EFFICIENT AND ACCURATE MULTIPLE-PHENOTYPES REGRESSION METHOD FOR HIGH DIMENSIONAL DATA CONSIDERING POPULATION STRUCTURE

Jong Wha Joo, Eun Yong Kang, Elin Org, Nick Furlotte, Brian Parks, Aldons J. Lusis, Eleazar Eskin

UCLA

Multiple-phenotypes analysis

- Typical GWAS examine the correlation of each phenotype and genotype pair one at a time, single-phenotype analysis.

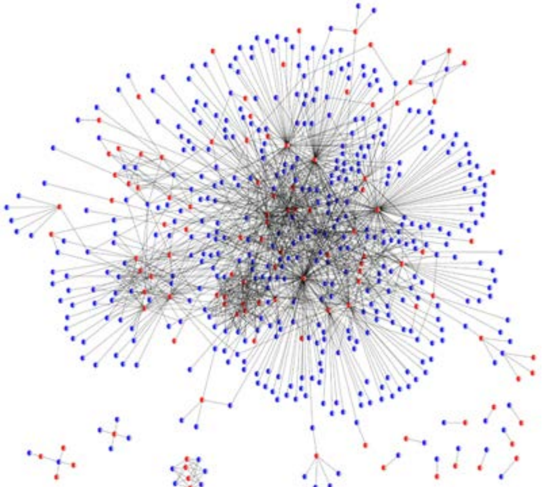
- Often it is very useful to analyze many phenotypes together. Especially, with the advent of high-throughput technology, high-dimensional multiple-phenotypes analysis is preferable.

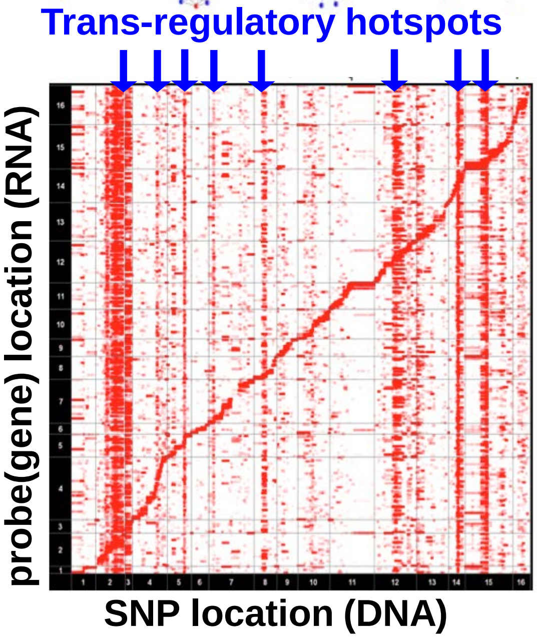
Multiple-phenotypes analysis

- □ Information can be borrowed across genes to improve variance estimates and thereby increase statistical power.

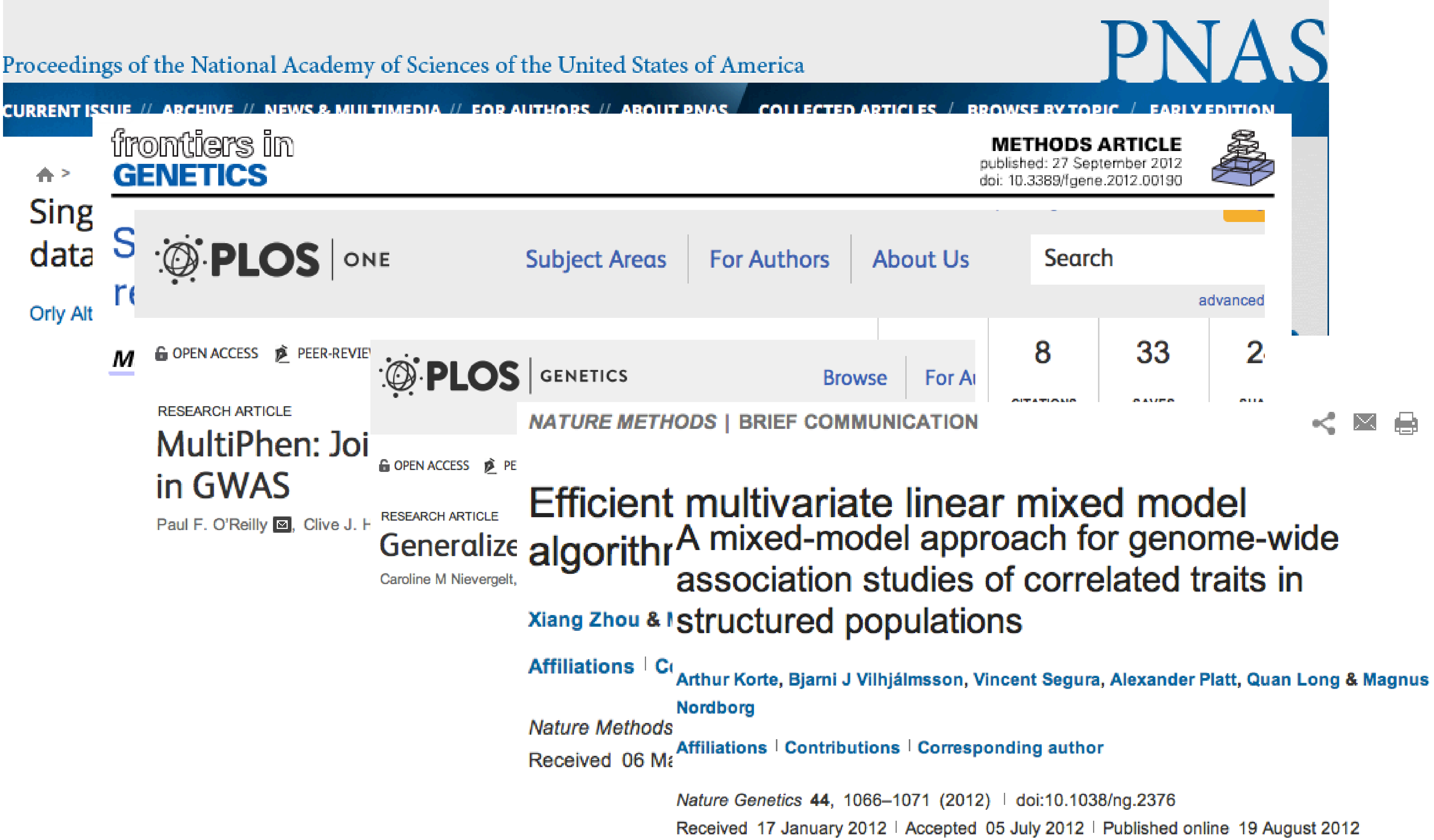
- □ Address overall state of a cell or tissue. Detect variants related to a profile of microbiota with tens of thousands species.

- □ Detecting regulatory hotspots in eQTL studies.





Previous methods



Proceedings of the National Academy of Sciences of the United States of America

PNAS

CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // FOR AUTHORS // ABOUT PNAS COLLECTED ARTICLES / BROWSE BY TOPIC / EARLY EDITION

Sing
data

Orly Alt

frontiers in GENETICS

METHODS ARTICLE
published: 27 September 2012
doi: 10.3389/fgene.2012.00190

S
re

PLOS ONE

Subject Areas | For Authors | About Us | Search

advanced

8 | 33 | 2

OPEN ACCESS PEER-REVIE

RESEARCH ARTICLE

MultiPhen: Joi
in GWAS

Paul F. O'Reilly, Clive J. H

PLOS GENETICS

Browse | For A

OPEN ACCESS PE

RESEARCH ARTICLE

Generalize

Caroline M Nievergelt,

NATURE METHODS | BRIEF COMMUNICATION

Efficient multivariate linear mixed model
algorith

Xiang Zhou & I

Affiliations | C

Nature Methods

Received 06 Ma

A mixed-model approach for genome-wide association studies of correlated traits in structured populations

Arthur Korte, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long & Magnus Nordborg

Affiliations | Contributions | Corresponding author

Nature Genetics **44**, 1066–1071 (2012) | doi:10.1038/ng.2376

Received 17 January 2012 | Accepted 05 July 2012 | Published online 19 August 2012

Previous methods

- MANOVA, multivariate regression analysis
  - Designed for use with a small number of variables. P<<N
  - MANOVA assumes MVN
  - Only can use Euclidean distances
- Data reduction methods - Cluster analysis, factor analysis, etc.
- mvLMMs (Furlotte and Eskin, Genetics 2015; Zhou *et al*., Nat Methods , 2012) , MTMM(Korte *et al*., Nat Genet , 2012) - LMM based approaches, computational costs scale quadratically with the number of phenotypes
- MDMR (Zapala *et al*., Front Genet, 2012)
  - **M**ultivariate **D**istance **M**atrix **R**egression analysis.
  - Form a statistic to test the effect of some covariates on all of the phenotypes by utilizing the similarity matrix that reflects the correlation of the samples with respect to the expression values over the genes.

$$\text{"Pseudo" F-statistics, } F = \frac{tr(\hat{Y}\hat{Y}')/(2-1)}{tr(\hat{R}\hat{R}')/(n-2)}$$

Population structure cause False Positives

- □ GWAS test the allele frequency differences between cases and controls to find SNPs correlated with a disease.
- □ Allele frequencies vary from population to population due to each population's unique genetic/social history.
- □ Not only disease-causing SNPs cause allele frequency difference but also SNPs influenced by ancestry may also cause allele frequency difference.

Human Genetics
June 2002, Volume 110, Issue 6, pp 553-560

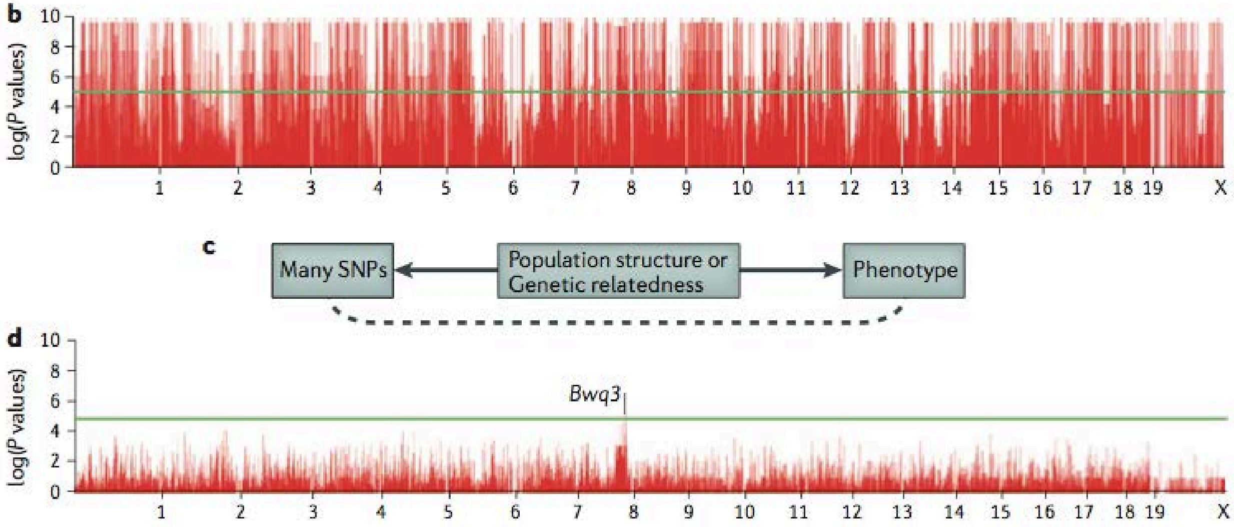
CYP3A4-V and p
Americans: caus

The Lancet, Volume 36
doi:10.1016/S0140-6

Copyright © 2003 Else

Population

Prof Lon R Cardon Ph



pean

xt Article >

ICS

Services Collections YeastBook

Factors That
re of Populations

Population structure cause False Positives

- GWAS test the allele frequency differences between cases and controls to find SNPs correlated with a disease.
- Allele frequencies vary from population to population due to each population’s unique genetic/social history.
- Not only disease-causing SNPs cause allele frequency difference but also SNPs influenced by ancestry may also cause allele frequency difference.
- This problem is even more serious when analyzing multiple-phenotypes because this bias in test statistics accumulates from each phenotype.
- Unfortunately, none of the previously mentioned multivariate methods are able to correct for the population structure and may cause a significant amount of false positive results.

Nature Reviews Genetics **13**, 807-817 (November 2012) | doi:10.1038/nrg3335

A typical single-SNP test

$$\mathbf{y} = \mu + X\beta + \mathbf{e}$$

y : phenotypes (size n)

X : A SNP to test

β: contribution from the SNP

e : (n × 1) random effect, Var(**e**) = $\sigma_e^2 I$

A ‘hypothetical’ true genetic model

$$\mathbf{y} = \mu + \sum_{i=1}^{m} X_i \beta_i + \mathbf{e}$$

y : phenotypes (size n)

X_i : i-th SNP to test

β_i: contribution from the i-th SNP

e : (n × 1) random effect, Var(**e**) = $\sigma_e^2 I$

True effect of a single SNP

$$\mathbf{y} = \mu + X_k\beta_k + \sum_{i \neq k} X_i\beta_i + \mathbf{e}$$

Actual test is simple

TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \sum_{i \neq k} X_i\beta_i + \mathbf{e}$$

SIMPLE LINEAR MODEL

$$\mathbf{y} = \hat{\mu} + X_k\hat{\beta}_k + \mathbf{e}$$

There are unmodeled genetic factors

TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \underline{\sum_{i\neq k} X_i\beta_i} + \mathbf{e}$$

UNMODELED FACTORS

SIMPLE LINEAR MODEL

$$\mathbf{y} = \hat{\mu} + X_k\hat{\beta}_k + \mathbf{e}$$

Unmodeled factors are not known

TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \underline{\boxed{\sum_{i\neq k} ?\, X_i\beta_i}} + \mathbf{e}$$

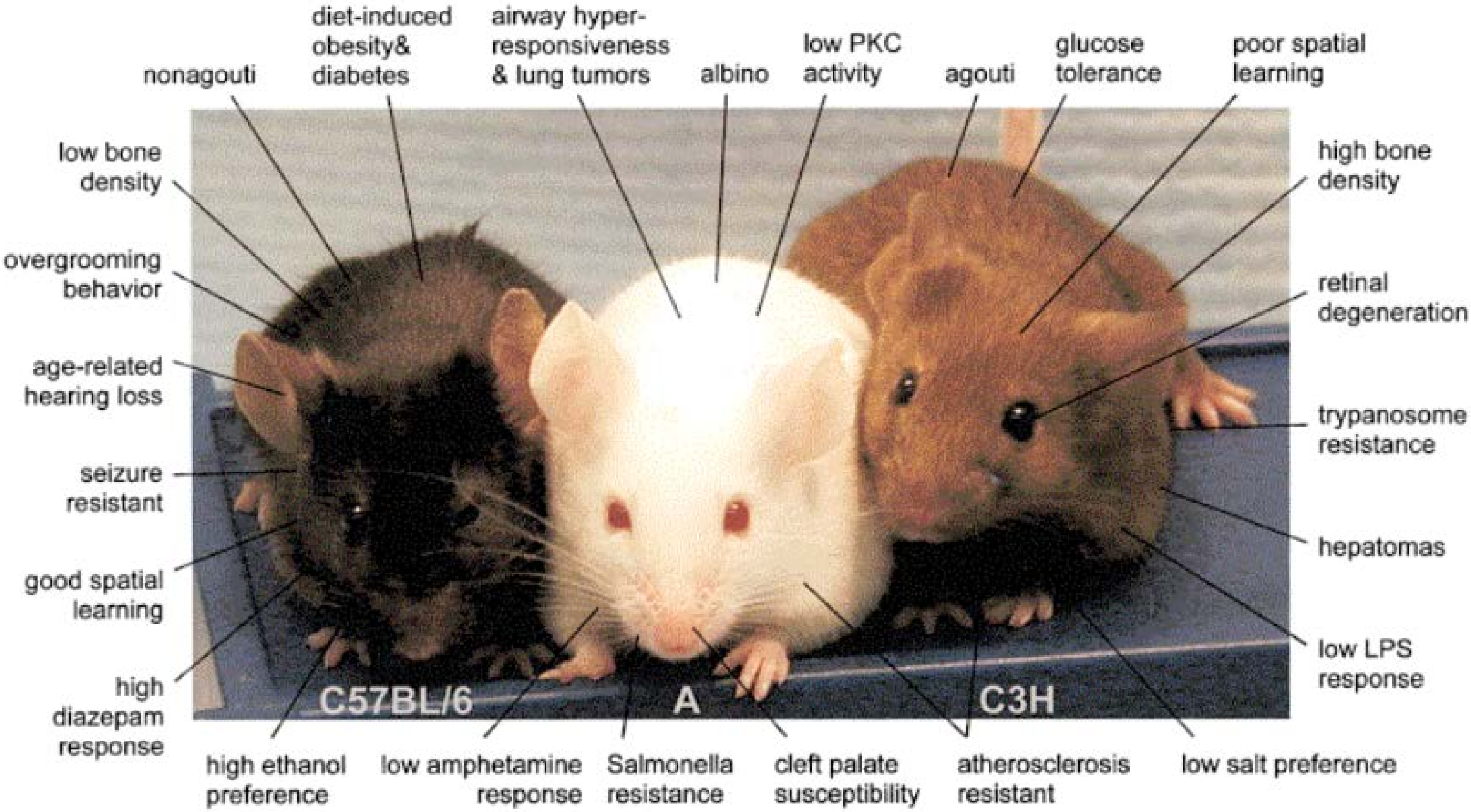
UNMODELED FACTORS

SIMPLE LINEAR MODEL

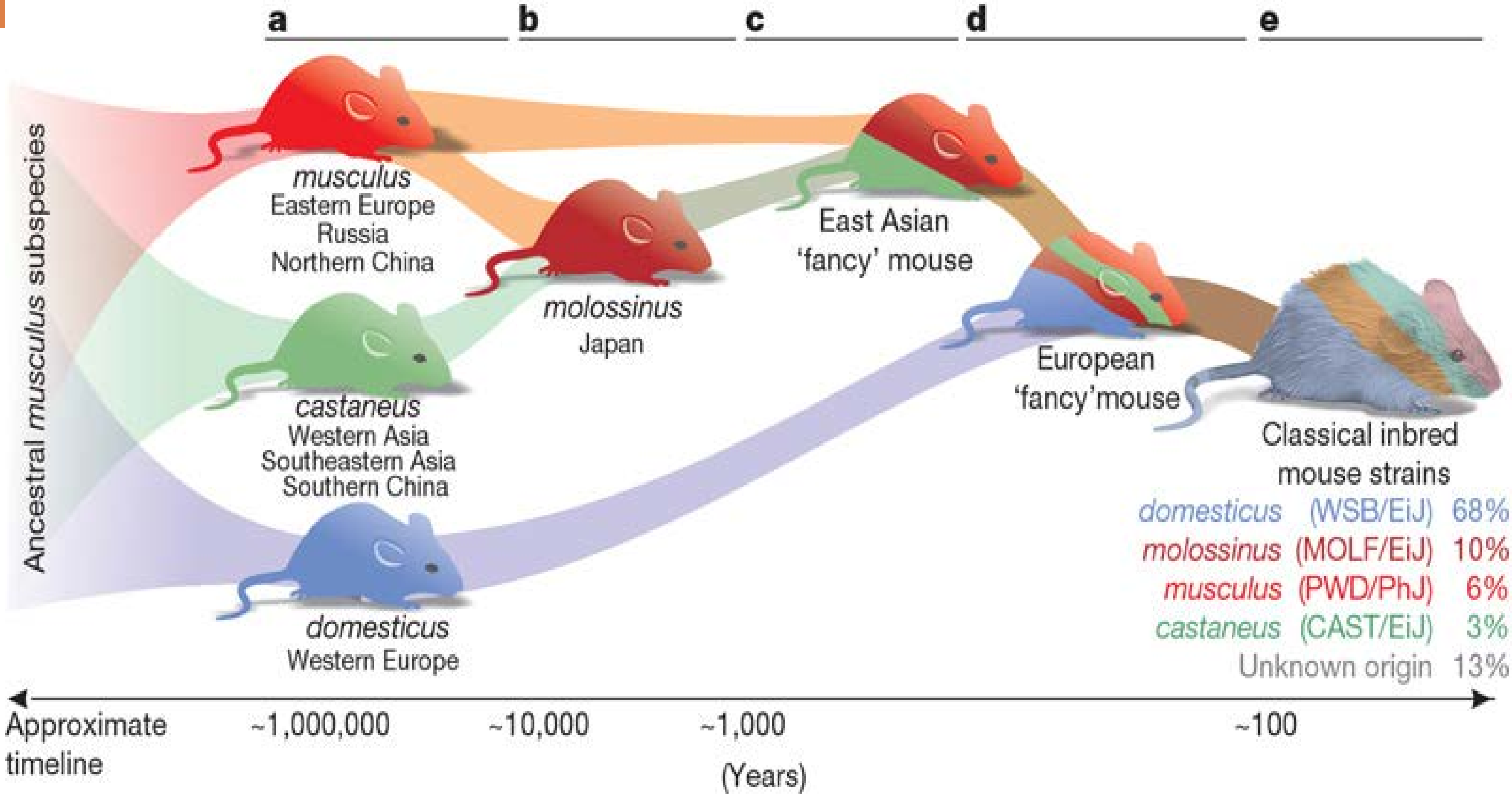
$$\mathbf{y} = \hat{\mu} + X_k\hat{\beta}_k + \mathbf{e}$$

Entering mouse genetics..

Inbred strains differ in common traits relevant to common diseases

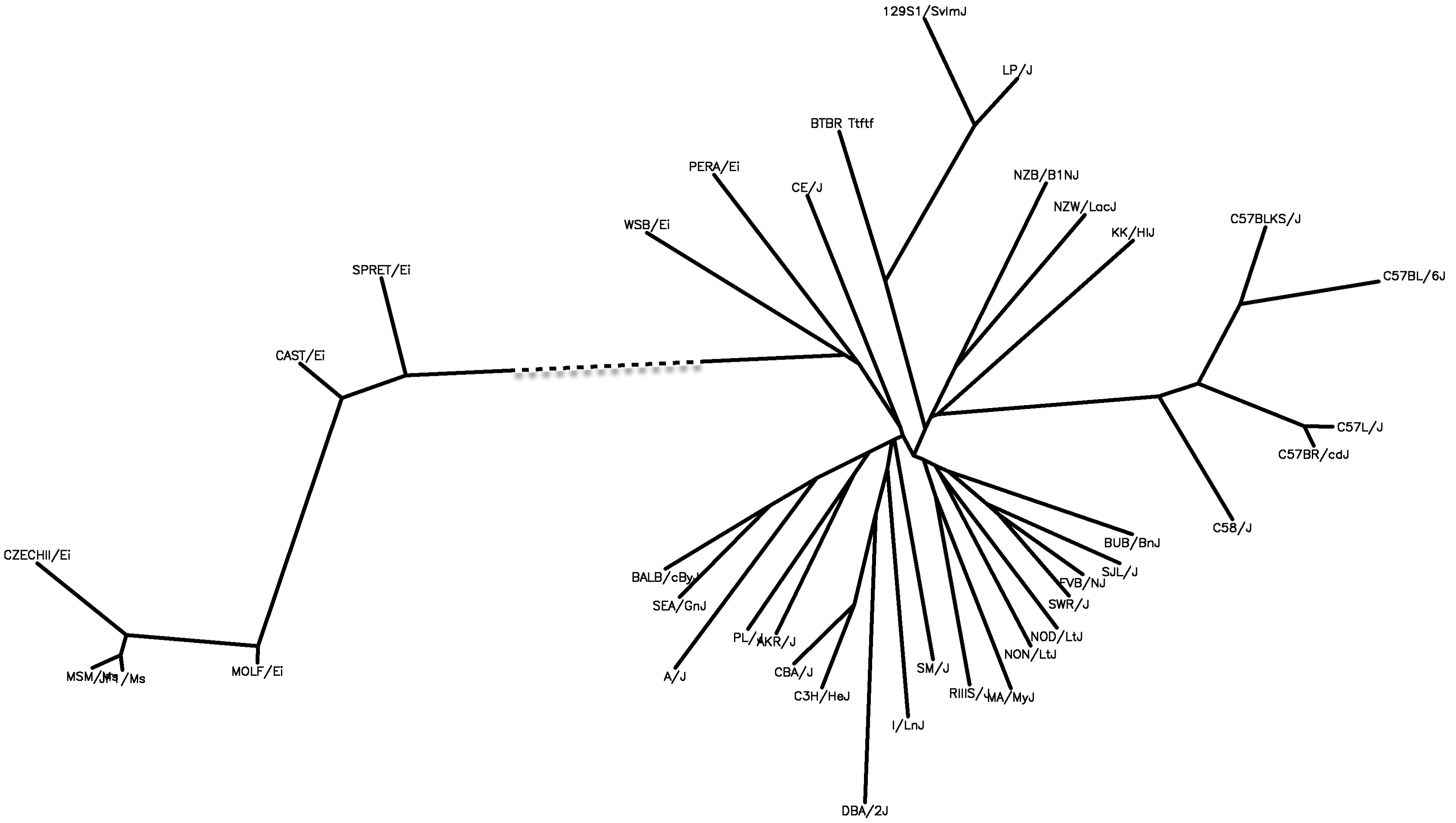


Classical inbred strains



Frazer KA, Eskin E, Kang HM et al. Nature. Aug 2007, 448(

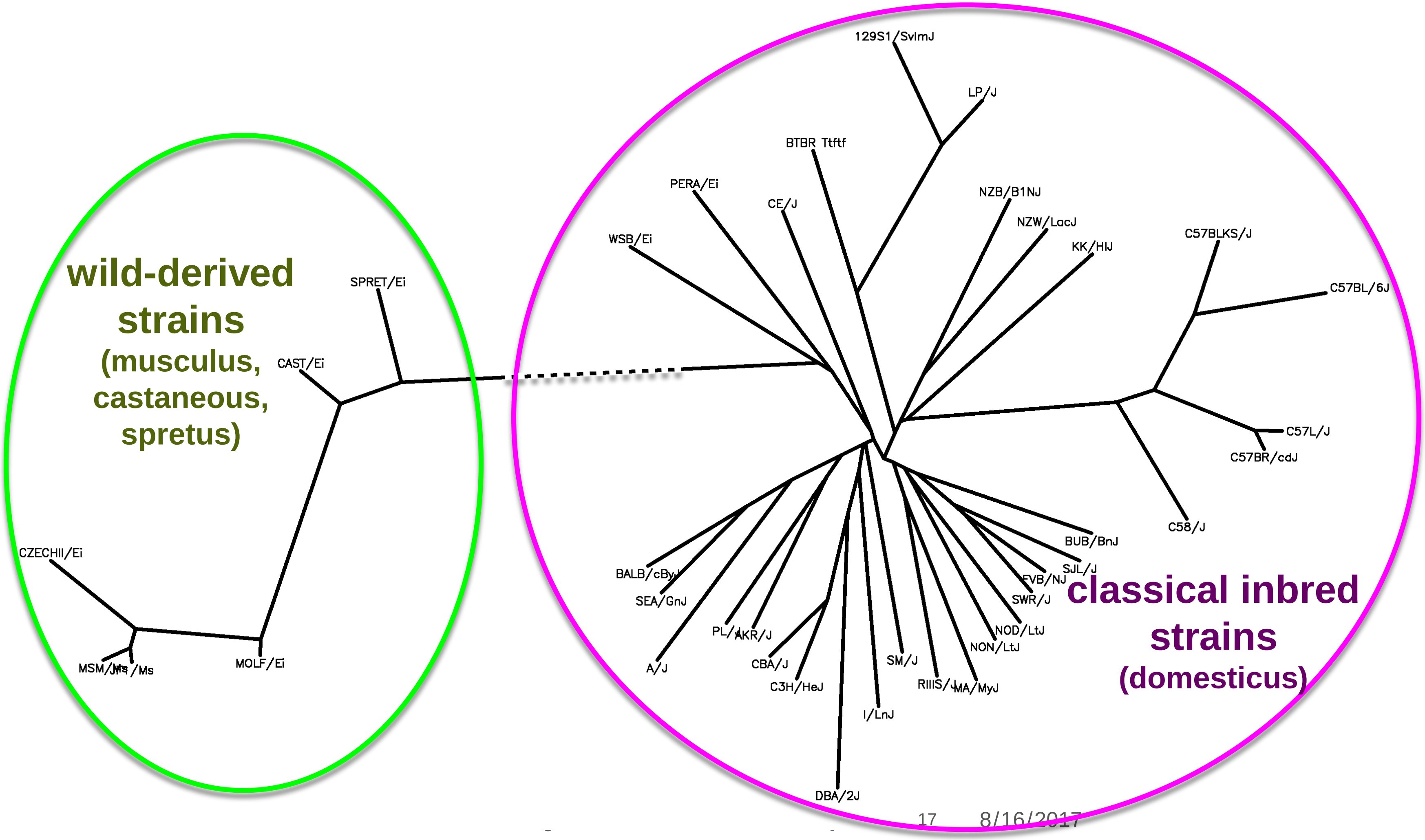
Complex genetic relatedness of lab strains



8/16/2017

Phylogeny of 38 inbred mouse strains using 140,000 mouse HapMap

Complex genetic relatedness of lab strains

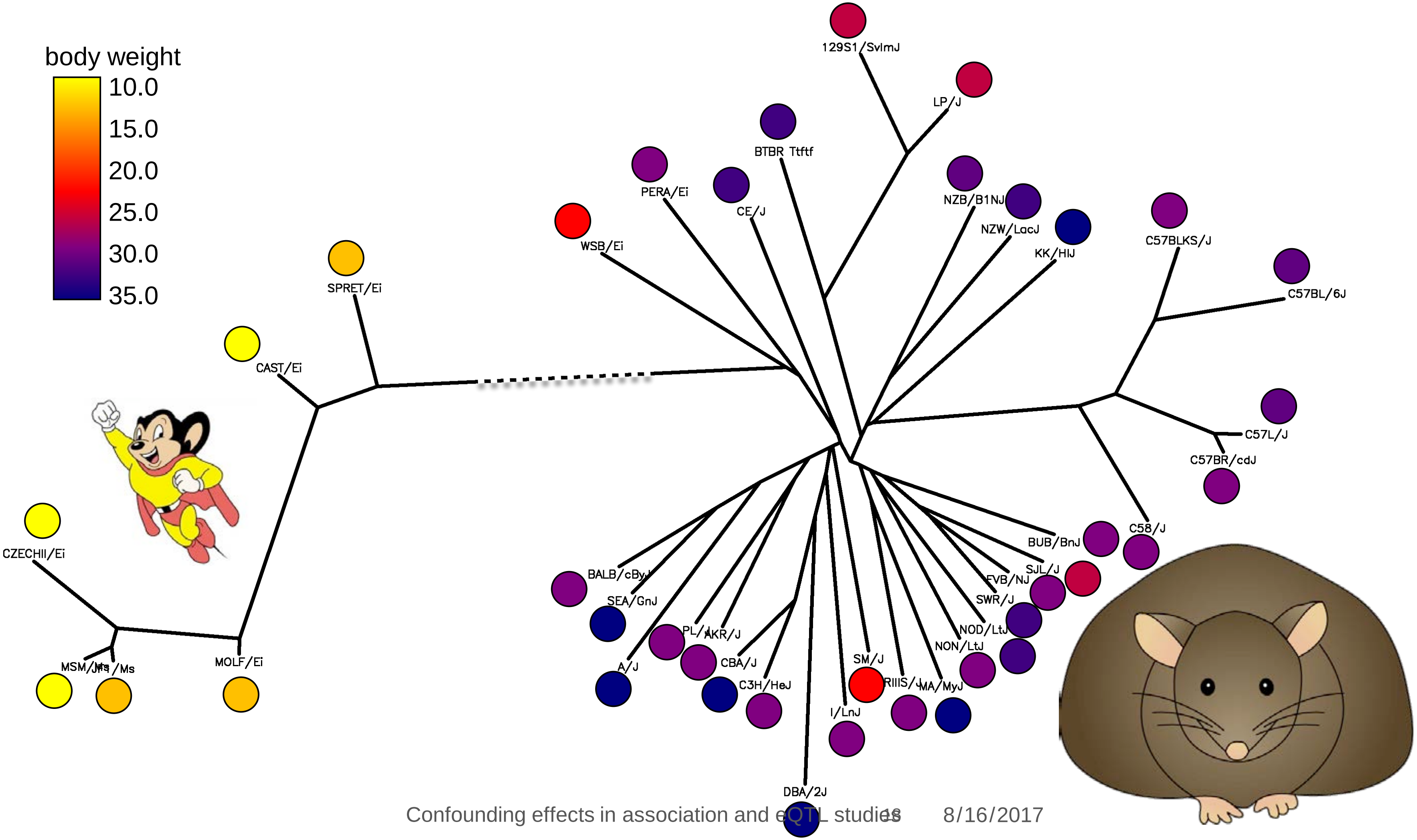


wild-derived strains
(musculus, castaneous, spretus)

classical inbred strains
(domesticus)

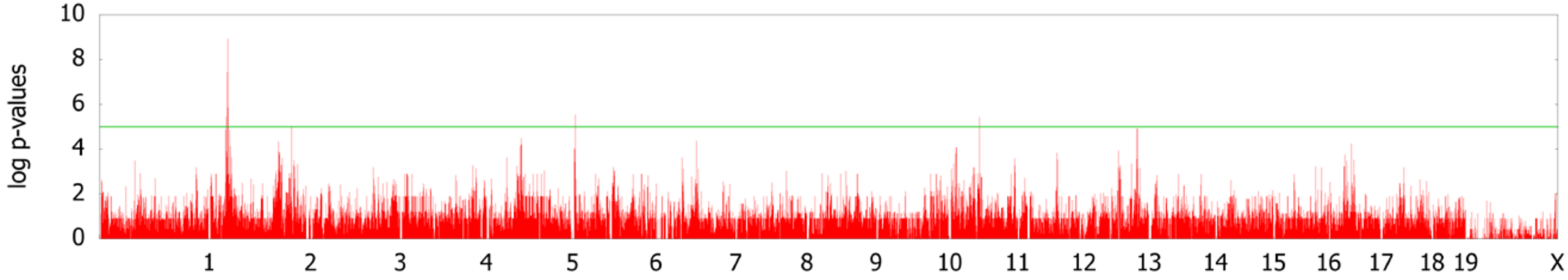
Phylogeny of 38 inbred mouse strains using 140,000 mouse HapMap S

Complex genetic relatedness of lab strains

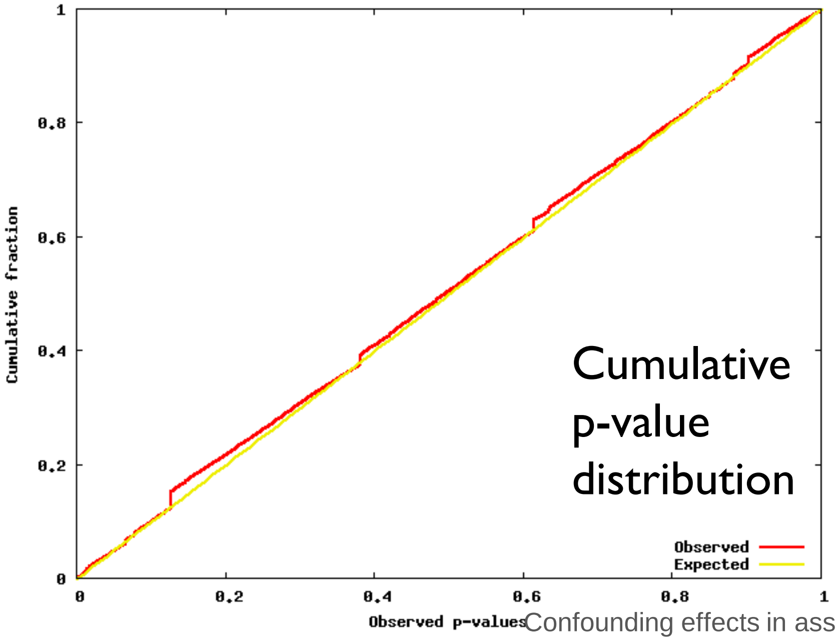


Body weight phenotypes of 38 inbred mouse strains from JAX MPD

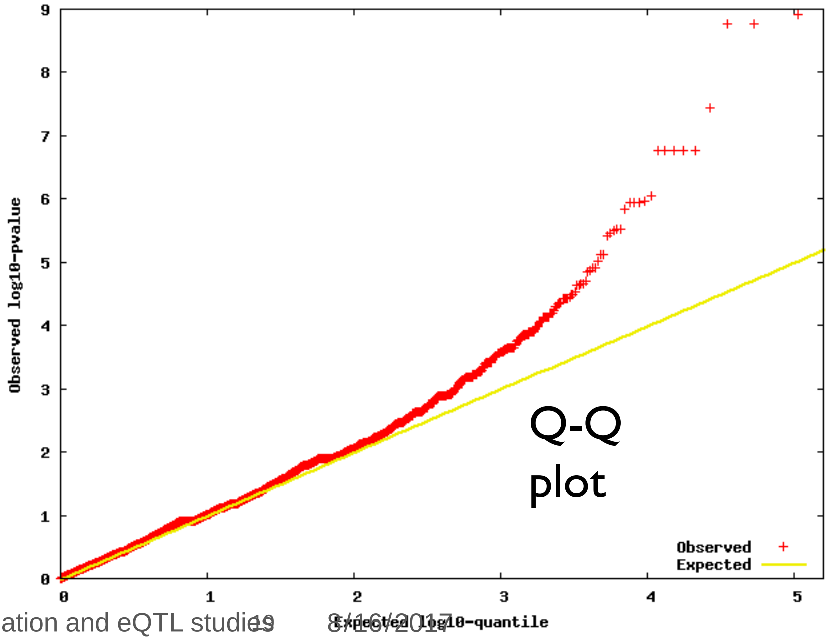
What we would expect



Genome-wide association map

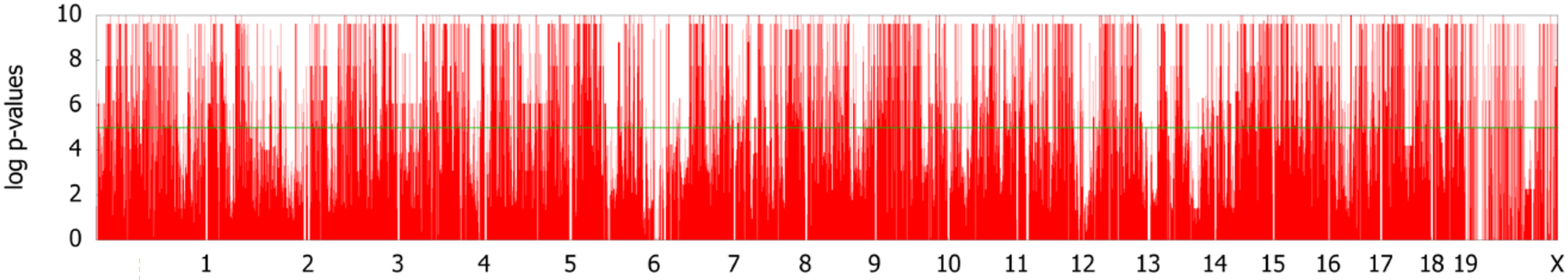


Cumulative p-value distribution

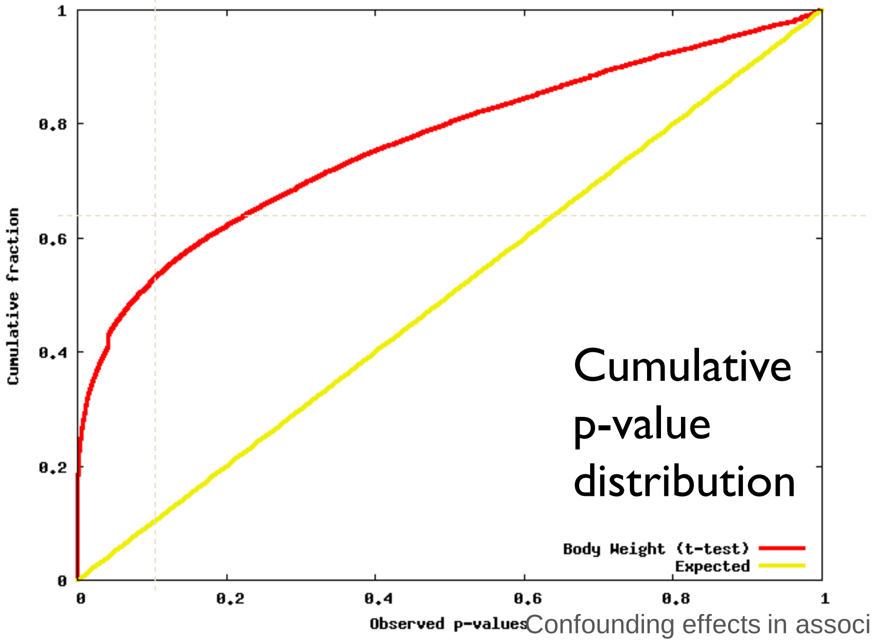


Q-Q plot

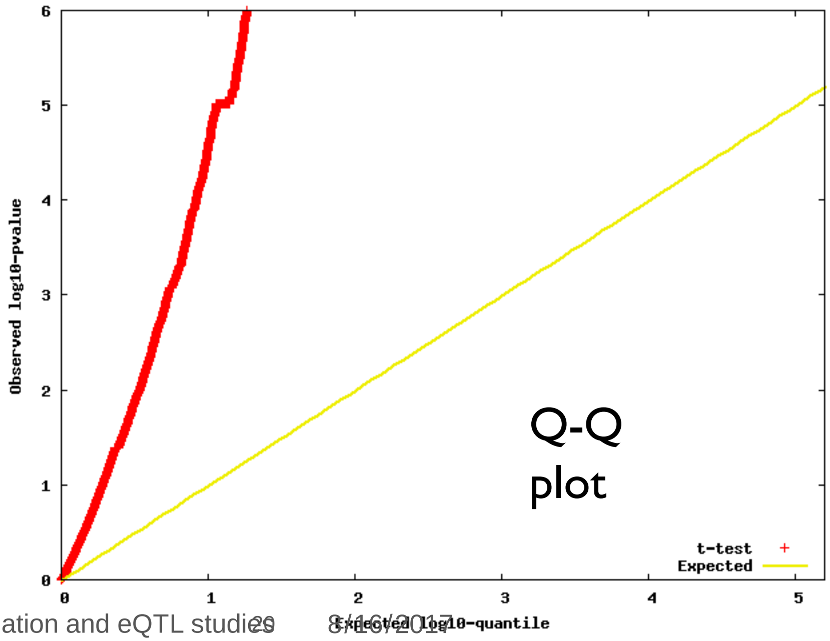
What we actually observed



Genome-wide association map

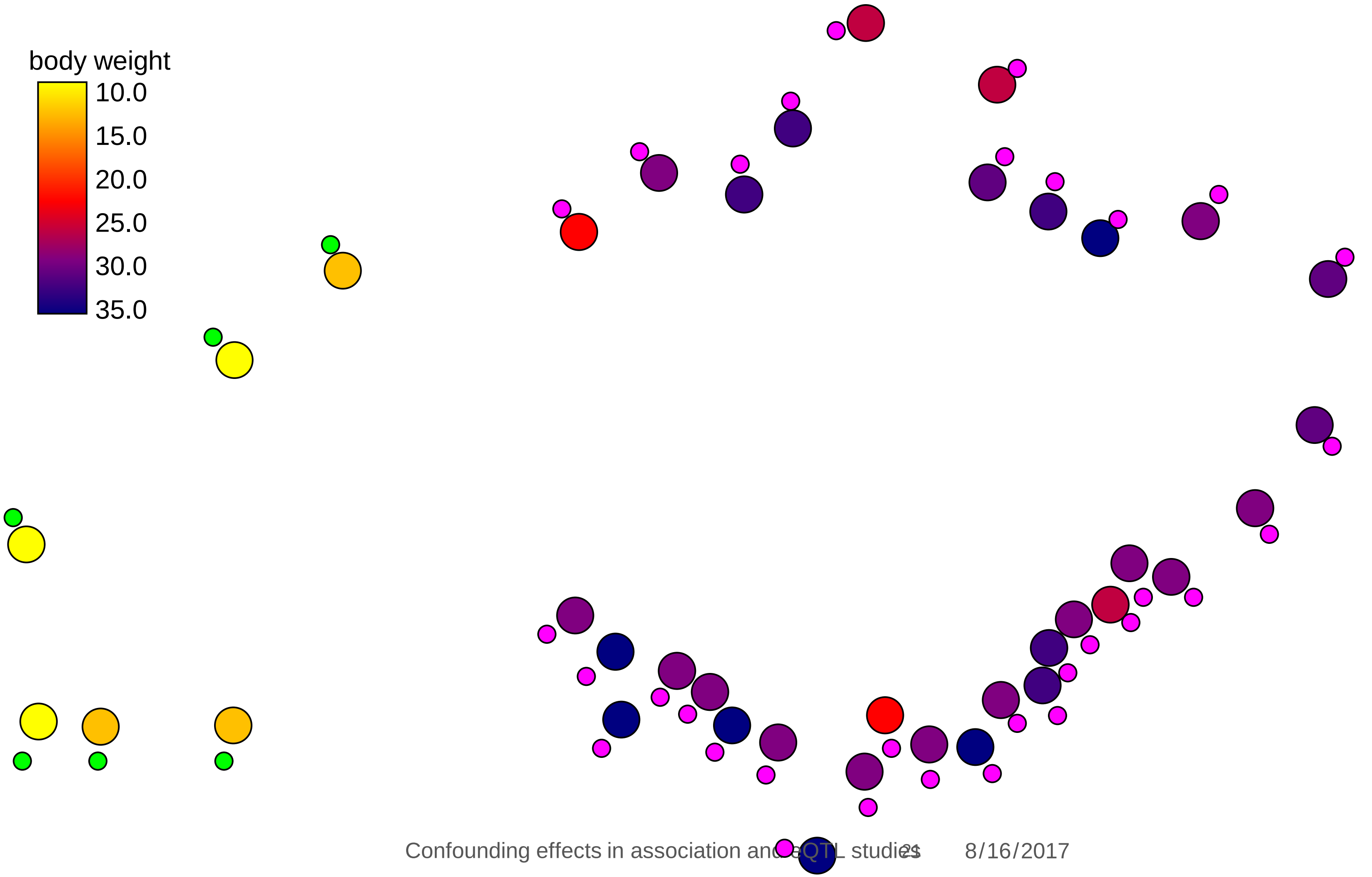


Cumulative p-value distribution



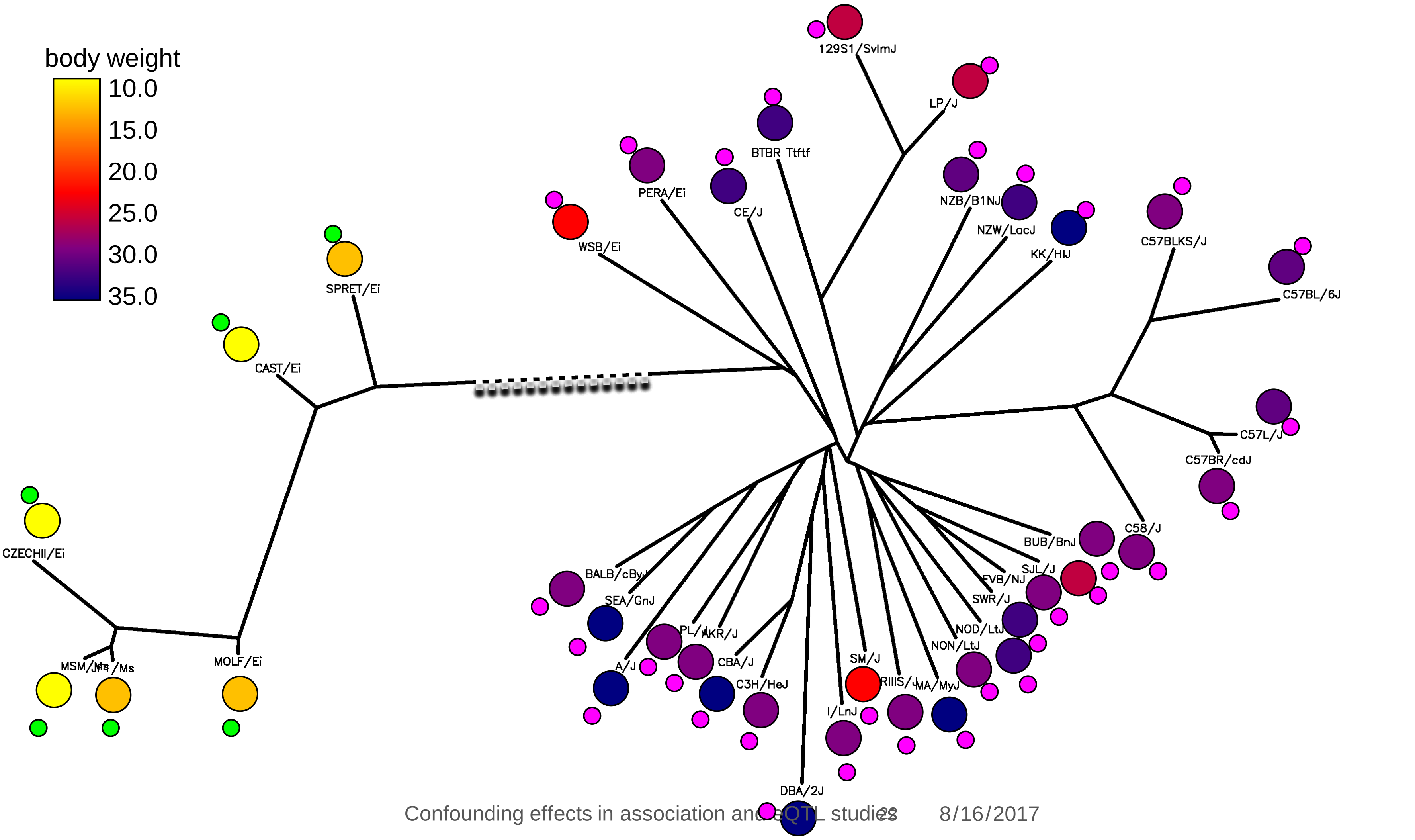
Q-Q plot

Example of spurious associations



Body weight phenotypes of 38 inbred mouse strains from JAX MPD

Example of spurious associations



Body weight phenotypes of 38 inbred mouse strains from JAX MPD

Source of spurious association

H_0: [Phenotype]$\perp$[SNP]

H_1: [Phenotype]~[SNP]

SNP

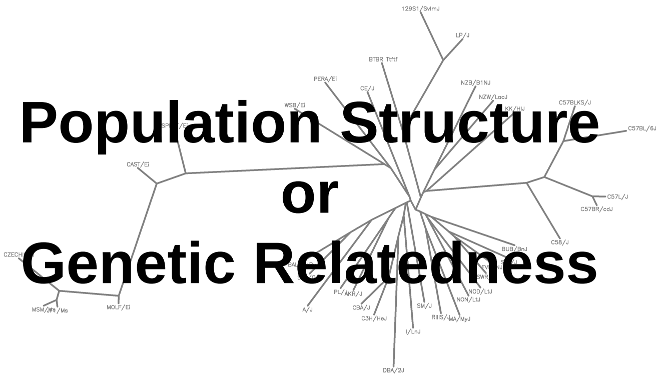
Phenotype

Source of spurious association

H_0: [Phenotype]⊥[SNP]

H_1: [Phenotype]~[SNP]

SNP



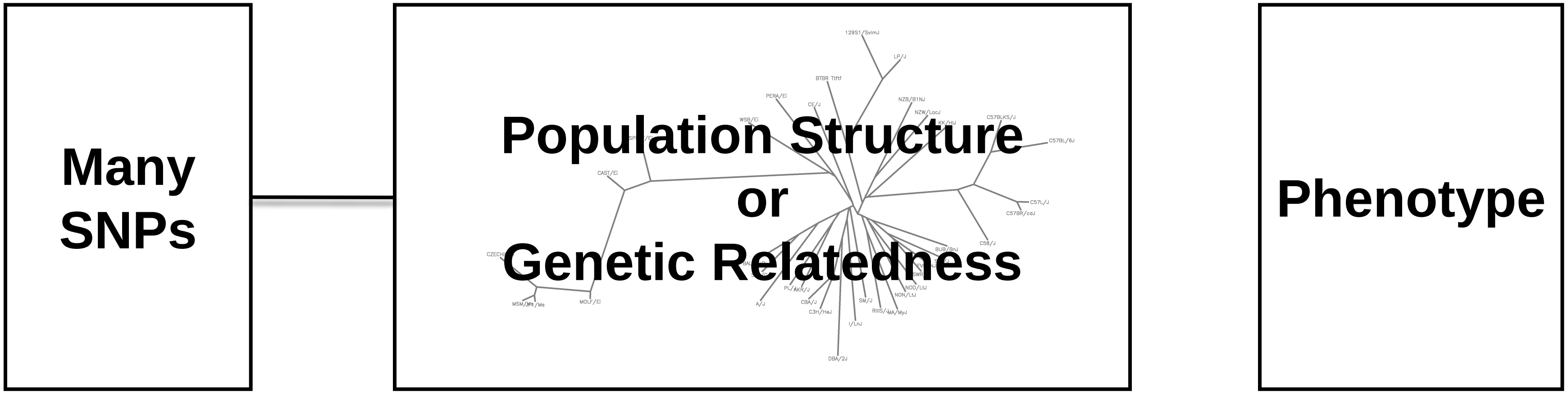
Population Structure
or
Genetic Relatedness

Phenotype

Many SNPs are
strongly correlated to the population structure

H_0: [Phenotype]⊥[SNP]

H_1: [Phenotype]~[SNP]



Many SNPs

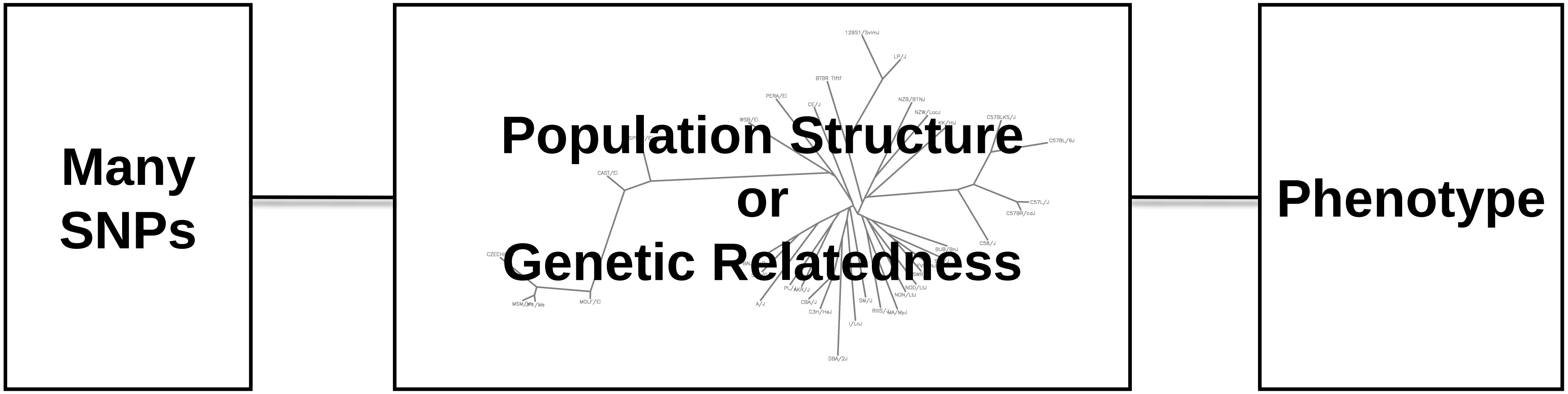
Population Structure
or
Genetic Relatedness

Phenotype

Some phenotypes are strongly correlated to population structure

H_0: [Phenotype]⊥[SNP]

H_1: [Phenotype]~[SNP]



Many SNPs

Population Structure or Genetic Relatedness

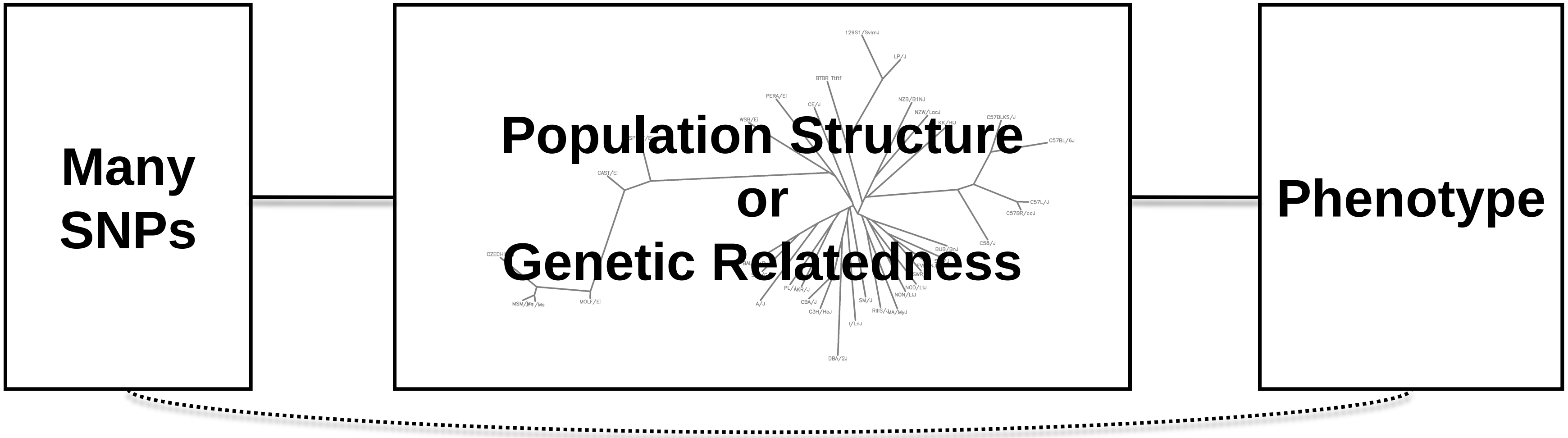
Phenotype

[redacted]e[redacted]pes become indirectly correlated

[redacted]⊥[redacted]

H_1: [Phenotype]~[SNP]

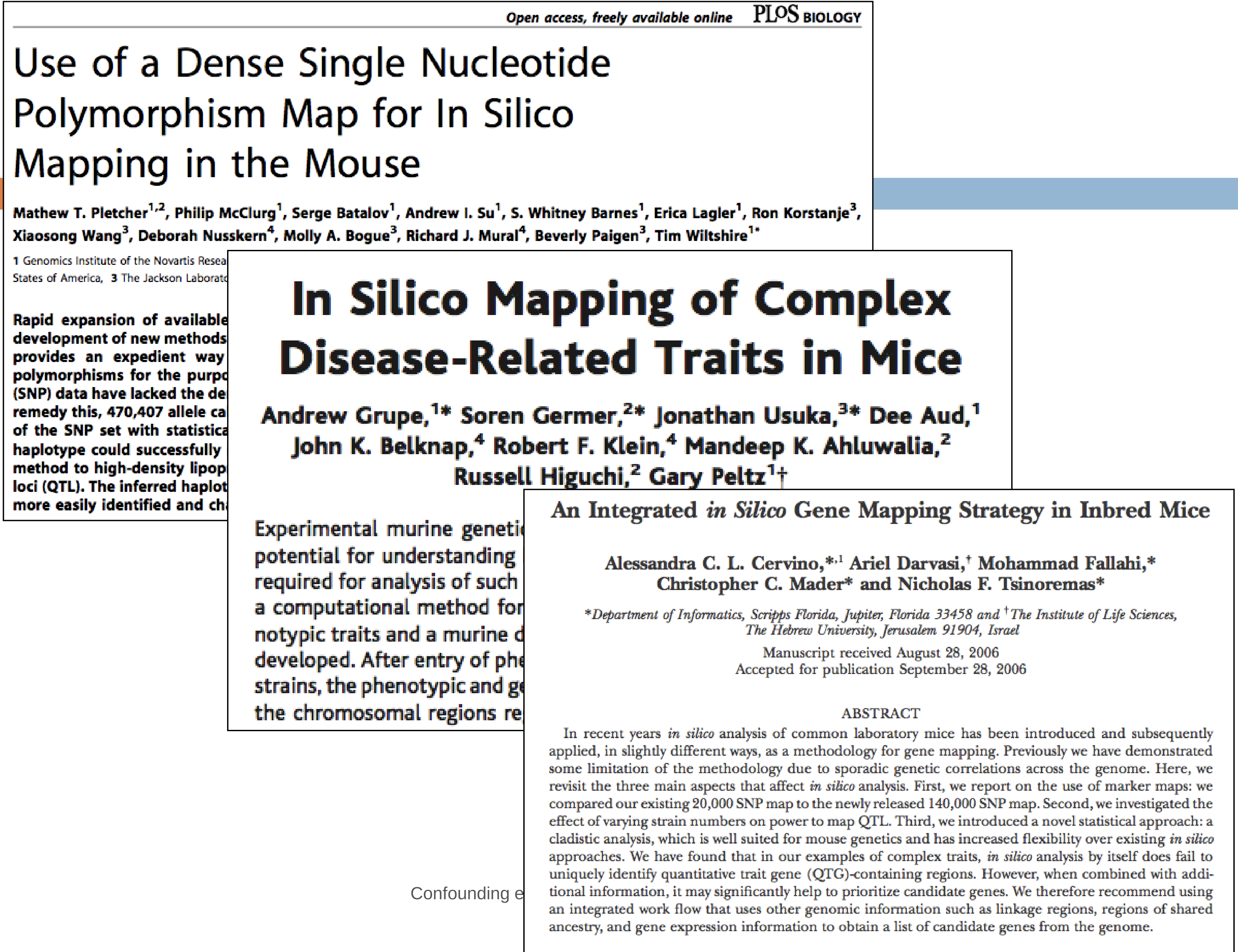
H_0: [Phenotype]~[SNP]



Many SNPs

Population Structure or Genetic Relatedness

Phenotype



Open access, freely available online PLOS BIOLOGY

Use of a Dense Single Nucleotide Polymorphism Map for In Silico Mapping in the Mouse

Mathew T. Pletcher[1,2], Philip McClurg[1], Serge Batalov[1], Andrew I. Su[1], S. Whitney Barnes[1], Erica Lagler[1], Ron Korstanje[3], Xiaosong Wang[3], Deborah Nusskern[4], Molly A. Bogue[3], Richard J. Mural[4], Beverly Paigen[3], Tim Wiltshire[1*]

1 Genomics Institute of the Novartis Resea
States of America, 3 The Jackson Laborato

Rapid expansion of available
development of new methods
provides an expedient way
polymorphisms for the purpo
(SNP) data have lacked the de
remedy this, 470,407 allele ca
of the SNP set with statistica
haplotype could successfully
method to high-density lipop
loci (QTL). The inferred haplot
more easily identified and ch

In Silico Mapping of Complex Disease-Related Traits in Mice

Andrew Grupe,[1*] Soren Germer,[2*] Jonathan Usuka,[3*] Dee Aud,[1] John K. Belknap,[4] Robert F. Klein,[4] Mandeep K. Ahluwalia,[2] Russell Higuchi,[2] Gary Peltz[1†]

Experimental murine geneti
potential for understanding
required for analysis of such
a computational method for
notypic traits and a murine d
developed. After entry of phe
strains, the phenotypic and ge
the chromosomal regions re

An Integrated *in Silico* Gene Mapping Strategy in Inbred Mice

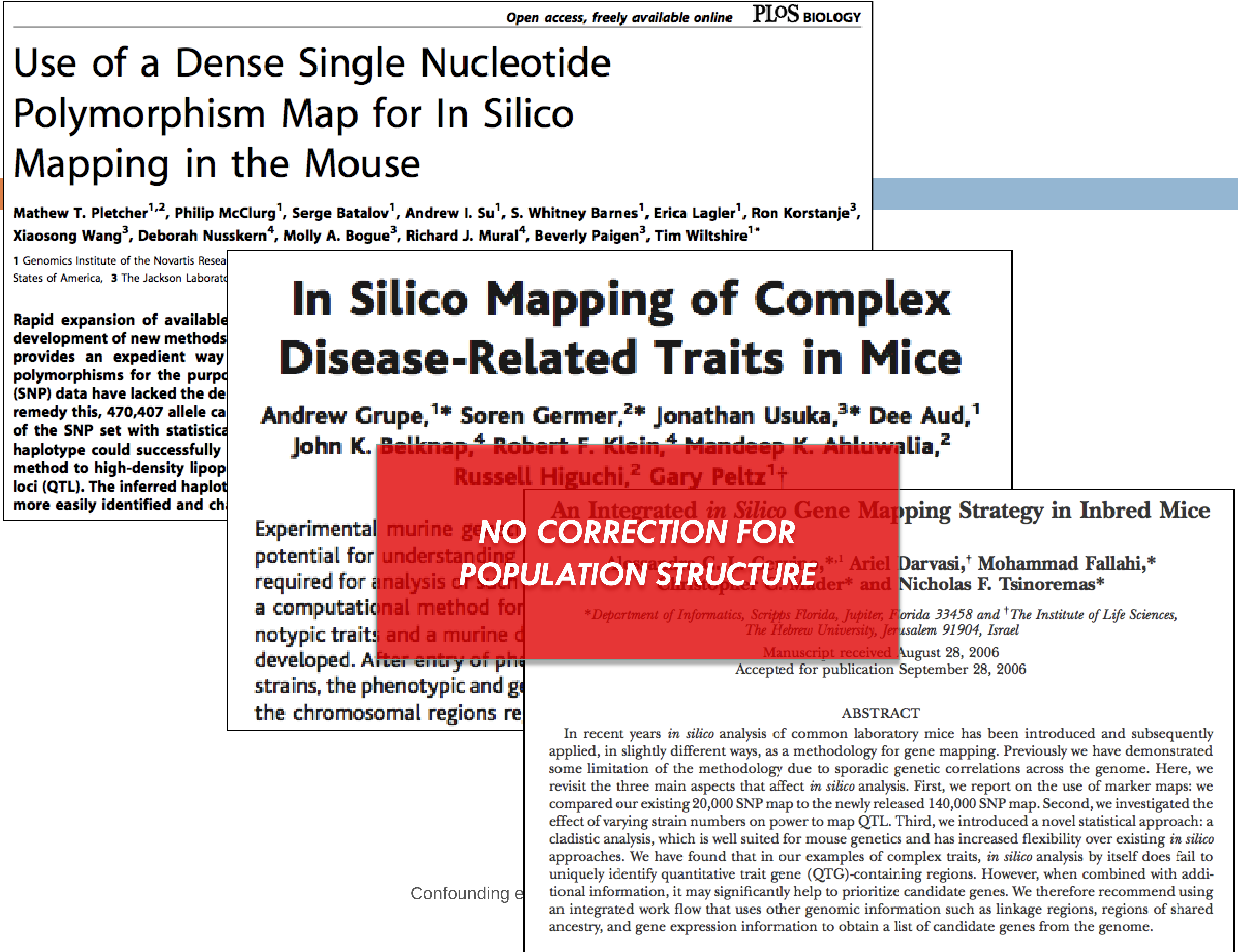
Alessandra C. L. Cervino,[*,1] Ariel Darvasi,[†] Mohammad Fallahi,* Christopher C. Mader* and Nicholas F. Tsinoremas*

*Department of Informatics, Scripps Florida, Jupiter, Florida 33458 and †The Institute of Life Sciences, The Hebrew University, Jerusalem 91904, Israel

Manuscript received August 28, 2006
Accepted for publication September 28, 2006

ABSTRACT

In recent years *in silico* analysis of common laboratory mice has been introduced and subsequently applied, in slightly different ways, as a methodology for gene mapping. Previously we have demonstrated some limitation of the methodology due to sporadic genetic correlations across the genome. Here, we revisit the three main aspects that affect *in silico* analysis. First, we report on the use of marker maps: we compared our existing 20,000 SNP map to the newly released 140,000 SNP map. Second, we investigated the effect of varying strain numbers on power to map QTL. Third, we introduced a novel statistical approach: a cladistic analysis, which is well suited for mouse genetics and has increased flexibility over existing *in silico* approaches. We have found that in our examples of complex traits, *in silico* analysis by itself does fail to uniquely identify quantitative trait gene (QTG)-containing regions. However, when combined with additional information, it may significantly help to prioritize candidate genes. We therefore recommend using an integrated work flow that uses other genomic information such as linkage regions, regions of shared ancestry, and gene expression information to obtain a list of candidate genes from the genome.



Open access, freely available online PLoS BIOLOGY

Use of a Dense Single Nucleotide Polymorphism Map for In Silico Mapping in the Mouse

Mathew T. Pletcher[1,2], Philip McClurg[1], Serge Batalov[1], Andrew I. Su[1], S. Whitney Barnes[1], Erica Lagler[1], Ron Korstanje[3], Xiaosong Wang[3], Deborah Nusskern[4], Molly A. Bogue[3], Richard J. Mural[4], Beverly Paigen[3], Tim Wiltshire[1*]

1 Genomics Institute of the Novartis Resea
States of America, 3 The Jackson Laborato

Rapid expansion of available
development of new methods
provides an expedient way
polymorphisms for the purpo
(SNP) data have lacked the de
remedy this, 470,407 allele ca
of the SNP set with statistica
haplotype could successfully
method to high-density lipop
loci (QTL). The inferred haplot
more easily identified and cha

In Silico Mapping of Complex Disease-Related Traits in Mice

Andrew Grupe,[1*] Soren Germer,[2*] Jonathan Usuka,[3*] Dee Aud,[1]
John K. Belknap,[4] Robert F. Klein,[4] Mandeep K. Ahluwalia,[2]
Russell Higuchi,[2] Gary Peltz[1]†

Experimental murine ge
potential for understanding
required for analysis of such
a computational method for
notypic traits and a murine d
developed. After entry of phe
strains, the phenotypic and ge
the chromosomal regions re

An Integrated *in Silico* Gene Mapping Strategy in Inbred Mice

Alexandra C. L. Cervino,*,[1] Ariel Darvasi,† Mohammad Fallahi,* Christopher C. Mader* and Nicholas F. Tsinoremas*

*Department of Informatics, Scripps Florida, Jupiter, Florida 33458 and †The Institute of Life Sciences, The Hebrew University, Jerusalem 91904, Israel

Manuscript received August 28, 2006
Accepted for publication September 28, 2006

ABSTRACT

In recent years *in silico* analysis of common laboratory mice has been introduced and subsequently applied, in slightly different ways, as a methodology for gene mapping. Previously we have demonstrated some limitation of the methodology due to sporadic genetic correlations across the genome. Here, we revisit the three main aspects that affect *in silico* analysis. First, we report on the use of marker maps: we compared our existing 20,000 SNP map to the newly released 140,000 SNP map. Second, we investigated the effect of varying strain numbers on power to map QTL. Third, we introduced a novel statistical approach: a cladistic analysis, which is well suited for mouse genetics and has increased flexibility over existing *in silico* approaches. We have found that in our examples of complex traits, *in silico* analysis by itself does fail to uniquely identify quantitative trait gene (QTG)-containing regions. However, when combined with additional information, it may significantly help to prioritize candidate genes. We therefore recommend using an integrated work flow that uses other genomic information such as linkage regions, regions of shared ancestry, and gene expression information to obtain a list of candidate genes from the genome.

NO CORRECTION FOR POPULATION STRUCTURE

Unmodeled factors are not known

TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \underline{\boxed{\sum_{i\neq k} ?\, X_i\beta_i}} + \mathbf{e}$$

UNMODELED FACTORS

SIMPLE LINEAR MODEL

$$\mathbf{y} = \hat{\mu} + X_k\hat{\beta}_k + \mathbf{e}$$

Unmodeled factors & population structure

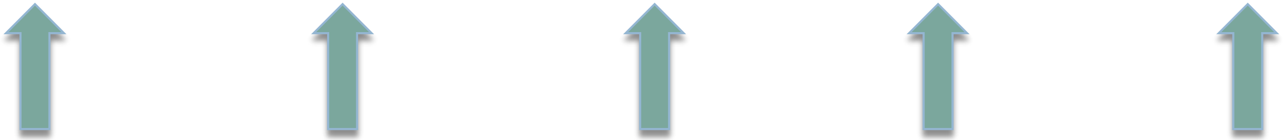
TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \underline{\sum_{i\neq k} X_i\beta_i} + \mathbf{e}$$

UNMODELED FACTORS

Strain	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
B6	A	C	C	G	T	A	A	G	C	T
C3H	A	C	C	G	A	A	A	G	C	T

CAUSAL SNPS



Unmodeled factors & population structure

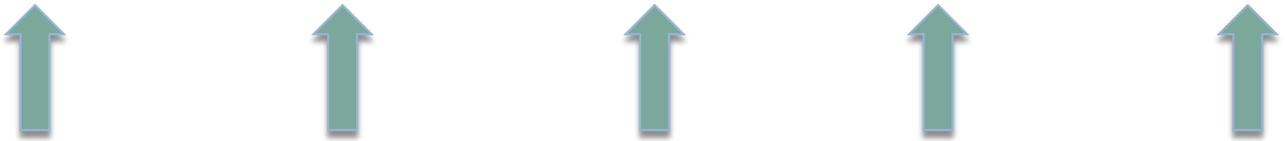
TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \underline{\sum_{i\neq k} X_i\beta_i} + \mathbf{e}$$

UNMODELED FACTORS

Strain	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
B6	A	C	C	G	T	A	A	G	C	T
CAST	T	G	T	C	A	C	A	A	T	G

CAUSAL SNPS



Unmodeled factors & population structure

TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \underbrace{\sum_{i\neq k} X_i\beta_i}_{\text{UNMODELED FACTORS}} + \mathbf{e}$$

Strain	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
B6	A	C	C	G	T	A	A	G	C	T
C3H	A	C	C	G	A	A	A	G	C	T
DBA	A	C	C	G	A	A	T	G	T	T
129S1	A	G	C	G	T	C	T	G	C	T
CAST	T	G	T	C	A	C	A	A	T	G

Unmodeled factors & population structure

TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \underline{\sum_{i\neq k} X_i\beta_i} + \mathbf{e}$$

UNMODELED FACTORS

# of shared SNPs (K)

B6		9	7	7	1
C3H	9		8	7	2
DBA	7	8		6	2
129S1	7	7	6		2
CAST	1	2	2	2	

Strain	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
B6	A	C	C	G	T	A	A	G	C	T
C3H	A	C	C	G	A	A	A	G	C	T
DBA	A	C	C	G	A	A	T	G	T	T
129S1	A	G	C	G	T	C	T	G	C	T
CAST	T	G	T	C	A	C	A	A	T	G

Dependency among unmodeled factors are ignored

TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \underline{\boxed{\sum_{i\neq k} ? X_i\beta_i}} + \mathbf{e}$$

UNMODELED FACTORS

# of shared SNPs (K)

B6		9	7	7	1
C3H	9		8	7	2
DBA	7	8		6	2
129S1	7	7	6		2
CAST	1	2	2	2	

SIMPLE LINEAR MODEL

$$\mathbf{y} = \hat{\mu} + X_k\hat{\beta}_k + \mathbf{e}$$

Mixed model accounts for the dependency

TRUE MODEL

$$\mathbf{y} = \mu + X_k\beta_k + \sum_{i\neq k} ? X_i\beta_i + \mathbf{e}$$

UNMODELED FACTORS

# of shared SNPs (K)

B6		9	7	7	1
C3H	9		8	7	2
DBA	7	8		6	2
129S1	7	7	6		2
CAST	1	2	2	2	

LINEAR MIXED MODEL

$$\mathbf{y} = \hat{\mu} + X_k\hat{\beta}_k + \mathbf{u} + \mathbf{e}$$

$$\mathbf{u} \sim N(0, \hat{\sigma}_g^2 K)$$

$$\mathbf{e} \sim N(0, \hat{\sigma}_e^2 I)$$

Linear Mixed Model (LMM)

- Recently, the LMM has become a popular approach for GWAS as it can correct for population structure.
- The LMM incorporates genetic similarities between all pairs of individuals, known as the kinship (**K**), into their model and corrects for population structure.

Linear Model

$$\mathbf{y} = \mu + X_i \beta_i + \mathbf{e}$$

LMM

$$\mathbf{y} = \mu + X_i \beta_i + \boxed{\mathbf{u}} + \mathbf{e}$$

$\mathrm{Var}(\mathbf{u}) = \sigma_g^2 \mathbf{K}$ $\mathrm{Var}(\mathbf{e}) = \sigma_e^2 \mathrm{I}$

$$\mathbf{y} \sim N(X\beta, \sigma_g^2 \mathrm{K} + \sigma_e^2 \mathrm{I})$$

Fraction of shared SNPs = IBS matrix

K

B6		.9	.7	.7	.1
C3H	.9		.8	.7	.2
DBA	.7	.8		.6	2
129S1	.7	.7	.6		2
CAST	.1	.2	.2	.2	

Previous methods

- MANOVA, multivariate regression analysis
  - Designed for use with a small number of variables. P<<N
  - MANOVA assumes MVN
  - Only can use Euclidean distances
- Data reduction methods - Cluster analysis, factor analysis, etc.
- mvLMMs (Furlotte and Eskin, Genetics, 2015; Zhou *et al*., Nat Methods , 2012) , MTMM(Korte *et al*., Nat Genet , 2012) - LMM based approaches, computational costs scale quadratically with the number of phenotypes
- MDMR (Zapala *et al*., Front Genet, 2012)
  - **M**ultivariate **D**istance **M**atrix **R**egression analysis.
  - Form a statistic to test the effect of some covariates on all of the phenotypes by utilizing the similarity matrix that reflects the correlation of the samples with respect to the expression values over the genes.

$$\text{"Pseudo" F-statistics, } F = \frac{tr(\hat{Y}\hat{Y}')/(2-1)}{tr(\hat{R}\hat{R}')/(n-2)}$$

Previous methods

- MANOVA, multivariate regression analysis
  - Designed for use with a small number of variables. P<<N
  - MANOVA assumes MVN
  - Only can use Euclidean distances
- Data reduction methods - Cluster analysis, factor analysis, etc.
- mvLMMs (Zhou *et al.*, Nat Methods , 2012) , MTMM(Korte *et al.*, Nat Genet , 2012) - LMM based approaches, computational costs scale quadratically with the number of phenotypes
- MDMR (Zapala *et al.*, Front Genet, 2012)
  - **M**ultivariate **D**istance **M**atrix **R**egression analysis.
  - Form a statistic to test the effect of some covariates on all of the phenotypes by utilizing the similarity matrix that reflects the correlation of the samples with respect to the expression values over the genes.

$$\text{"Pseudo" F-statistics, } F = \frac{tr(\hat{Y}\hat{Y}')/(2-1)}{tr(\hat{R}\hat{R}')/(n-2)}$$

Univariate-phenotypes analysis

- Traditional univariate analysis for snp i and phenotype j

RSS_i : Sum of squares stimates of model i

p_i : Number of parameters of model i

n : Number of samples

$$y_j = X_i\beta_j + e_j$$

$$\hat{y}_j = X_i\hat{\beta}_j = X_i(X_i'X_i)^{-1}X_i'y_j$$

$$\hat{r}_j = y_j - \hat{y}_j = y_j - X_i(X_i'X_i)^{-1}X_i'y_j$$

- Hypothesis testing

$$\begin{cases} H_0 : \beta_j = 0 \\ H_A : \beta_j \neq 0 \end{cases} \quad \begin{cases} \text{Model 1} : y_j = e_j \\ \text{Model 2} : y_j = X_i\beta_j + e_j \end{cases}$$

$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(n - p_2)} = \frac{\hat{y}_j'\hat{y}_j \ /(2-1)}{\hat{r}_j'\hat{r}_j \ /(n-2)}$$

Multiple-phenotypes analysis

- □ Extend to multivariate case for snp i and m number of phenotypes

$$Y = X_i\beta + E$$

$$\hat{Y} = X_i(X_i'X_i)^{-1}X_i'Y$$

$$\hat{R} = Y - \hat{Y}$$

- □ Hypothesis testing

$$F = \frac{tr(\hat{Y}'\hat{Y})/(2-1)}{tr(\hat{R}'\hat{R})/(n-2)}$$

- □ Caveat: Since Y is not independent, F does not follow F distribution

Linear Mixed Model

$$y_j = X_i\beta_j + u_j + e_j \qquad y_j \sim N(X_i\beta_j, \Sigma_j) \qquad \Sigma = \sigma_g^2 K + \sigma_e^2 I$$

$$\hat{\Sigma}^{-1/2} y_j \sim N(\hat{\Sigma}^{-1/2} X_i\beta_j, \Sigma_j)$$

$$\tilde{X}_i = \hat{\Sigma}^{-1/2} X_{ii}$$

$$\tilde{y}_i = \hat{\Sigma}^{-1/2} y_j$$

$$\hat{\tilde{y}}_j = \tilde{X}_i(\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'\tilde{y}_j$$

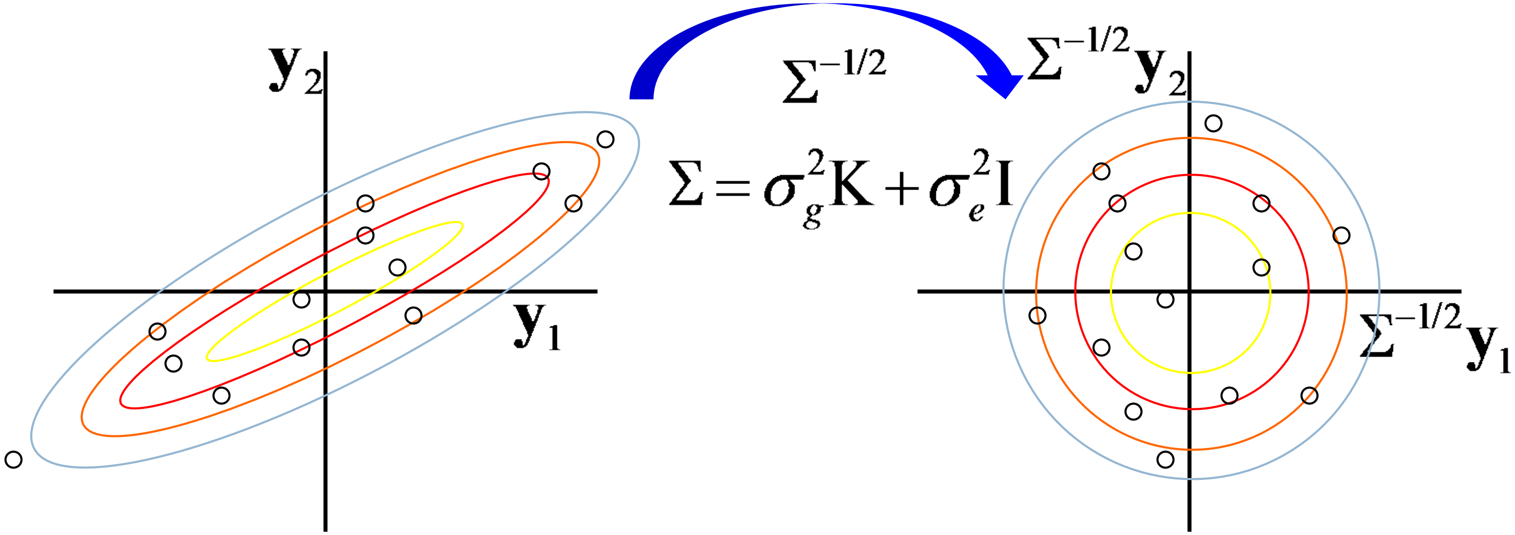
$$\hat{\tilde{r}}_j = \tilde{y}_j - \hat{\tilde{y}}_j$$

$$F = \frac{\hat{\tilde{y}}_j'\hat{\tilde{y}}_j \,/\, (2-1)}{\hat{\tilde{r}}_j'\hat{\tilde{r}}_j \,/\, (n-2)}$$

GAMMA

(Generalized Analysis of Molecular variance for Mixed model Analysis)

□ Use LMM to de-correlate the correlation structure between the individuals (population structure) by rotating the genotype and phenotype space with their variance.



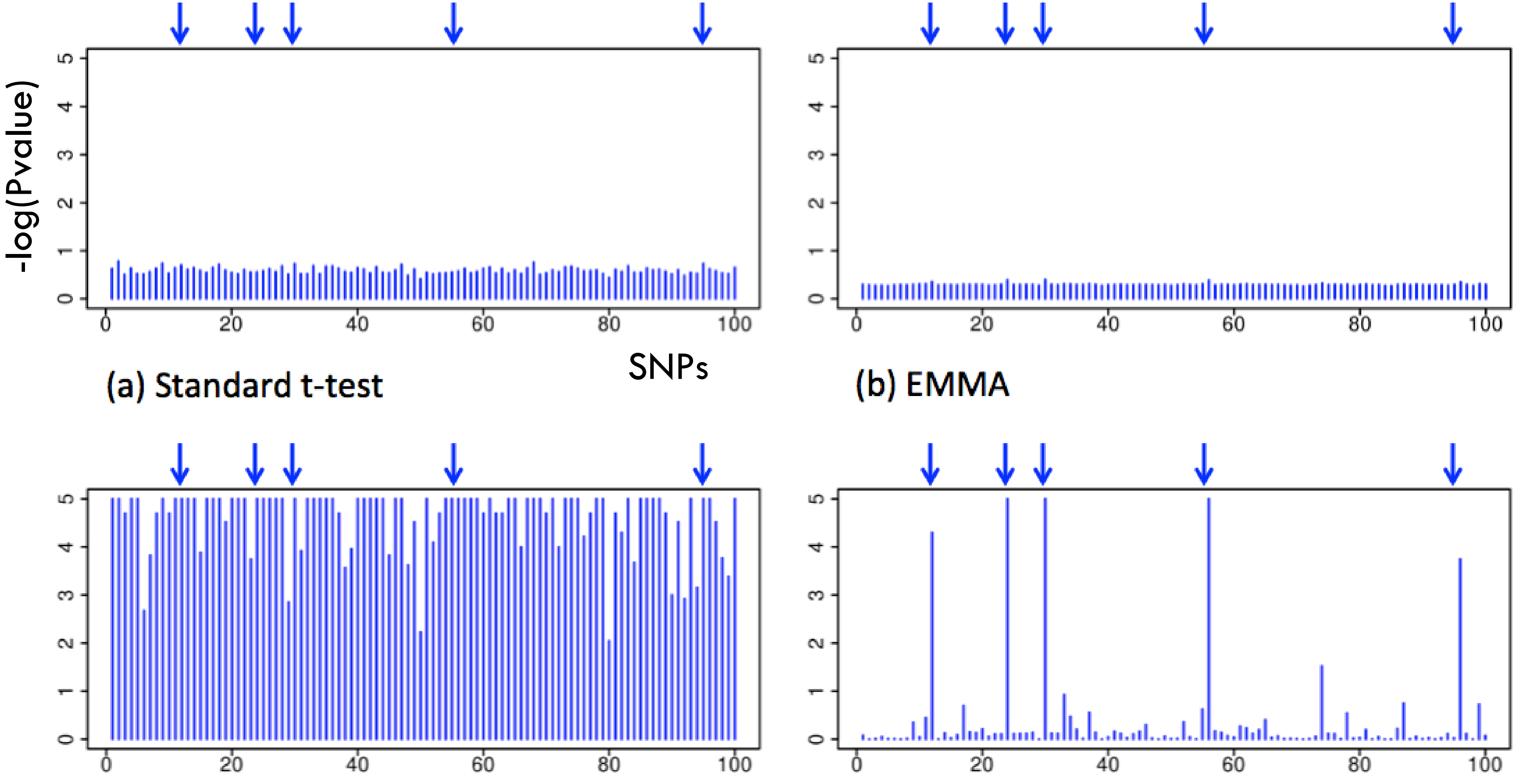
$$\Sigma = \sigma_g^2 \mathrm{K} + \sigma_e^2 \mathrm{I}$$

$$\mathbf{y} \sim N(X\beta, \sigma_g^2 \mathrm{K} + \sigma_e^2 \mathrm{I}) \qquad \Sigma^{-1/2}\mathbf{y} \sim N(\Sigma^{-1/2} X\beta, \mathrm{I})$$

□ Then apply multivariate regression method (MDMR) to form a statistic to test the effect of covariates on multiple phenotypes.

Simulated Study

True trans regulatory hotspots



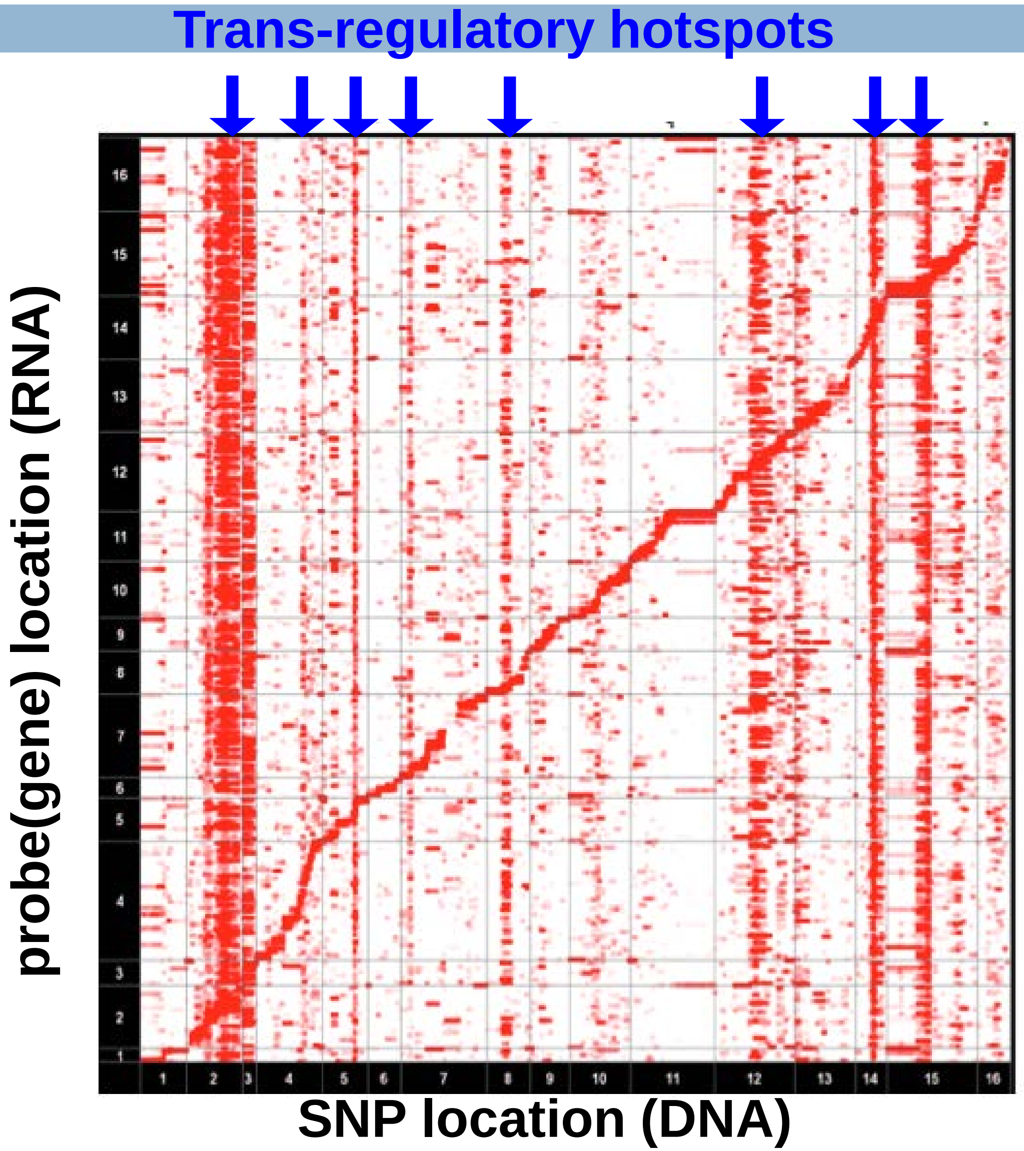
(a) Standard t-test

(b) EMMA

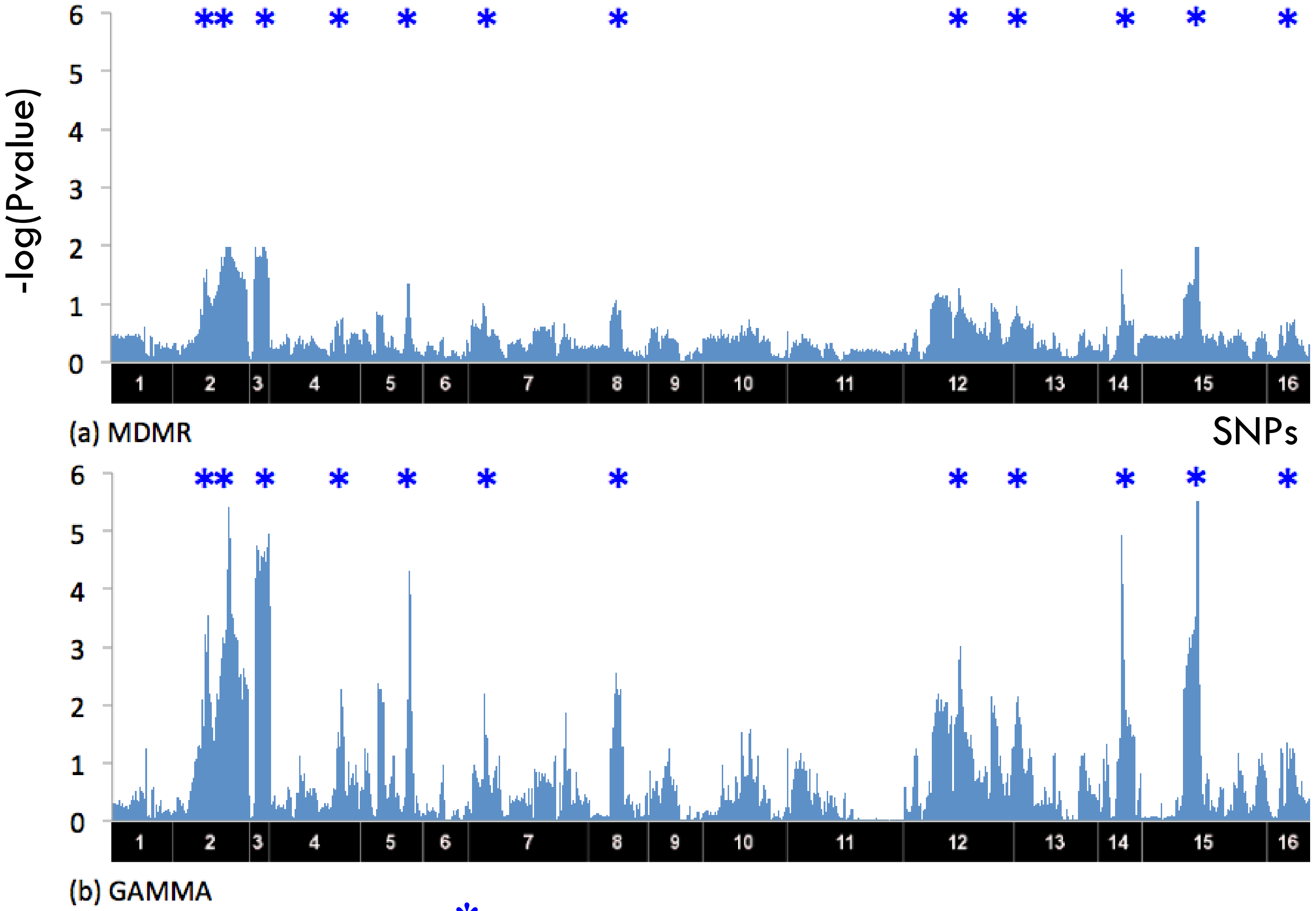
(c) MDMR

(d) GAMMA

Yeast dataset

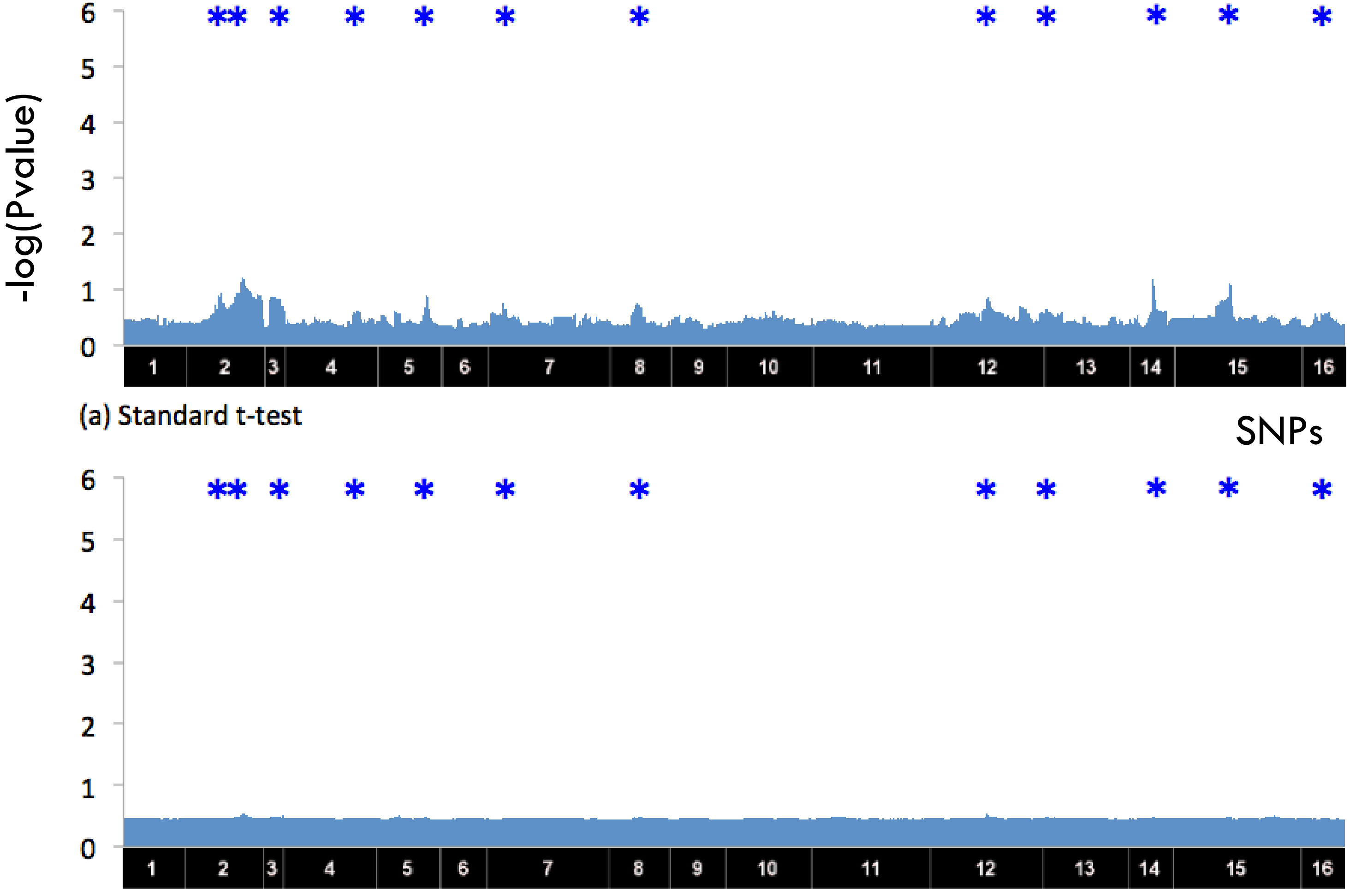


Yeast dataset



* Putative hotspots identified from NICE (GenomeBiol. Joo et al, 2014)

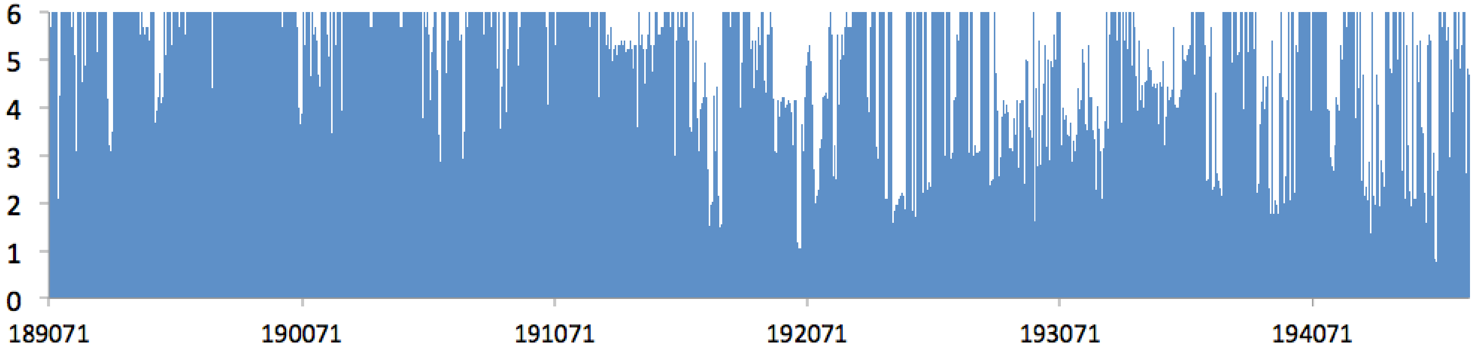
Yeast dataset



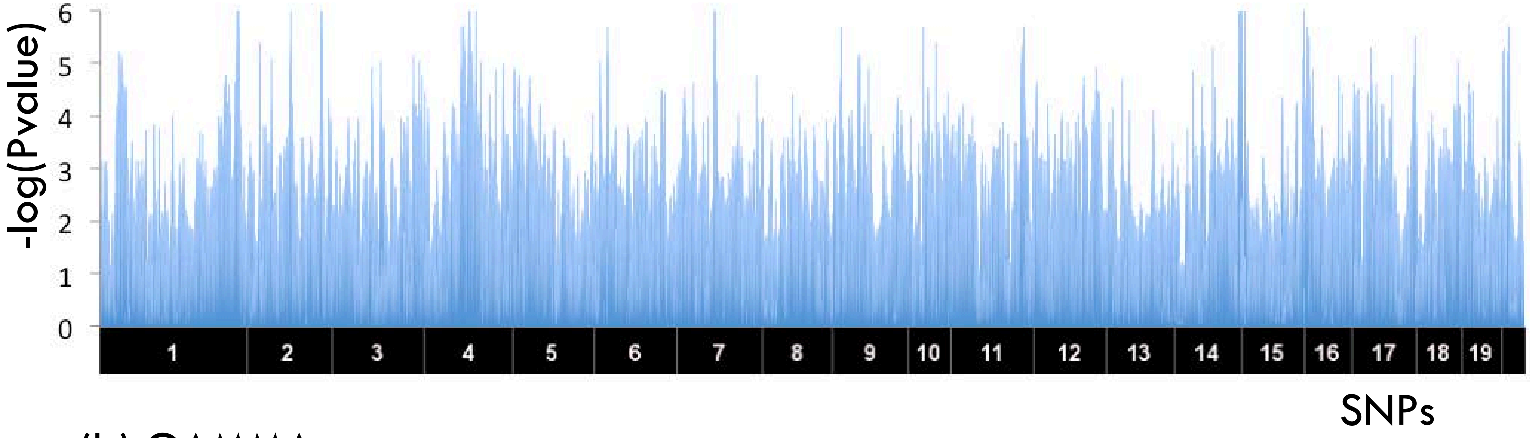
(a) Standard t-test

(b) EMMA

Gut microbiome dataset

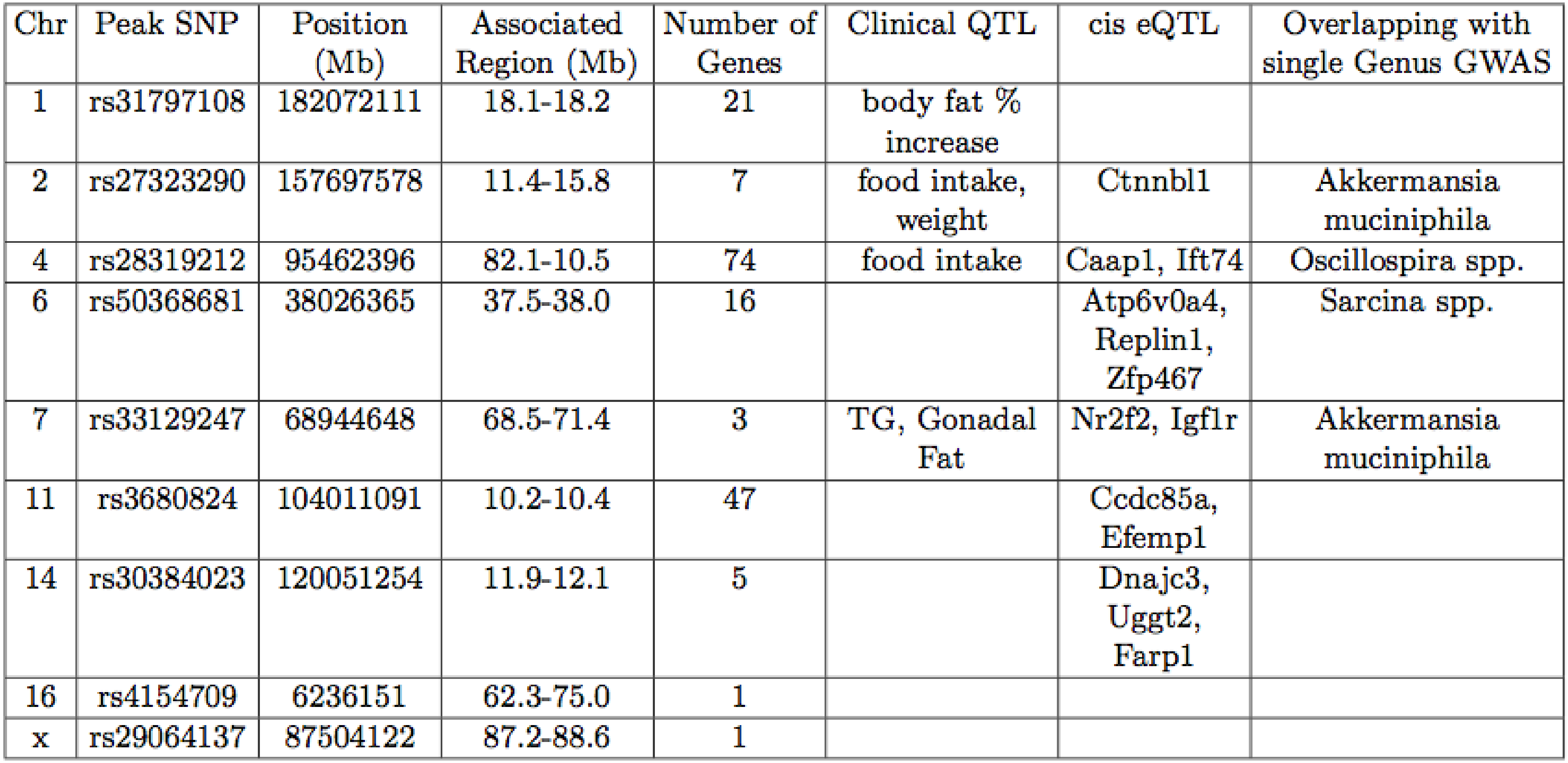


(a) MDMR on Chromosome 19



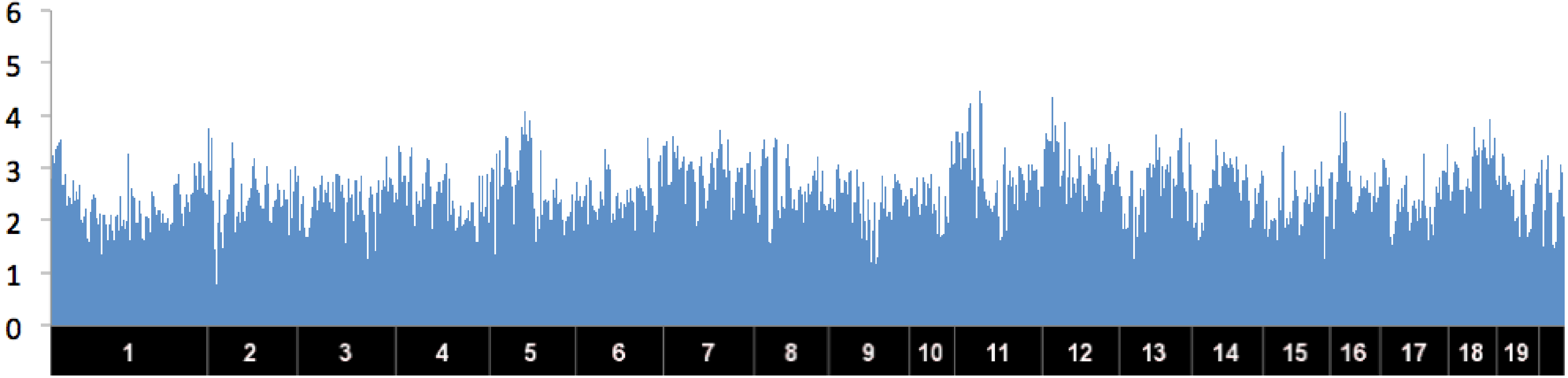
(b) GAMMA

Signals detected by GAMMA

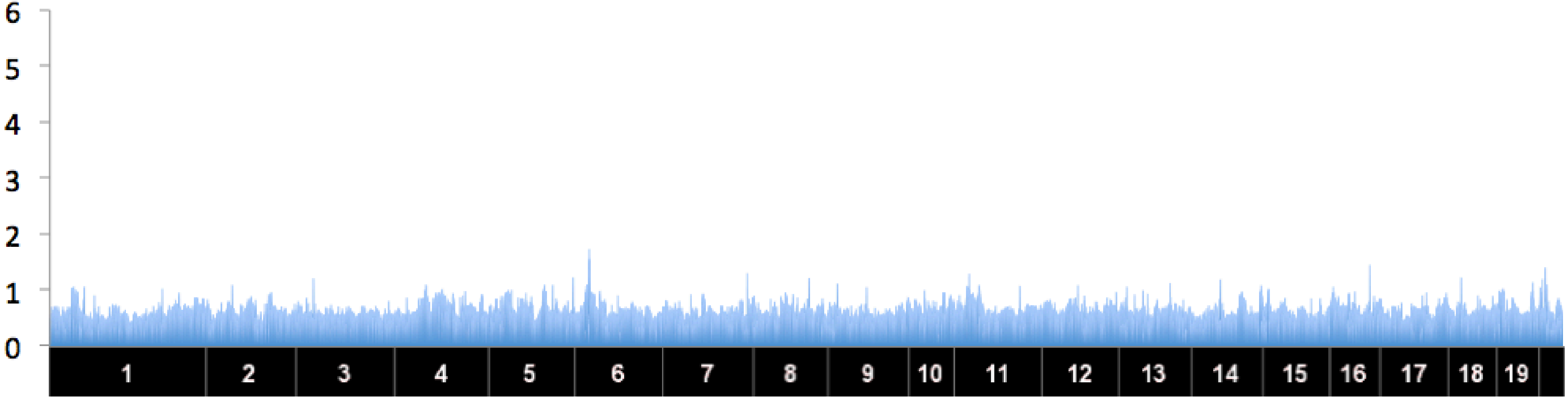


Chr	Peak SNP	Position (Mb)	Associated Region (Mb)	Number of Genes	Clinical QTL	cis eQTL	Overlapping with single Genus GWAS
1	rs31797108	182072111	18.1-18.2	21	body fat % increase		
2	rs27323290	157697578	11.4-15.8	7	food intake, weight	Ctnnbl1	Akkermansia muciniphila
4	rs28319212	95462396	82.1-10.5	74	food intake	Caap1, Ift74	Oscillospira spp.
6	rs50368681	38026365	37.5-38.0	16		Atp6v0a4, Replin1, Zfp467	Sarcina spp.
7	rs33129247	68944648	68.5-71.4	3	TG, Gonadal Fat	Nr2f2, Igf1r	Akkermansia muciniphila
11	rs3680824	104011091	10.2-10.4	47		Ccdc85a, Efemp1	
14	rs30384023	120051254	11.9-12.1	5		Dnajc3, Uggt2, Farp1	
16	rs4154709	6236151	62.3-75.0	1			
x	rs29064137	87504122	87.2-88.6	1			

gut microbiome dataset



(a) Standard t-test



(b) EMMA

Thank you ! – zarlab.cs.ucla.edu

Eun Yong Kang

Nick Furlotte

Elin Org

Brian Parks

Aldons J. Lusis

Eleazar Eskin*