Analysis of Complex Disease Association Studies

Juan R González
juanr.gonzalez@isglobal.org
http://brge.isglobal.org
BRGE – Bioinformatic Research Group in Epidemiology
Center for Research in Environmental Epidemiology (CREAL)
Department of Mathematics, Universidad Autònoma de Barcelona (UAB)

General issues

- **Three lectures**
  - Single SNP association analysis
  - Haplotype and GWAS
    - population stratification and multiple comparisons
  - CNV association analysis

- **Each lecture**
  - ~1h describing the main statistical approaches
  - ~30'illustrating how to use R to analyze real data
  - ~1h practical exercises

- **Material**
  - R package including required libraries
  - Slides
  - Selected papers
  - R code (illustrating how to analyze real data)
  - Real data

Material

- **Slides**
- **R code**
- **Selected papers**
  - **Chapter book:** Sole X, Gonzalez JR, Moreno V. Analysis of population–based genetic association studies applied to cancer susceptibility and prognosis. In **Computational Biology:** Issues and Applications in Oncology, 2009
  - **GWAS data analysis of common diseases:** The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nat Genet, 2007
  - **General test to asses association:** Gonzalez JR, et al. Maximizing association statistics over genetic models. Genet Epidemiol, 2008
  - **Software:** Gonzalez et al. SNPassoc: an R package to perform whole genome association studies. Bioinformatics, 2007

Outline

- Introduction
- Statistical Methods
  - Association analysis: models of inheritance
  - GWAS
  - Stratification
  - Multiple comparisons
  - Haplotype analysis
  - GxG and GxE interaction
- Software: SNPassoc & snpStats (formerly snpMatrix)
- LocusZoom – Plots for genomic data

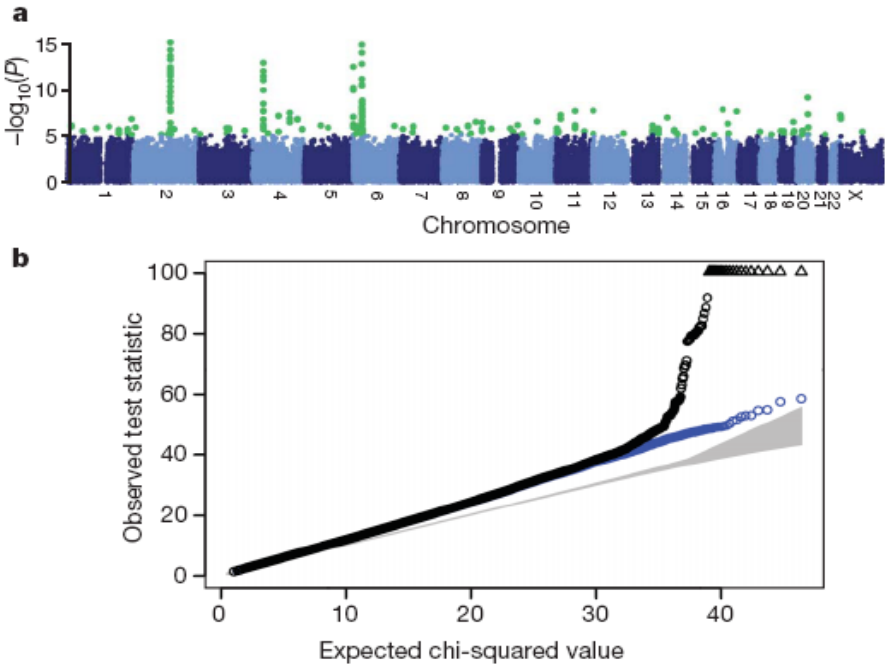
Introduction

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*



Introduction

Table 2 | Evidence for signal of association at previously robustly replicated loci

Collection	Gene	Chromosome	Reported SNP	WTCCC SNP	HapMap r^2	Trend *P* value	Genotypic *P* value
CAD	*APOE*	19q13	*	rs4420638	-	1.7×10^{-01}	1.7×10^{-01}
CD	*NOD2*	16q12	rs2066844	rs17221417	0.23	9.4×10^{-12}	4.0×10^{-11}
CD	*IL23R*	1p31	rs11209026	rs11805303	0.01	6.5×10^{-13}	5.9×10^{-12}
RA	*HLA-DRB1*	6p21	*	rs615672	-	2.6×10^{-27}	7.5×10^{-27}
RA	*PTPN22*	1p13	rs2476601	rs6679677	0.75	4.9×10^{-26}	5.6×10^{-25}
T1D	*HLA-DRB1*	6p21	*	rs9270986	-	4.0×10^{-116}	2.3×10^{-122}
T1D	*INS*	11p15	rs689	†	-	-	-
T1D	*CTLA4*	2q33	rs3087243	rs3087243	1	2.5×10^{-05}	1.8×10^{-05}
T1D	*PTPN22*	1p13	rs2476601	rs6679677	0.75	1.2×10^{-26}	5.4×10^{-26}
T1D	*IL2RA*	10p15	rs706778	rs2104286	0.25	8.0×10^{-06}	4.3×10^{-05}
T1D	*IFIH1*	2q24	rs1990760	rs3788964	0.26	1.9×10^{-03}	7.6×10^{-03}
T2D	*PPARG*	3p25	rs1801282	rs1801282	1	1.3×10^{-03}	5.4×10^{-03}
T2D	*KCNJ11*	11p15	rs5219	rs5215	0.9	1.3×10^{-03}	5.6×10^{-03}
T2D	*TCF7L2*	10q25	rs7903146	rs4506565	0.92	5.7×10^{-13}	5.1×10^{-12}

Where information on the strength of association at a particular SNP had been previously published and replicated we tabulated the *P* value of both the trend and genotype test at the same SNP (if in our study), or the best tag SNP (defined to be the SNP with highest r^2 with the reported SNP, calculated in the CEU sample of the HapMap project). Positions are in NCBI build-35 coordinates. *Previous reports relate to haplotypes rather than single SNPs. †Not well tagged by SNPs that pass the quality control, see main text.

Introduction

Table 3 | Regions of the genome showing the strongest association signals

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
					Standard analysis								
BD	16p12	23.3–23.62	rs420259	2.19×10^{-04}	6.29×10^{-08}	1.96	4.79	A	G	2.08 (1.60–2.71)	2.07 (1.6–2.69)	0.282	0.248
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	11.66	11.19	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554
CD	1p31	67.3–67.48	rs11805303	6.45×10^{-13}	5.85×10^{-12}	10.07	9.41	T	T	1.39 (1.22–1.58)	1.86 (1.54–2.24)	0.317	0.391
CD	2q37	233.92–234	rs10210302	7.10×10^{-14}	5.26×10^{-14}	11.11	11.28	T	C	1.19 (1.01–1.41)	1.85 (1.56–2.21)	0.481	0.402
CD	3p21	49.3–49.87	rs9858542	7.71×10^{-07}	3.58×10^{-08}	4.24	5.22	A	A	1.09 (0.96–1.24)	1.84 (1.49–2.26)	0.282	0.331
CD	5p13	40.32–40.66	rs17234657	2.13×10^{-13}	1.99×10^{-12}	10.41	9.89	G	G	1.54 (1.34–1.76)	2.32 (1.59–3.39)	0.125	0.181
CD	5q33	150.15–150.31	rs1000113	5.10×10^{-08}	3.15×10^{-07}	5.36	5.01	T	T	1.54 (1.31–1.82)	1.92 (0.92–4.00)	0.067	0.098
CD	10q21	64.06–64.31	rs10761659	2.68×10^{-07}	1.75×10^{-06}	4.69	4.13	G	A	1.23 (1.05–1.45)	1.55 (1.3–1.84)	0.461	0.406
CD	10q24	101.26–101.32	rs10883365	1.41×10^{-08}	5.82×10^{-08}	5.91	5.48	G	G	1.2 (1.03–1.39)	1.62 (1.37–1.92)	0.477	0.537
CD	16q12	49.02–49.4	rs17221417	9.36×10^{-12}	3.98×10^{-11}	8.93	8.47	G	G	1.29 (1.13–1.46)	1.92 (1.58–2.34)	0.287	0.356
CD	18p11	12.76–12.91	rs2542151	4.56×10^{-08}	2.03×10^{-08}	5.42	5.00	G	G	1.3 (1.14–1.48)	2.01 (1.46–2.76)	0.163	0.208
RA	1p13	113.54–114.16	rs6679677	4.90×10^{-26}	5.55×10^{-25}	22.36	21.99	A	A	1.98 (1.72–2.27)	3.32 (1.93–5.69)	0.096	0.168
RA	6	MHC	rs6457617*	3.44×10^{-76}	5.18×10^{-75}	74.84	73.18	T	T	2.36 (1.97–2.84)	5.21 (4.31–6.30)	0.489	0.685
T1D	1p13	113.54–114.16	rs6679677	1.17×10^{-26}	5.43×10^{-26}	23.07	22.83	A	A	1.82 (1.59–2.09)	5.19 (3.15–8.55)	0.096	0.169
T1D	6	MHC	rs9272346*	2.42×10^{-134}	5.47×10^{-134}	141.9	142.2	A	G	5.49 (4.83–6.24)	18.52 (27.03–12.69)	0.387	0.150
T1D	12q13	54.64–55.09	rs11171739	1.14×10^{-11}	9.71×10^{-11}	8.89	8.24	C	C	1.34 (1.17–1.54)	1.75 (1.48–2.06)	0.423	0.493
T1D	12q24	109.82–111.49	rs17696736	2.17×10^{-15}	1.51×10^{-14}	12.53	11.88	G	G	1.34 (1.16–1.53)	1.94 (1.65–2.29)	0.424	0.506
T1D	16p13	10.93–11.37	rs12708716	9.24×10^{-08}	4.92×10^{-07}	5.15	4.70	A	G	1.19 (0.97–1.45)	1.55 (1.27–1.89)	0.350	0.297
T2D	6p22	20.63–20.84	rs9465871	1.02×10^{-06}	3.34×10^{-07}	4.15	3.98	C	C	1.18 (1.04–1.34)	2.17 (1.6–2.95)	0.178	0.218
T2D	10q25	114.71–114.81	rs4506565	5.68×10^{-13}	5.05×10^{-12}	10.14	9.43	T	T	1.36 (1.2–1.54)	1.88 (1.56–2.27)	0.324	0.395
T2D	16q12	52.36–52.41	rs9939609	5.24×10^{-08}	1.91×10^{-07}	5.35	5.05	A	A	1.34 (1.17–1.52)	1.55 (1.3–1.84)	0.398	0.453
					Multi-locus analysis								
T1D	4q27	123.26–123.92	rs6534347	4.48×10^{-07}	1.83×10^{-06}	5.15	4.69	A	A	1.30 (1.10–1.55)	1.49 (1.25–1.78)	0.351	0.402
T1D	12p13	9.71–9.86	rs3764021	7.19×10^{-05}	5.08×10^{-08}	2.12	4.55	C	T	1.57 (1.38–1.79)	1.48 (1.25–1.75)	0.467	0.426
					Sex differentiated analysis								
RA	7q32	130.80–130.84	rs11761231	3.91×10^{-07}	1.37×10^{-06}	-	-	G	A	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
					Combined cases								
RA+T1D	10p15	6.07–6.17	rs2104286	5.92×10^{-08}	2.52×10^{-07}	5.26	4.45	T	C	1.35 (1.11–1.65)	1.62 (1.34–1.97)	0.286	0.245

Introduction

How do we measure genetic susceptibility?

- Variation in one concrete position of the genome is known as polymorfism
- It is normally used as a genetic marker that allow us to identify those allele that are related to disease

Genetic markers

- Microsatellites: VNTR ("Variable Number Tandem Repeat")

CGTACCTGCACACACATGCGAACT

- SNP ("Single Nucleotide Polymorphism")

ATTGATC

ATTCATC G->C

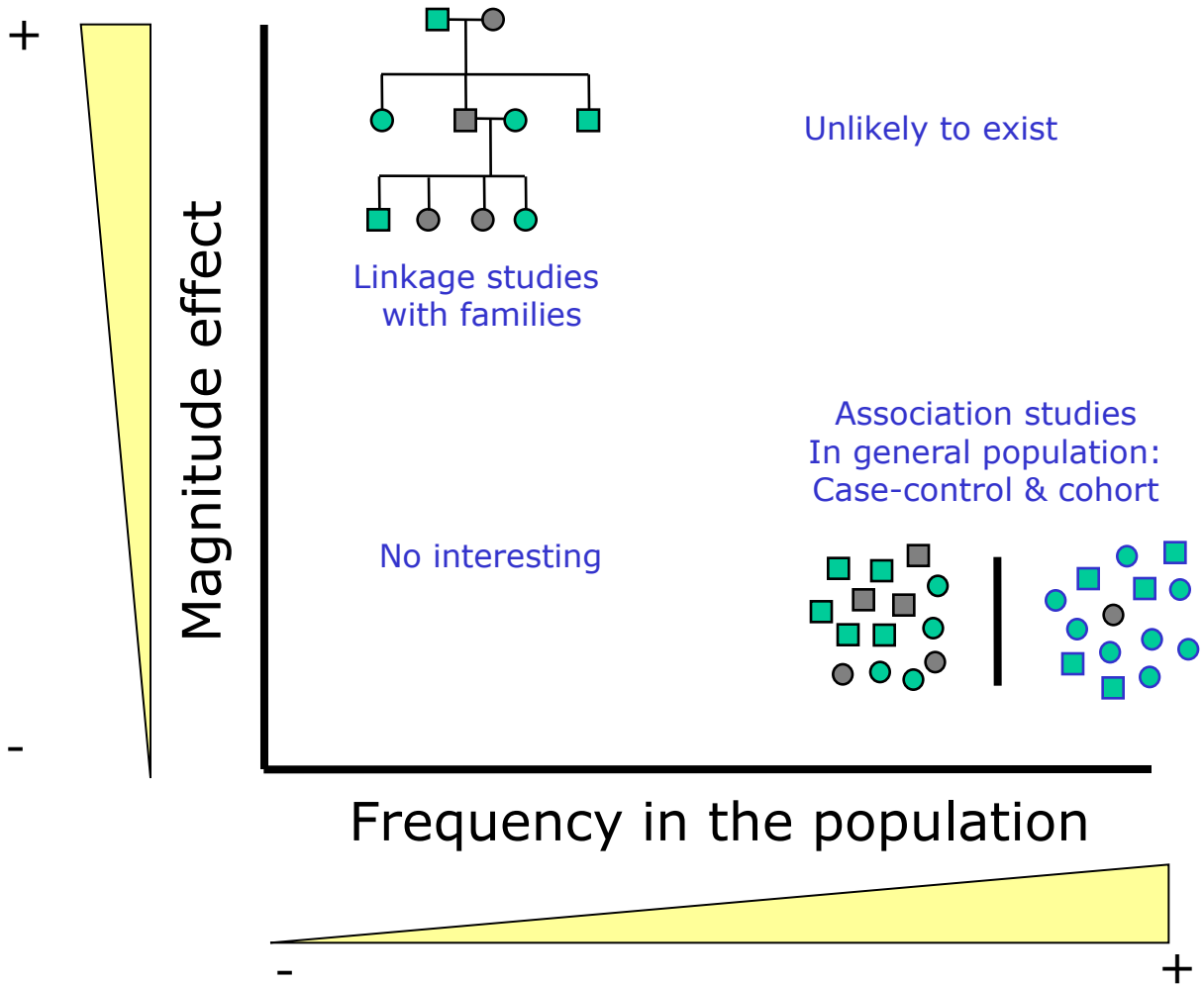
- CNV ("Copy number variant")

Aim

- To establish the role that genes and environment play in the variability (inter-individual differences) with regard to disease and complex traits

- To find those susceptibility genes

Designs


+
-
Magnitude effect
Linkage studies
with families
Unlikely to exist
Association studies
In general population:
Case-control & cohort
No interesting
Frequency in the population
-
+

Introduction

Case-control studies – Why?

Case-control Association studies

Association studies

- Candidate Polimorfism
  - Known functional impact
- Candidate Gen
  - A selection of SNPs (5-15) are scanned
  - Maybe none of them is causal (LD existence?)
- Candidate Region
  - Identified using linkage studies
  - 10-100 SNPs are scanned
- Whole genome scan
  - 300.000, 500.000, 1.000.000, … SNPs are analyzed

Association studies

- There are several reason to find association
  - Causal: genetic diseases due to the marker
  - Linkage disequilibrium. True associated marker is near the studied one
  - Population stratification: confusion due to mixed population with different allele frequencies

Before assessing association:
Hardy-Weinberg test

- This hypothesis must be tested only in controls since it may be disequilibrium when allele is related to the disease.

Assume that *f* denotes the minor allelic frequency (MAF), then

Frequency of genotype AA is $(1-f)^2$
Frequency of genotype Aa is $2f(1-f)$
Frequency of genotype aa is f^2

HWE is tested using a chi-square test (goodness-of-fit)

Types of association analyses

- Each polymorphism at time
  - Information obtained from genotype platforms (Single association analysis)

- Combination of polymorphisms: Haplotypes
  - This information is not available. It has to be estimated using complex statistical methods.

Single association analysis

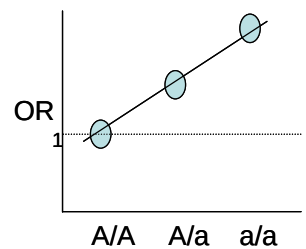
codominant

SNP

disease

	A/A	A/a	a/a	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

additive



χ^2 trend test

Cochran-Armitage trend test

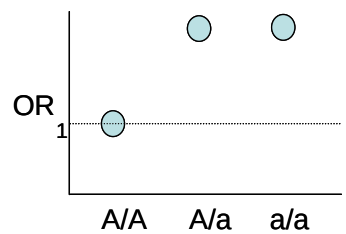
$OR_{A/A} < OR_{A/a} < OR_{a/a}$ (linear trend)

dominant

SNP

disease

	A/A	A/a + a/a	Total
Cases	r_0	$r_1 + r_2$	R
Controls	s_0	$s_1 + s_2$	S
Total	n_0	$n_1 + n_2$	N



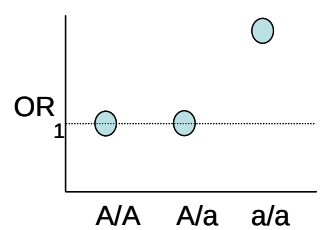
χ^2 test (Fisher test)

recessive

SNP

disease

	A/A + A/a	a/a	Total
Cases	$r_0 + r_1$	r_2	R
Controls	$s_0 + s_1$	s_2	S
Total	$n_0 + n_1$	n_2	N



χ^2 test (Fisher test)

Association analysis

- Regression models
  - $\Pr(D|G,Z) = f(G,Z)$

- Logistic regression

$$\log\{p/(1-p)\} = \alpha + \beta G + \gamma Z$$

Genotype

other covariates

Association analysis

- Lineal regresion:

$$Y = \alpha + \beta G + \gamma Z$$

Quantitative trait

Genotype

Other covariates

- Y must be Gaussian. If not -> transformation (logarithm)

Association analysis

- Goodness-of-fit: Likelihood (L)

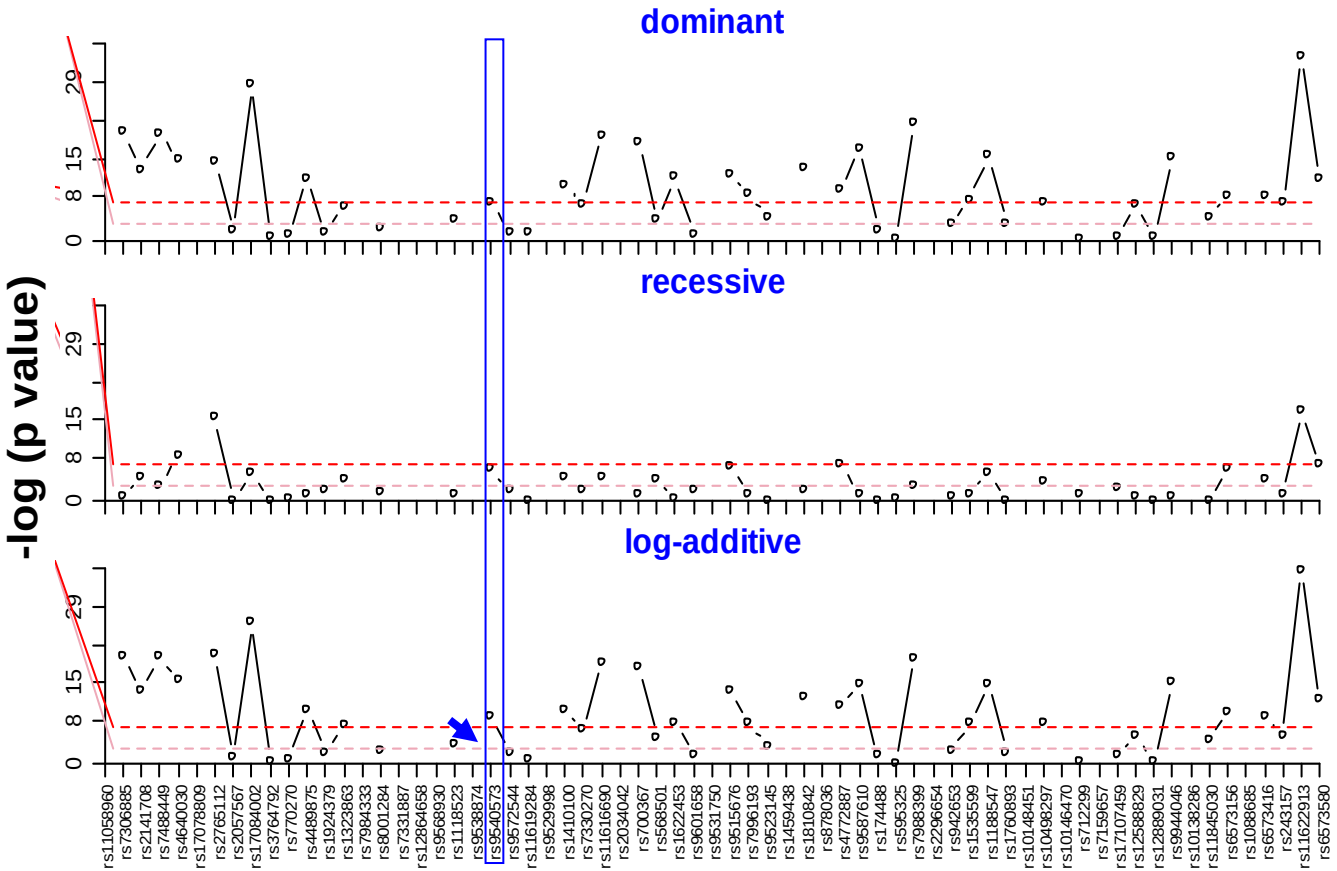
Mod_0: Basal model (null or adjusted)
Mod_1: Model including SNP

- Likelihood ratio test (LRT)

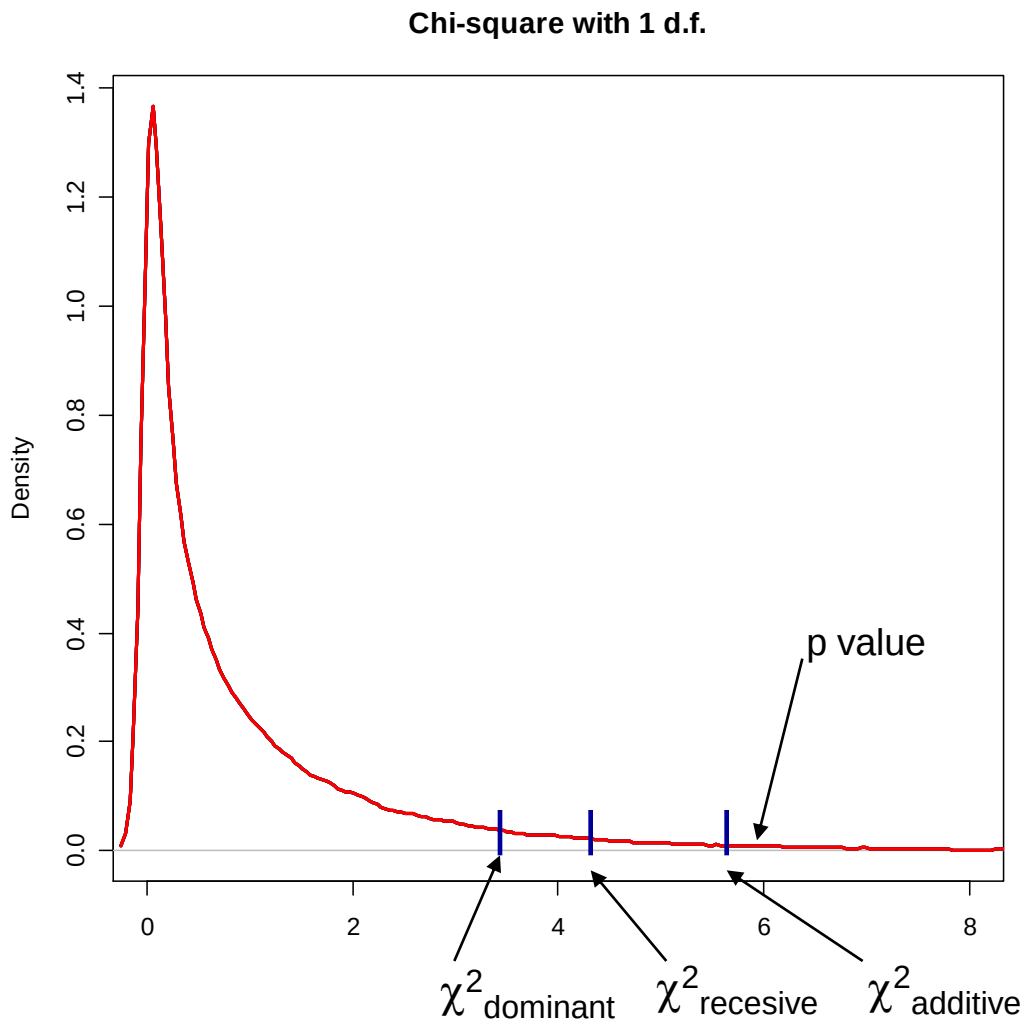
$$\frac{L_{\mathrm{mod}_1}}{L_{\mathrm{mod}_0}} = \begin{cases} > 1 \text{ if } \mathrm{mod}_1 \text{ more 'likely' than } \mathrm{mod}_0 \\ = 1 \text{ if both models are equal} \\ < 1 \text{ if } \mathrm{mod}_0 \text{ more 'likely' than } \mathrm{mod}_1 \end{cases}$$

$$-2\log\left(\frac{L_{\mathrm{mod}_1}}{L_{\mathrm{mod}_0}}\right) \sim \chi^2_{g.l.}$$

Max-statistic



Max-statistic



$$\chi^2_{\max} = \max\{\chi^2_D, \chi^2_R, \chi^2_A\}$$

Max-statistic

Vol 445 | 22 February 2007 | doi:10.1038/nature05616 nature

ARTICLES

A genome-wide association study identifies novel risk loci for type 2 diabetes

Robert Sladek[1,2,4], Ghislain Rocheleau[1]*, Johan Rung[4]*, Christian Dina[5]*, Lishuang Shen[1], David Serre[1], Philippe Boutin[5], Daniel Vincent[4], Alexandre Belisle[4], Samy Hadjadj[6], Beverley Balkau[7], Barbara Heude[7], Guillaume Charpentier[8], Thomas J. Hudson[4,9], Alexandre Montpetit[4], Alexey V. Pshezhetsky[10], Marc Prentki[10,11], Barry I. Posner[2,12], David J. Balding[13], David Meyre[5], Constantin Polychronakos[1,3] & Philippe Froguel[5,14]

A max statistic was formed across these to select the strongest obtainable association for any of the three models.

$$X^2_{\max,i} = \max\left\{X^2_{A,i}, X^2_{D,i}, X^2_{R,i}\right\} \qquad \textbf{E6}$$

P-values were calculated for the observed test statistic against the null distribution for the genetic model giving the strongest association. Also, since the distribution for the max statistic itself under the null hypothesis is not known, we establish such p-values by permutation testing. To obtain these, N_{perm} permutations of the disease state vector were done

Max-statistic

Genetic Epidemiology 32: 246–254 (2008)

Maximizing Association Statistics Over Genetic Models

Juan R. González,[1–3*] Josep L. Carrasco,[3] Frank Dudbridge,[4] Lluís Armengol,[2,5] Xavier Estivill,[2,5] and Victor Moreno[6]

[1]*Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain*
[2]*CIBER de Epidemiología y Salud Pública (CIBERESP), Spain*
[3]*Biostatistic Unit, Department of Public Health, University of Barcelona, Spain*
[4]*MRC Biostatistics Unit, Cambridge, United Kingdom*
[5]*Genes and Disease Program, Center for Genomic Regulation, Barcelona, Spain*
[6]*IDIBELL, Catalan Institute of Oncology, Barcelona, Spain*

GWAS

GWAS

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

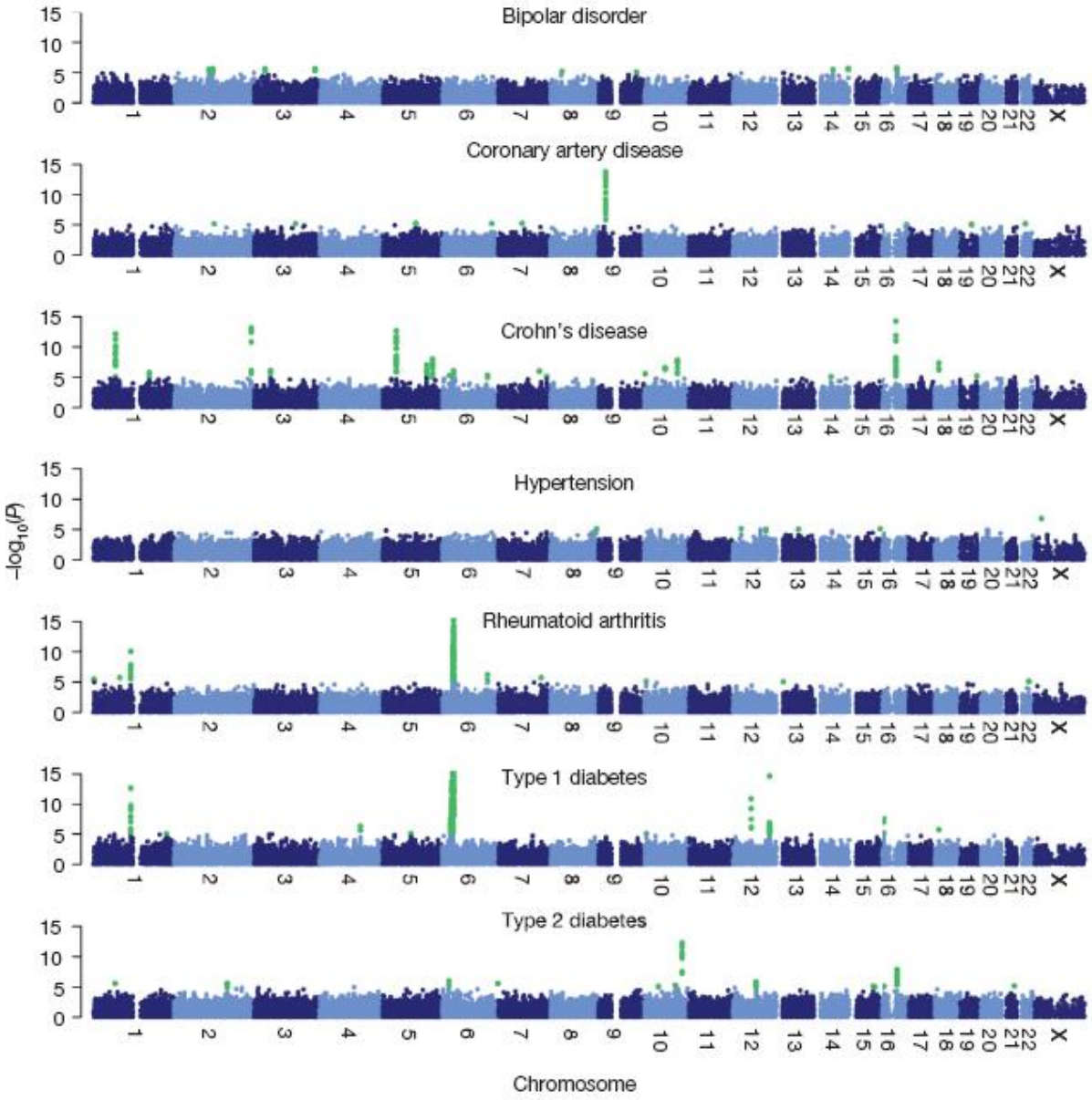
ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these

GWAS


Bipolar disorder
Coronary artery disease
Crohn's disease
Hypertension
Rheumatoid arthritis
Type 1 diabetes
Type 2 diabetes
$-\log_{10}(P)$
Chromosome

GWAS

Table 3 | Regions of the genome showing the strongest association signals

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
					Standard analysis								
BD	16p12	23.3–23.62	rs420259	2.19×10^{-04}	6.29×10^{-08}	1.96	4.79	A	G	2.08 (1.60–2.71)	2.07 (1.6–2.69)	0.282	0.248
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	11.66	11.19	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554
CD	1p31	67.3–67.48	rs11805303	6.45×10^{-13}	5.85×10^{-12}	10.07	9.41	T	T	1.39 (1.22–1.58)	1.86 (1.54–2.24)	0.317	0.391
CD	2q37	233.92–234	rs10210302	7.10×10^{-14}	5.26×10^{-14}	11.11	11.28	T	C	1.19 (1.01–1.41)	1.85 (1.56–2.21)	0.481	0.402
CD	3p21	49.3–49.87	rs9858542	7.71×10^{-07}	3.58×10^{-08}	4.24	5.22	A	A	1.09 (0.96–1.24)	1.84 (1.49–2.26)	0.282	0.331
CD	5p13	40.32–40.66	rs17234657	2.13×10^{-13}	1.99×10^{-12}	10.41	9.89	G	G	1.54 (1.34–1.76)	2.32 (1.59–3.39)	0.125	0.181
CD	5q33	150.15–150.31	rs1000113	5.10×10^{-08}	3.15×10^{-07}	5.36	5.01	T	T	1.54 (1.31–1.82)	1.92 (0.92–4.00)	0.067	0.098
CD	10q21	64.06–64.31	rs10761659	2.68×10^{-07}	1.75×10^{-06}	4.69	4.13	G	A	1.23 (1.05–1.45)	1.55 (1.3–1.84)	0.461	0.406
CD	10q24	101.26–101.32	rs10883365	1.41×10^{-08}	5.82×10^{-08}	5.91	5.48	G	G	1.2 (1.03–1.39)	1.62 (1.37–1.92)	0.477	0.537
CD	16q12	49.02–49.4	rs17221417	9.36×10^{-12}	3.98×10^{-11}	8.93	8.47	G	G	1.29 (1.13–1.46)	1.92 (1.58–2.34)	0.287	0.356
CD	18p11	12.76–12.91	rs2542151	4.56×10^{-08}	2.03×10^{-07}	5.42	5.00	G	G	1.3 (1.14–1.48)	2.01 (1.46–2.76)	0.163	0.208
RA	1p13	113.54–114.16	rs6679677	4.90×10^{-26}	5.55×10^{-25}	22.36	21.99	A	A	1.98 (1.72–2.27)	3.32 (1.93–5.69)	0.096	0.168
RA	6	MHC	rs6457617*	3.44×10^{-76}	5.18×10^{-75}	74.84	73.18	T	T	2.36 (1.97–2.84)	5.21 (4.31–6.30)	0.489	0.685
T1D	1p13	113.54–114.16	rs6679677	1.17×10^{-26}	5.43×10^{-26}	23.07	22.83	A	A	1.82 (1.59–2.09)	5.19 (3.15–8.55)	0.096	0.169
T1D	6	MHC	rs9272346*	2.42×10^{-134}	5.47×10^{-134}	141.9	142.2	A	G	5.49 (4.83–6.24)	18.52 (27.03–12.69)	0.387	0.150
T1D	12q13	54.64–55.09	rs11171739	1.14×10^{-11}	9.71×10^{-11}	8.89	8.24	C	C	1.34 (1.17–1.54)	1.75 (1.48–2.06)	0.423	0.493
T1D	12q24	109.82–111.49	rs17696736	2.17×10^{-15}	1.51×10^{-14}	12.53	11.88	G	G	1.34 (1.16–1.53)	1.94 (1.65–2.29)	0.424	0.506
T1D	16p13	10.93–11.37	rs12708716	9.24×10^{-08}	4.92×10^{-07}	5.15	4.70	A	G	1.19 (0.97–1.45)	1.55 (1.27–1.89)	0.350	0.297
T2D	6p22	20.63–20.84	rs9465871	1.02×10^{-06}	3.34×10^{-07}	4.15	3.98	C	C	1.18 (1.04–1.34)	2.17 (1.6–2.95)	0.178	0.218
T2D	10q25	114.71–114.81	rs4506565	5.68×10^{-13}	5.05×10^{-12}	10.14	9.43	T	T	1.36 (1.2–1.54)	1.88 (1.56–2.27)	0.324	0.395
T2D	16q12	52.36–52.41	rs9939609	5.24×10^{-08}	1.91×10^{-07}	5.35	5.05	A	A	1.34 (1.17–1.52)	1.55 (1.3–1.84)	0.398	0.453
					Multi-locus analysis								
T1D	4q27	123.26–123.92	rs6534347	4.48×10^{-07}	1.83×10^{-06}	5.15	4.69	A	A	1.30 (1.10–1.55)	1.49 (1.25–1.78)	0.351	0.402
T1D	12p13	9.71–9.86	rs3764021	7.19×10^{-05}	5.08×10^{-08}	2.12	4.55	C	T	1.57 (1.38–1.79)	1.48 (1.25–1.75)	0.467	0.426
					Sex differentiated analysis								
RA	7q32	130.80–130.84	rs11761231	3.91×10^{-07}	1.37×10^{-06}	-	-	G	A	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
					Combined cases								
RA+T1D	10p15	6.07–6.17	rs2104286	5.92×10^{-08}	2.52×10^{-07}	5.26	4.45	T	C	1.35 (1.11–1.65)	1.62 (1.34–1.97)	0.286	0.245

GWAS

Table 3 | Regions of the genome showing the strongest association signals

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
Standard analysis													
BD	16p12	23.3–23.62	rs420259	2.19×10^{-04}	6.29×10^{-08}	1.96	4.79	A	G	2.08 (1.60–2.71)	2.07 (1.6–2.69)	0.282	0.248
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	11.66	11.19	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554
CD	1p31	67.3–67.48	rs11805303	6.45×10^{-13}	5.85×10^{-12}	10.07	9.41	T	T	1.39 (1.22–1.58)	1.86 (1.54–2.24)	0.317	0.391
CD	2q37	233.92–234	rs10210302	7.10×10^{-14}	5.26×10^{-14}	11.11	11.28	T	C	1.19 (1.01–1.41)	1.85 (1.56–2.21)	0.481	0.402
CD	3p21	49.3–49.87	rs9858542	7.71×10^{-07}	3.58×10^{-08}	4.24	5.22	A	A	1.09 (0.96–1.24)	1.84 (1.49–2.26)	0.282	0.331
CD	5p13	40.32–40.66	rs17234657	2.13×10^{-13}	1.99×10^{-12}	10.41	9.89	G	G	1.54 (1.34–1.76)	2.32 (1.59–3.39)	0.125	0.181
CD	5q33	150.15–150.31	rs1000113	5.10×10^{-08}	3.15×10^{-07}	5.36	5.01	T	T	1.54 (1.31–1.82)	1.92 (0.92–4.00)	0.067	0.098
CD	10q21	64.06–64.31	rs10761659	2.68×10^{-07}	1.75×10^{-06}	4.69	4.13	G	A	1.23 (1.05–1.45)	1.55 (1.3–1.84)	0.461	0.406
T2D	10q25	114.71–114.81	rs4506565	5.68×10^{-13}	5.05×10^{-12}	10.14	9.43	T	T	1.36 (1.2–1.54)	1.88 (1.56–2.27)	0.324	0.395
T2D	16q12	52.36–52.41	rs9939609	5.24×10^{-08}	1.91×10^{-07}	5.35	5.05	A	A	1.34 (1.17–1.52)	1.55 (1.3–1.84)	0.398	0.453
Multi-locus analysis													
T1D	4q27	123.26–123.92	rs6534347	4.48×10^{-07}	1.83×10^{-06}	5.15	4.69	A	A	1.30 (1.10–1.55)	1.49 (1.25–1.78)	0.351	0.402
T1D	12p13	9.71–9.86	rs3764021	7.19×10^{-05}	5.08×10^{-08}	2.12	4.55	C	T	1.57 (1.38–1.79)	1.48 (1.25–1.75)	0.467	0.426
Sex differentiated analysis													
RA	7q32	130.80–130.84	rs11761231	3.91×10^{-07}	1.37×10^{-06}	-	-	G	A	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
Combined cases													
RA+T1D	10p15	6.07–6.17	rs2104286	5.92×10^{-08}	2.52×10^{-07}	5.26	4.45	T	C	1.35 (1.11–1.65)	1.62 (1.34–1.97)	0.286	0.245

Trend *P* value | Genotypic *P* value | $\log_{10}$(BF), additive | $\log_{10}$(BF), general | Risk allele | Minor allele | Heterozygote odds ratio | Homozygote

Quality Control

At SNP level:

-> Missing rate (95%, 99%)
-> HWE ($p < 0.001$)
-> MAF (5%)

At sample level:

-> Call rate (e.g. missing) (95%, 99%)
-> Sex discrepancies
-> Heterozygosity rate
-> Duplicated or related individuals
-> Divergent ancestry

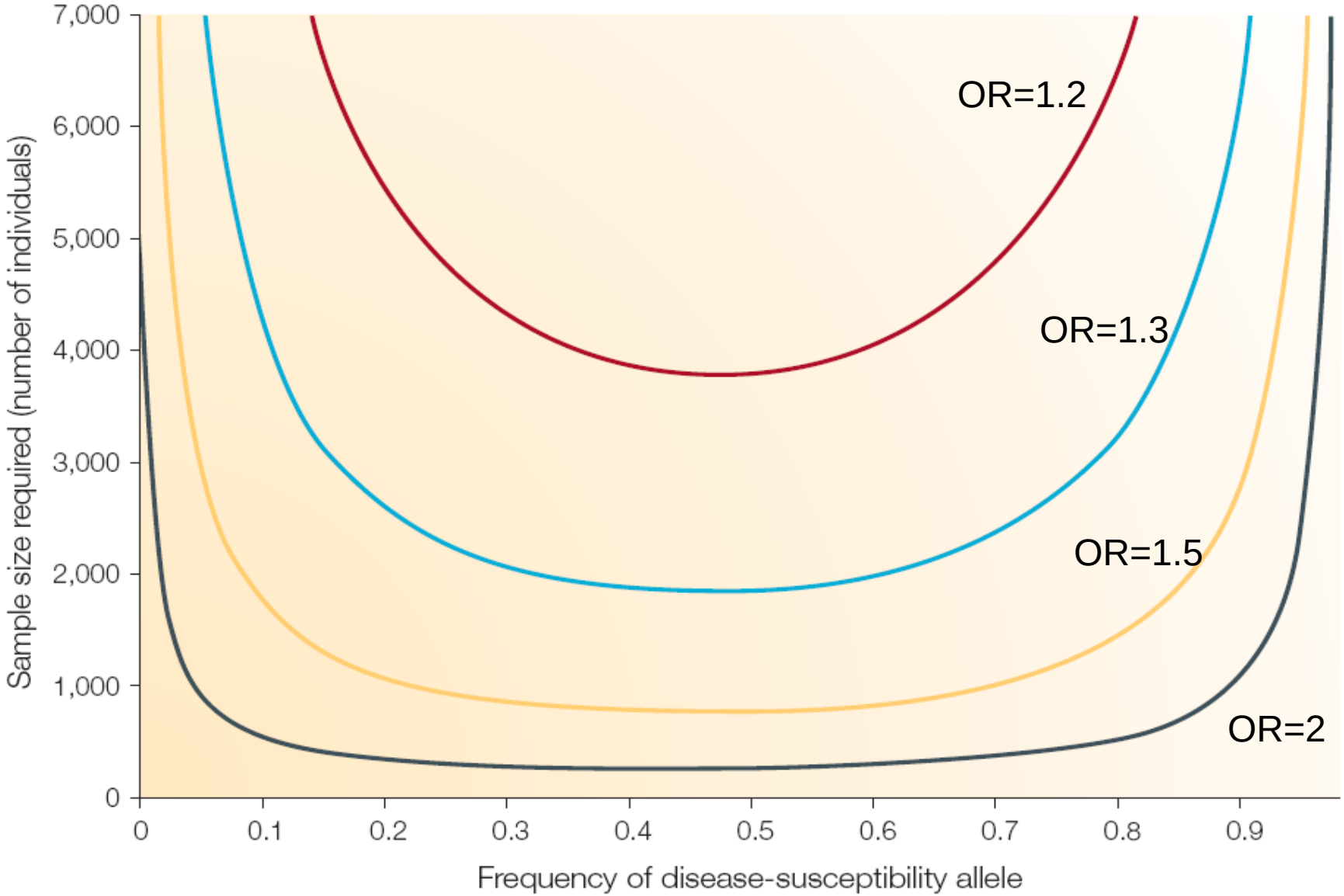
GWAS

Problem 1: Sample size
-> Multi-stage approach

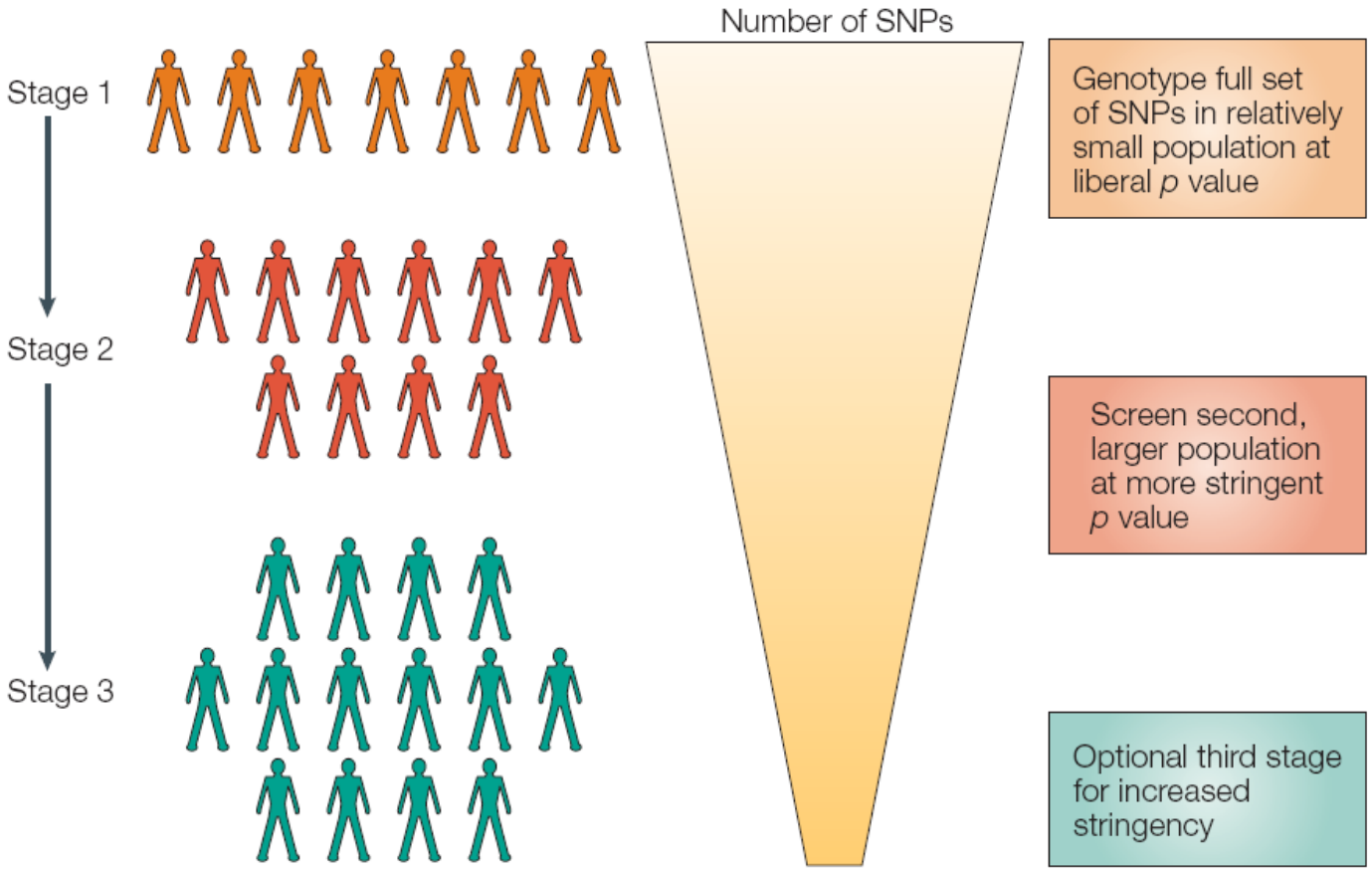
Problem 2: Population Stratification
-> Principal Component Analysis

Problem 2: Multiple comparisons
-> Bonferroni Correction
-> Permutation test

Sample size (β=80% α=10^{-6})



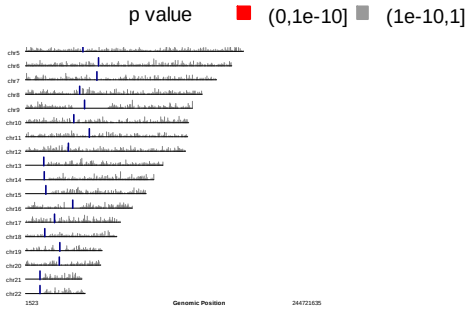
Multi-stage approach

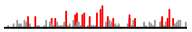


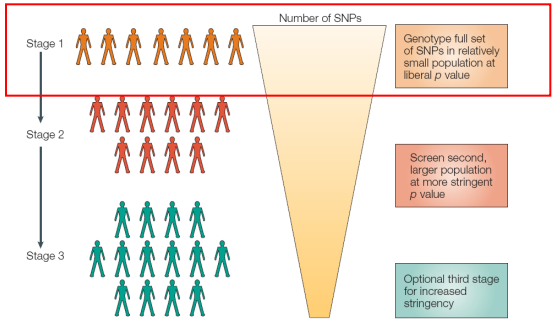
Hirschhorn and Daly (2005), Nature Reviews

Multi-stage approach

Genetic model: log-additive



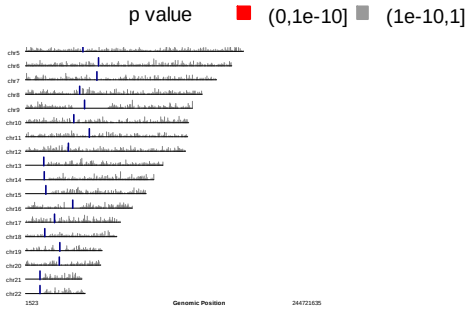


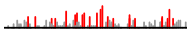


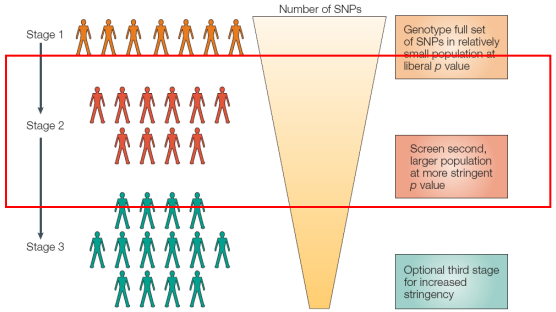
Adaptep from González et al, Bioinformatics, 2007

Multi-stage approach

Genetic model: log-additive

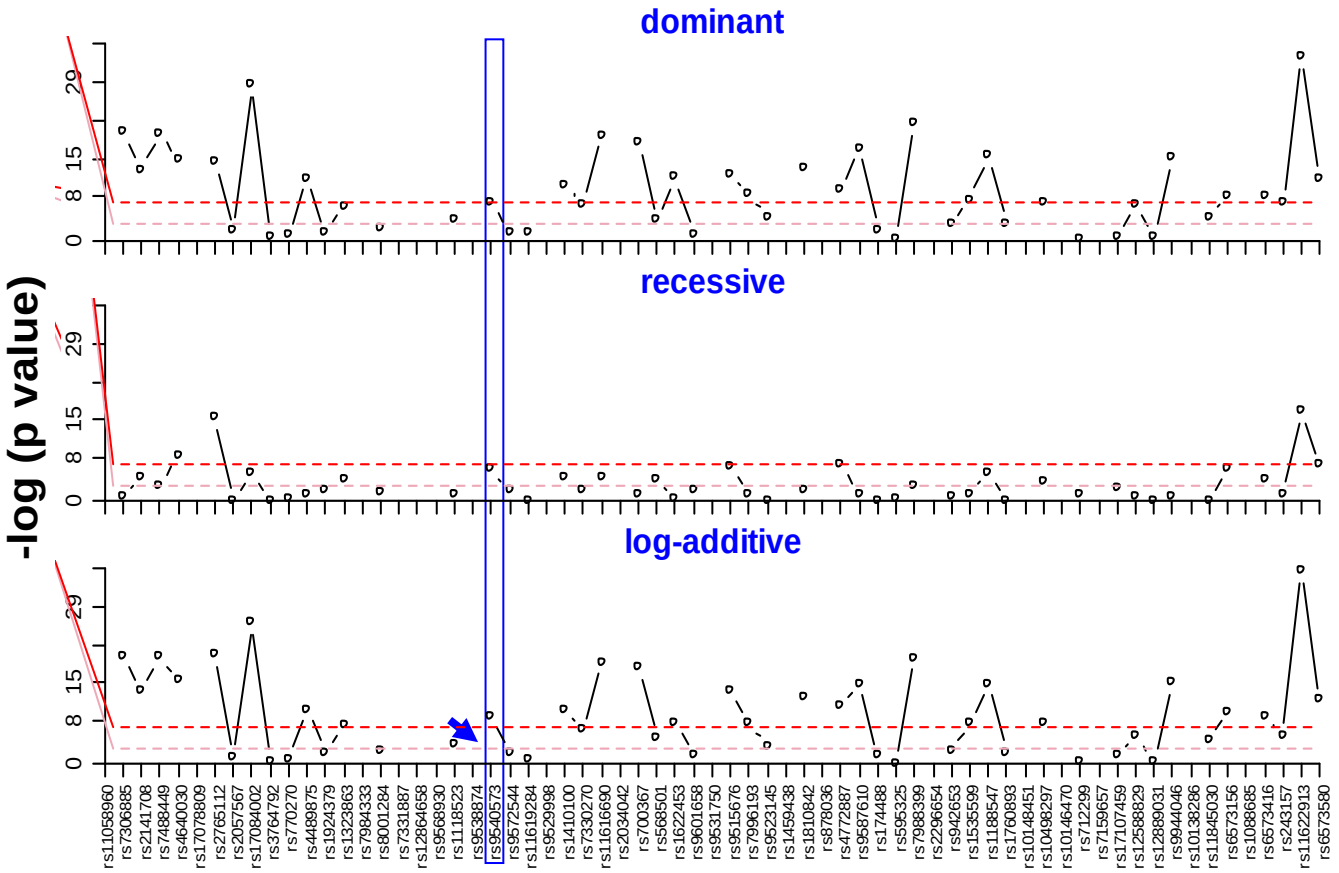






Adaptep from González et al, Bioinformatics, 2007

Association



Multi-stage approach

- To evaluate association using an unique inheritance model (additive or max-statistic)

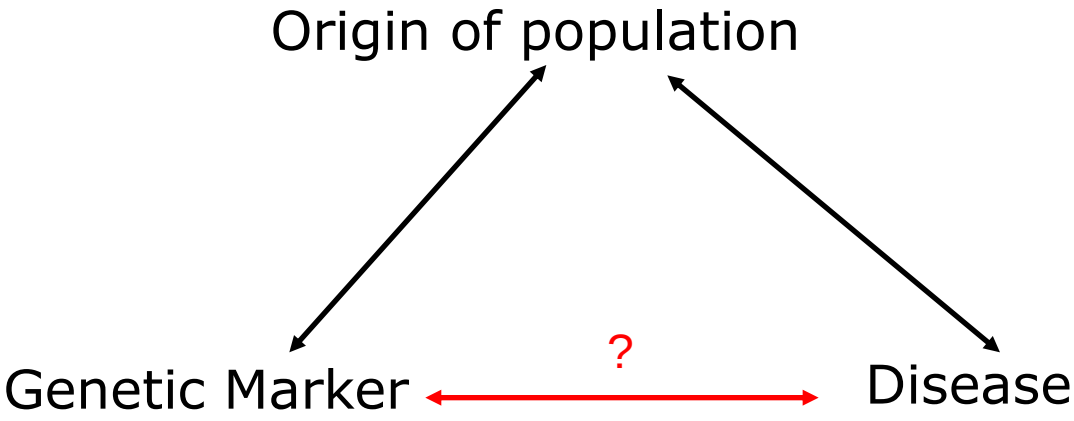
- Re-genotype using another sample (large) those SNPs that are significant using a non-stringent p-value (p=0.1) and assess association using these SNPs

- To validate association using an independent sample only for those SNPs that are statistically significant using a more stringent p-value (Bonferroni)

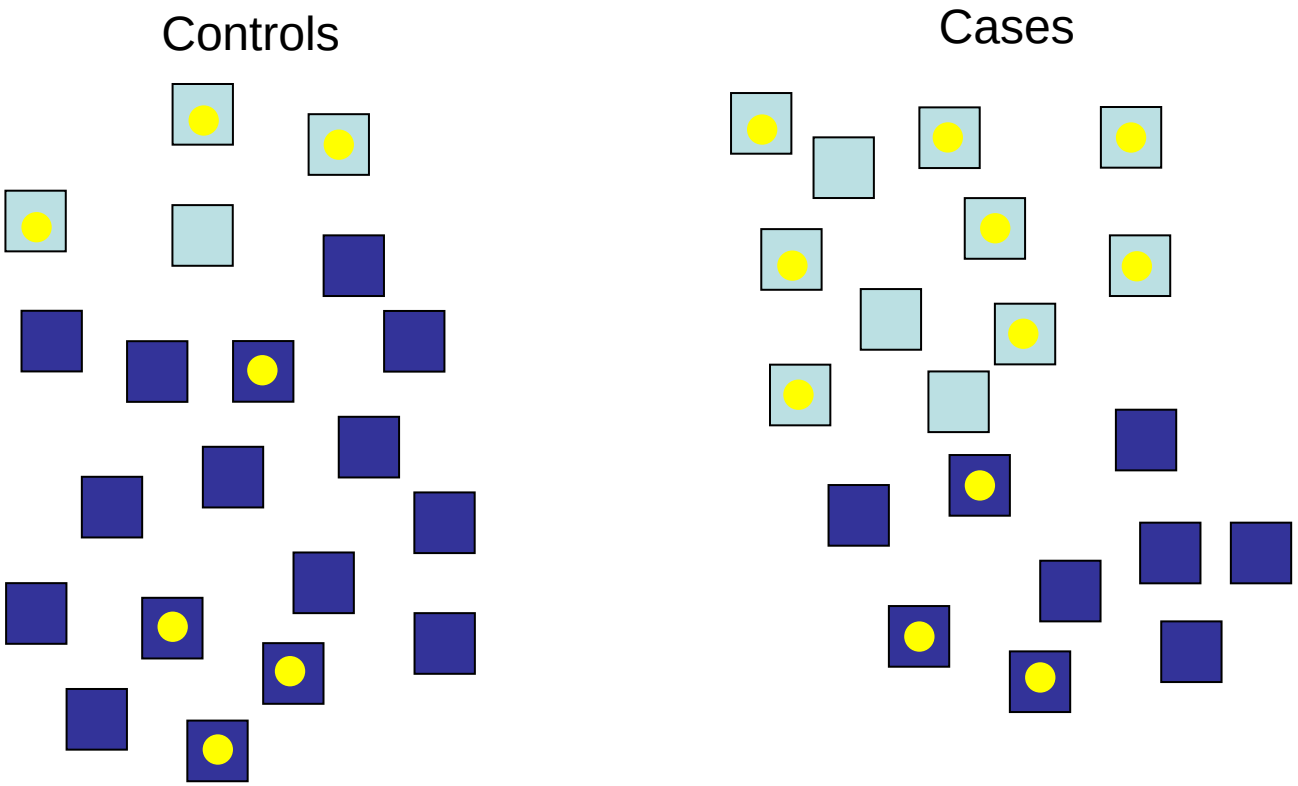
- To estimate OR and CI05% using different inheritance models

Population stratification

Stratification



Stratification



Susceptibility allele

Population A

Population B

Cases 11/20 vs Controls: 7/20 -> OR=2.27

Cases 8/11 vs Controls: 3/4 -> OR=0.89

Cases 3/9 vs Controls: 4/16 -> OR=1.08

Stratification

How to address the problem

- At design
  - Race-matched design
  - Relatives as controls (under-powered)

- Statistical methods
  - Devlin's method (genomic control)
  - Clayton's approach
  - EIGENSTRAT
  - STRUCTURE

Stratification

- If stratification is presetn, chi-square of association is inflated -> λ
- Algorithm:
  - To estimate λ
  - Re-compute chi-square test statistic
  - To compute corrected p-value
- Two methods
  - Devlin's method (genomic control)
  - Clayton's approach

Stratification

Devlin and Roeder, Biometrics, 1999

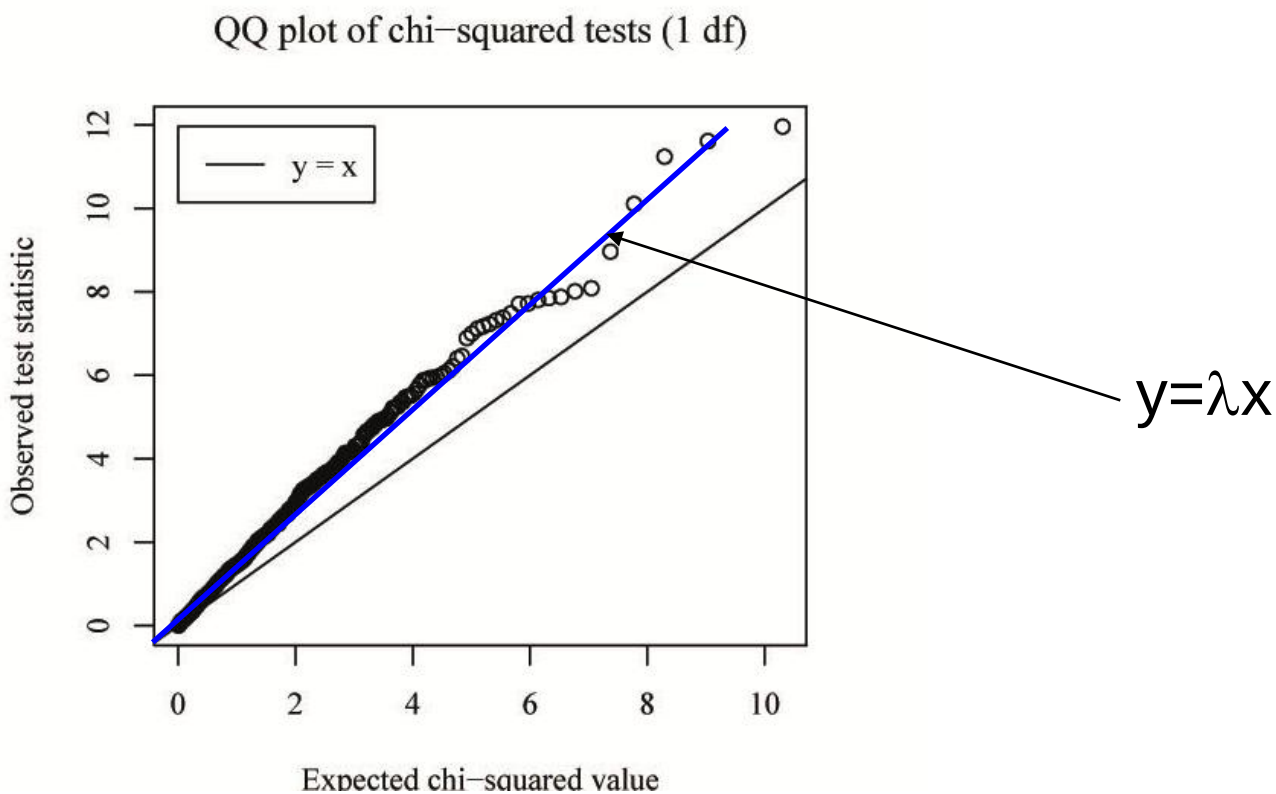
- Inflation factor, λ, can be estimated using:

$$\hat{\lambda} = \{\text{median}(X_1, X_2, \ldots, X_n)/0.675\}^2$$

Stratification

Clayton et al, Nature Genetics, 2005

- Inflation factor, λ can be estimated using a q-q plot

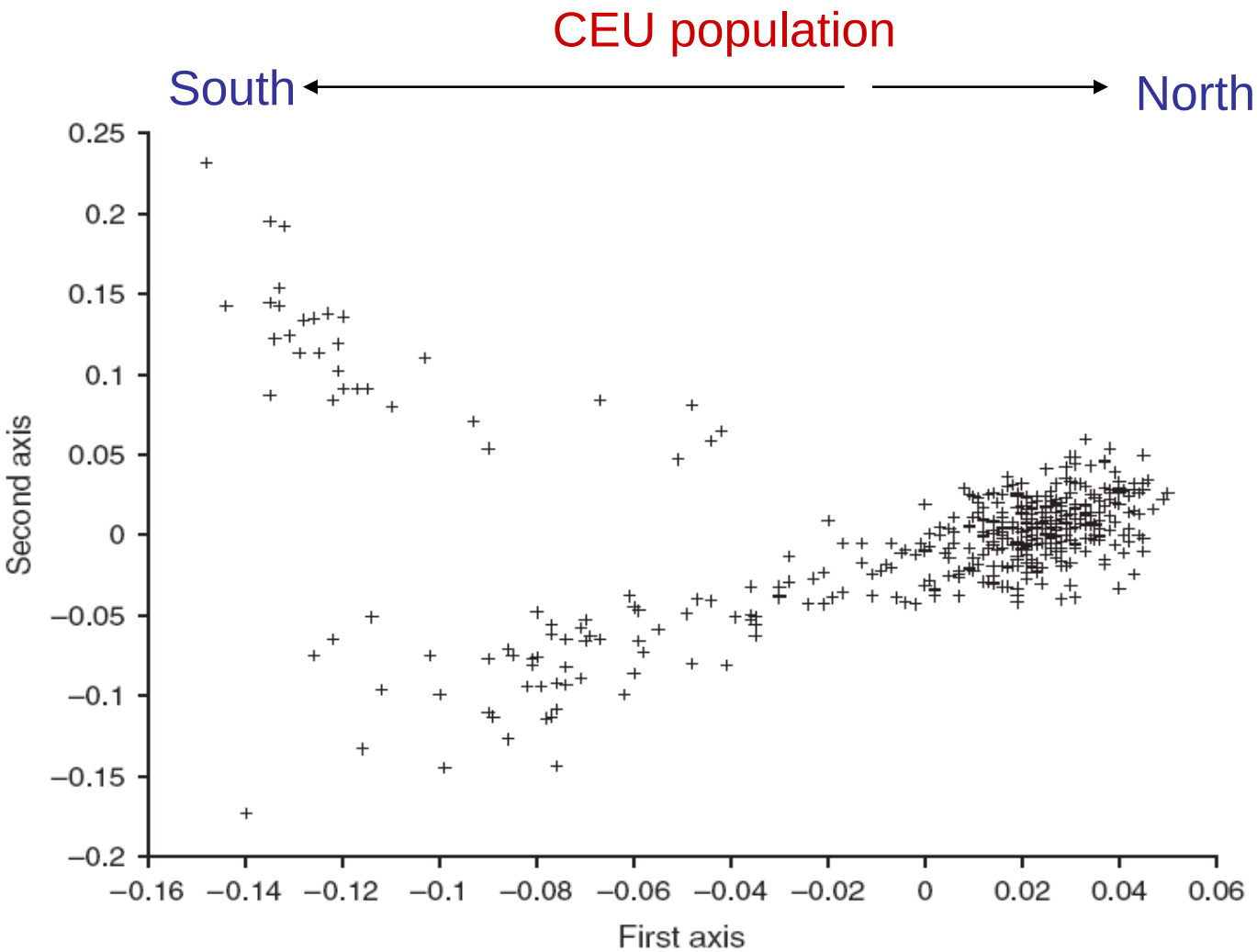


Stratification

Price et al., Nature Genetics, 2006 (EIGENSTRAT)

- Step A: To use principal component analysis (eigen values) using genotypes to infer continuous axes that account for genetic variation
- Step B: To adjust fenotipe and genotypes using attributable fraction of each axis
- Step C: To assess association adjusting for this fraction (i.e., logistic regression adjusting for the 3rd and 4th principal components)

Stratification



Multiple comparisons

How does it appear?

- When more than 1 phenotype is analyzed
- When more than one technique is used to analyze the same dataset. For instance, association with a SNP and with haplotype
- When using different tests
  - Dominant, recessive and additive
  - Crude and adjusted analysis
- When assessing association with more than 1 SNP **-> GWAS**

Multiple testing: approaches

- Bonferroni / Sidak's correction
- False Discovery Rate (q-value)
- Permutation Testing

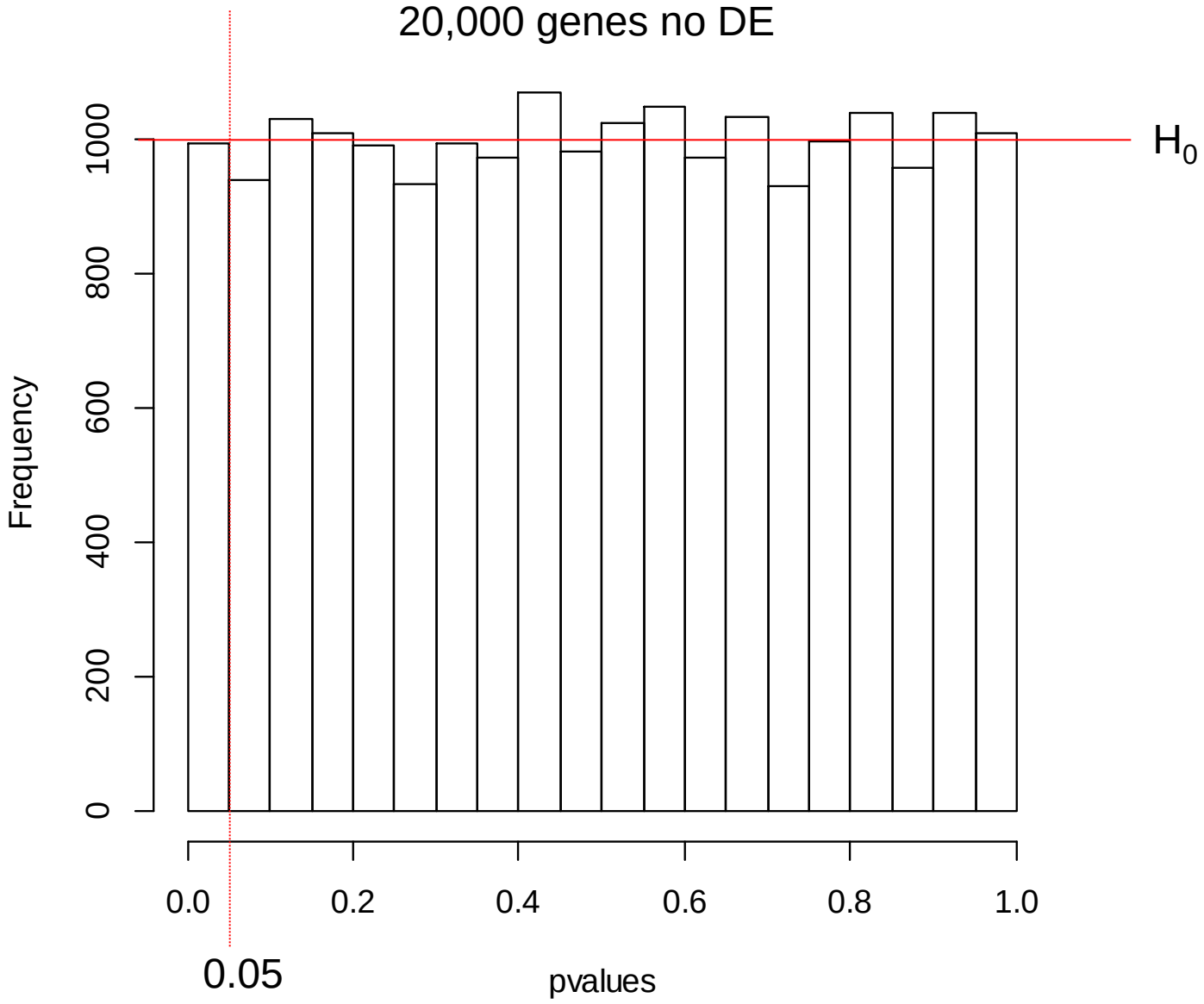
Bonferroni correction

- Testing 500,000 SNPs …
  - 5,000 are expected to be significant at $p<0.01$ level
  - 500 are expected to be significant at $p<0.001$ level
  - ……
  - 5 are expected to be significant at $p<10^{-5}$ level

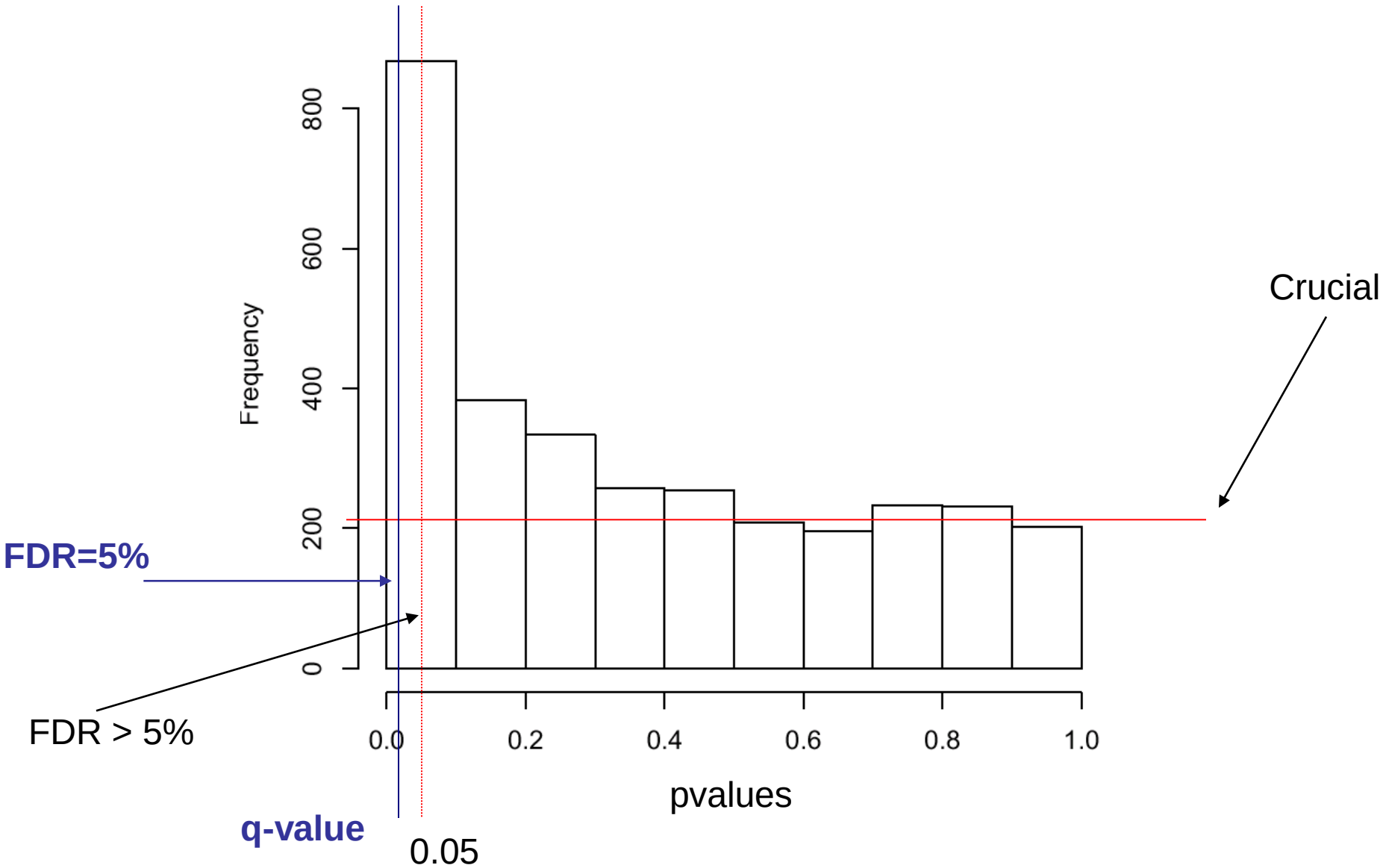
- Bonferroni correction when testing *m* markers
  - Consider significant level $\alpha = 0.05 / m$
- Sidak's correction when testing *m* markers
  - $\alpha^* = 1-(1-\alpha)^m$

- *See Risch and Merikangas 1999

False Discovery rate (q-value)



False Discovery rate (q-value)



Multiple comparisons

	Error control for		Appropriate for	
	Whole study	Single test	Association study	Expression study
Family wise error, strong	Yes (1)	Yes (1)	No	No
Family wise error, weak	Yes (1)	No	Yes	Yes
Minimum P-value	Yes (1)	Yes (1)	Somewhat	No
Truncated P-value product	Yes (1)	No	Yes	Possibly
Random gene effects model	Yes (1)	Yes (2)	Possibly	Yes
False discovery rate	Yes (3)	No	No	Yes
Q-value	Yes (3)	Some (3)	No	Yes
Local FDR	Yes (2)	Yes (2)	Yes	Yes
False positive reporting probability	Yes (3)	Some (3)	Yes	Yes

Source: Dudbridge, et al. (2006) Human Genomics

LocusZoom: http://csg.sph.umich.edu/locuszoom/

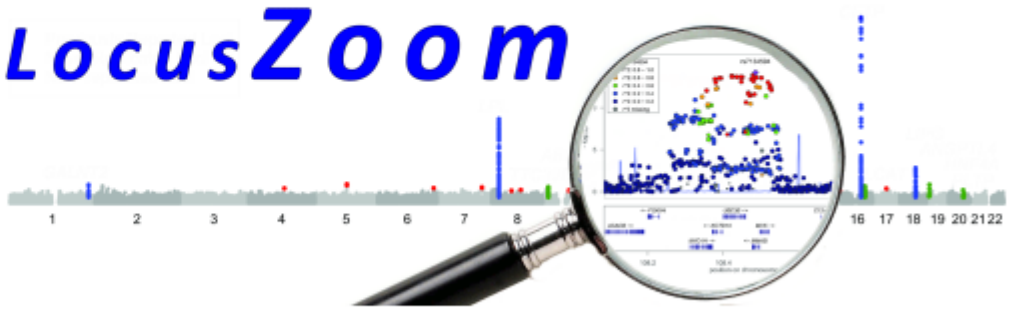
Login/Logout

UNIVERSITY OF MICHIGAN
CENTER FOR STATISTICAL GENETICS

Main
- CSG Home
- CSG News
- Members
- Photo Album
- Seminars
- Positions
- Documentation
- Symposia
- Links

Genome Science Training Program
- Overview
- Application Information
- Handbook 2012
- Members

Software
- Authors
- GOLD
- HAPLOTYPER
- LocusZoom
- MERLIN
- CaTS (Power Calculator)
- QTDT
- RELPAIR
- SIBMED
- Other



LocusZoom is a tool to plot regional association results from genome-wide association scans or candidate gene studies. This is Version 1.1

Report problems to cristen@umich.edu

We are pleased to announce that our paper on *LocusZoom* has been published. [ABSTRACT | PDF]

REFERENCE:
Pruim RJ*, Welch RP*, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. (2010) LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics 2010 September 15; 26(18): 2336.2337.

Count of Succe
- This week
- Year **2013**
- Jan
- Year **2012**
- Year **2011**
- Year **2010**

Links

- Plot Using Your Data

LocusZoom: http://csg.sph.umich.edu/locuszoom/

LocusZoom

LocusZoom - Plot with Your Data

Plot Your Data

Depending on the size of your data, runs can require 30-60 seconds to generate a plot

Provide Details for Your Data

Path to Your File: [ ] Examinar...

File will sent to server and used for plotting (Maximum 200MB)

[Help]

P-Value Column Name: [ ] Default is P.value

Set for PLINK data or WikiGWA data

Marker Column Name: [ ] Default is MarkerName

Column Delimiter: Tab Default is tab

Specify Region to Display

Required: Fill in Only ONE of These Three

SNP	SNP Reference Name	+/-	400 Kb Flanking Size	
Gene	Gene Reference Name	+/-	200 Kb Flanking Size	Optional Index SNP Default=lowest p-value
Region	Chr: None	Starting Chr Position: Mb	through Ending Chr Position: Mb	Optional Index SNP Default=lowest p-value

LocusZoom: http://csg.sph.umich.edu/locuszoom/