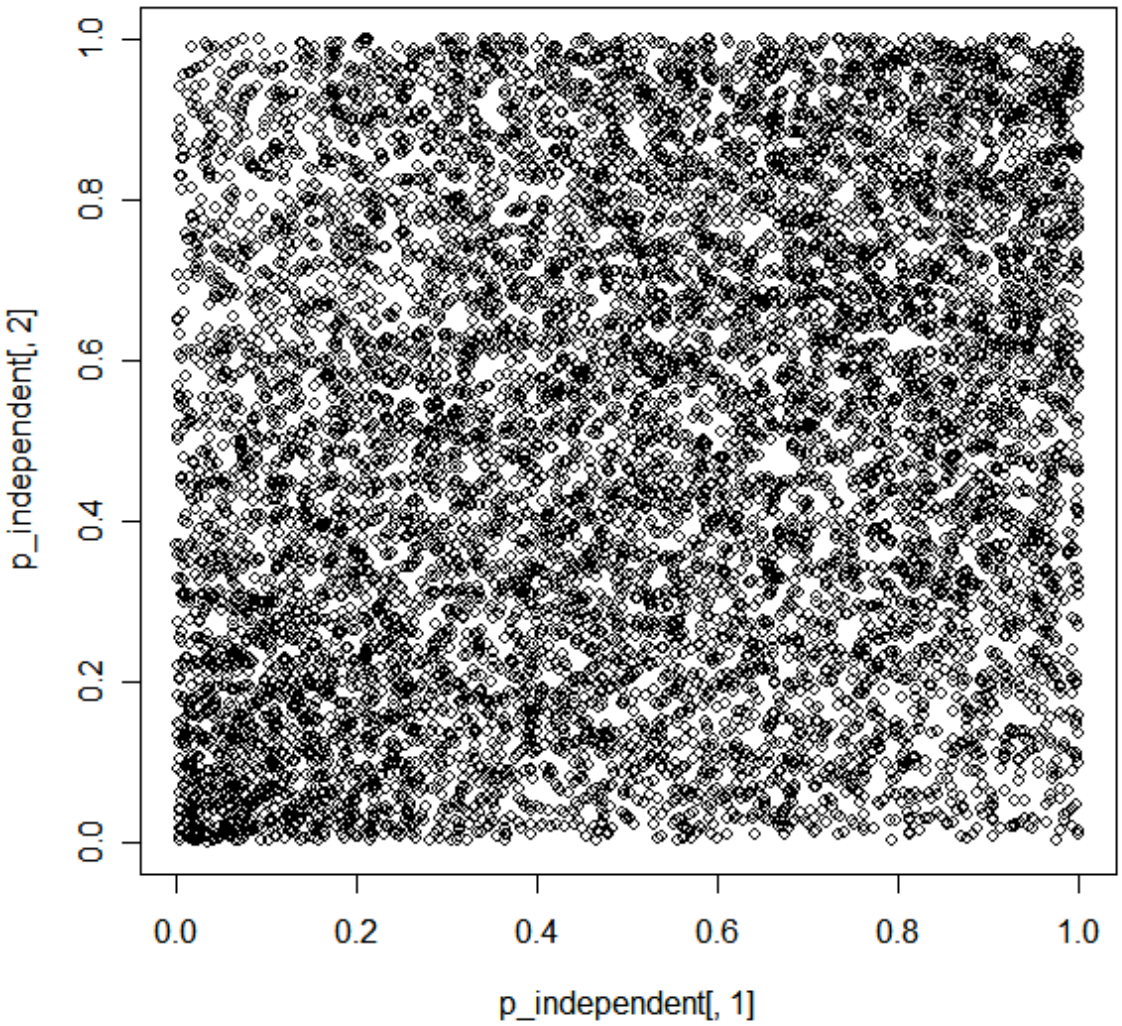
Original data and 9 replicates (permutations)

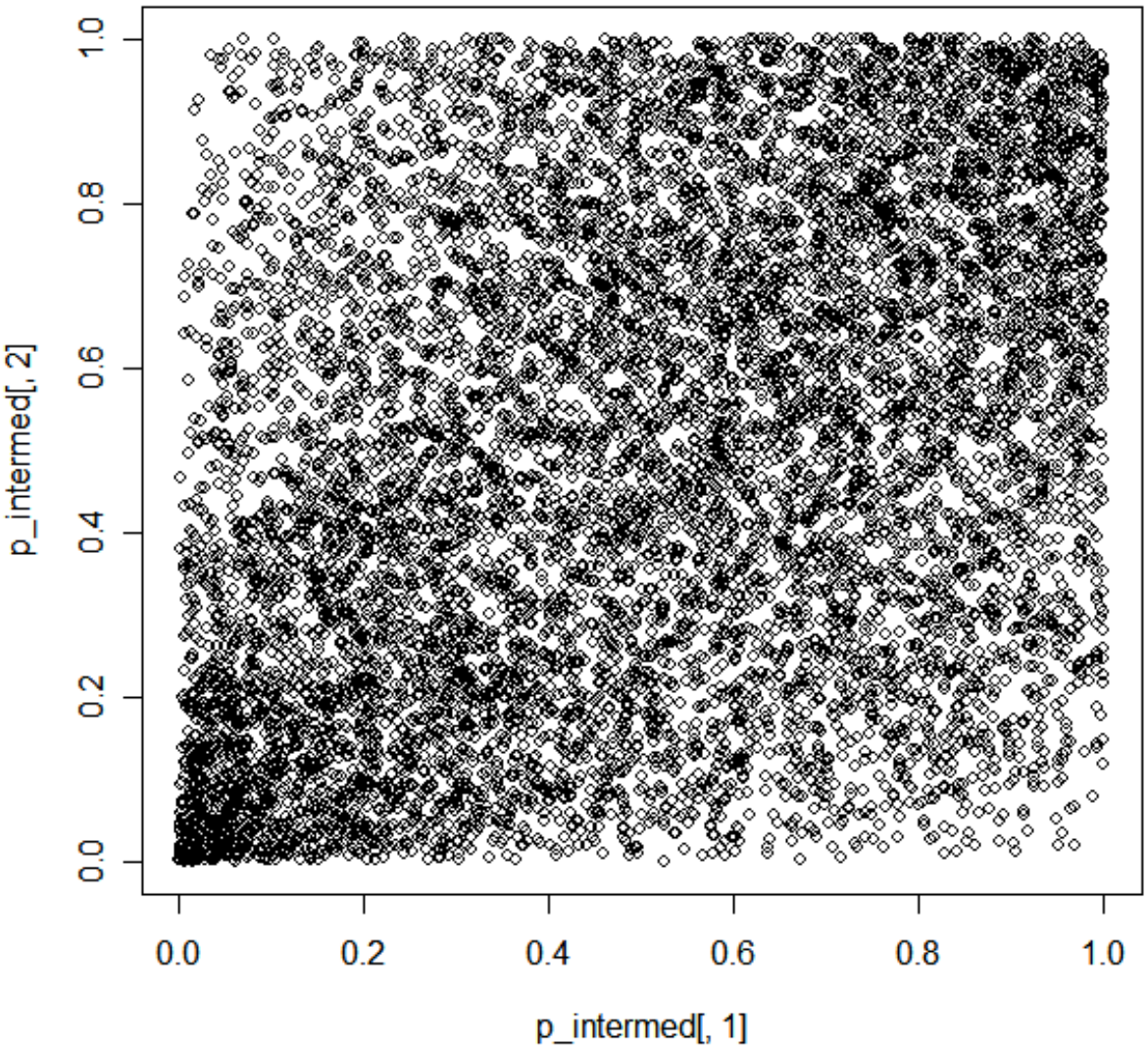
1		1	1	1	1	1	0	1	0	0
0		0	1	1	0	0	1	1	0	1
0		1	0	1	0	1	1	1	1	1
1		0	1	0	1	0	0	1	1	1
1		1	0	0	1	1	1	0	1	0
1		0	1	1	1	1	1	0	0	1
0		1	0	0	0	0	0	0	1	0

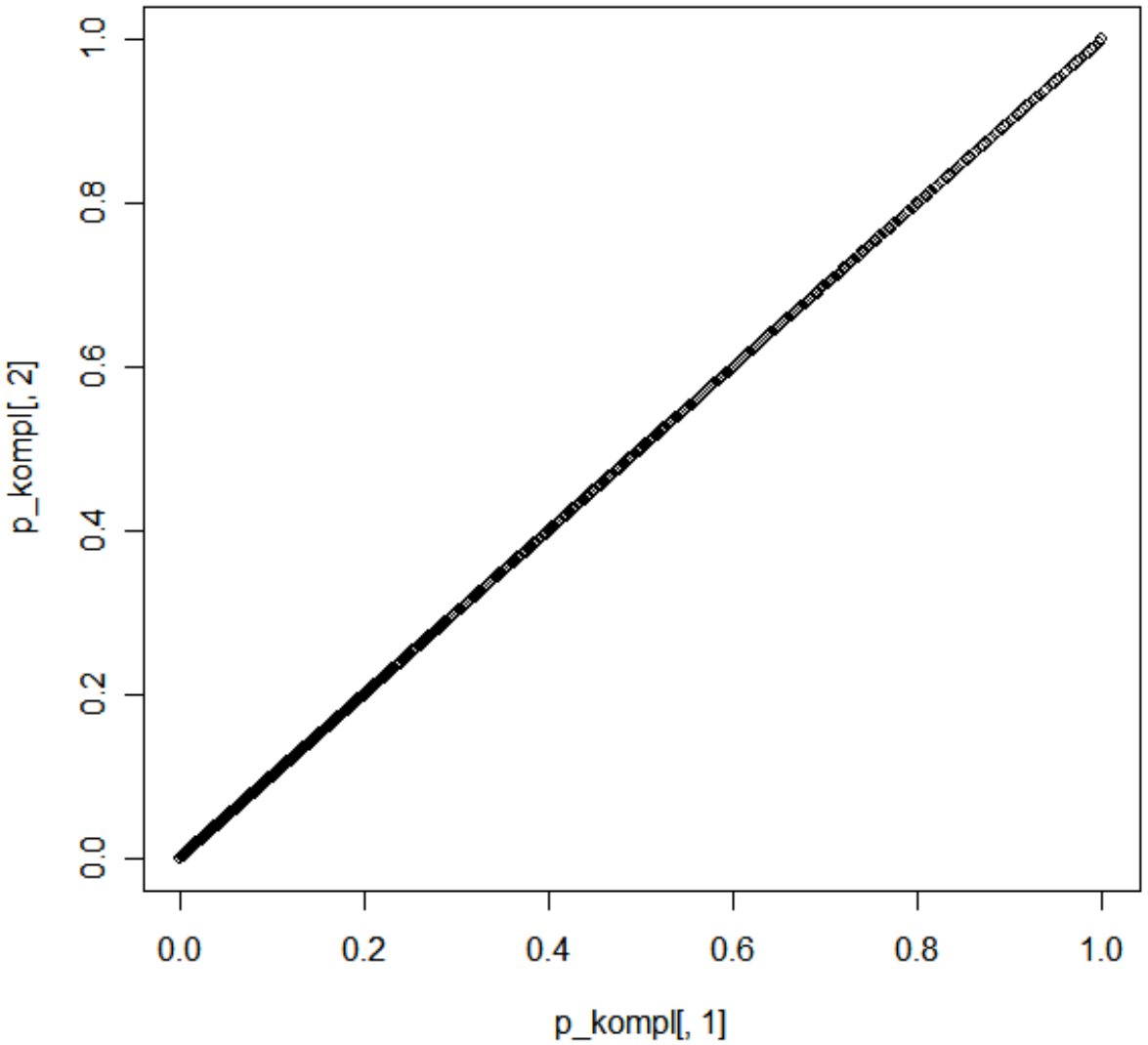
Permutation in principle

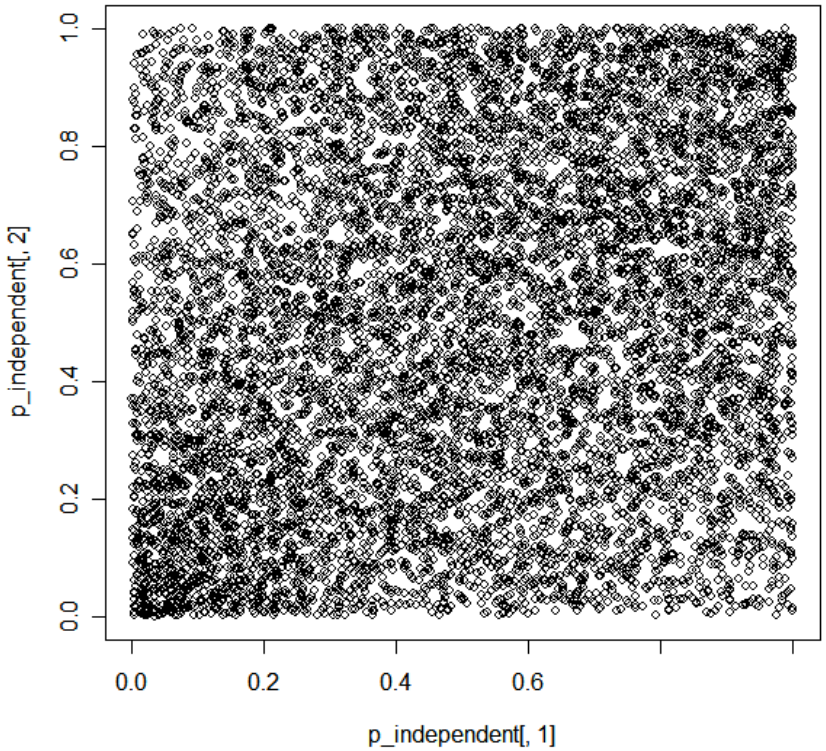
- Permute order of outcomes, keeping order of independent variables
- Calculate p-values for each of the permutations
- Store p-values per permutation

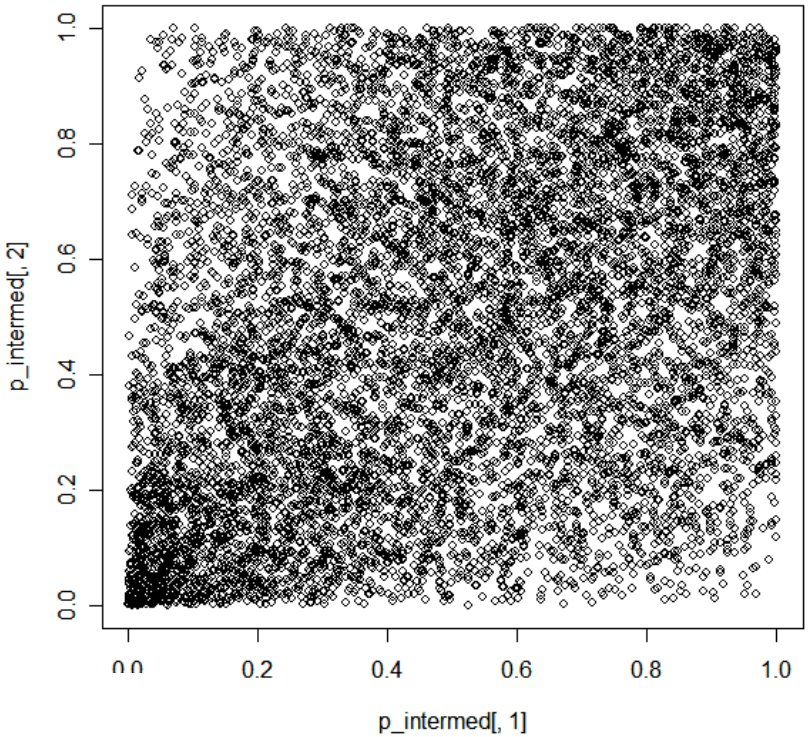
- Compute and store minimum of replicate-wise p-values
  -> test distribution of test statistic
- Compare p-value found against distribution of minimum p-values (equivalent to maximum of test statistic)
  – > correct for multiple testing

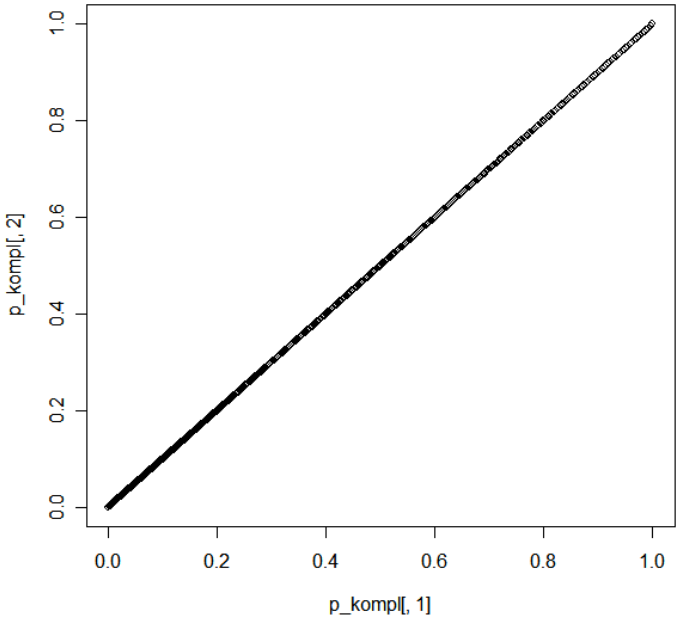

p_independent[, 2]
0.0
0.2
0.4
0.6
0.8
1.0
p_independent[, 1]
0.0
0.2
0.4
0.6
0.8
1.0

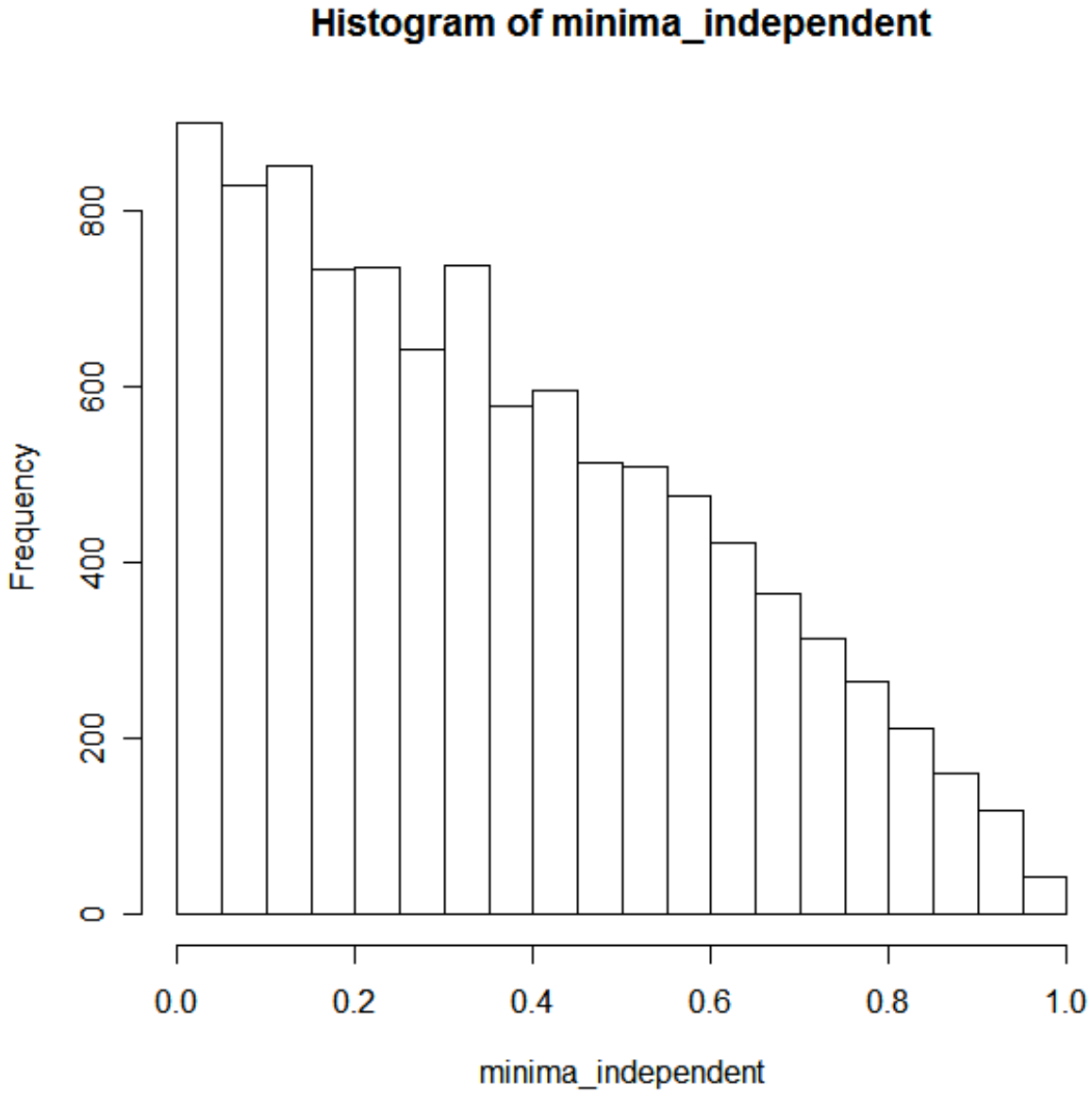

p_intermed[, 2]
0.0
0.2
0.4
0.6
0.8
1.0
p_intermed[, 1]
0.0
0.2
0.4
0.6
0.8
1.0

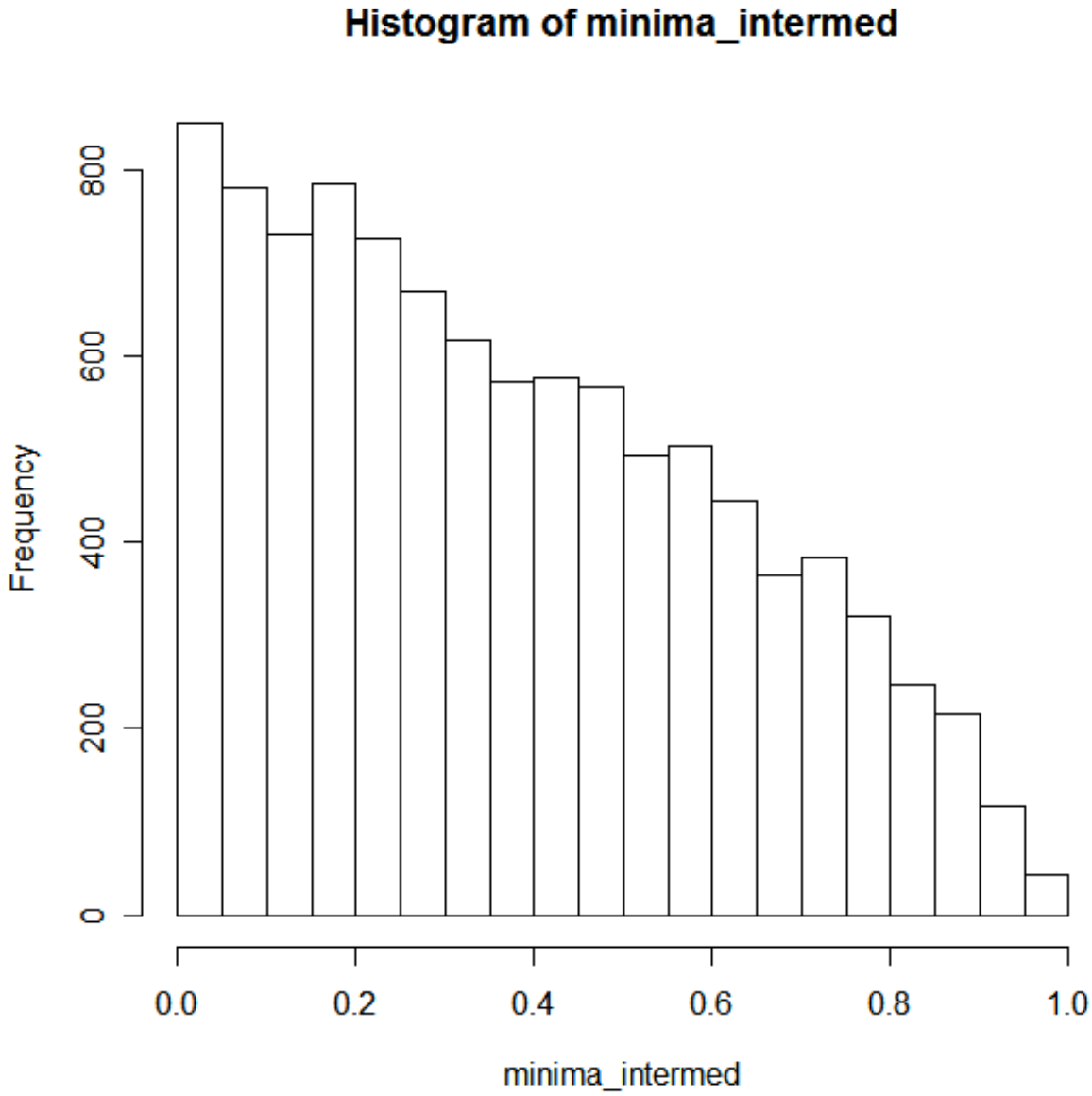

p_kompl[, 2]
0.0
0.2
0.4
0.6
0.8
1.0
p_kompl[, 1]
0.0
0.2
0.4
0.6
0.8
1.0

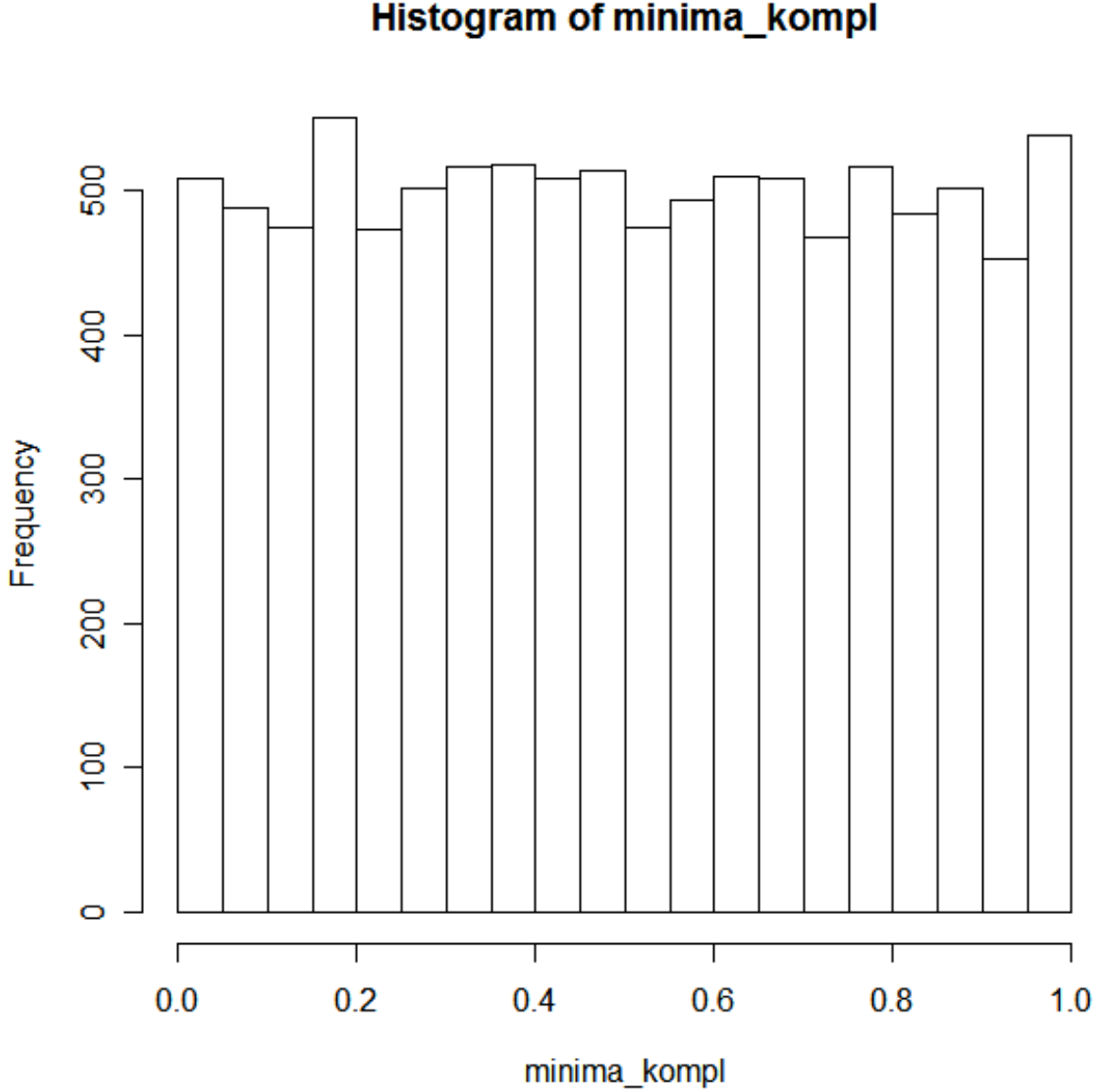

p_independent[, 2]
p_independent[, 1]
0.0
0.2
0.4
0.6
0.8
1.0


p_intermed[, 2]
p_intermed[, 1]
0.0
0.2
0.4
0.6
0.8
1.0


p_kompl[, 2]
p_kompl[, 1]
0.0
0.2
0.4
0.6
0.8
1.0


Histogram of minima_independent
Frequency
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
minima_independent


Histogram of minima_intermed
Frequency
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
minima_intermed


Histogram of minima_kompl
Frequency
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
minima_kompl

Quantile of distributions

1%	5%	10%	
0.006	0.027	0.055	ex1
0.006	0.029	0.060	ex2
0.009	0.049	0.100	ex3

Two technical notes

- Compute p-value
  - P-value found = p_f
  - N replications, of which M <= p_f
  - $P_{\text{post permut}}$ = (M+1)/(N+1)
- How many permutations for value of α to be tested
  - Recommendations:
  - N = 10/(α)
  - E.g. a = 0.05, then N = 10/0.05 = 10/(1/20) = 200

Papers on the way

Am. J. Hum. Genet. 74:765–769, 2004

A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other

Dale R. Nyholt

Papers on the way

Am. J. Hum. Genet. 75:424–435, 2004

Efficient Computation of Significance Levels for Multiple Associations in Large Studies of Correlated Data, Including Genomewide Association Studies

Frank Dudbridge and Bobby P. C. Koeleman

Papers on the way

Hum Hered 2005;60:19–25

Evaluation of Nyholt's Procedure for Multiple Testing Correction

Daria Salyakina
Shaun R. Seaman
Brian L. Browning
Frank Dudbridge
Bertram Müller-Myhsok

Genetic Epidemiology 32: 381–385 (2008)

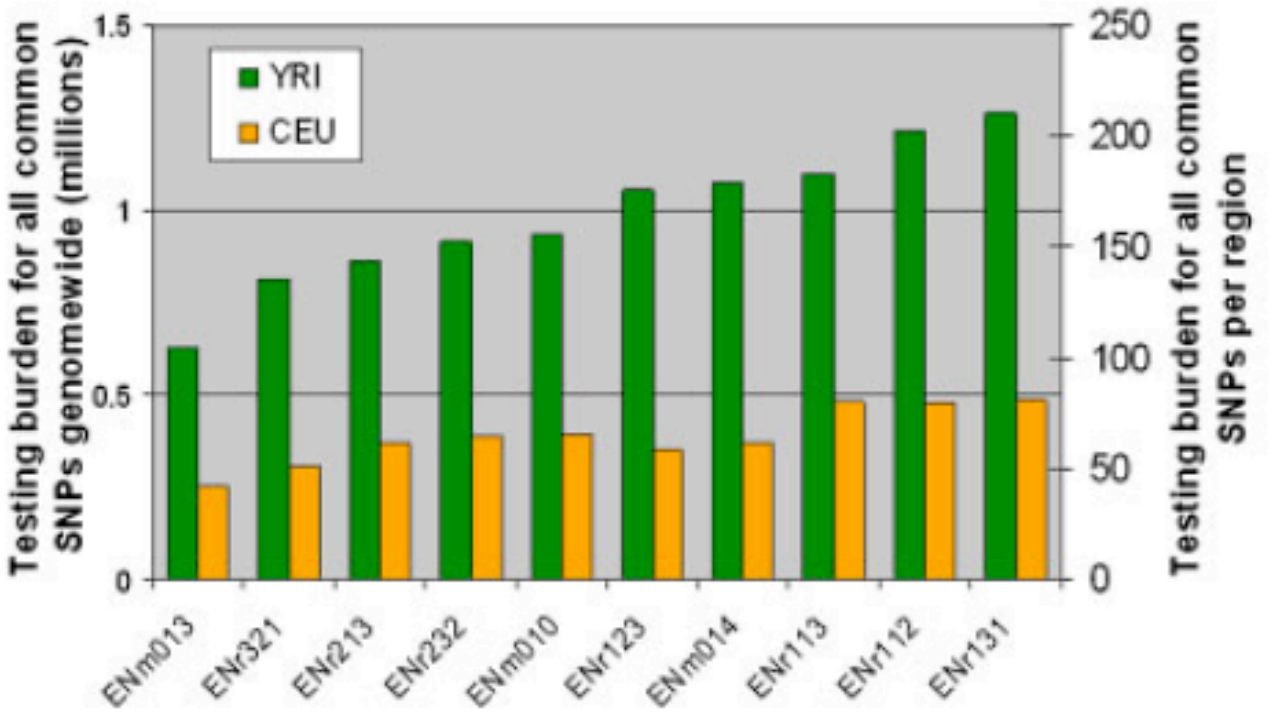
Brief Report

Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants

Itsik Pe'er,[1] Roman Yelensky,[2–4] David Altshuler,[2,3,5–7] and Mark J. Daly[2,5,8*]

[1]*Department of Computer Science, Columbia University, New York*
[2]*Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts*
[3]*Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts*
[4]*Harvard-M.I.T. Division of Health Sciences and Technology, Cambridge, Massachusetts*
[5]*Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts*
[6]*Broad Institute of M.I.T. and Harvard, Cambridge, Massachusetts*
[7]*Department of Genetics, Harvard Medical School, Boston, Massachusetts*
[8]*Department of Medicine, Harvard Medical School, Boston, Massachusetts*

(a)


Testing burden for all common SNPs genomewide (millions)
1.5
1
0.5
0
YRI
CEU
250
200
150
100
50
0
Testing burden for all common SNPs per region
ENm013
ENr321
ENr213
ENr232
ENm010
ENr123
ENm014
ENr113
ENr112
ENr131

b)

Genetic Epidemiology 32: 227–234 (2008)

Estimation of Significance Thresholds for Genomewide Association Scans

Frank Dudbridge* **and Arief Gusnanto**

MRC Biostatistics Unit, Institute for Public Health, Cambridge, United Kingdom

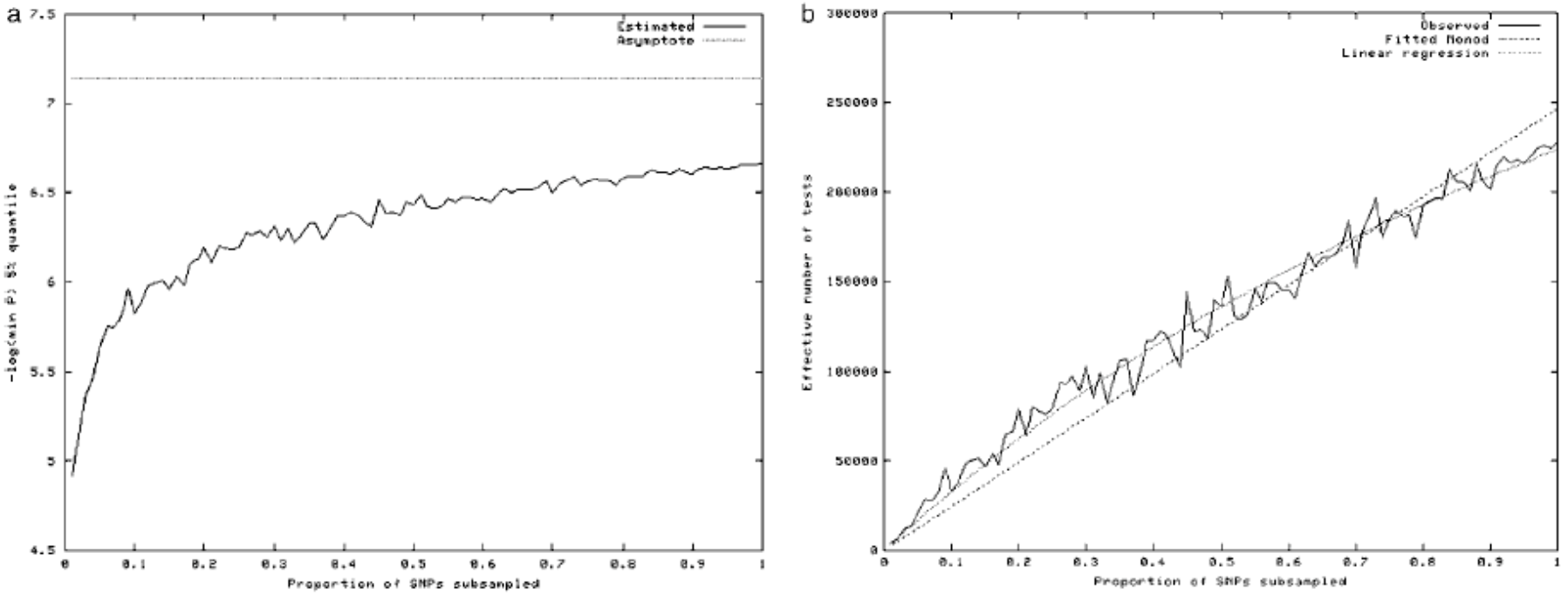


Fig. 1. (a) Significance threshold as a function of marker density in combined NBS and 58BC sample from permutation procedure. At current density (359K single nucleotide polymorphisms typed) the significance threshold is about 2.2×10^{-7}. The dotted line shows the estimated asymptote of 7.2×10^{-8}. (b) Fitted Monod function to the effective number of tests associated with the significance threshold. At infinite density the number of tests is estimated at 693,138 giving the asymptote in (a).

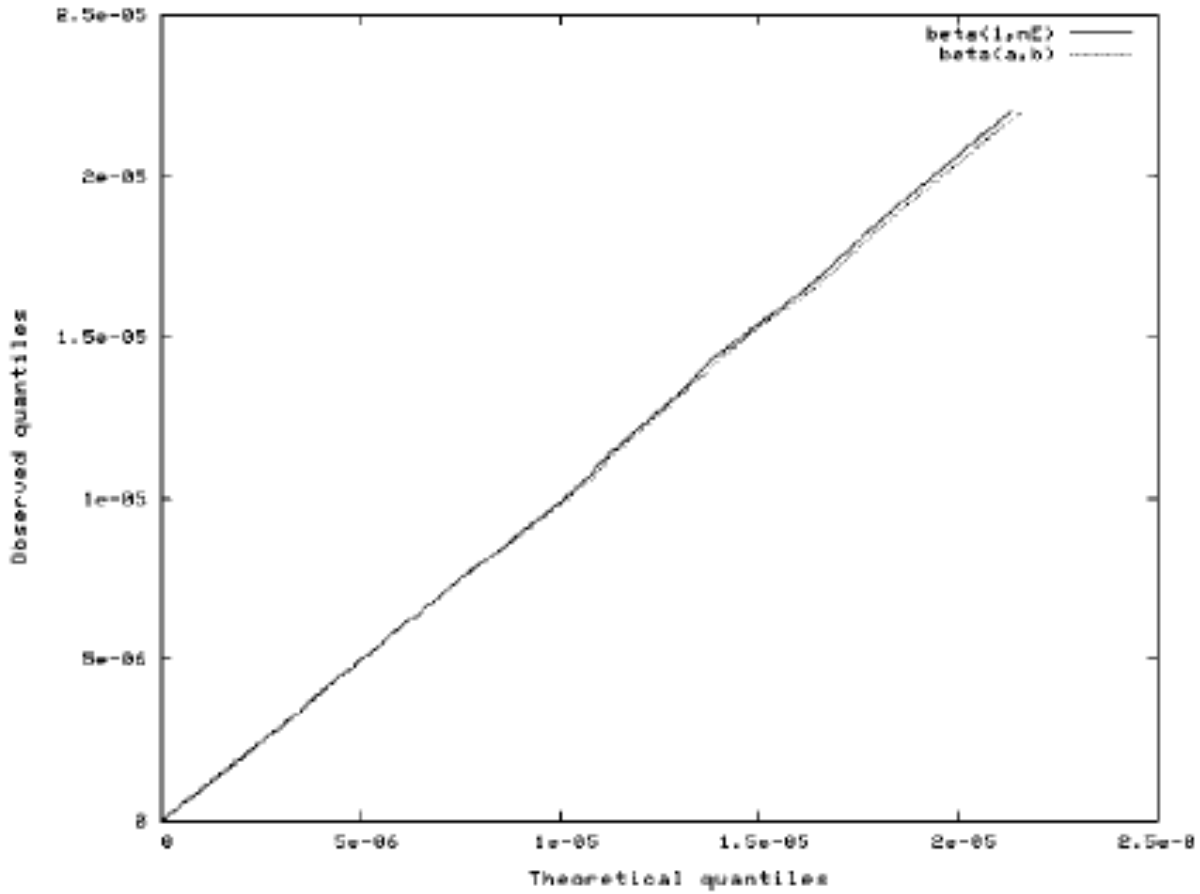


Fig. 3. Quantile–quantile plot comparing fitted Beta distributions with minimum *P*-values from permutation replicates.