MRC | Epidemiology Unit

Some Perspectives of Graphical Methods for Genetic Data

Zhao JH, Q Tan, S Li, J Luan, W Qian, RJF Loos, NJ Wareham

jinghua.zhao@mrc-epid.cam.ac.uk
http://www.mrc-epid.cam.ac.uk/~jinghua.zhao

14 August 2008, Dortmund, Germany

useR! 2008

Outline

- Background
- Case studies
- Examples from R
- General discussion

Background

- This can be seen as an addition to a useR!2007 presentation.
  - ctv for genetics
  - identity, powerpkg, multic, lodplot, qtl
  - gap, genetics, haplo.stats (hapassoc,…), GenABEL, pbatR, SNPassoc, snpMatrix
- The general context is the promise of genetic analysis of complex traits (useR!2008 Tutorials) due to recent genotyping technology and characterization of human genome:
  - HapMap, http://www.hapmap.org
  - One thousand genome project

Consortium

- Wellcome Trust Case-Control Consortium (WTCCC): >17000 individuals on BD, CAD, CD, HT, RA, T1D, T2D
- DIAbetes Genetic Replication And Meta-analysis (DIAGRAM), >50000 individuals on T2D
- Genetic Investigation of ANthropometric Traits (GIANT): >32000 individuals followed by >58000 on obesity, weight, height and central adiposity
- Meta-Analysis of Glucose- and Insulin-related traits Consortium (MAGIC), >45000 individuals

Steps in Positional Cloning

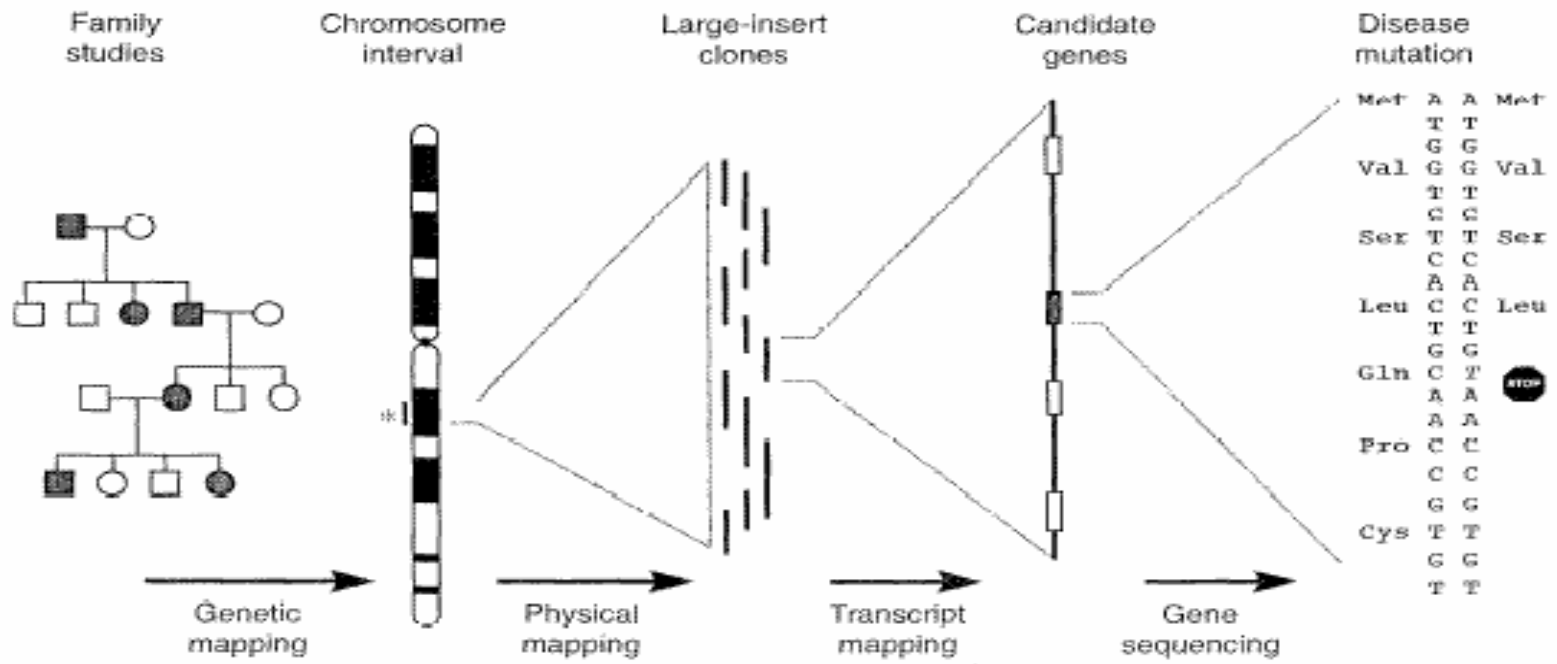
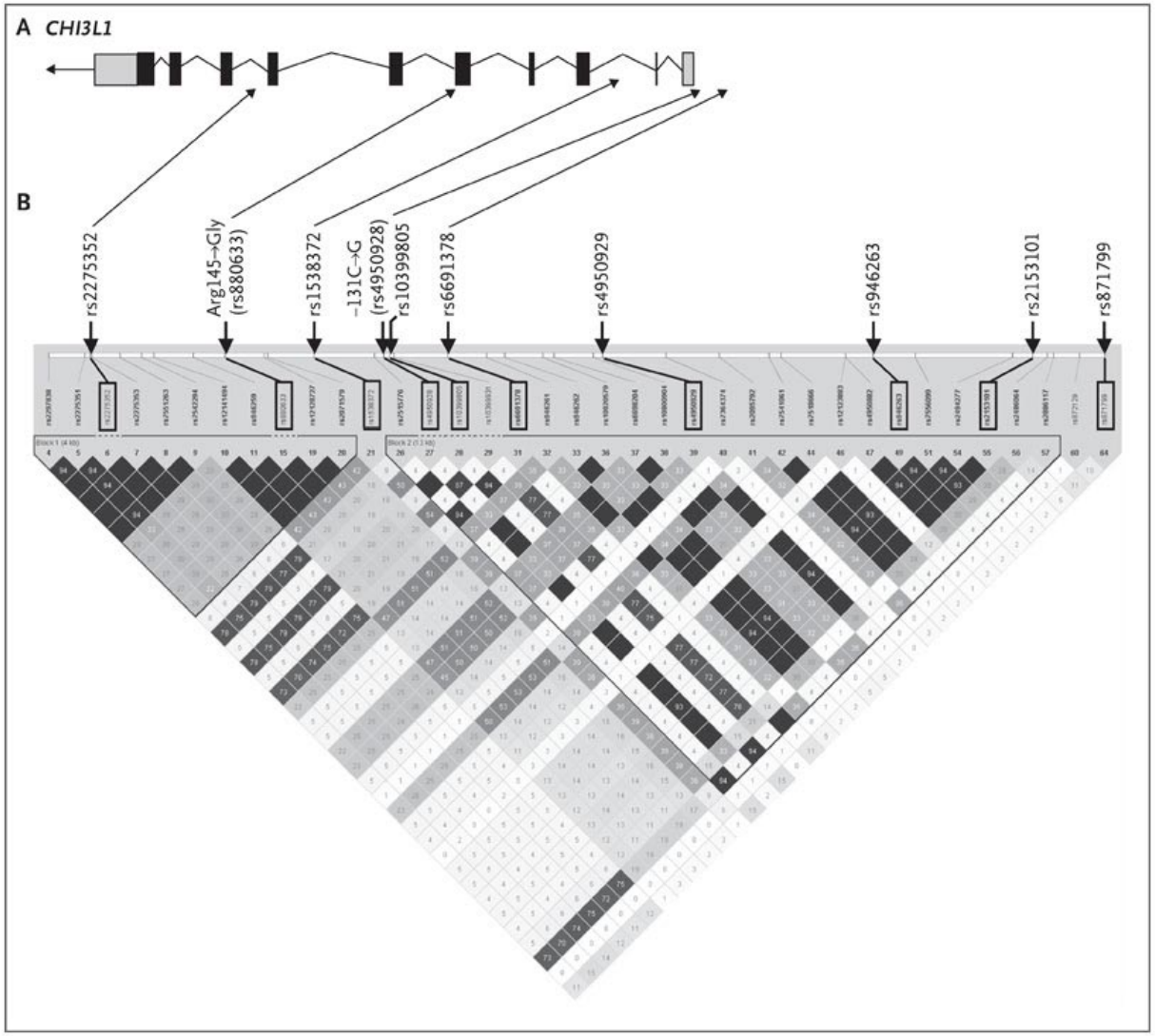


Fig. 1. Steps in positional cloning. Positioning of disease loci to chromosomal regions with genetic markers has become increasingly straightforward, particularly given the recent release of the Généthon genetic map containing 5264 markers (*17*). However, identification and evaluation of the genes within the implicated region remains a major stumbling block.

Schuler (1996) *Science*

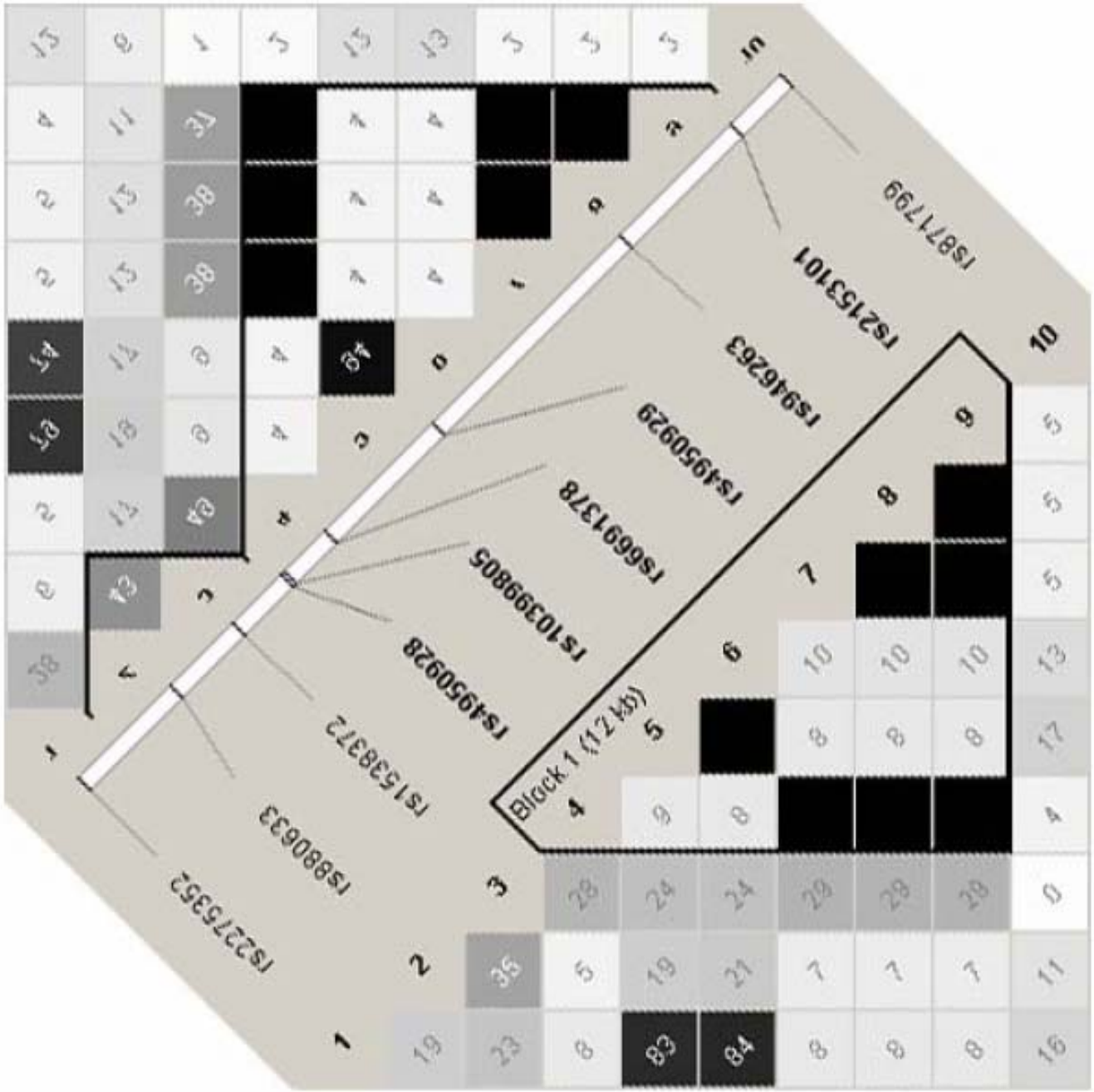
Aspects in need of graphical representation

- Phenotypic data
  - Individual data, e.g., two-way plot, conditional plot
  - Summary statistics
  - Specific features, e.g., pedigree diagram
- Genotypic data
  - Genome level, regional level, functional level
- Genotype-phenotype correlation
  - Q-Q plot
  - Manhattan plot
  - Regional plot
  - Forest plot
  - Receiver-operating-characteristic (ROC) curve



Single-Nucleotide polymorphisms (SNPs) in CHI3L1 and its upstream region on chromosome 1q32.1

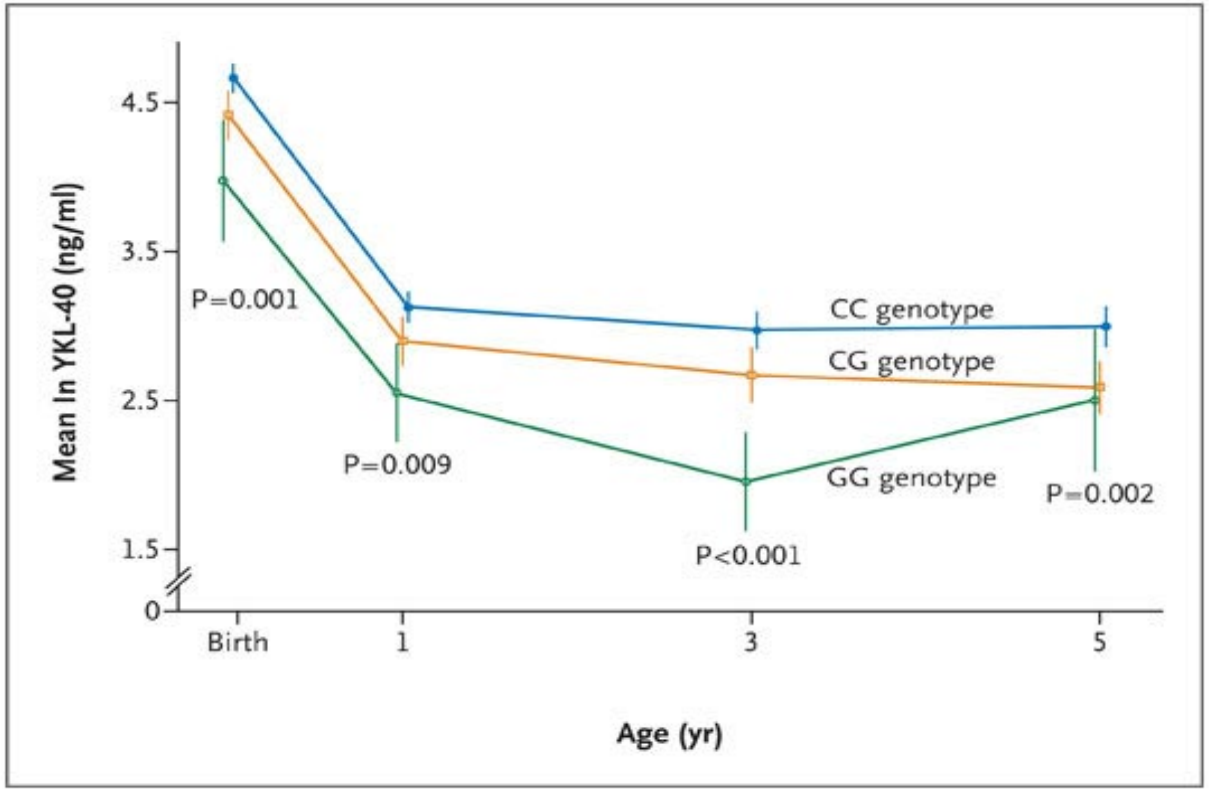
Ober et al. NEJM 2008



LD (r2) between

10 SNPs of CHI3L1 in Europeans (UL) and Hutterites (LR)

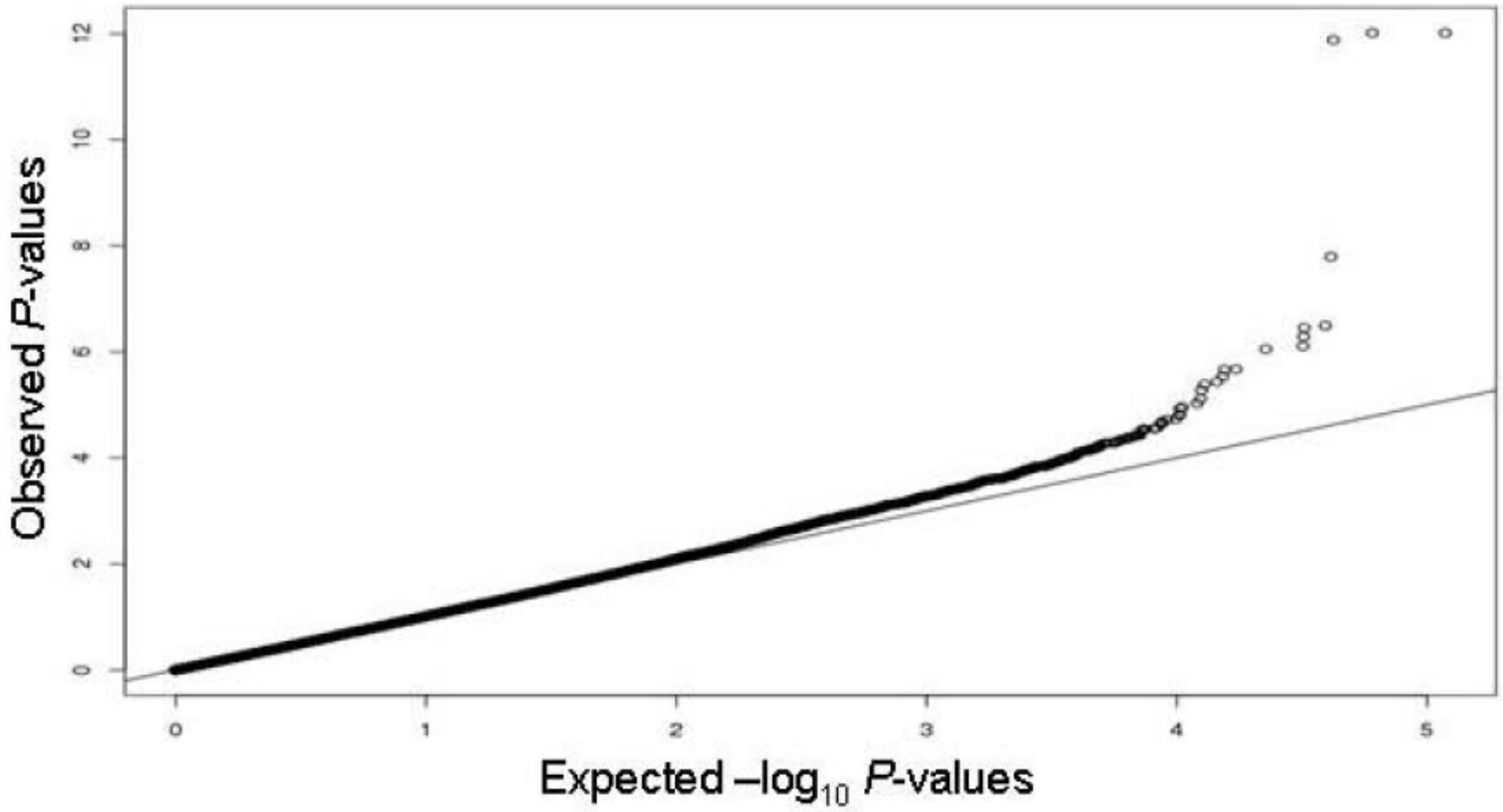
Ober et al. *NEJM* 2008



Mean serum YKL-40 levels in Asthma

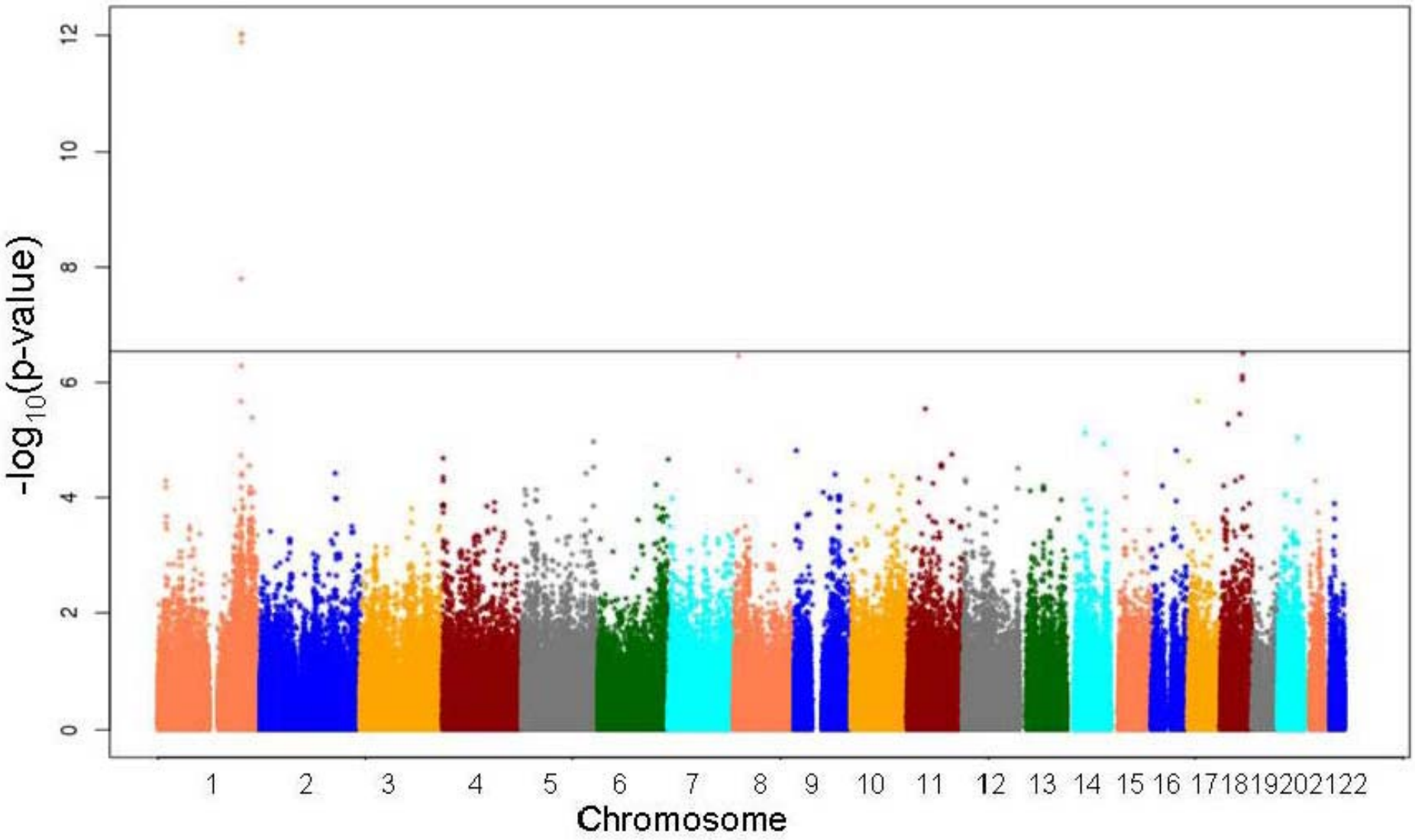
Ober et al. *NEJM* 2008

	CC Genotype			CG Genotype			GG Genotype			*P*-value
	N	Mean	SE	N	Mean	SE	N	Mean	SE	
Cord Blood	82	4.66	0.048	39	4.41	0.082	4	3.98	0.207	0.0010
Year 1	82	3.12	0.054	39	2.90	0.081	4	2.55	0.166	0.0089
Year 3	82	2.97	0.063	39	2.67	0.092	4	1.95	0.169	0.00025
Year 5	71	2.99	0.067	30	2.59	0.090	4	2.50	0.240	0.0016



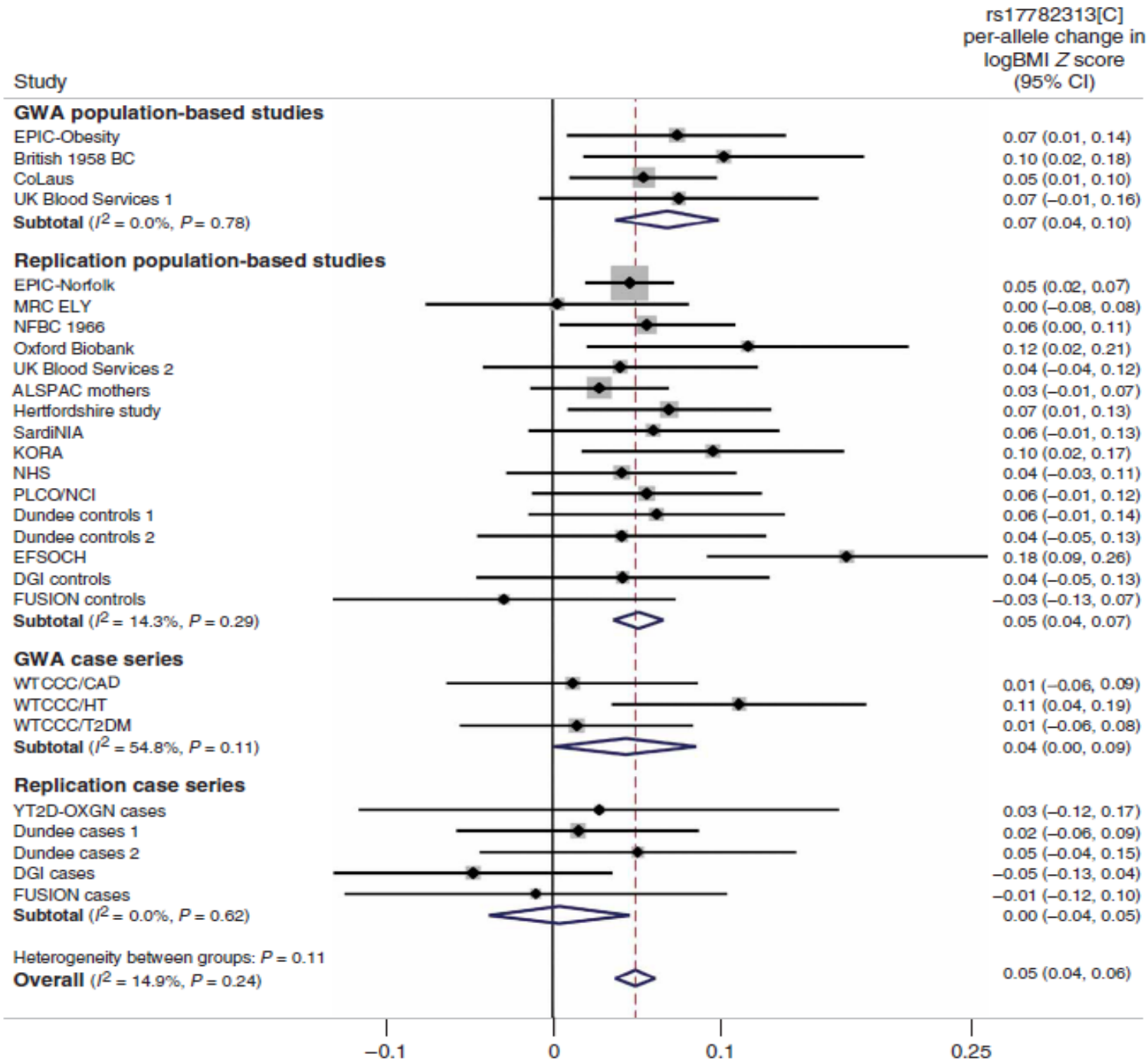
Q-Q Plot of the genome-wide P-values

Ober et al. *NEJM* 2008

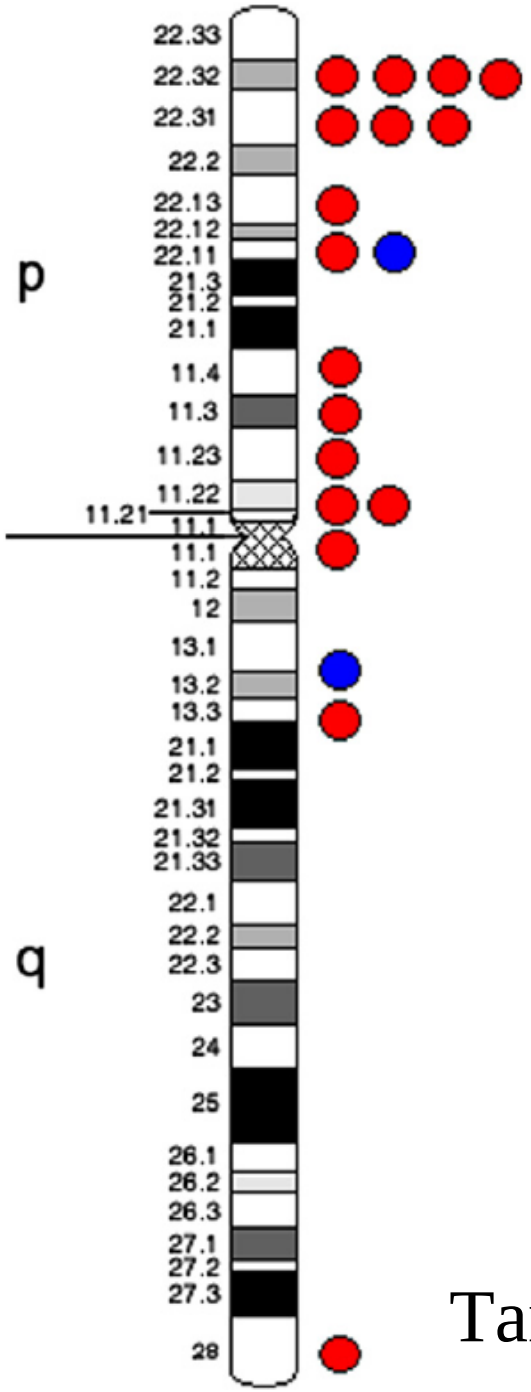


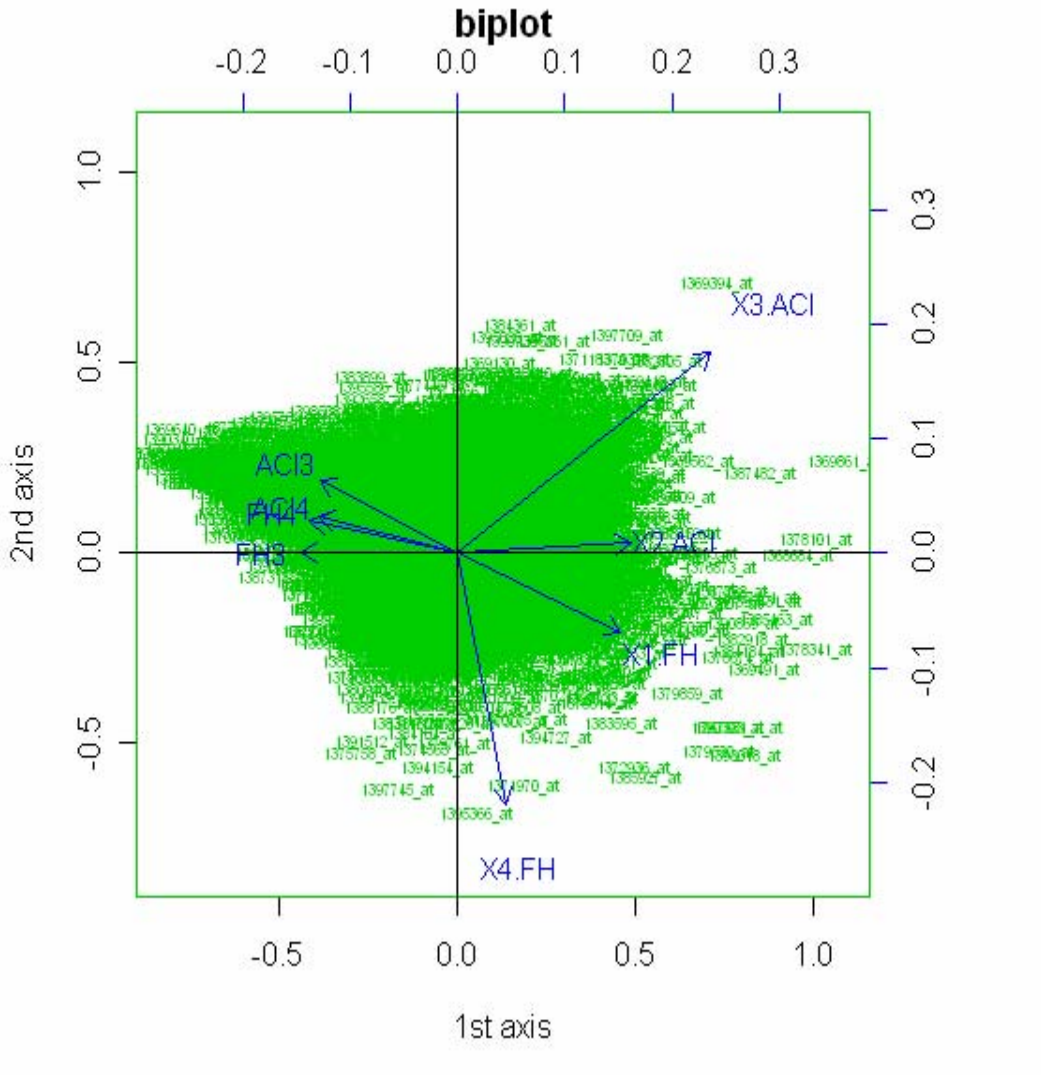
Genome-wide P-values and serum YKL-40 levels.

Ober et al. *NEJM* 2008

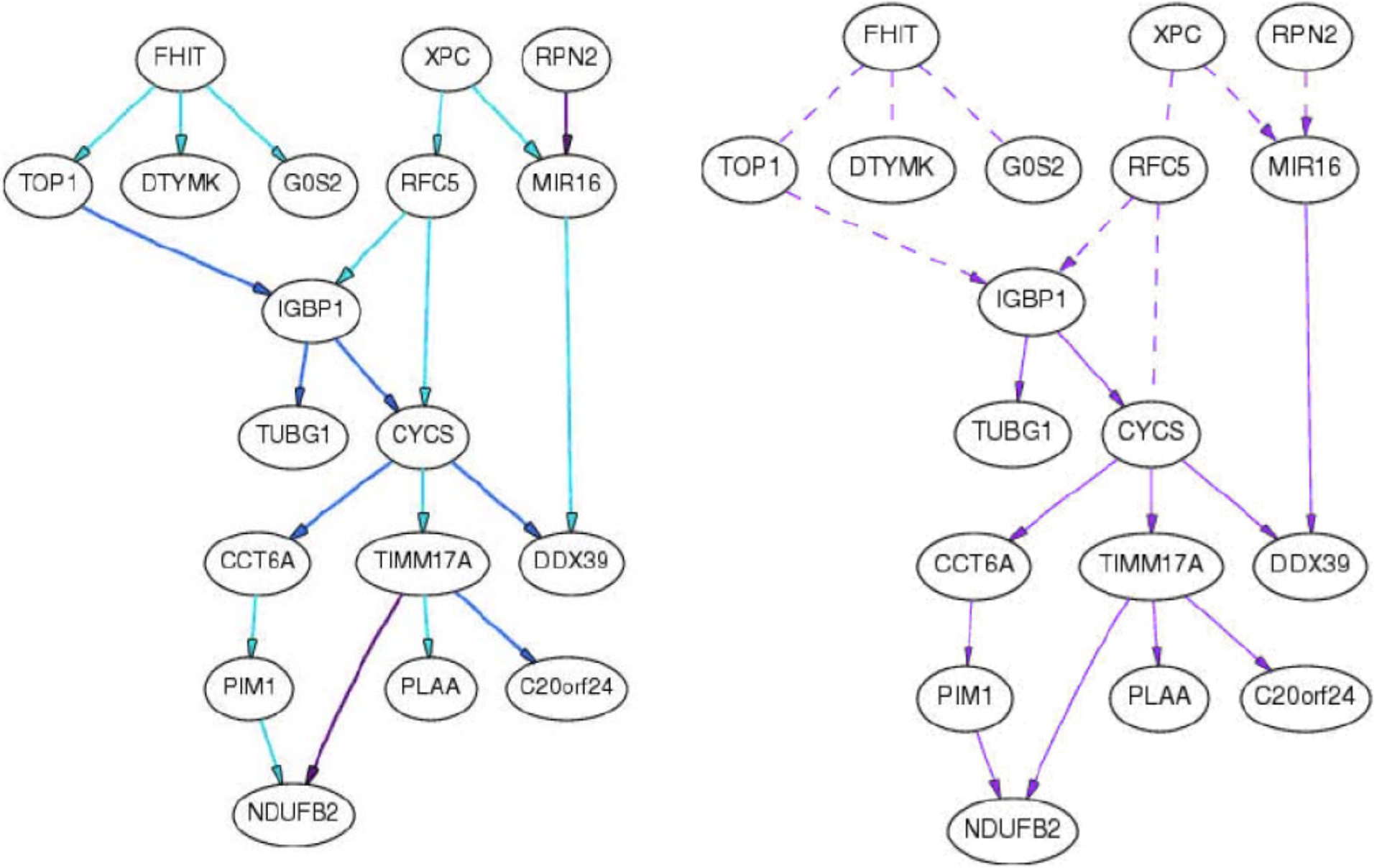


Loos et al.
Nat Genet
2008



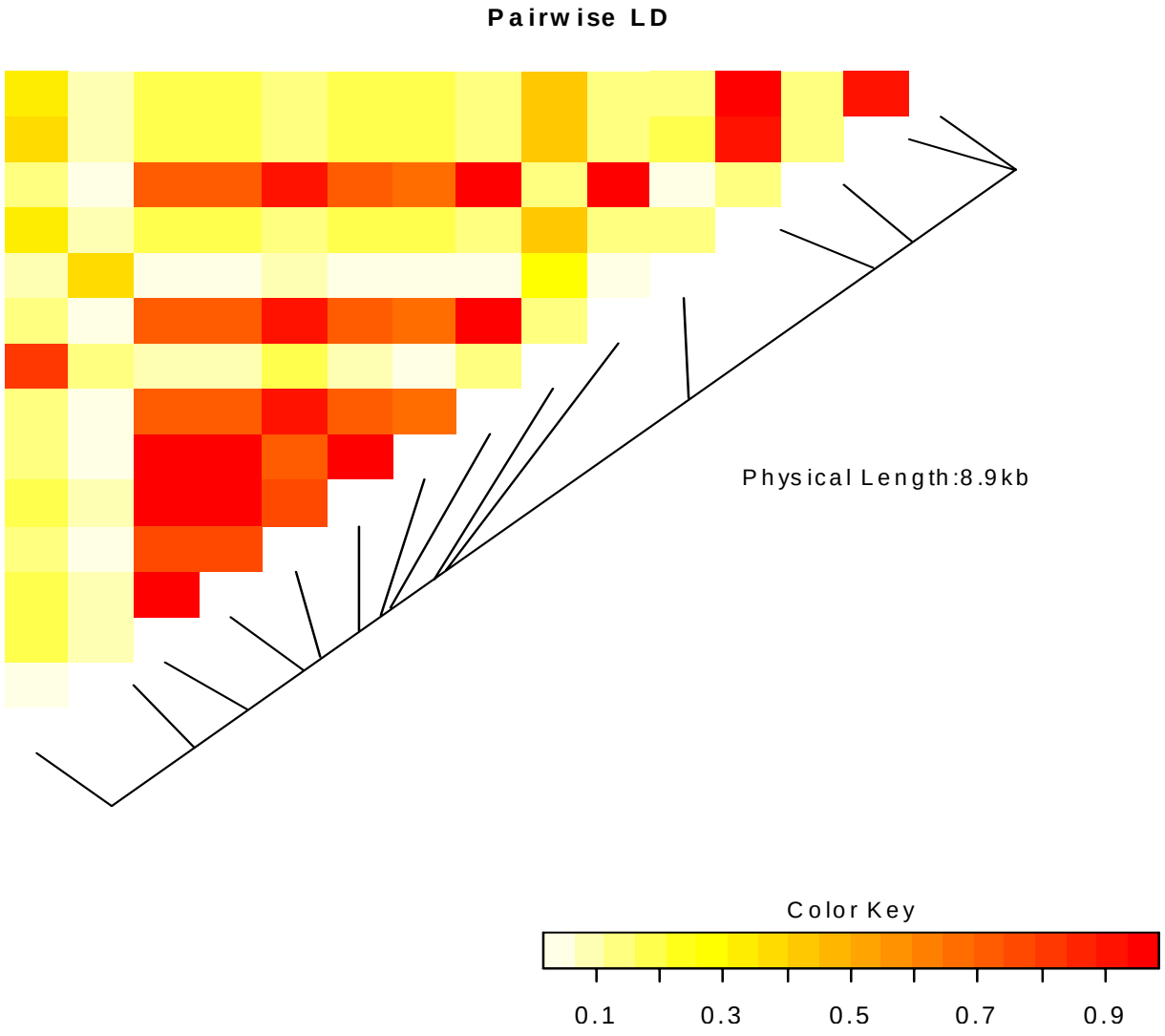


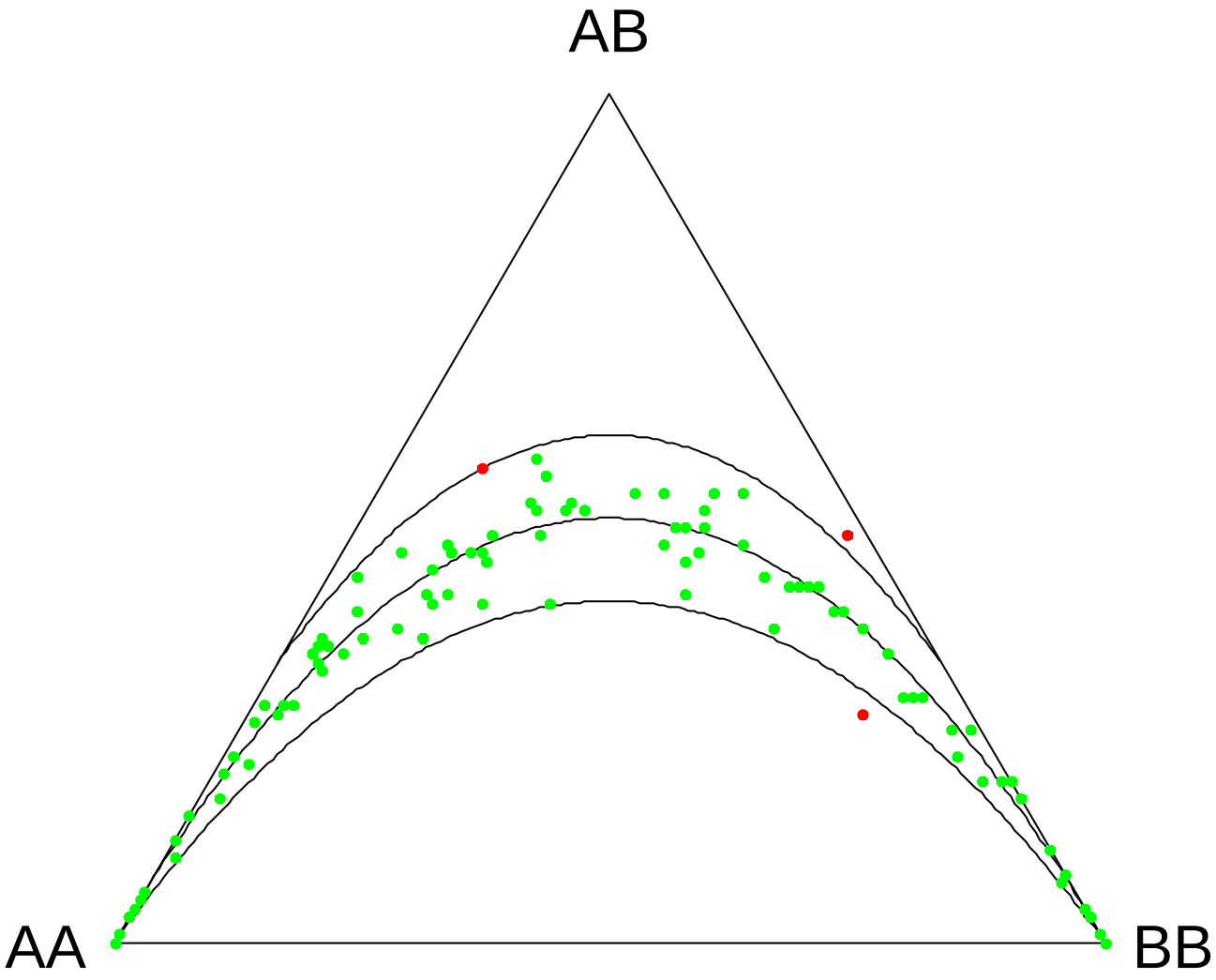
Tan et al. *Genomics* 2008 (and unpublished)



Zhao et al. *BMC Proc* 2007

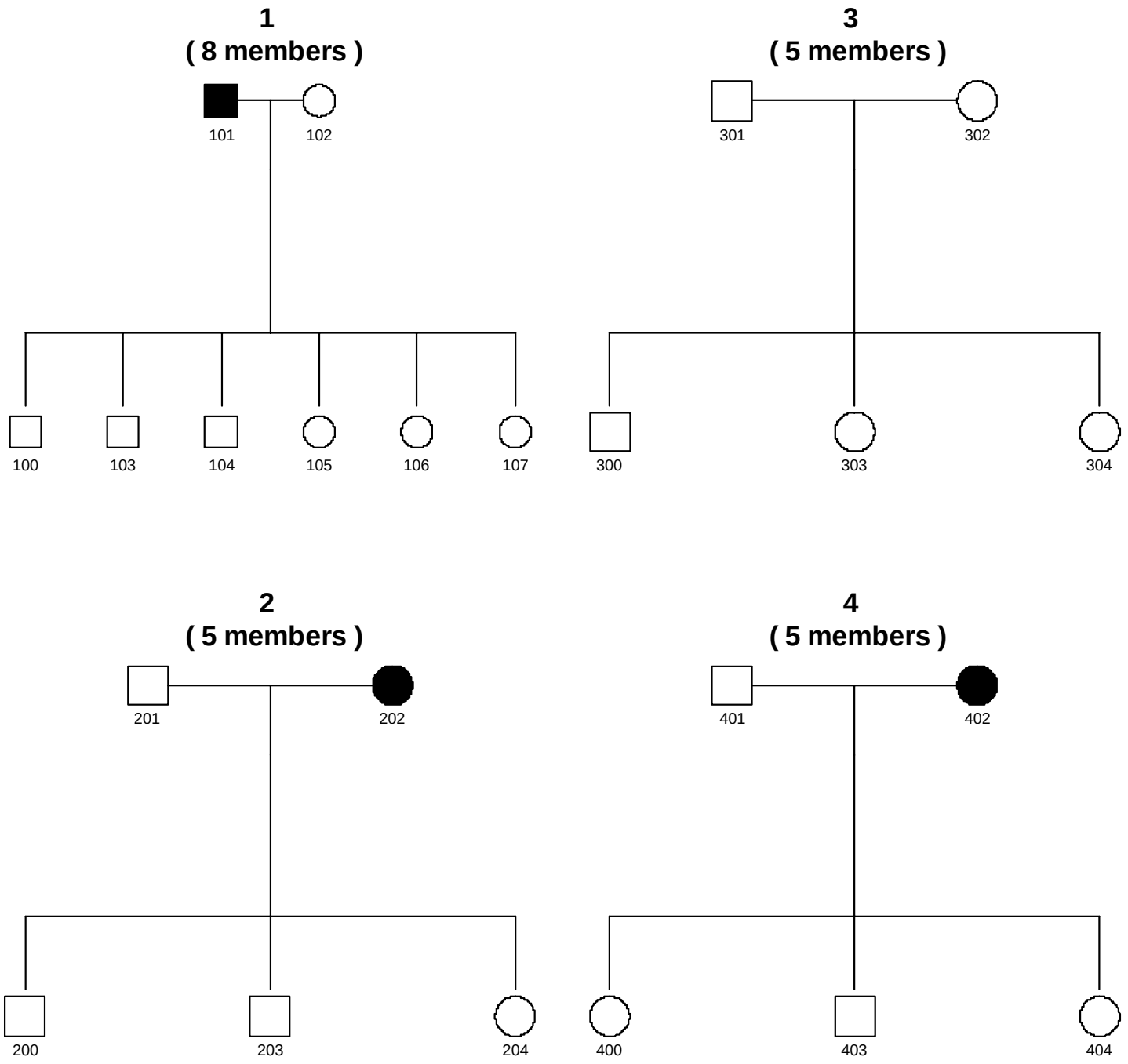
LD heatmap

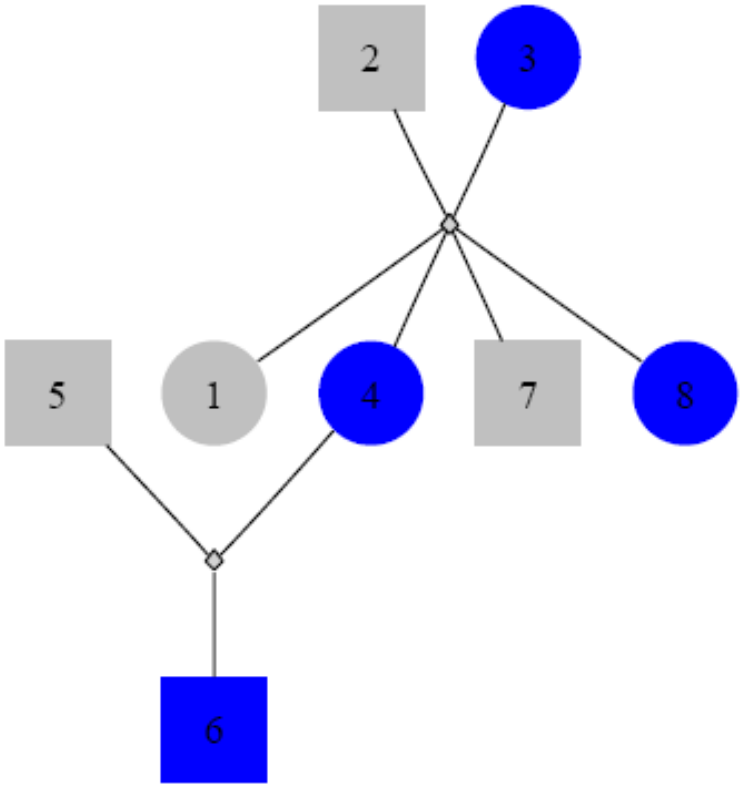




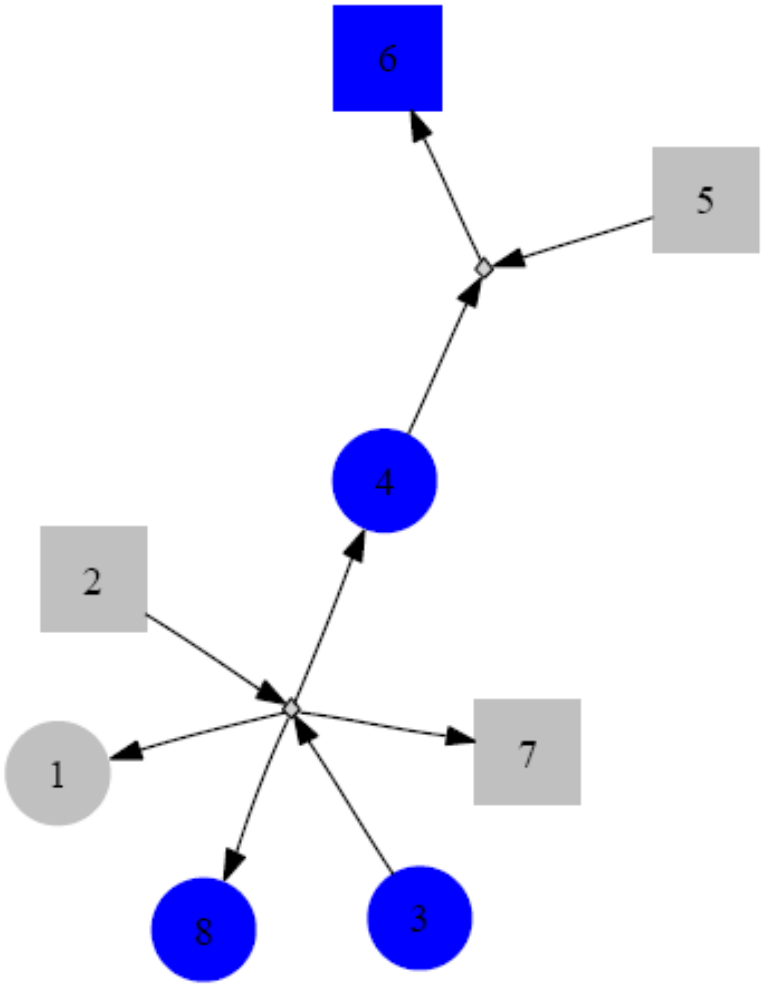
Ternary plot showing distributions of 100 markers for 100 SNPs

Graffelman & Morales-Camarena *Hum Hered* 2008

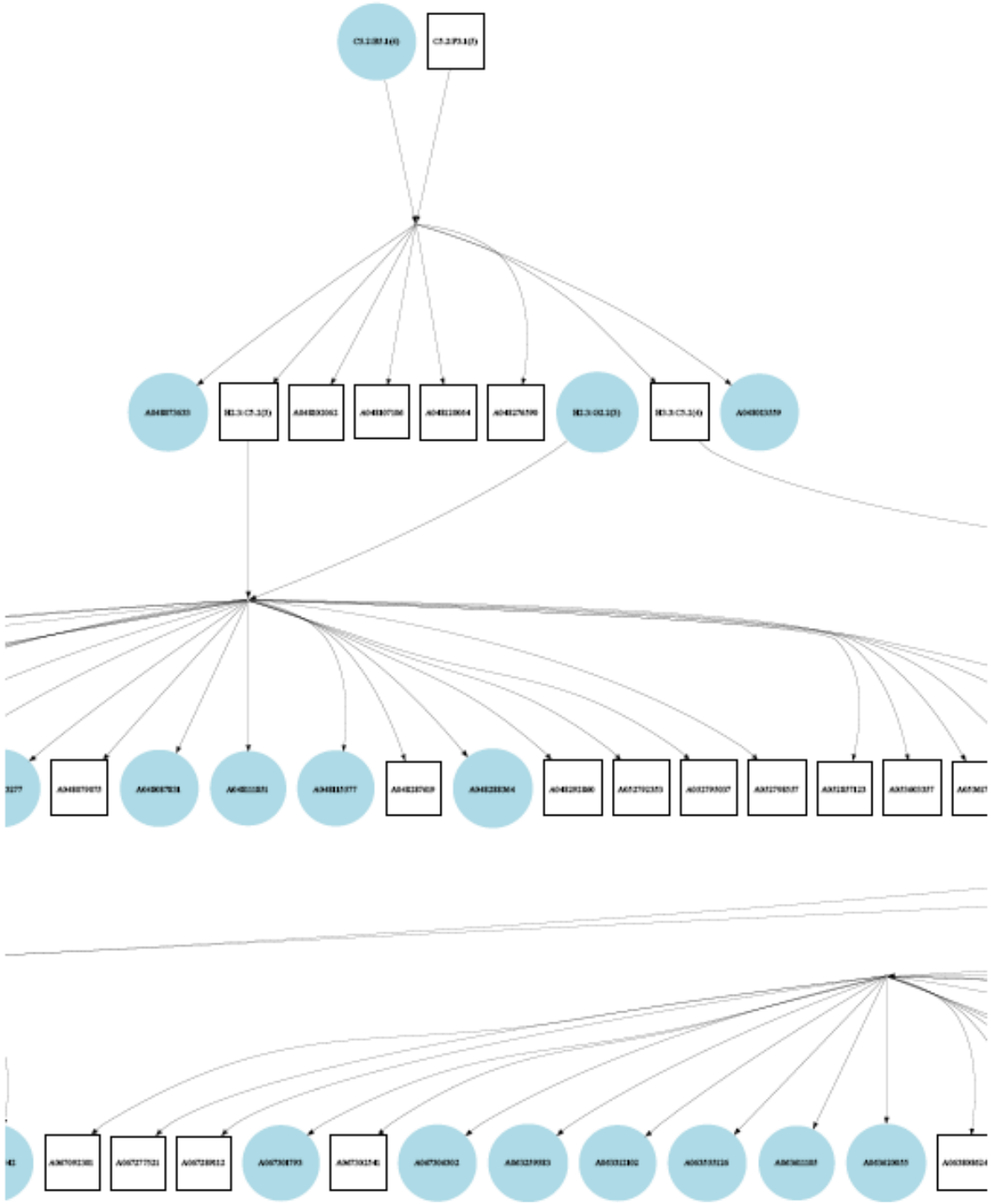

1
(8 members)
101
102
100
103
104
105
106
107
3
(5 members)
301
302
300
303
304
2
(5 members)
201
202
200
203
204
4
(5 members)
401
402
400
403
404



pedigree 10084

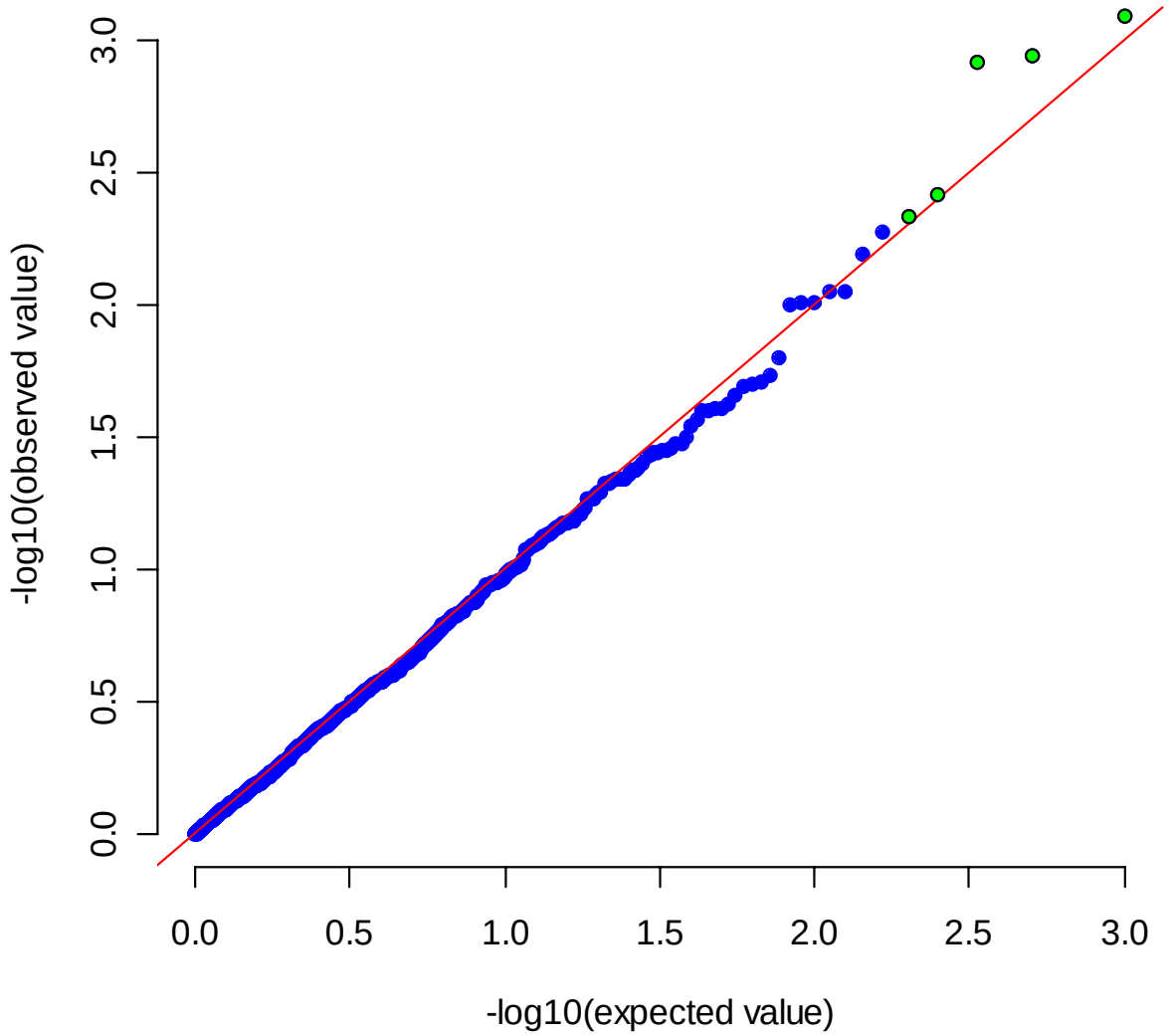


pedigree 10084

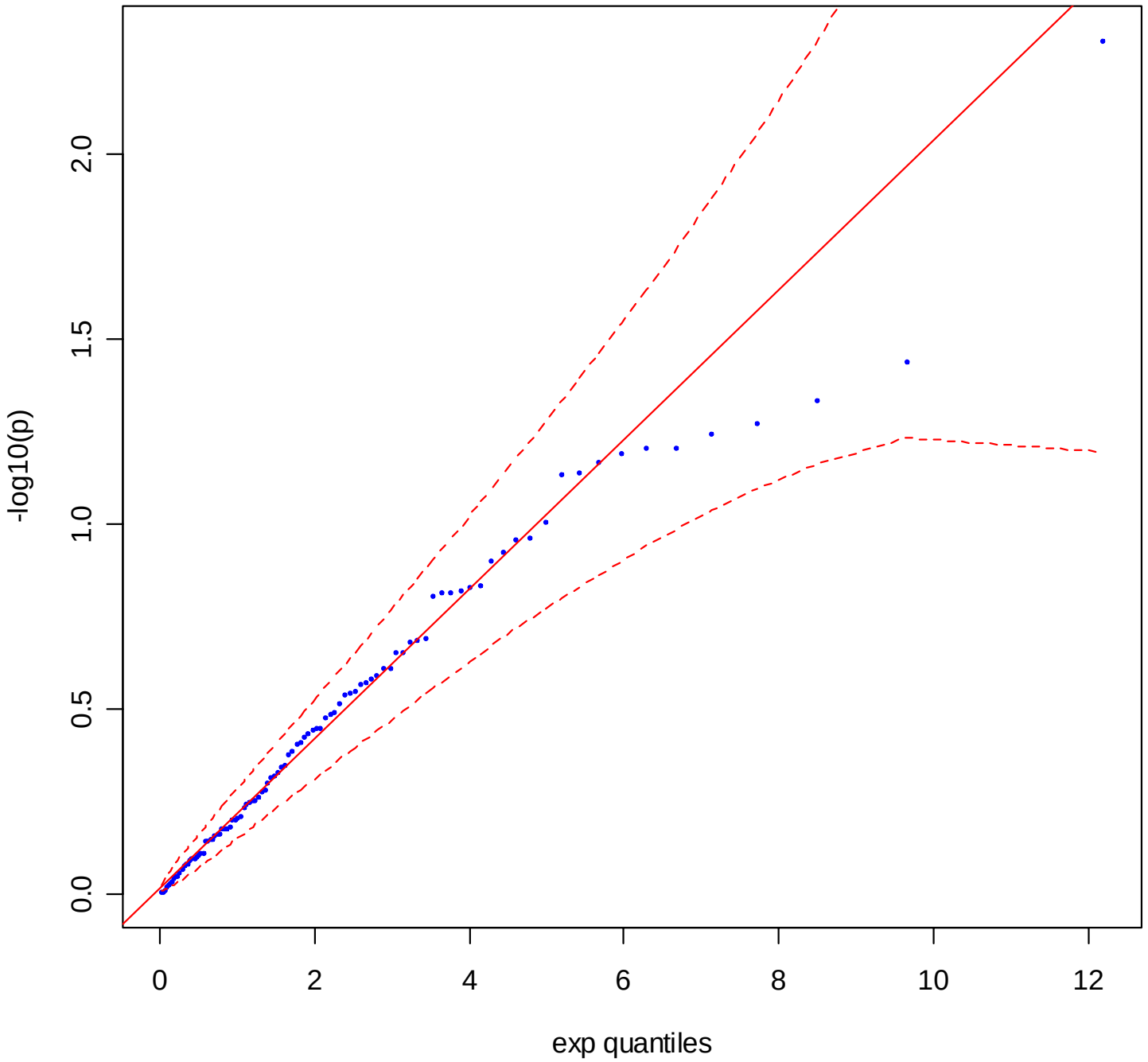


Part of the mouse pedigree from Richard Mott

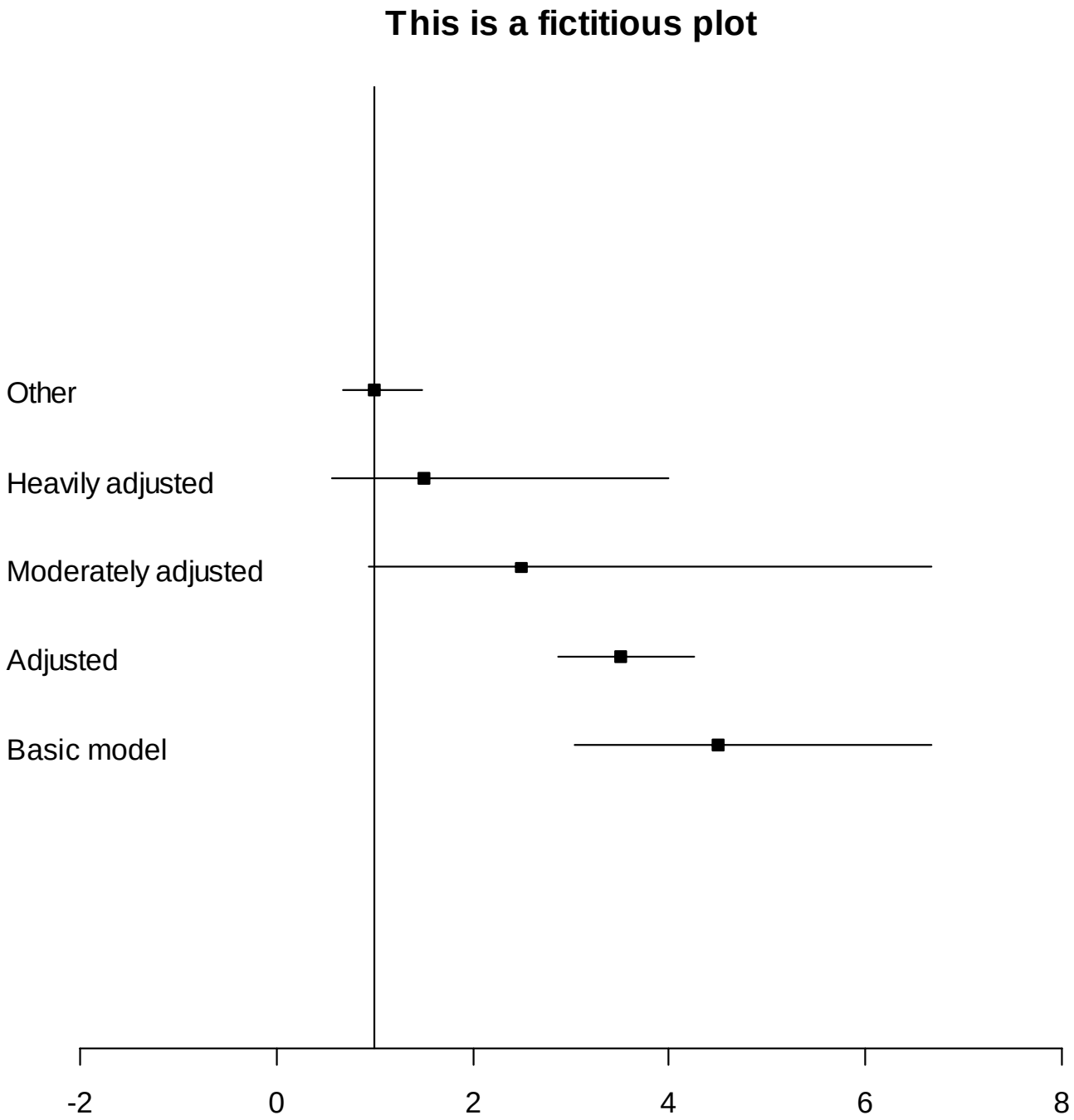
Similar functionality exists in Rgraphviz package but ideally it can also accept .dot file directly



This is unlike *qq.plot*, *qqmath*, the former uses robust statistics, but with information such as population substructure

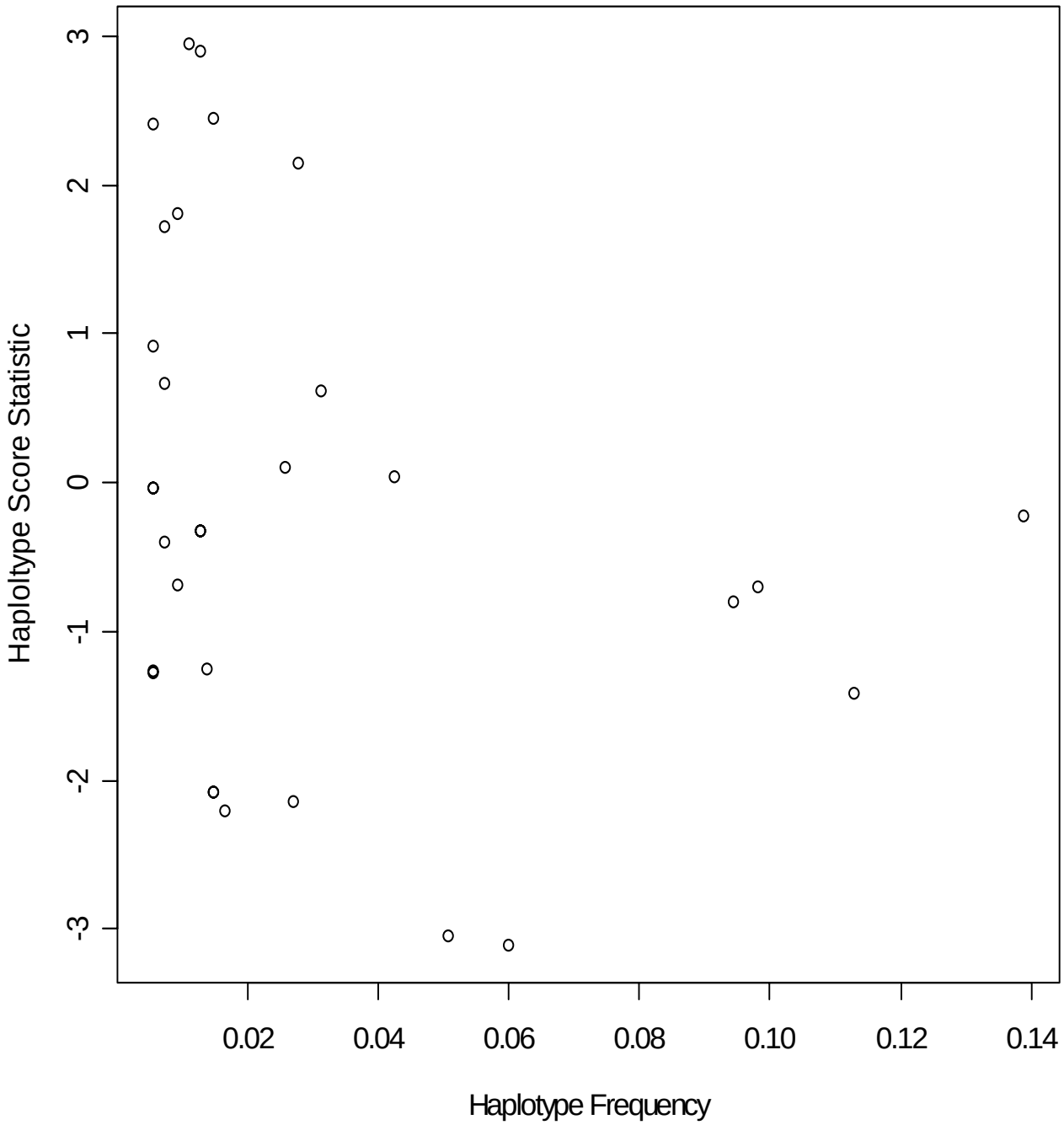


A 95%CI is added, based generally on the order statistics

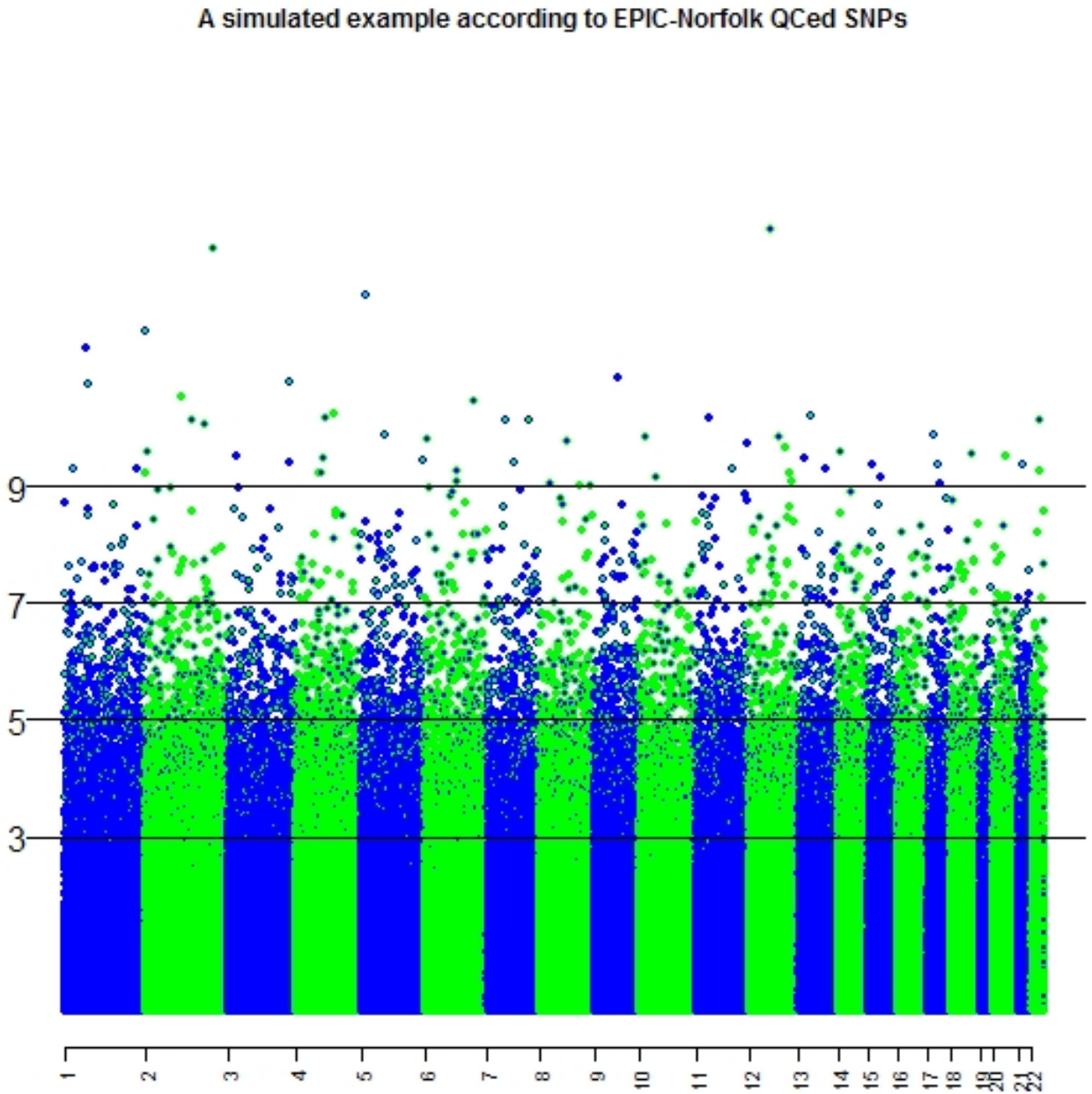


A way of effect-size visualisation

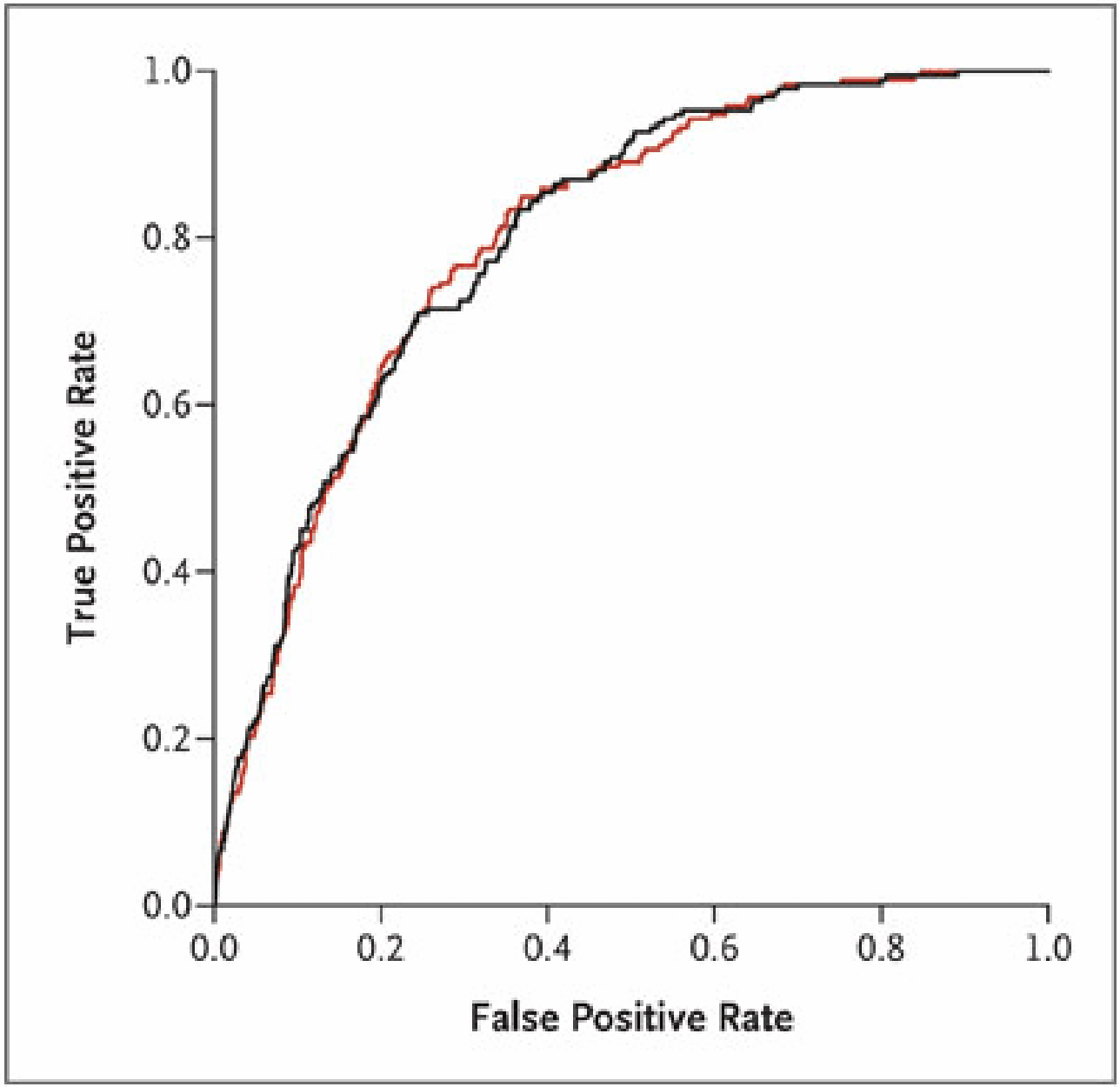
Not unlike forest plot in meta-analysis



The graph is used to identify particular haplotype with strong effect on phenotype

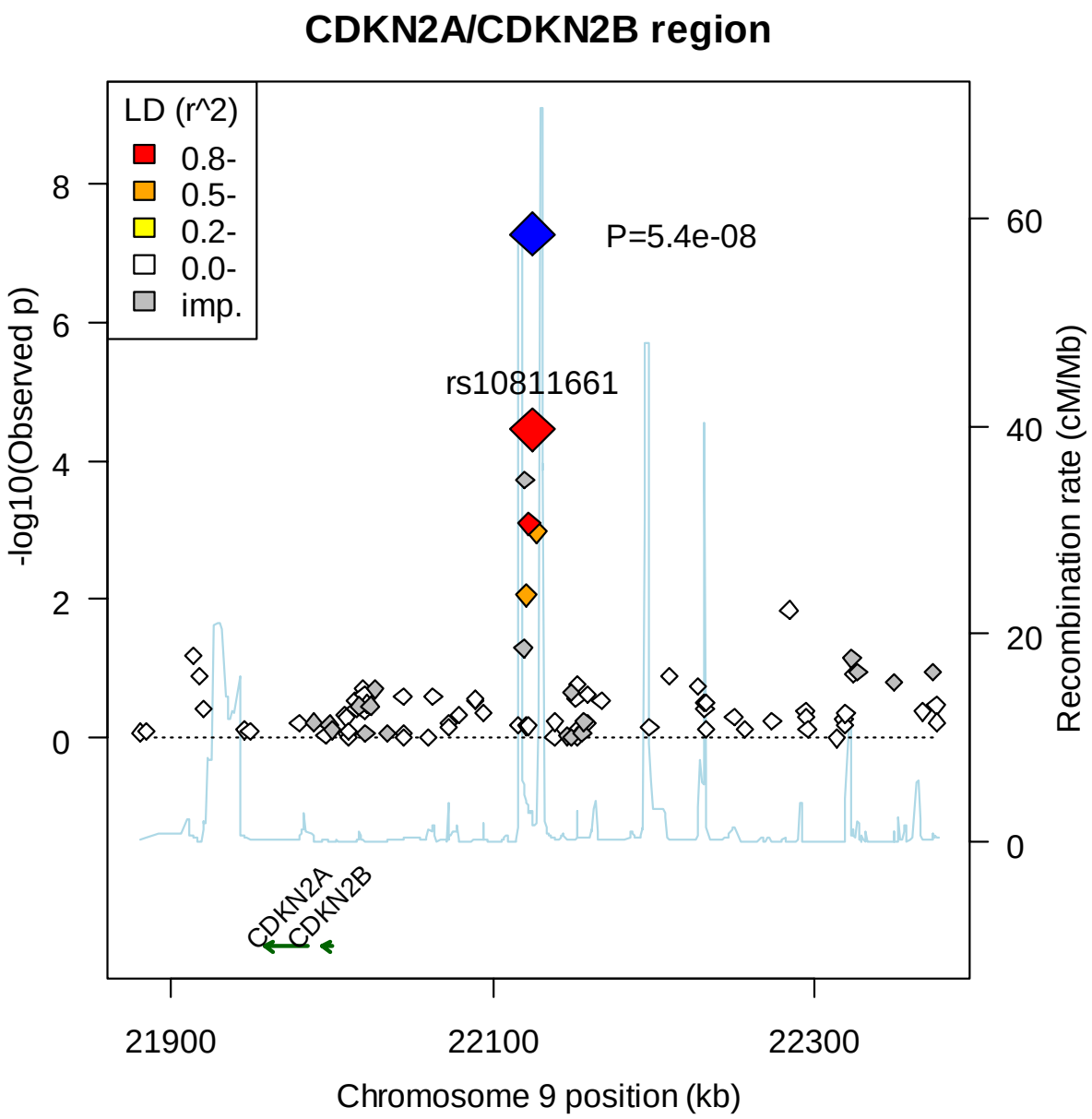


A random colour scheme can be used, highlight or identify points of interests



ROC curves for MI, stroke and death with (black)/without (red) genotype.

Kathiresan et al. *NEJM 2008*



It requires the recombination map, chromosomal position, both available from HapMap, and correlation (r^2) between (observed and imputed) SNPs associated with the top-hit SNP

R packages used

- HardyWeinberg
- LDheatmap
- kinship
  - *plot.pedigree*
- gap
  - *pedtodot*
  - *qqunif, qqfun, plot.hap.score*
  - *esplot, asplot*
- ROCR

Summary

- The use of summary statistics and graphics is classic technique for descriptive analysis.
- Graphical representation is one of the major driving forces for using R.
- There is still a gap between specialized program and a need for more rigorous work in R, e.g., HaploView and a number of R packages (genetics, snpMatrix, LDheatmap). It would be great to have some dynamic flavour, e.g.,
  - To implement in rggobi?, optional from spRay?
  - To modify code under GPL for R (e.g., HaploView)?
- This hopes to be a call for more inputs from the R community, perhaps as motivated from familiarity with both practices.