Statistical models for complex trait architecture

Berlin Summer Meeting: From Systems Biology to Systems Medicine

Sayan Mukherjee

Departments of Statistical Science, Computer Science, Mathematics
Institute for Genome Sciences & Policy, Duke University
Joint work with:
Part I – DE. Runcie (UC Davis)
Part II – M. Weiser (UNC), T. Furey (UNC)
Part III – K. Turner (U Chicago) D. Boyer(Duke)
www.stat.duke.edu/∼sayan

June 13, 2014

Three parts

(1) Bayesian sparse factor model to estimate genetic covariance.

Three parts

(1) Bayesian sparse factor model to estimate genetic covariance.

(2) Finding distal eQTLs.

Three parts

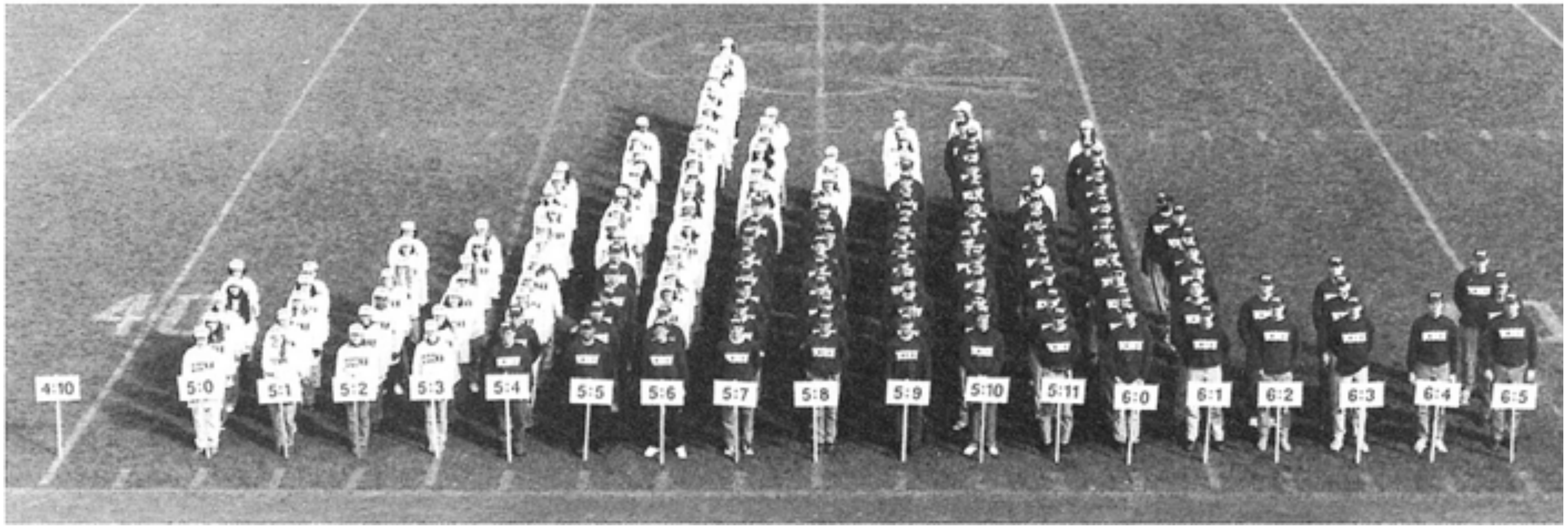
(1) Bayesian sparse factor model to estimate genetic covariance.

(2) Finding distal eQTLs.

(3) Quantitative genetics of shapes.

Genetics of multiple traits

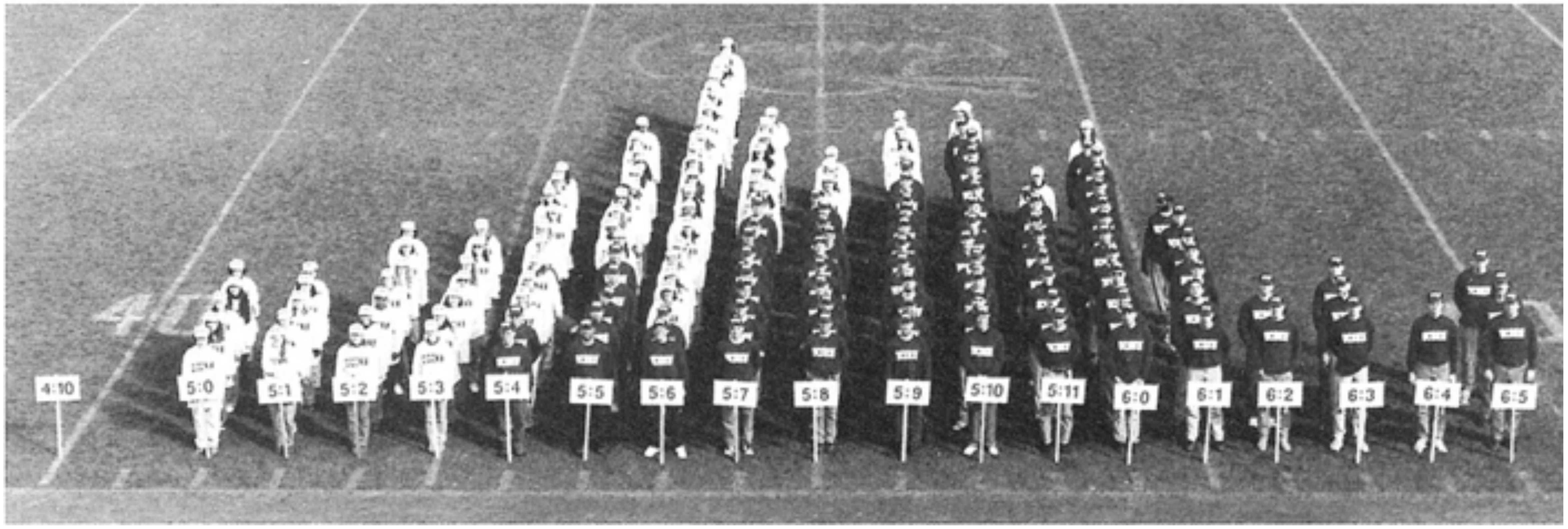
Phenotypic traits are often considered individually



Linda Strausbaugh (Genetics 147:5, 1997)

Genetics of multiple traits

Phenotypic traits are often considered individually



Linda Strausbaugh (Genetics 147:5, 1997)

Important phenotypes often involve many traits

BBC

Some objectives in quantitative genetics

Partition total phenotypic (trait) variation into genetic and environmental components.

$$\mathbf{P} = \mathbf{G} + \mathbf{E}.$$

G-matrix: matrix of genetic covariance among traits, **G**.
E-matrix: matrix covariance among traits due to environment **E**.

Some objectives in quantitative genetics

Partition total phenotypic (trait) variation into genetic and environmental components.

$$\mathbf{P} = \mathbf{G} + \mathbf{E}.$$

G-matrix: matrix of genetic covariance among traits, **G**.
E-matrix: matrix covariance among traits due to environment **E**.

Broad-sense heritability = genetic effects on phenotype, can be further partitioned into additive, dominant, and interaction effects.

Lande’s equation

Focus on additive effects: narrow-sense heritability, h^2

Fisher’s fundamental theorem (1930):
”The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time.”

Lande's equation

Focus on additive effects: narrow-sense heritability, h^2

Fisher's fundamental theorem (1930):
"The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time."

Lande or breeder's equation:

$$R = h^2 s,$$

R - response to selection, S - selection differential.

Multivariate Lande’s equation

G: matrix of additive genetic covariance among traits, **G**

Multivariate Lande's equation

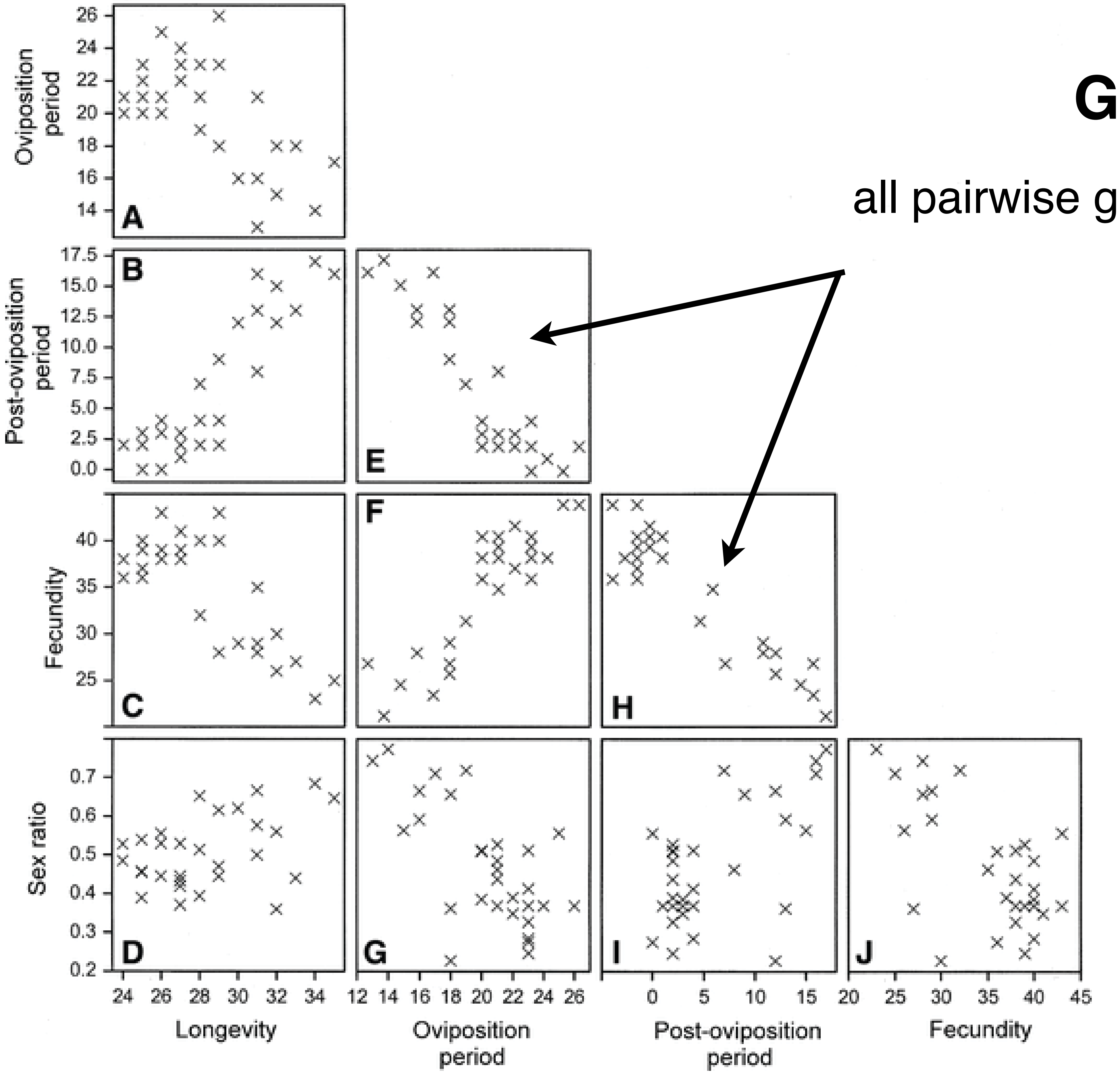
G: matrix of additive genetic covariance among traits, **G**

Lande or breeder's equation:

$$\Delta \mathbf{y} = \mathbf{G}\mathbf{s}$$

$\mathbf{Y} \sim \mathrm{N}_p$: traits are multivariate normal
$\mathbf{s} = \frac{\partial F(\bar{\mathbf{Y}})}{\partial \mathbf{y}}$: selection gradient.

Genetics of multiple traits

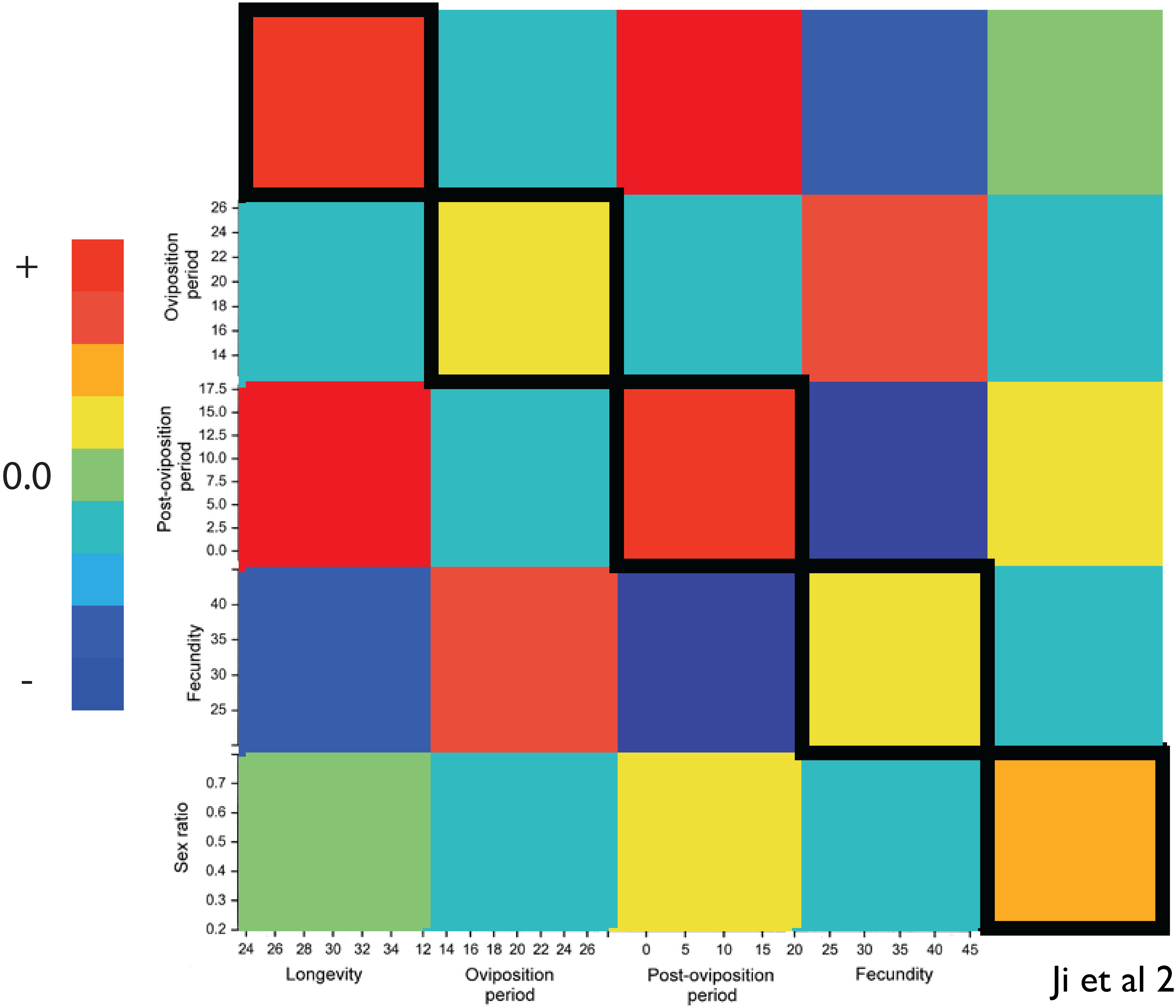


G matrix

all pairwise genetic covariances

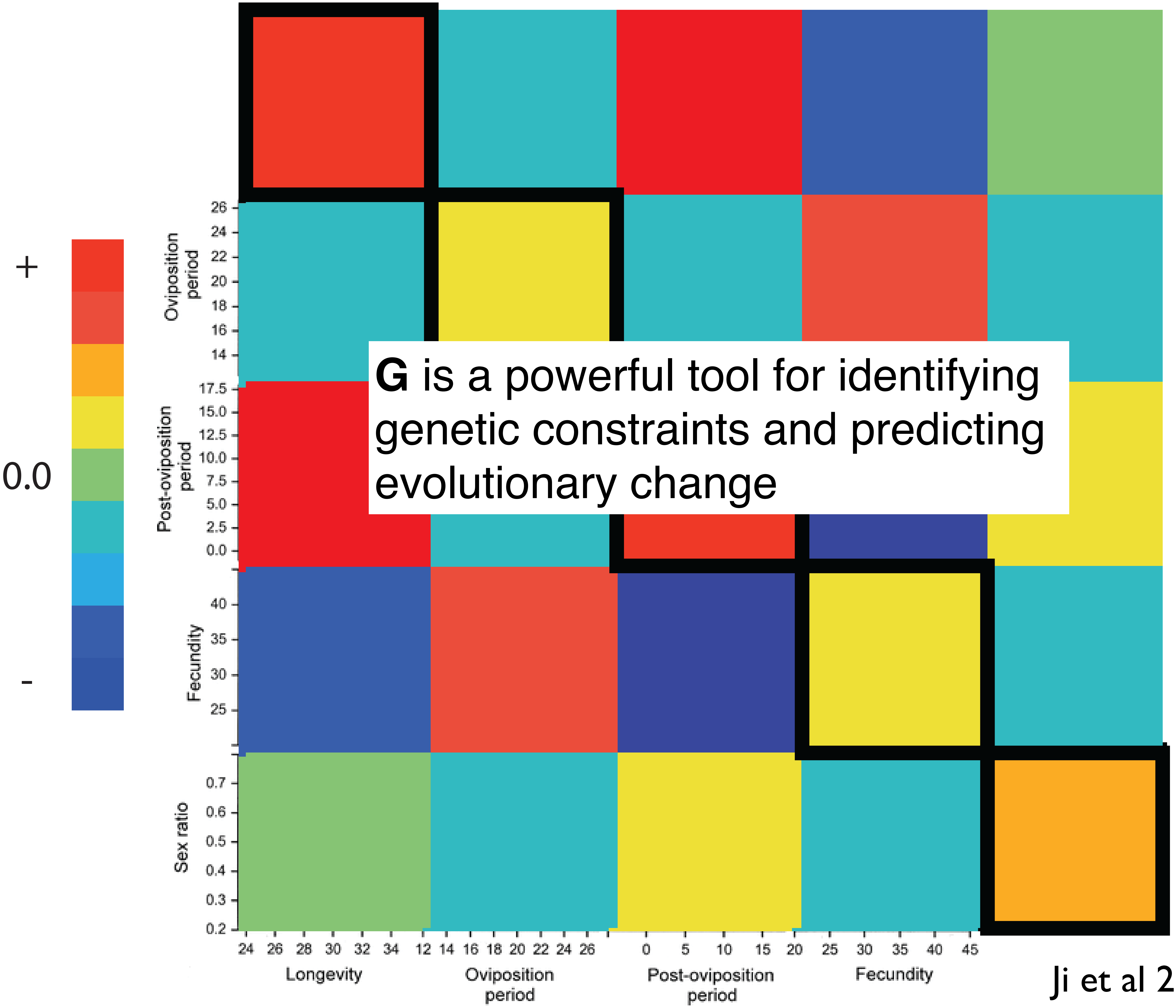
Ji et al 2007

Genetics of multiple traits



Ji et al 2007

Genetics of multiple traits



G is a powerful tool for identifying genetic constraints and predicting evolutionary change

Ji et al 2007

Genetics of many traits

Today we can measure thousands of traits simultaneously

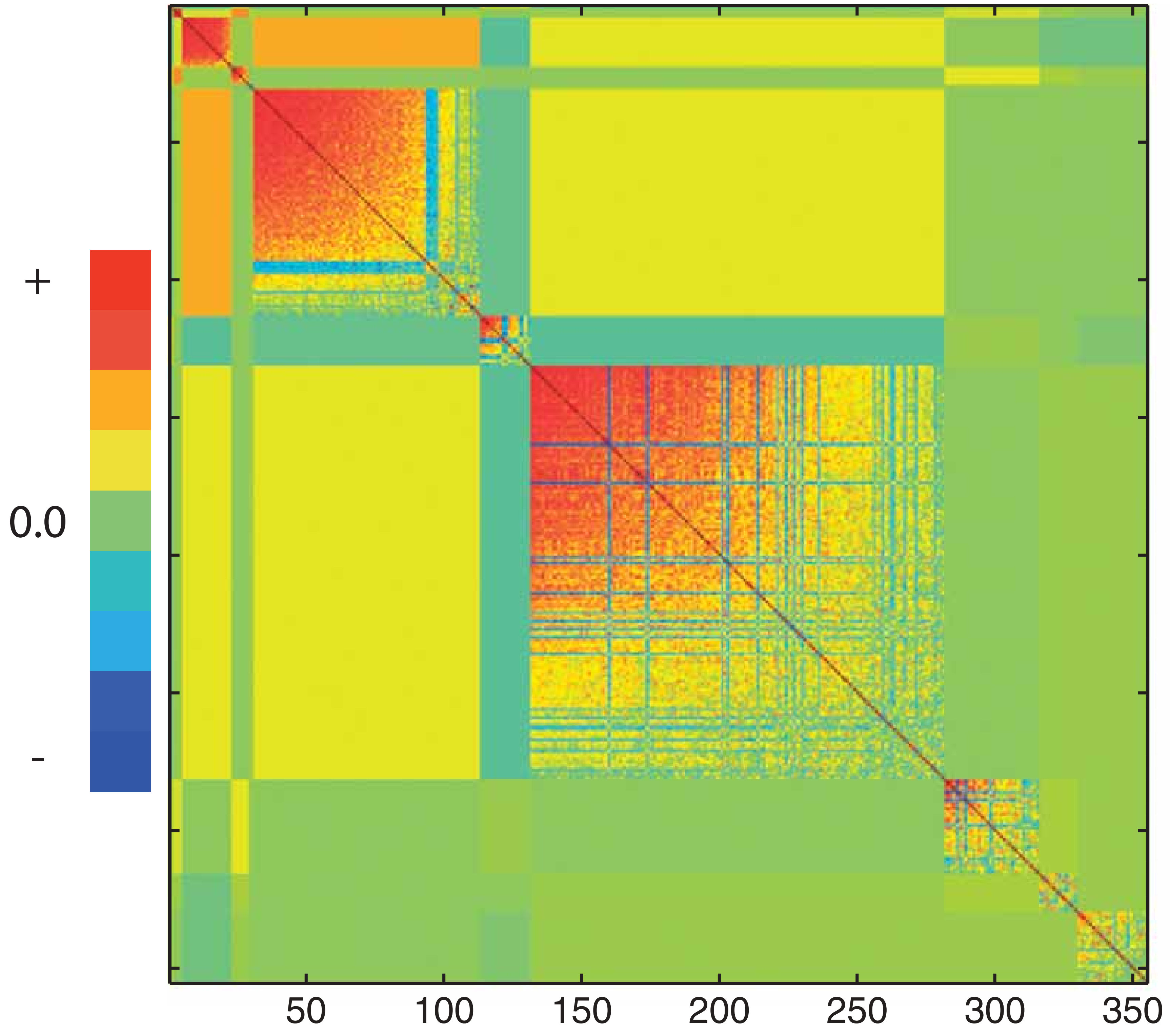
Genome-wide gene expression

Proteomics / metabolomics

morphometrics

genotype-environment interactions

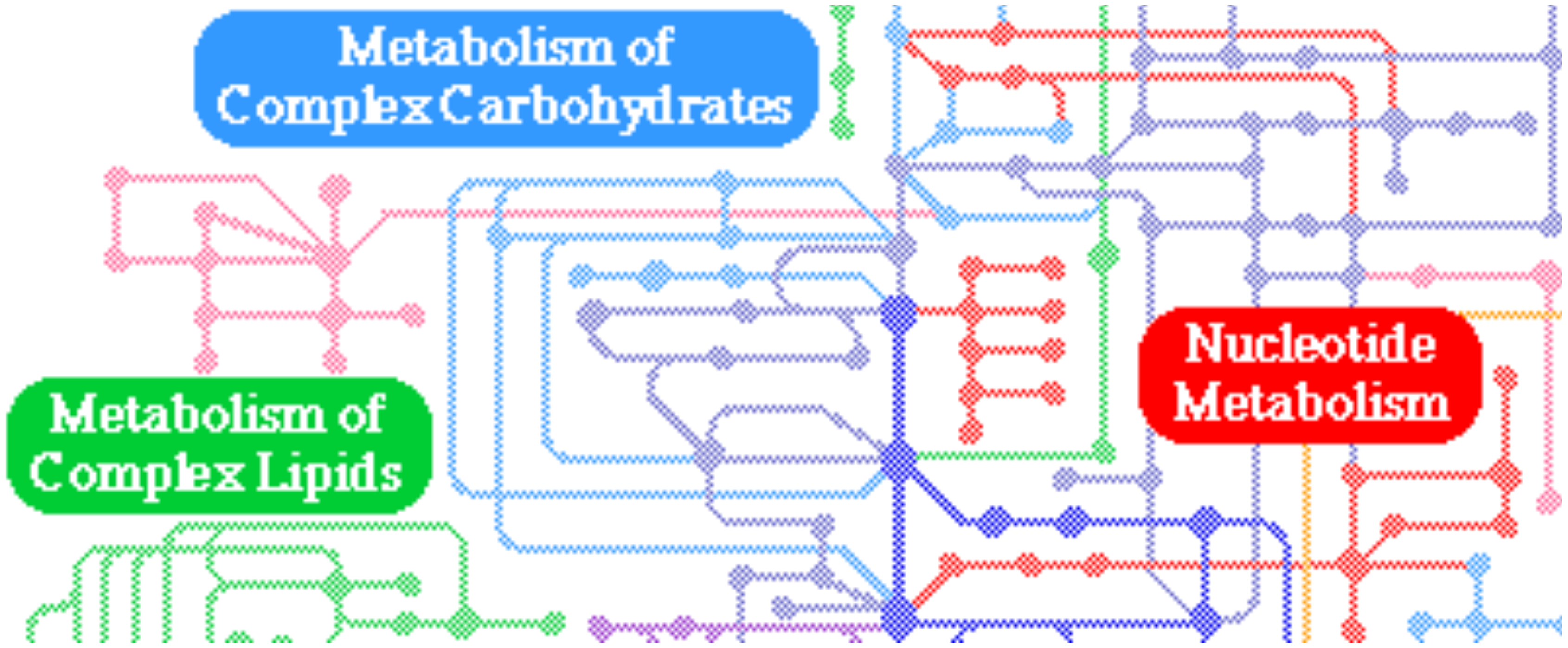
Drosophila gene expression from Ayroles et al 2009



New methods are necessary to take advantage of these data

Quantitative Genetics of Gene Expression

Gene expression is a readout of cellular activities


Metabolism of Complex Carbohydrates
Metabolism of Complex Lipids
Nucleotide Metabolism

Metabolism, and cell-signaling activity is difficult to measure
but may be key determinants of fitness

Bayesian genetic sparse factor model

Goal:

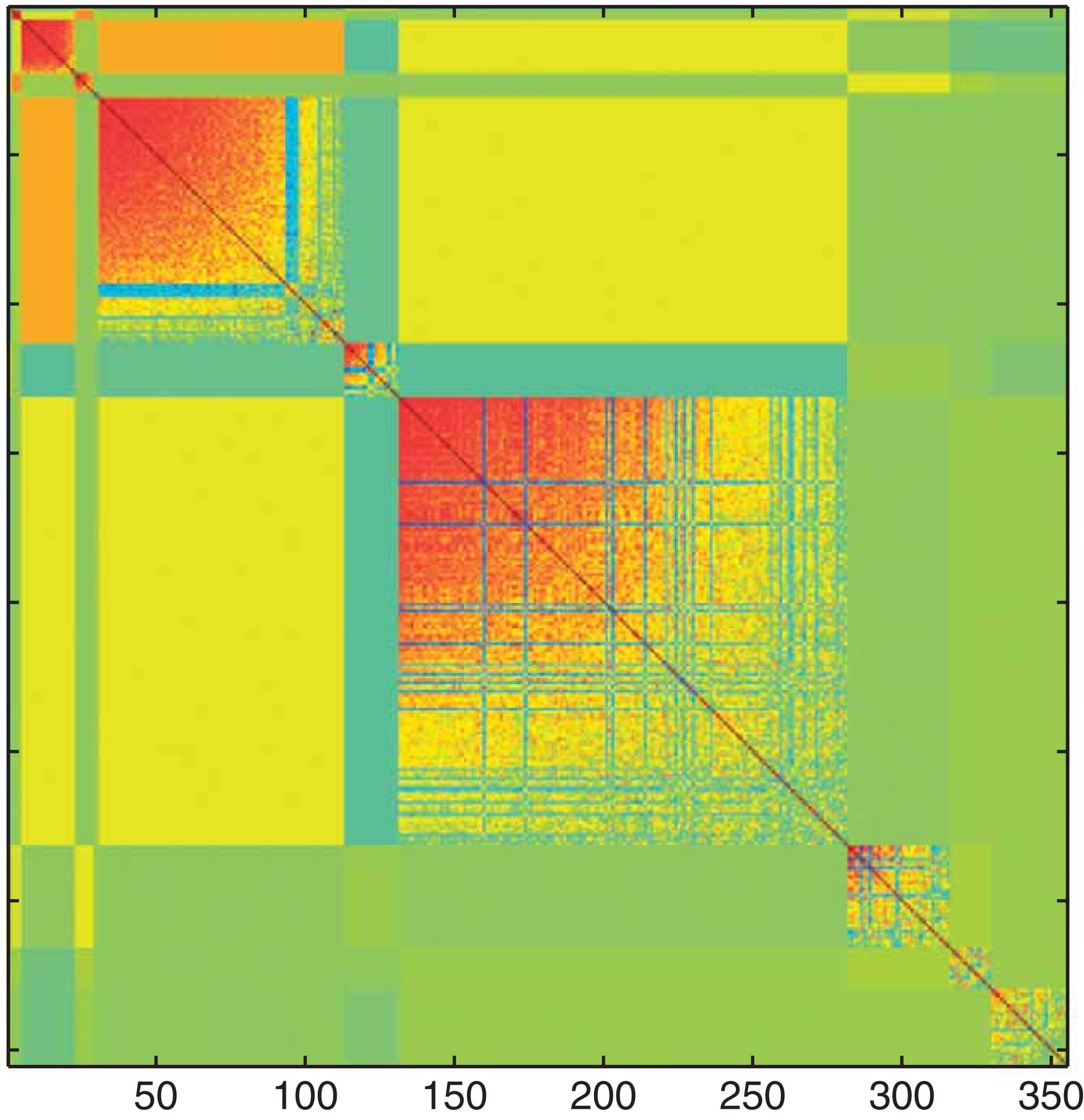
Reduce high-dimensional data to its underlying structure

Estimate evolutionary parameters

Handle complicated experimental designs or complex pedigrees

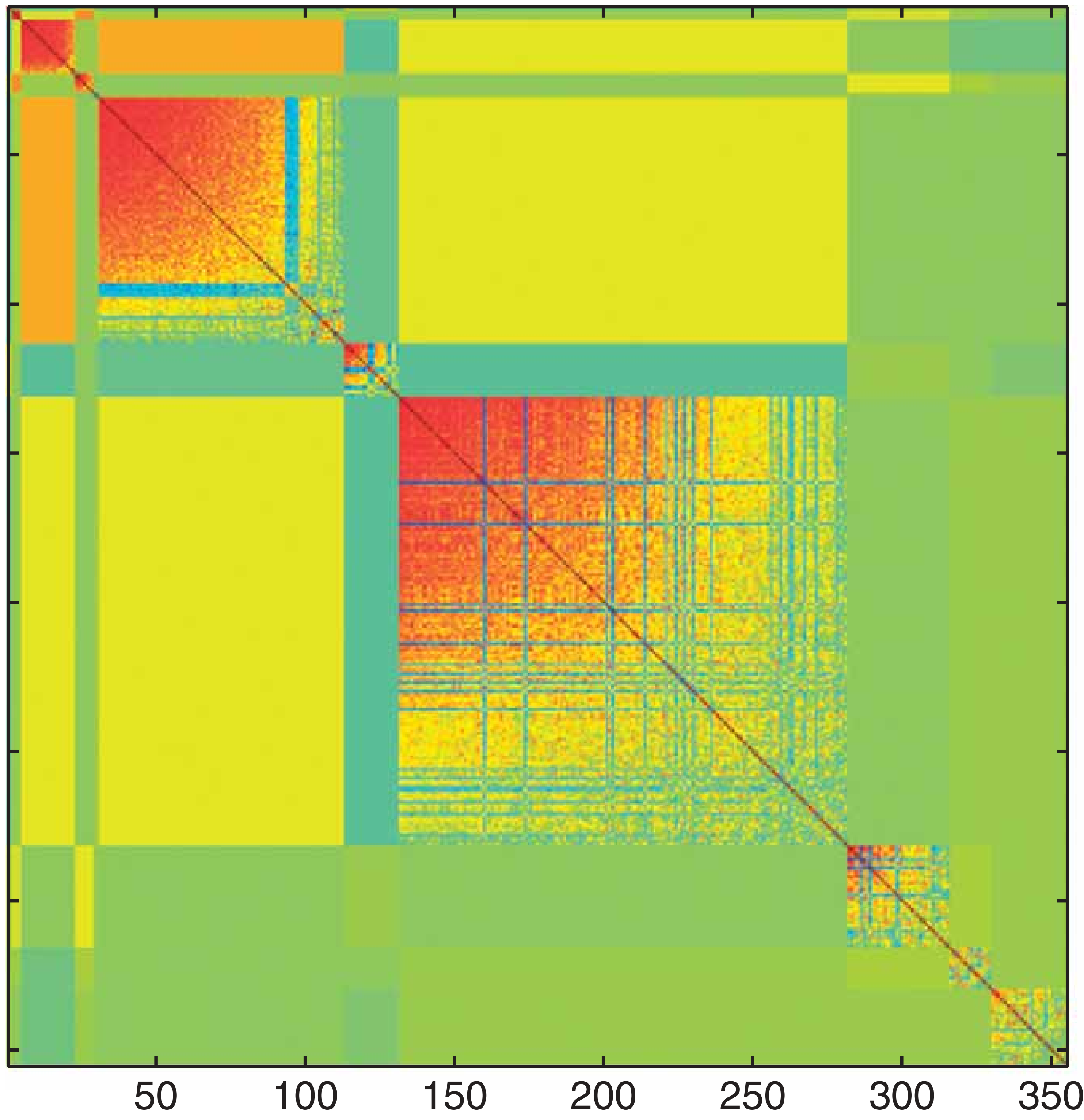
Be scalable to large numbers of traits

Ayroles et al 2009



Bayesian genetic sparse factor model

Ayroles et al 2009



Goal:

Reduce high-dimensional data to its underlying structure

Estimate evolutionary parameters

Handle complicated experimental designs or complex pedigrees

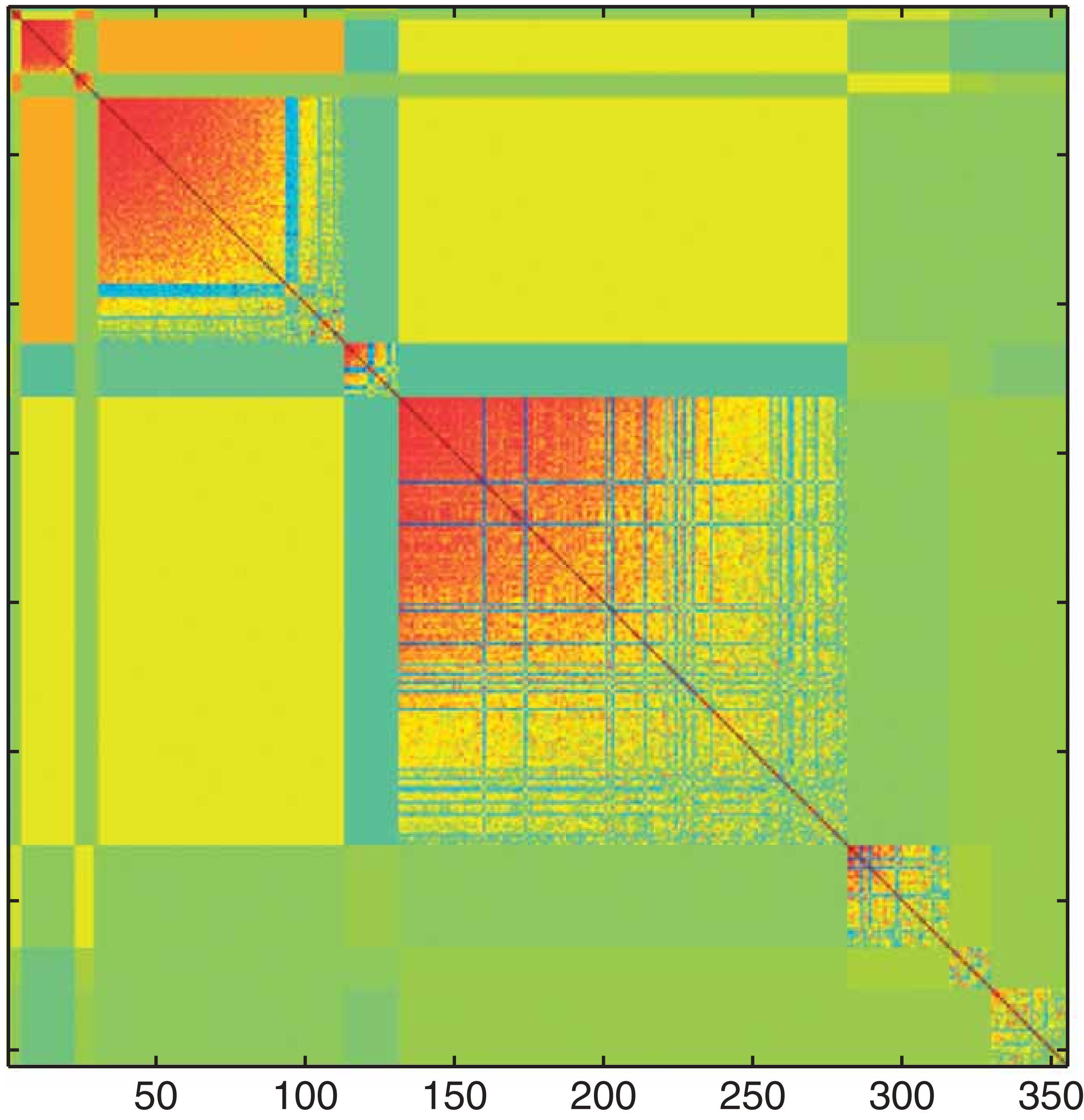
Be scalable to large numbers of traits

Methods: Bayesian dimension reduction

Sparse estimation of the **G** matrix based on an animal model

Bayesian genetic sparse factor model

Ayroles et al 2009



Goal:

Reduce high-dimensional data to its underlying structure

Estimate evolutionary parameters

Handle complicated experimental designs or complex pedigrees

Be scalable to large numbers of traits

Methods: Bayesian dimension reduction

Sparse estimation of the **G** matrix based on an animal model

Case study:

An application to *Drosophila* gene expression data

A factor model for G

Animal model for multiple traits

$$\overset{\substack{\text{phenotypes}\\ \searrow}}{\mathbf{y}_i} = \mathbf{x}_i^T \overset{\substack{\text{fixed effects}\\ \downarrow}}{\mathbf{b}_i} + \overset{\substack{\text{Random}\\ \text{genetic effects}\\ \swarrow}}{\mathbf{u}_i} + \overset{\substack{\text{residual error}\\ \swarrow}}{\epsilon_i}$$

A factor model for G

Animal model for multiple traits

phenotypes

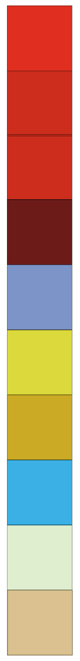
fixed effects

Random genetic effects

residual error

$$\mathbf{y}_i = \mathbf{x}_i^T \mathbf{b}_i + \mathbf{u}_i + \epsilon_i$$

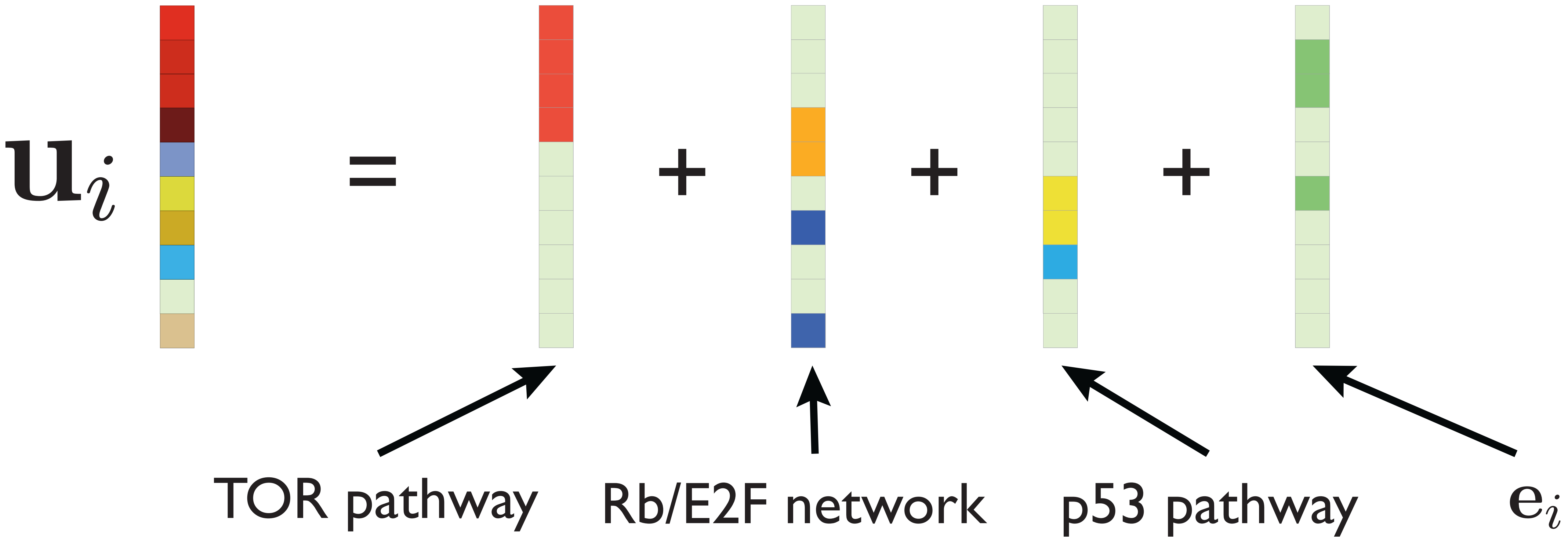
genes



$$\mathbf{u}_i \sim \mathrm{N}\left(\mathbf{0}, \mathbf{G}\right)$$

A factor model for G

Model **u** as output of development



$\mathbf{u}_i$ = + + +

TOR pathway

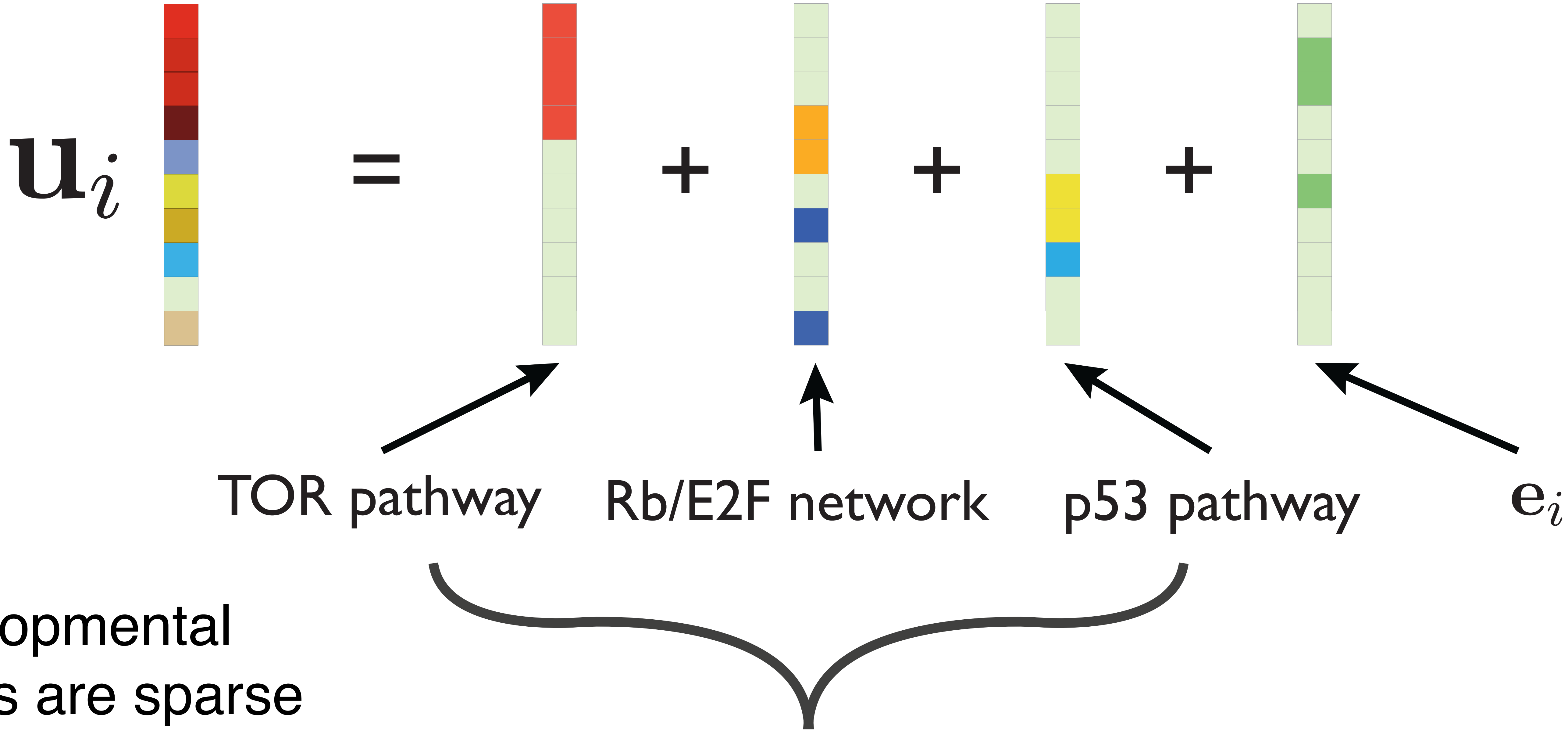
Rb/E2F network

p53 pathway

$\mathbf{e}_i$

A factor model for G

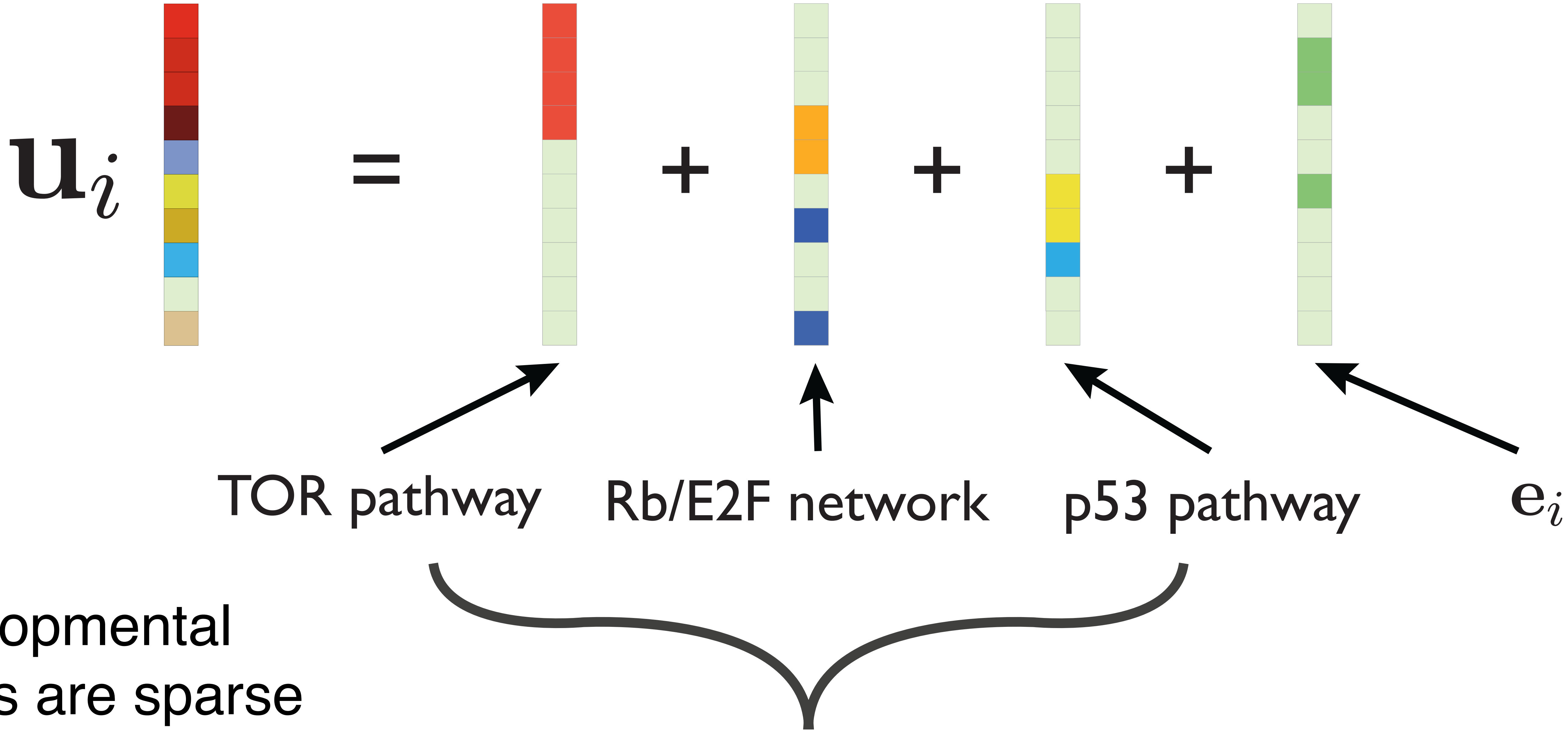
Model **u** as output of development



Developmental effects are sparse

A factor model for G

Model **u** as output of development



$\mathbf{u}_i$ = + + +

TOR pathway

Rb/E2F network

p53 pathway

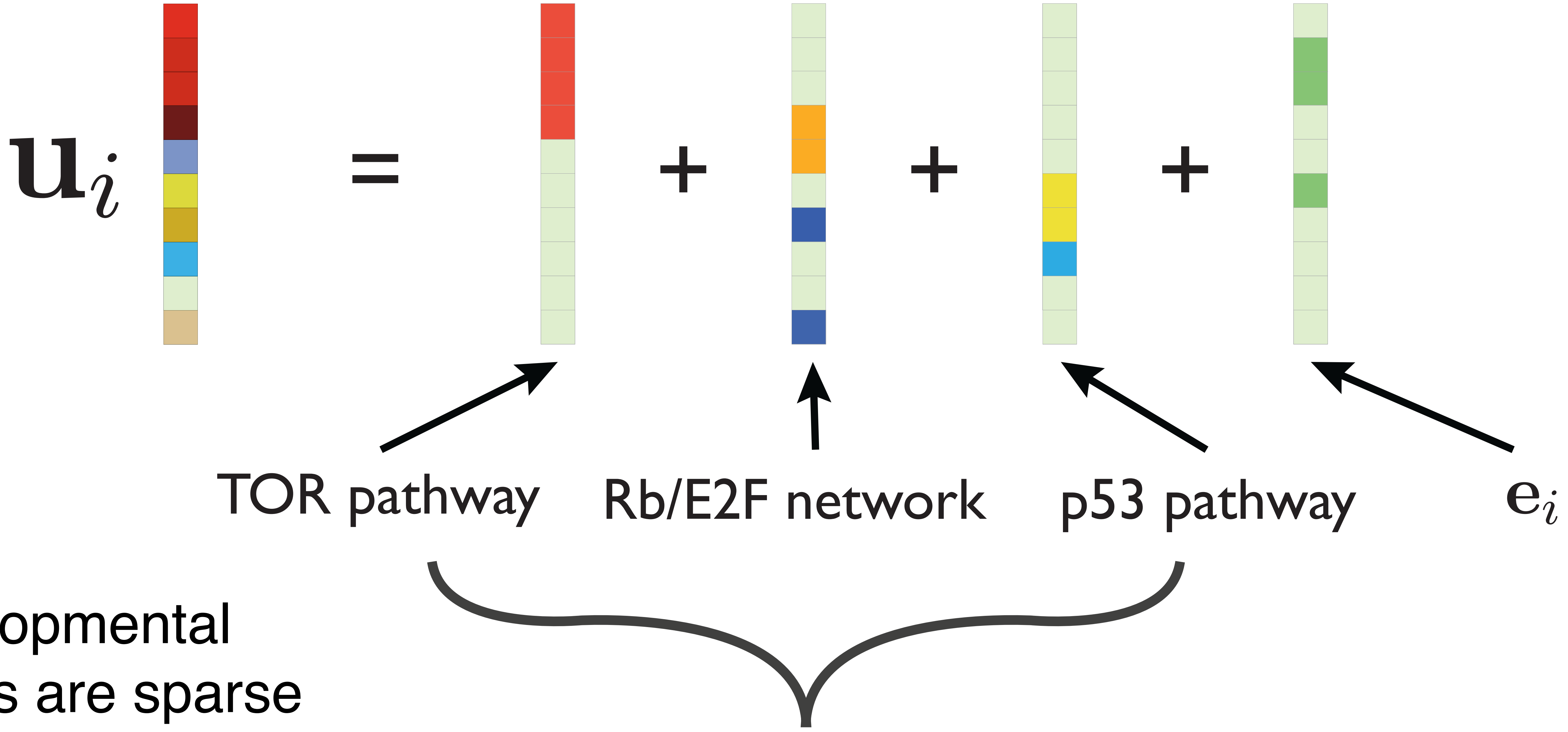
$\mathbf{e}_i$

Developmental effects are sparse

1) Few underlying developmental pathways are genetically variable

A factor model for G

Model **u** as output of development



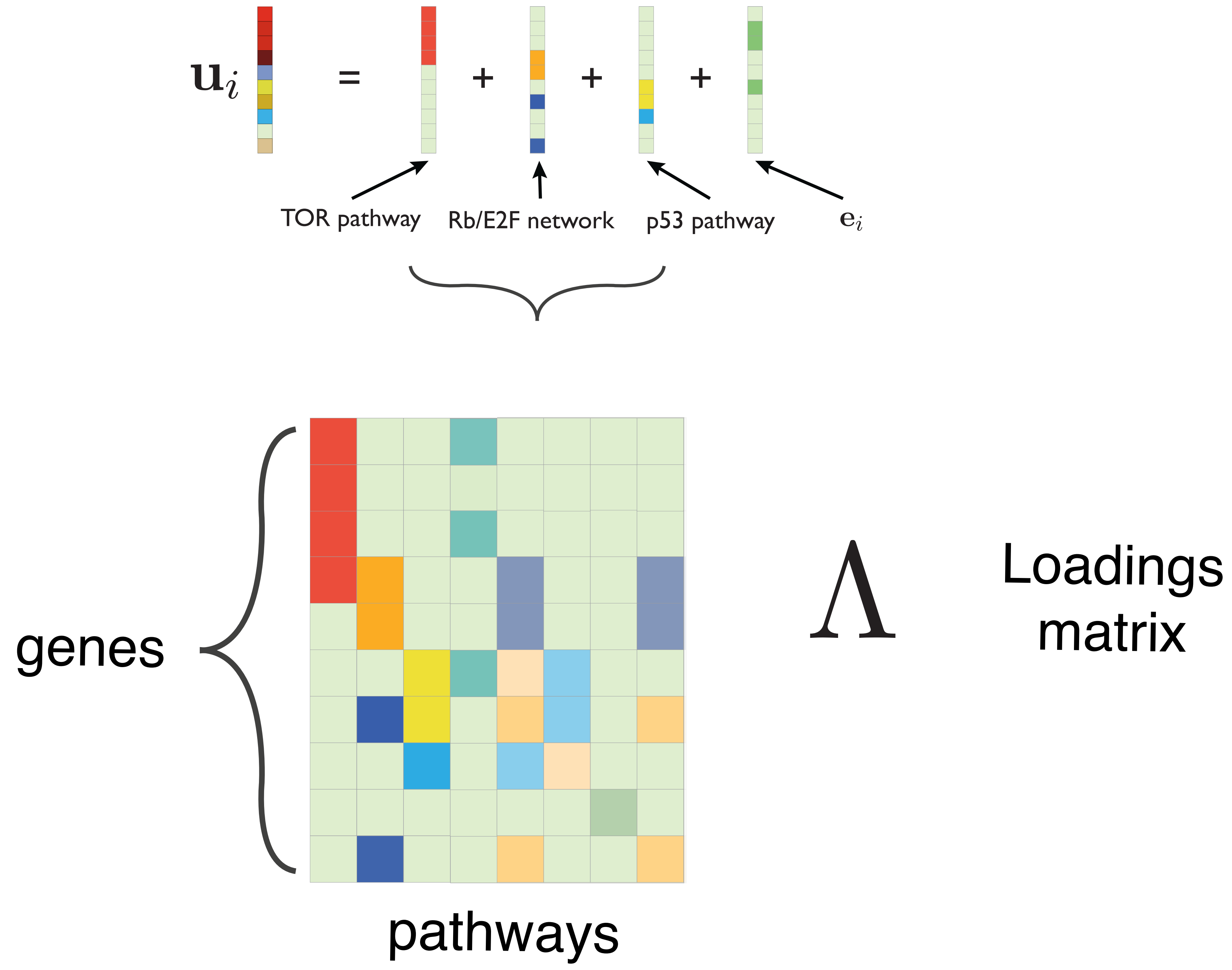
Developmental effects are sparse

1) Few underlying developmental pathways are genetically variable

2) Each pathway affects a low number of genes

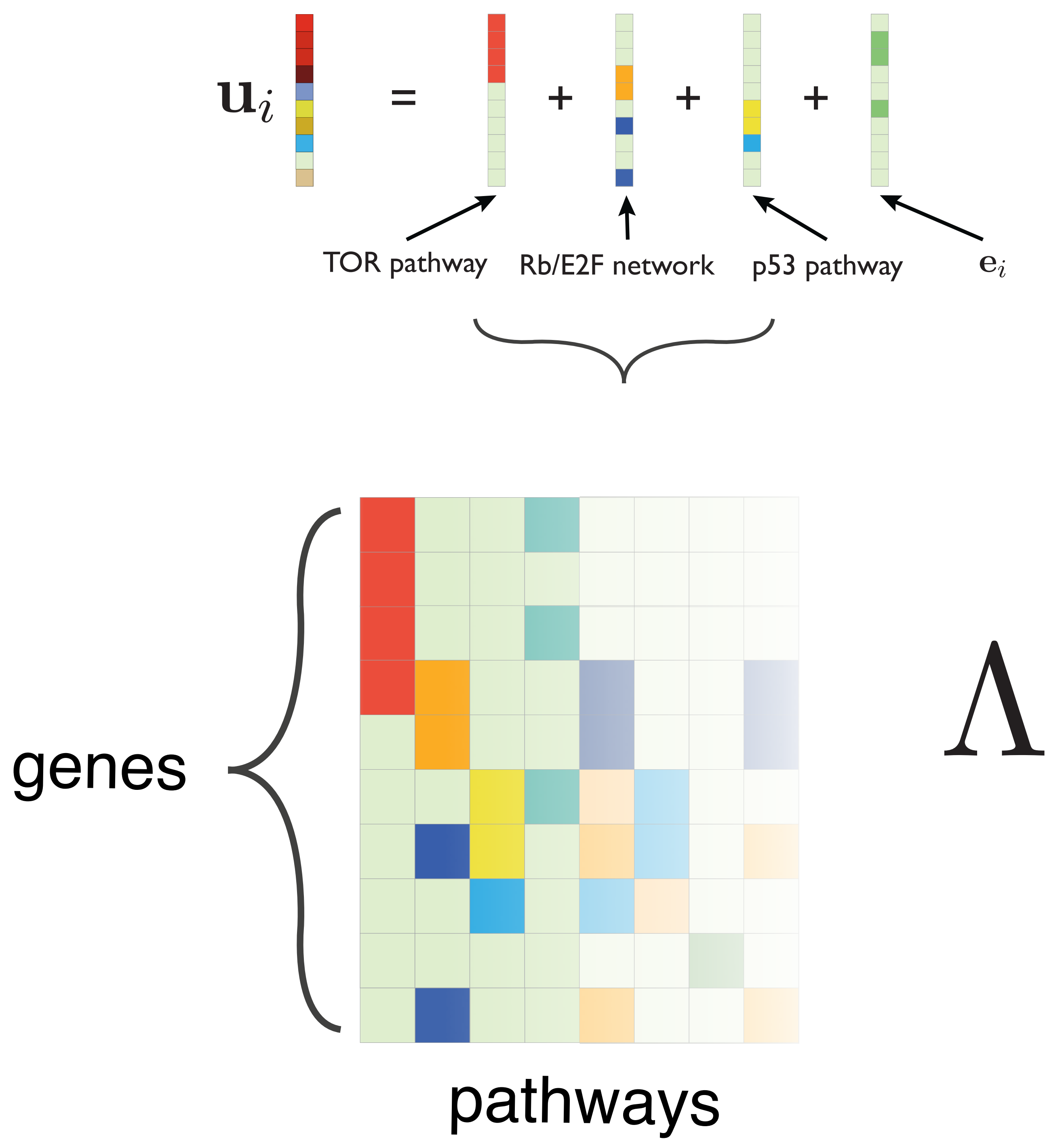
A factor model for G

Sparsity assumptions are key for high-dimensional data



A factor model for G

Sparsity assumptions are key for high-dimensional data

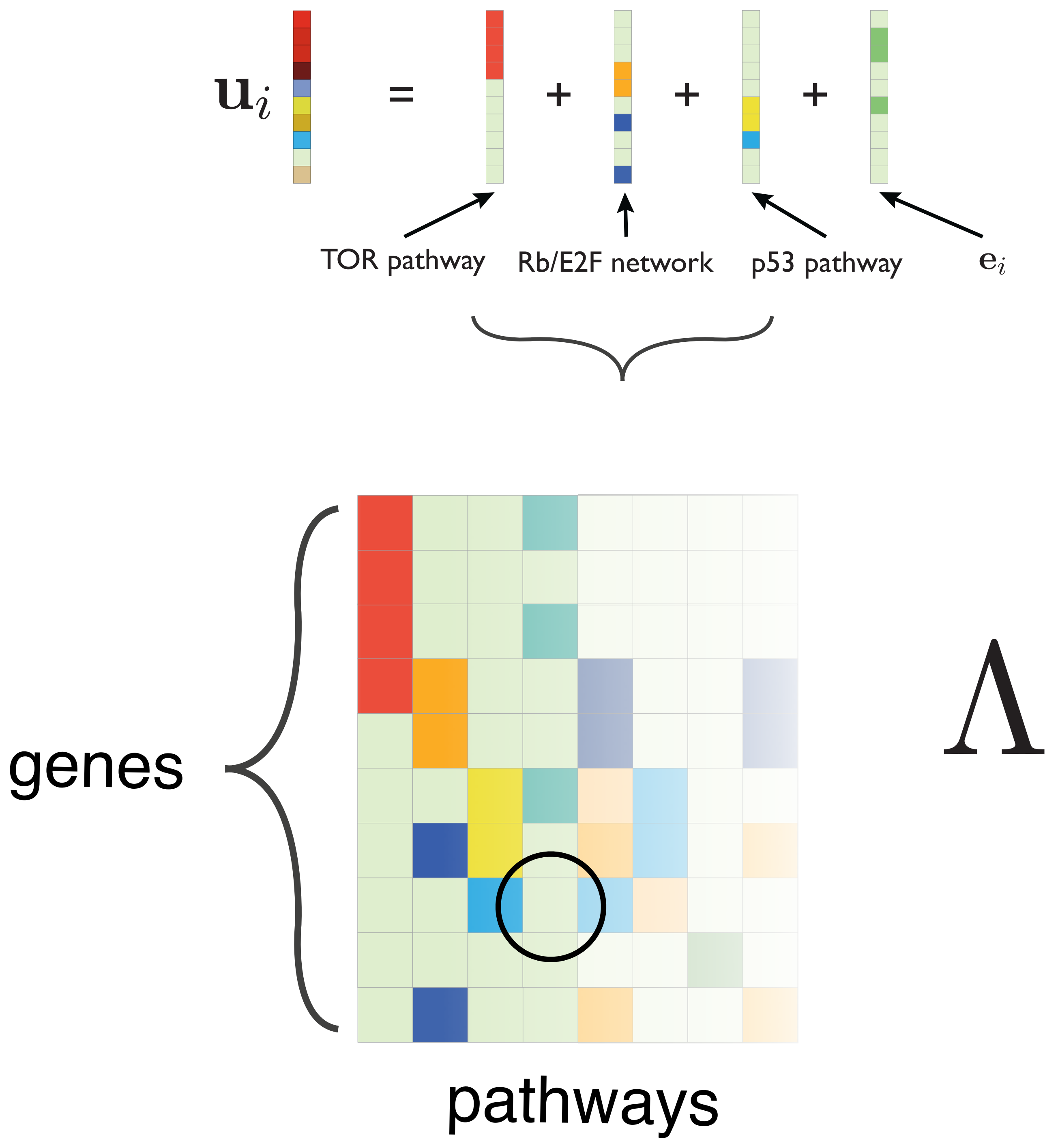


Loadings matrix

Few underlying pathways = few parameters to estimate

A factor model for G

Sparsity assumptions are key for high-dimensional data



Loadings matrix

Few underlying pathways = few parameters to estimate

Few effects per pathway = pathways are robust and interpretable

A factor model for G

genetic effects

measured traits

underlying traits

$$\mathbf{u}_i = \Lambda\,\mathbf{f}_i + \mathbf{e}_i$$

Loadings matrix

A factor model for G

genetic effects

measured traits

underlying traits

$$\mathbf{u}_i = \Lambda\,\mathbf{f}_i + \mathbf{e}_i$$

Loadings matrix

Residual covariance

$$\mathbf{G} = \Lambda\Lambda^T + \Sigma_{\mathbf{e}}$$

Genetic covariances

Bayesian genetic sparse factor model

Bayes' Theorem

Posterior Likelihood Prior

$$\underbrace{p(\mathbf{G} \mid \mathbf{Y})}_{\text{Posterior}} = \frac{\underbrace{p(\mathbf{Y} \mid \mathbf{G})}_{\text{Likelihood}}\underbrace{\pi(\mathbf{G})}_{\text{Prior}}}{p(\mathbf{Y})}$$

Bayesian genetic sparse factor model

Posterior Likelihood Prior

Bayes' Theorem

$$p(\mathbf{G} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathbf{G})\pi(\mathbf{G})}{p(\mathbf{Y})}$$

Animal model likelihood

$p(\mathbf{Y} \mid \mathbf{G})$

$$\mathbf{y}_i \sim \mathrm{N}\left(\mathbf{x}_i\mathbf{b} + \mathbf{u}_i, \mathbf{R}\right)$$

Bayesian genetic sparse factor model

Posterior Likelihood Prior

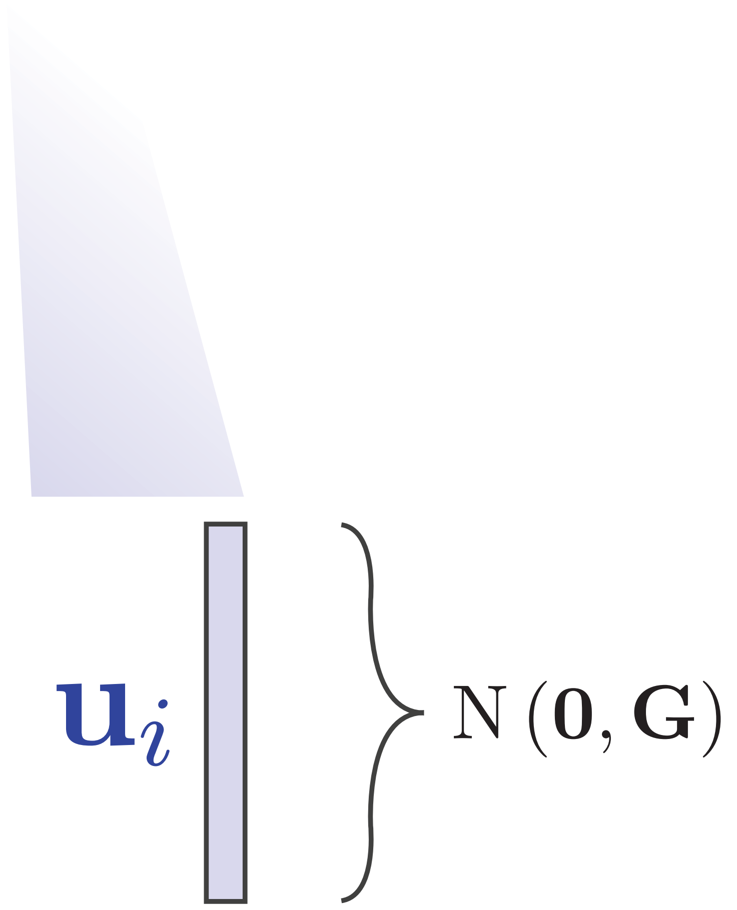
Bayes' Theorem

$$p(\mathbf{G} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathbf{G})\pi(\mathbf{G})}{p(\mathbf{Y})}$$

Animal model likelihood

$p(\mathbf{Y} \mid \mathbf{G})$

$$\mathbf{y}_i \sim \mathrm{N}\left(\mathbf{x}_i\mathbf{b} + \mathbf{u}_i, \mathbf{R}\right)$$



Bayesian genetic sparse factor model

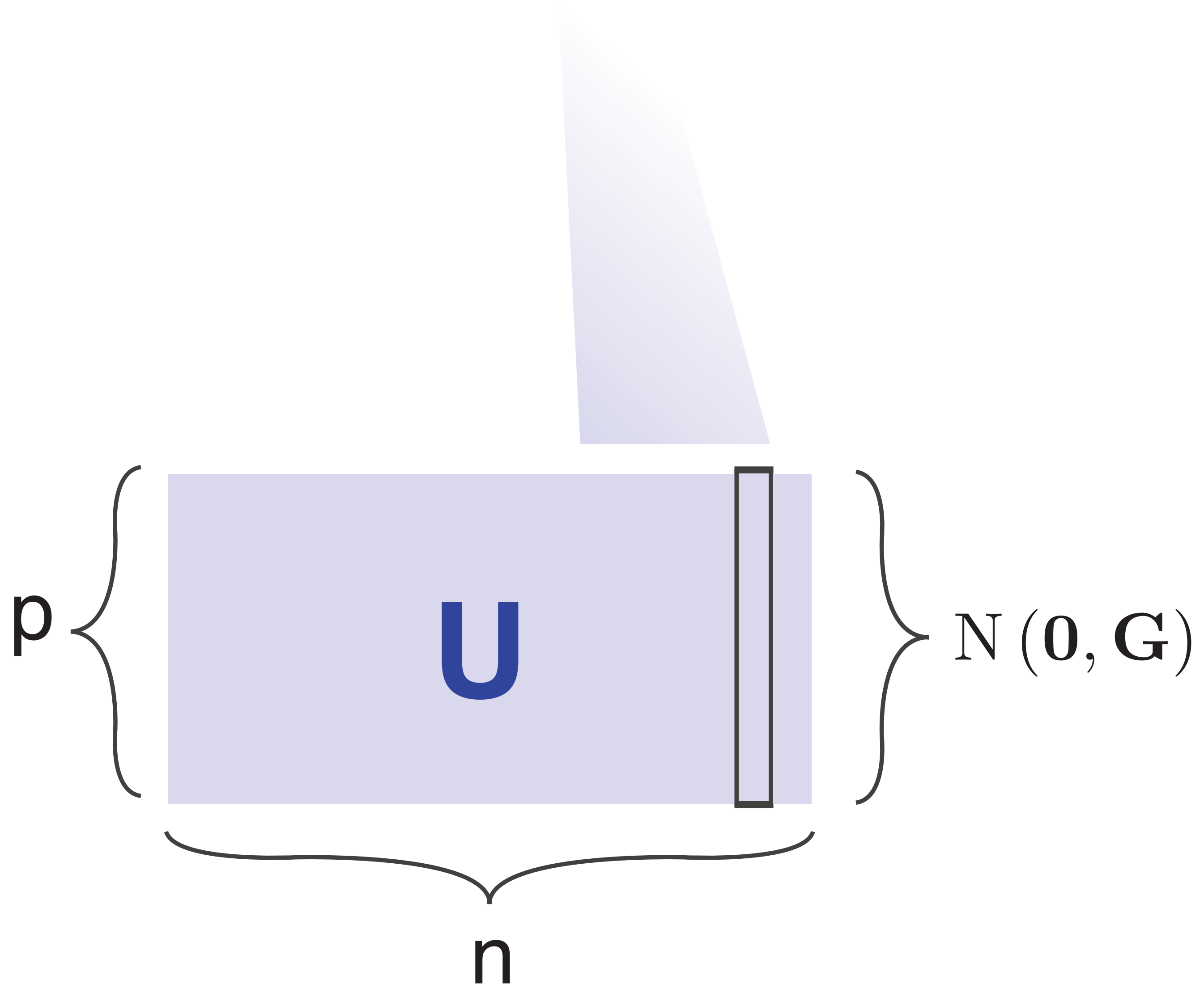
Posterior Likelihood Prior

Bayes' Theorem

$$p(\mathbf{G} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathbf{G})\pi(\mathbf{G})}{p(\mathbf{Y})}$$

Animal model likelihood

$$p(\mathbf{Y} \mid \mathbf{G}) \qquad \mathbf{y}_i \sim \mathrm{N}\left(\mathbf{x}_i\mathbf{b} + \mathbf{u}_i, \mathbf{R}\right)$$



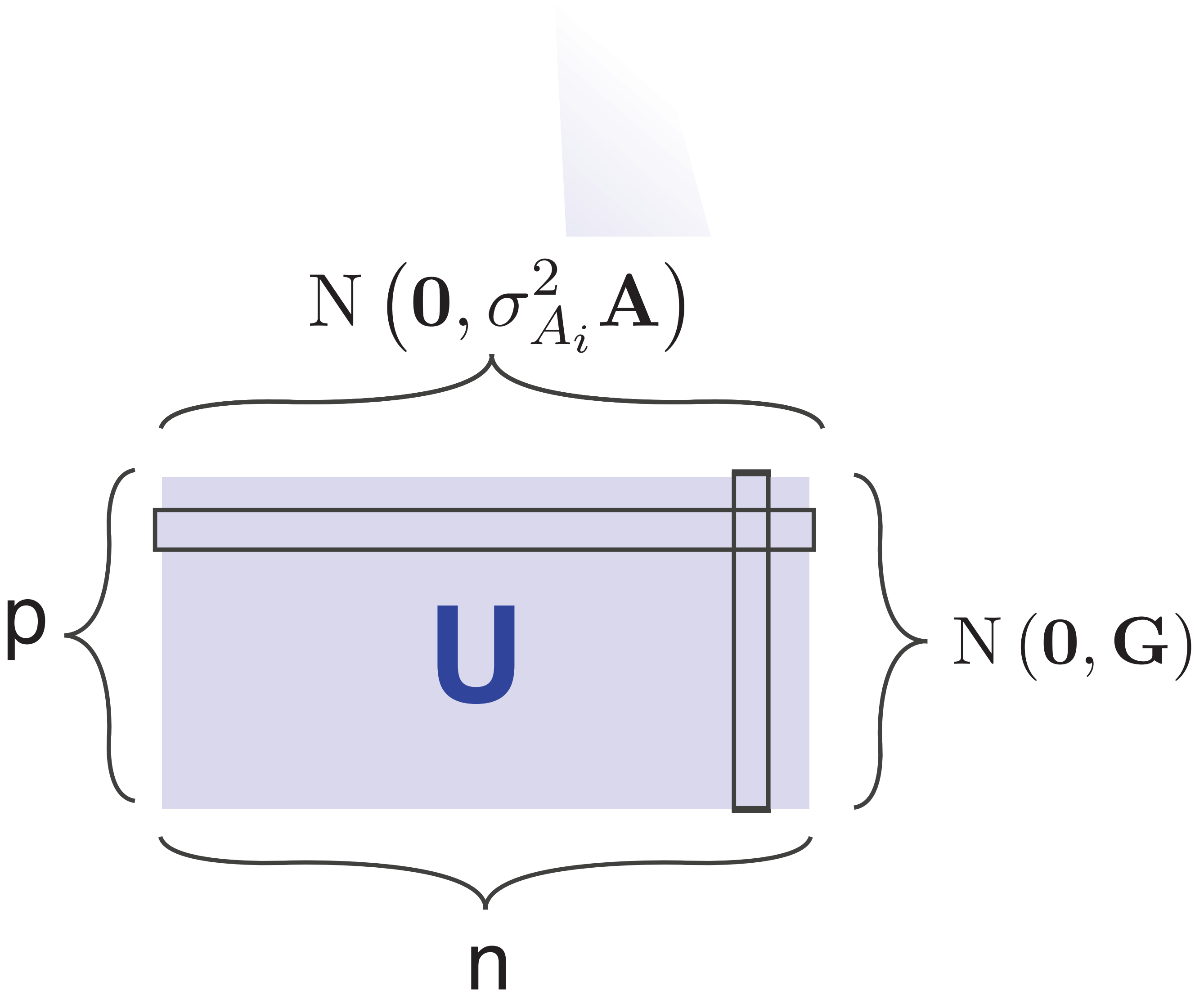
Bayesian genetic sparse factor model

Bayes' Theorem

$$\overset{\text{Posterior}}{p(\mathbf{G} \mid \mathbf{Y})} = \frac{\overset{\text{Likelihood}}{p(\mathbf{Y} \mid \mathbf{G})}\overset{\text{Prior}}{\pi(\mathbf{G})}}{p(\mathbf{Y})}$$

Animal model likelihood

$$p(\mathbf{Y} \mid \mathbf{G}) \qquad \mathbf{y}_i \sim \mathrm{N}\left(\mathbf{x}_i\mathbf{b} + \mathbf{u}_i, \mathbf{R}\right)$$


$\mathrm{N}\left(\mathbf{0}, \sigma^2_{A_i}\mathbf{A}\right)$
p
U
$\mathrm{N}\left(\mathbf{0}, \mathbf{G}\right)$
n

Bayesian genetic sparse factor model

Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

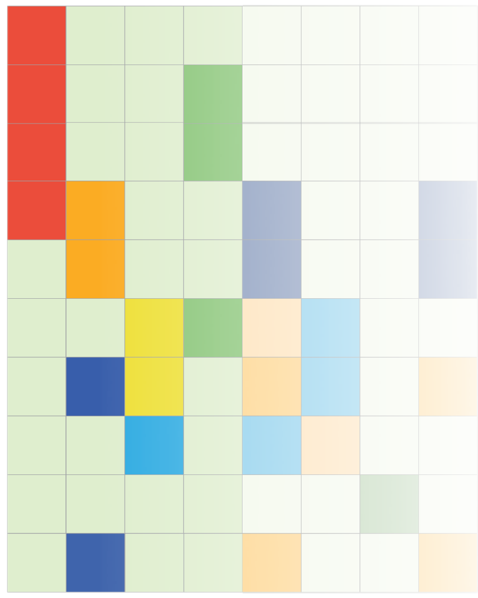
$$\pi(\mathbf{G}) \ = \pi\left(\Lambda\Lambda^T + \Sigma_{\mathbf{e}}\right)$$

Bayesian genetic sparse factor model

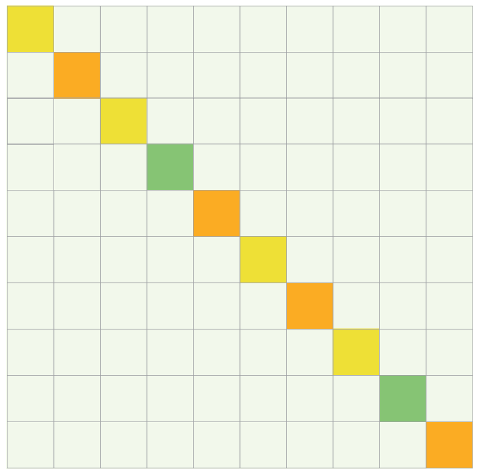
Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

$$\pi(\mathbf{G}) \; = \pi\left(\Lambda\Lambda^T + \Sigma_\mathbf{e}\right)$$

Λ



$\Sigma_\mathbf{e}$

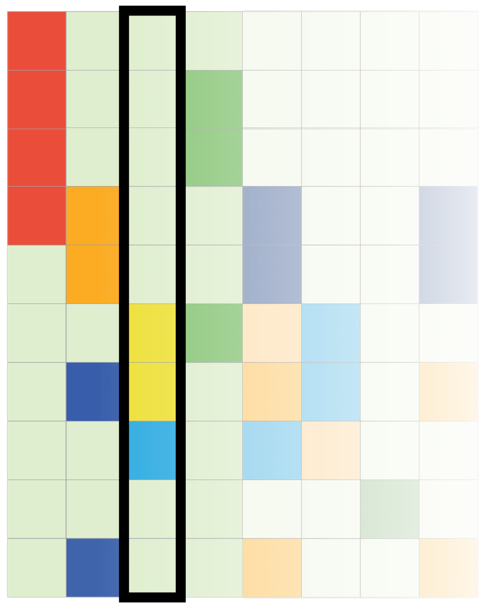


Bayesian genetic sparse factor model

Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

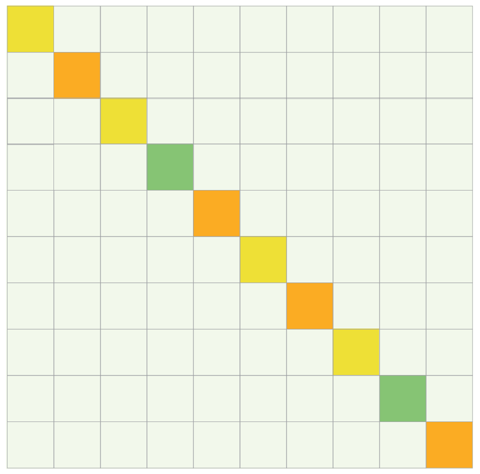
$$\pi(\mathbf{G}) \; = \pi\left(\Lambda\Lambda^T + \Sigma_\mathbf{e}\right)$$

Λ



factors

$\Sigma_\mathbf{e}$

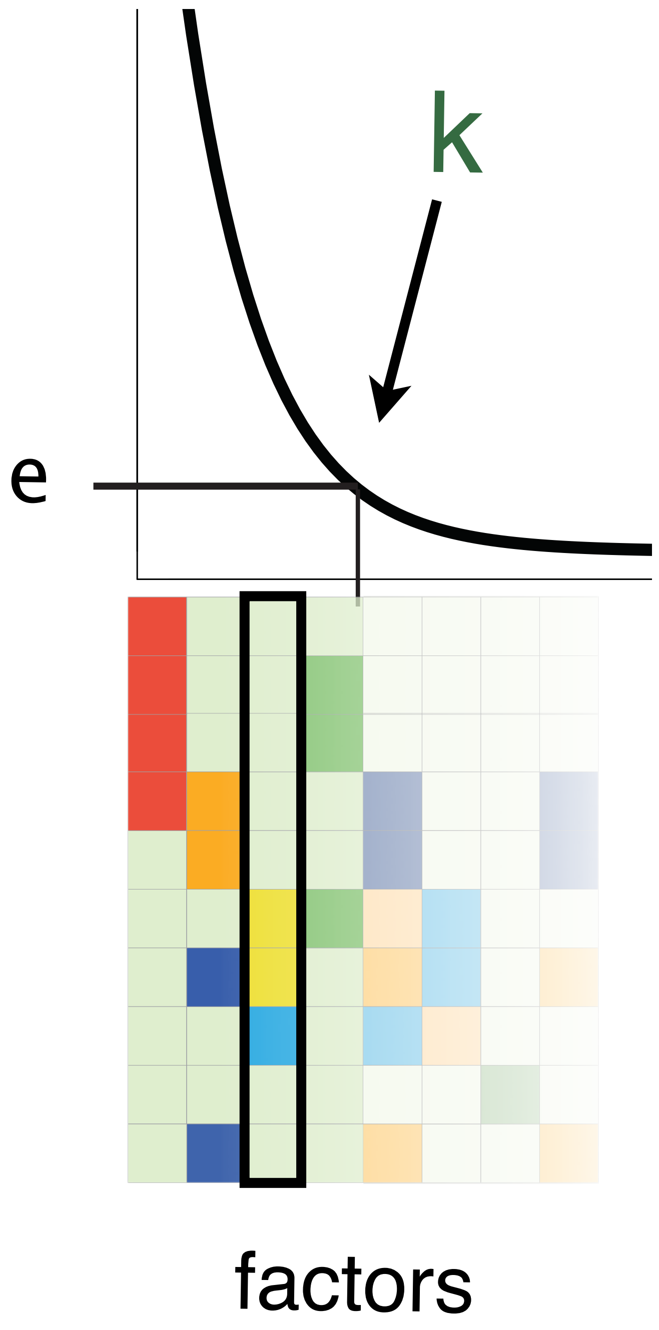


Bayesian genetic sparse factor model

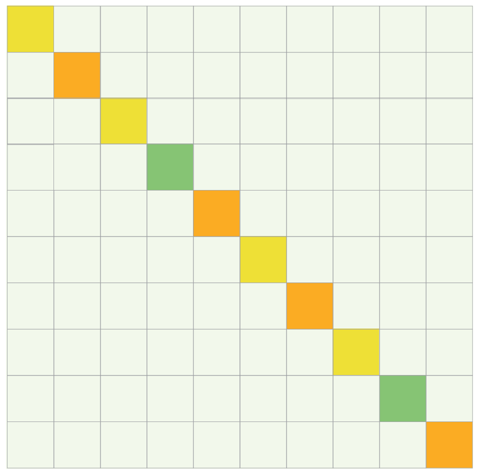
Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

$$\pi(\mathbf{G}) = \pi\left(\Lambda\Lambda^T + \Sigma_\mathbf{e}\right)$$

Λ



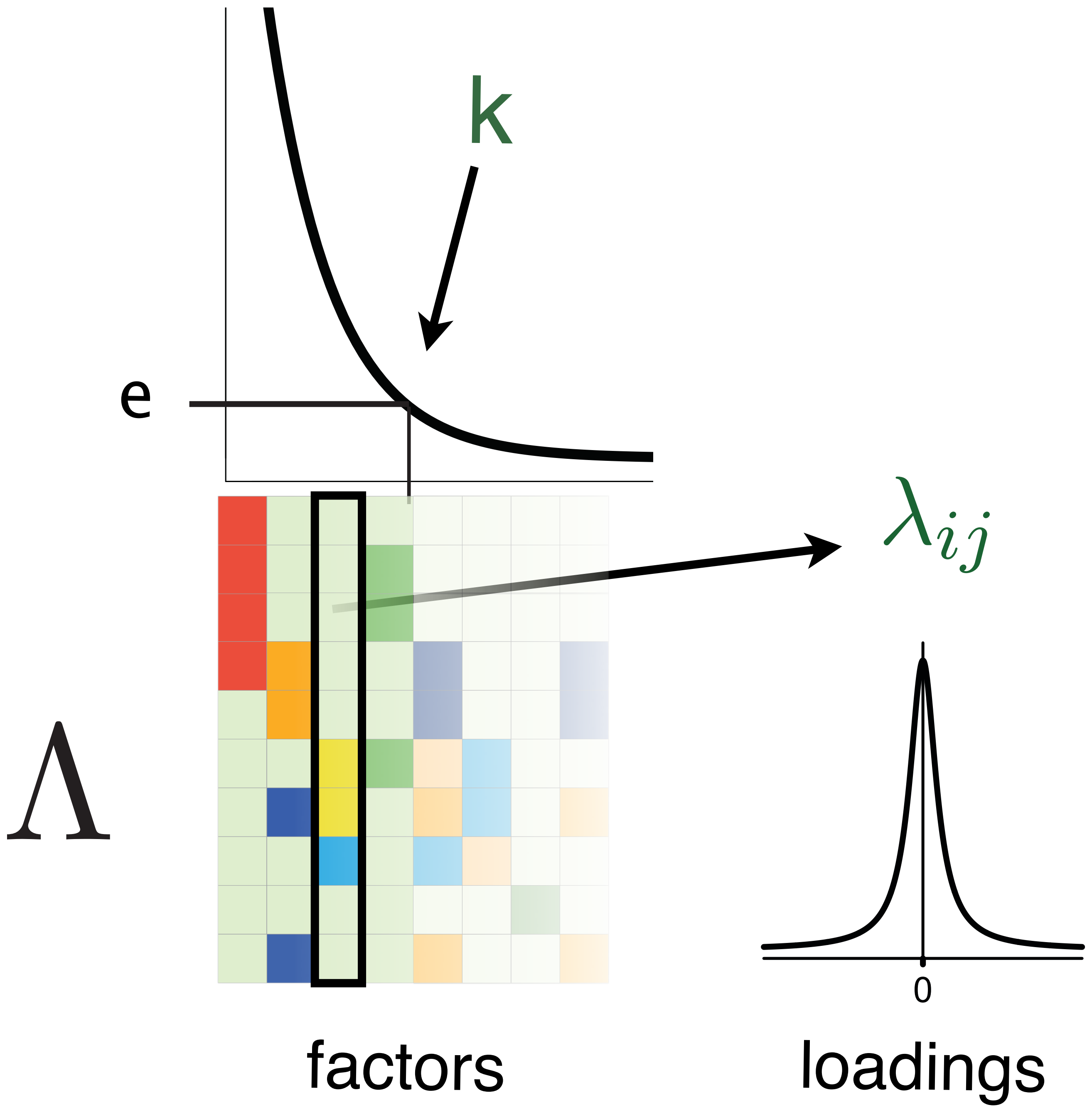
$\Sigma_\mathbf{e}$

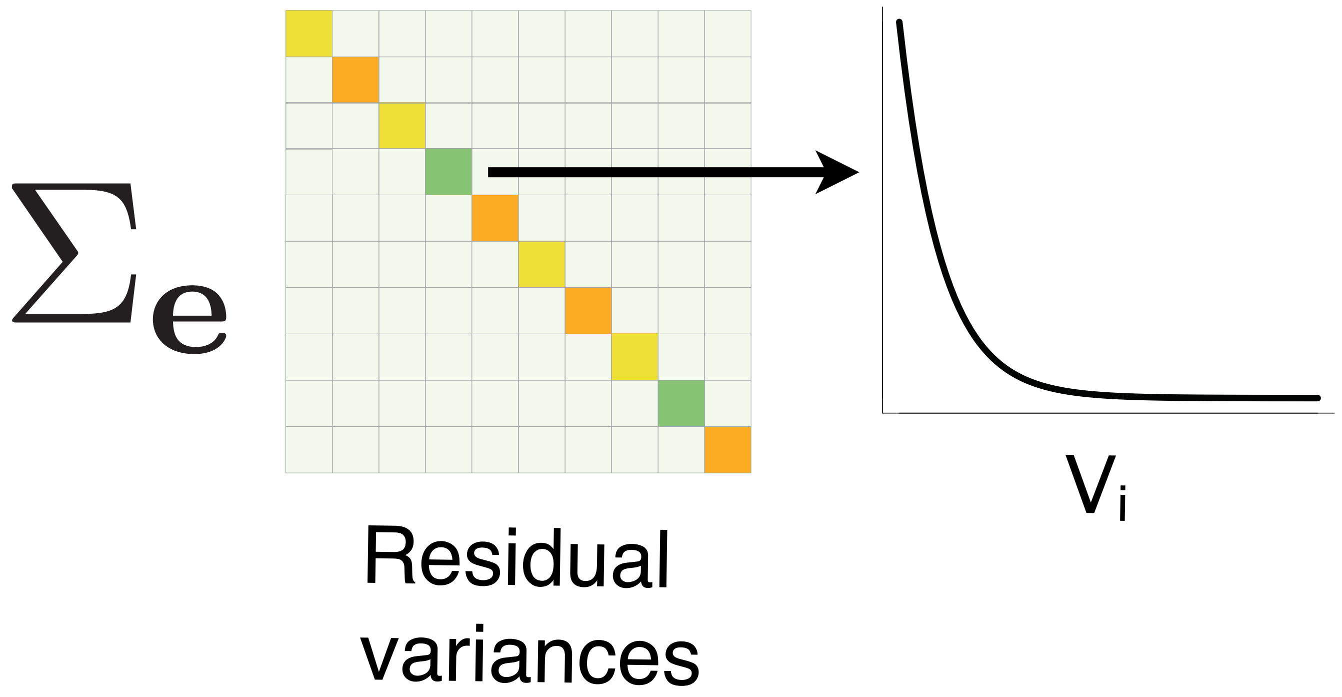


Bayesian genetic sparse factor model

Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

$$\pi(\mathbf{G}) = \pi\left(\Lambda\Lambda^T + \Sigma_{\mathbf{e}}\right)$$





Prior specification on $\mathbf{\Lambda}$

Based on (Bhattacharya and Dunson, 2011)

$$\begin{aligned}
\lambda_{im} \mid \phi_{im}, \tau_m &\sim \mathrm{N}\left(0, \phi_{im}^{-1}\tau_m^{-1}\right) \\
\phi_{im} &\sim \mathrm{Ga}(\nu/2, \nu/2), \\
\tau_m &= \prod_{\ell=1}^{m} \delta_\ell, \\
\delta_1 &\sim \mathrm{Ga}(a_1, b_1), \\
\delta_\ell &\sim \mathrm{Ga}(a_2, b_2) \text{ for } \ell = 2, ..., k.
\end{aligned}$$

Heritability prior (Zhou and Stephens, pers. comm.)

$$\pi(h_i^2 = \ell/n_h) = 1/n_h, \text{ where } \ell = 0 \ldots (n_h - 1).$$

Advantages

Scalable

- Can estimate **G** with $n << p$
- Adding genes doesn’t necessarily increase the number of factors
- More genes can actually improve the estimation of the factors

Advantages

Scalable

Can estimate **G** with n << p

Adding genes doesn't necessarily increase the number of factors

More genes can actually improve the estimation of the factors

Regularized

The sparsity prior on λ_{ij} provides shrinkage to reduce the impact of noise in the high dimensional space

Advantages

Scalable

Can estimate **G** with n << p

Adding genes doesn't necessarily increase the number of factors

More genes can actually improve the estimation of the factors

Regularized

The sparsity prior on λ_{ij} provides shrinkage to reduce the impact of noise in the high dimensional space

Interpretable

The latent factors can inform the *cause* of genetic correlations among genes

Advantages

Scalable

Can estimate **G** with n << p

Adding genes doesn't necessarily increase the number of factors

More genes can actually improve the estimation of the factors

Regularized

The sparsity prior on λ_{ij} provides shrinkage to reduce the impact of noise in the high dimensional space

Interpretable

The latent factors can inform the *cause* of genetic correlations among genes

Bayesian

Calculate posterior distributions of evolutionary parameters:

breeding values, heritability, genetic covariances, dimensionality of **G**

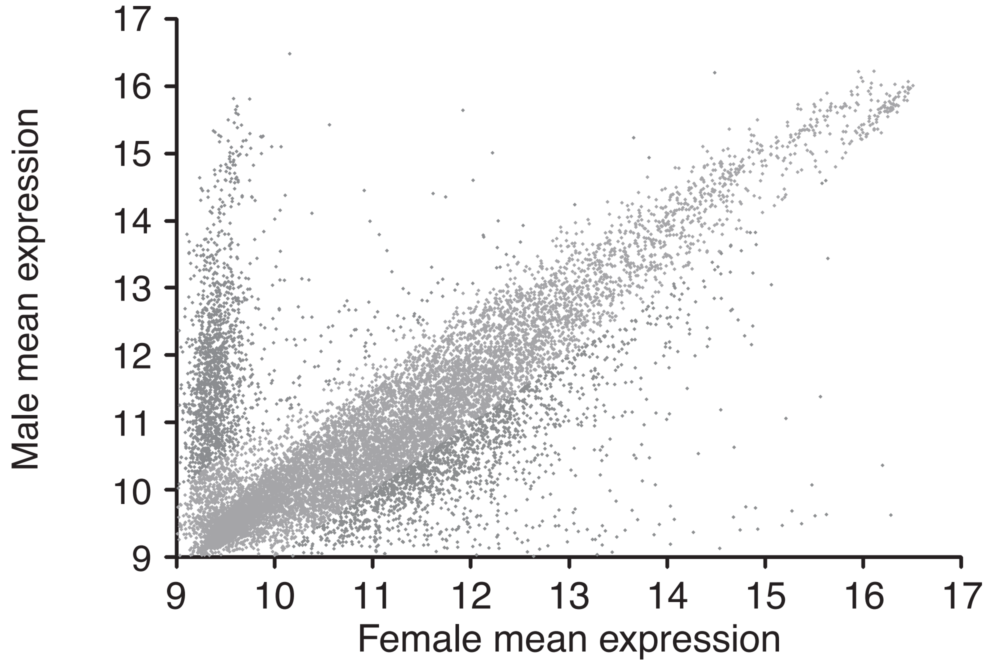
Case study: *Drosophila* gene expression

As a demonstration, we collected gene expression from:

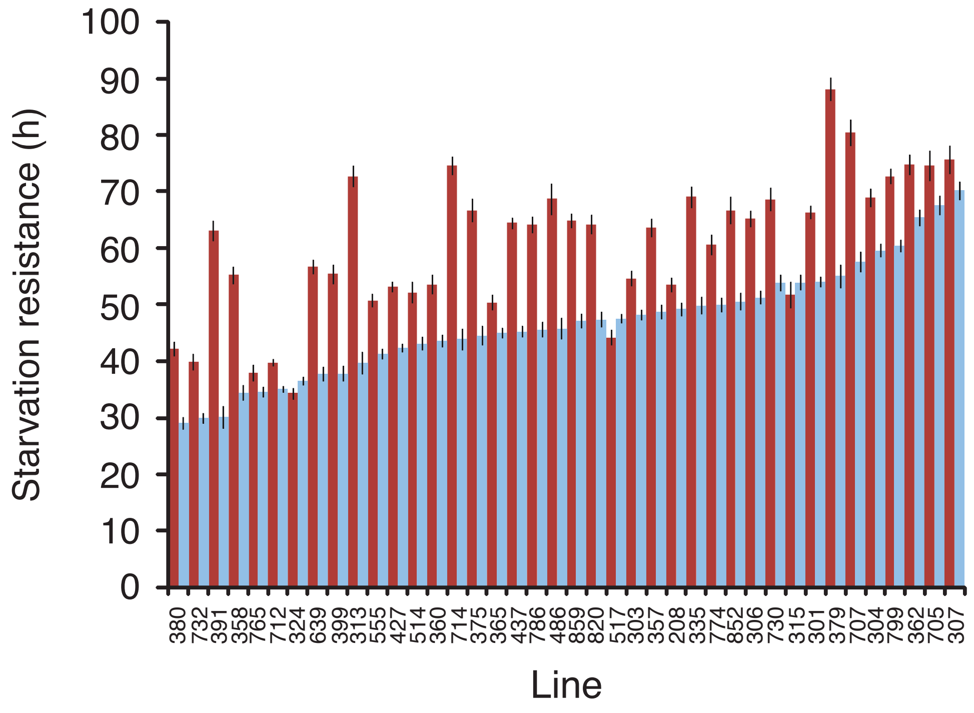
Ayroles et al (2009) Systems genetics of complex traits in *Drosophila melanogaster*. Nat Genet, 41, 299–307.

DGRP

40 lines of *D. melanogaster*

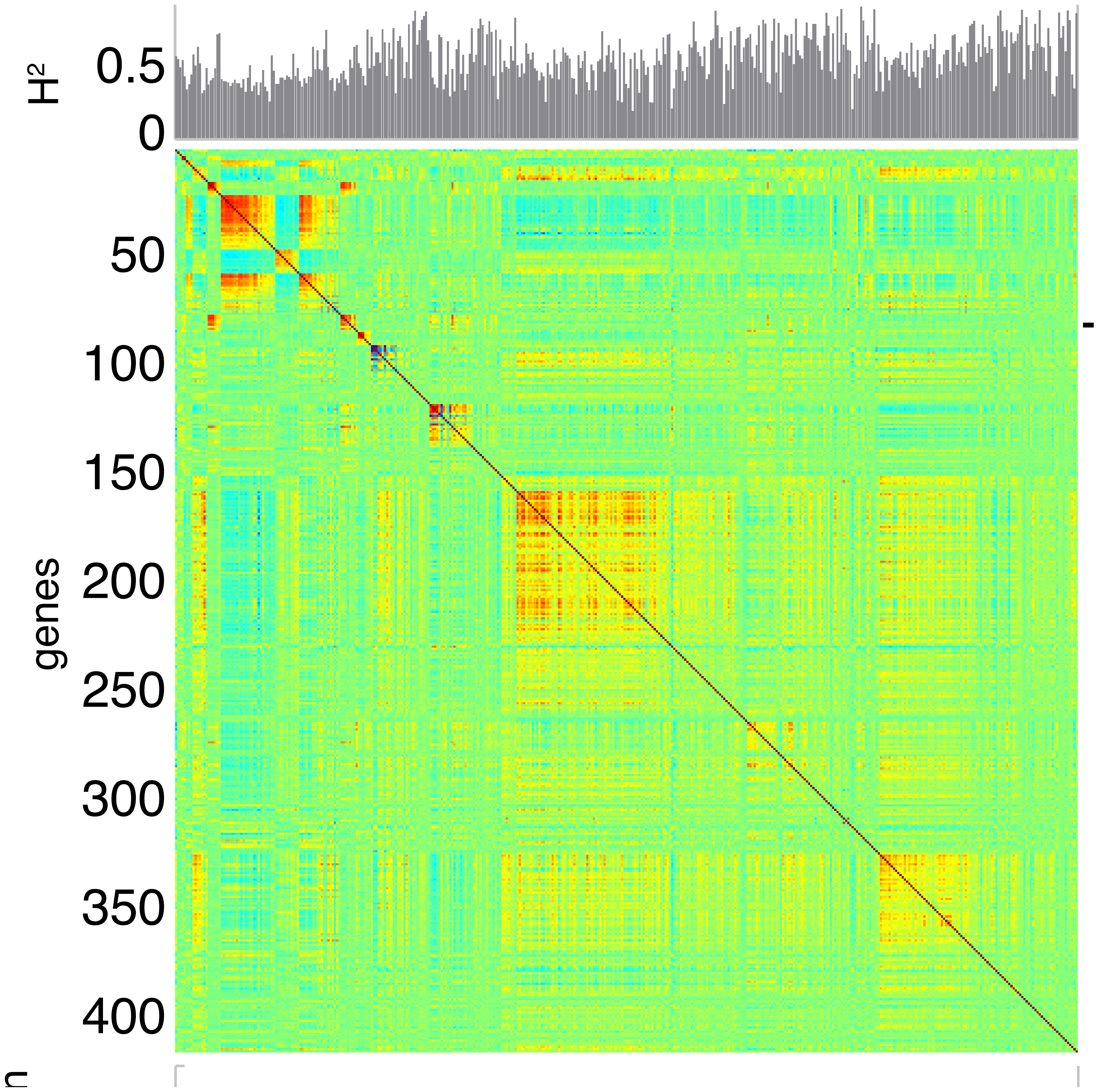


gene expression of >10,000 genes



Phenotype data on 7 fitness-related traits

Case study: *Drosophila* gene expression



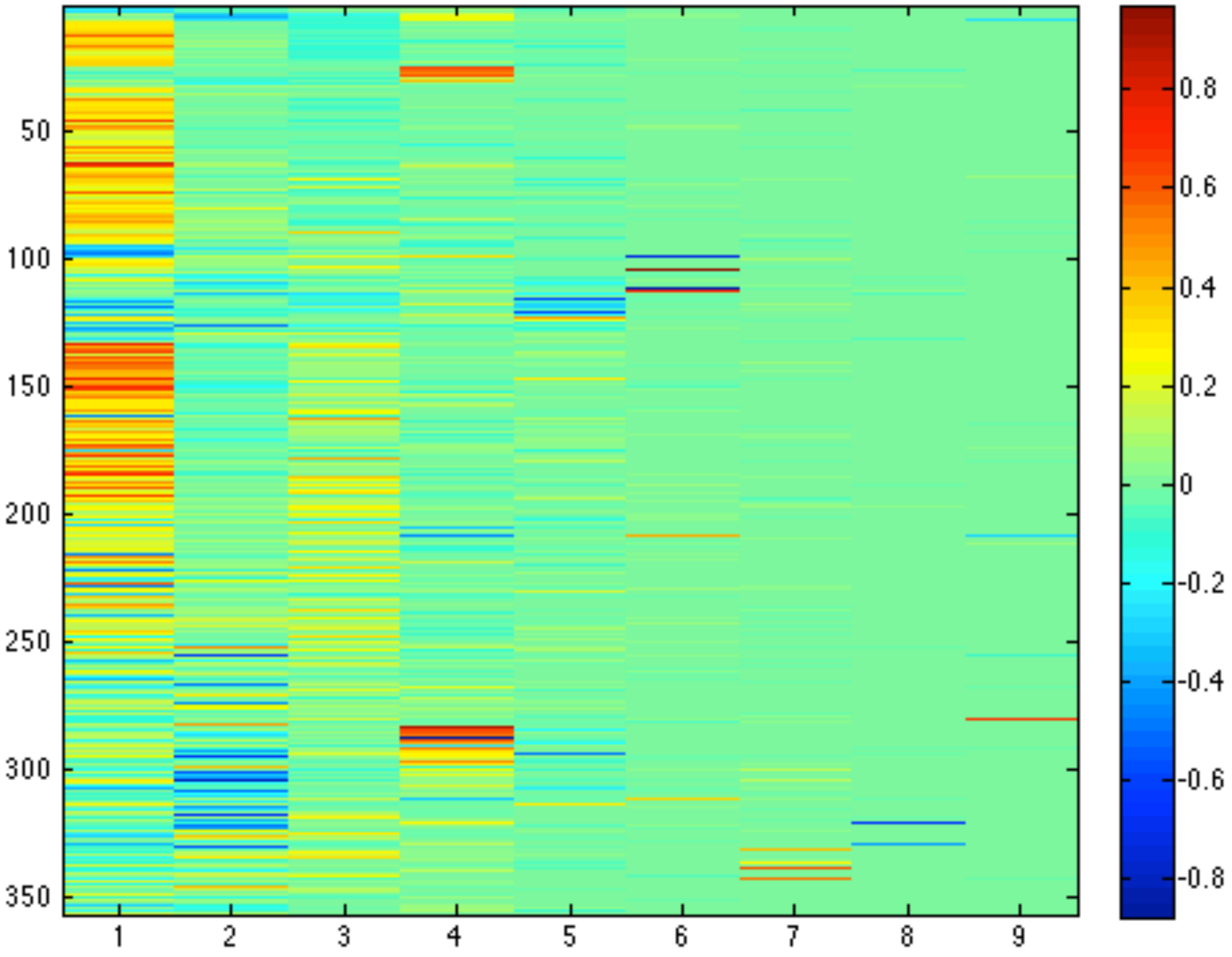
Case study: *Drosophila* gene expression

We estimate that the genetic covariation in expression could be explained by 9 factors

Factor 1 is dense but the remainder are very sparse.

Λ

Loadings matrix



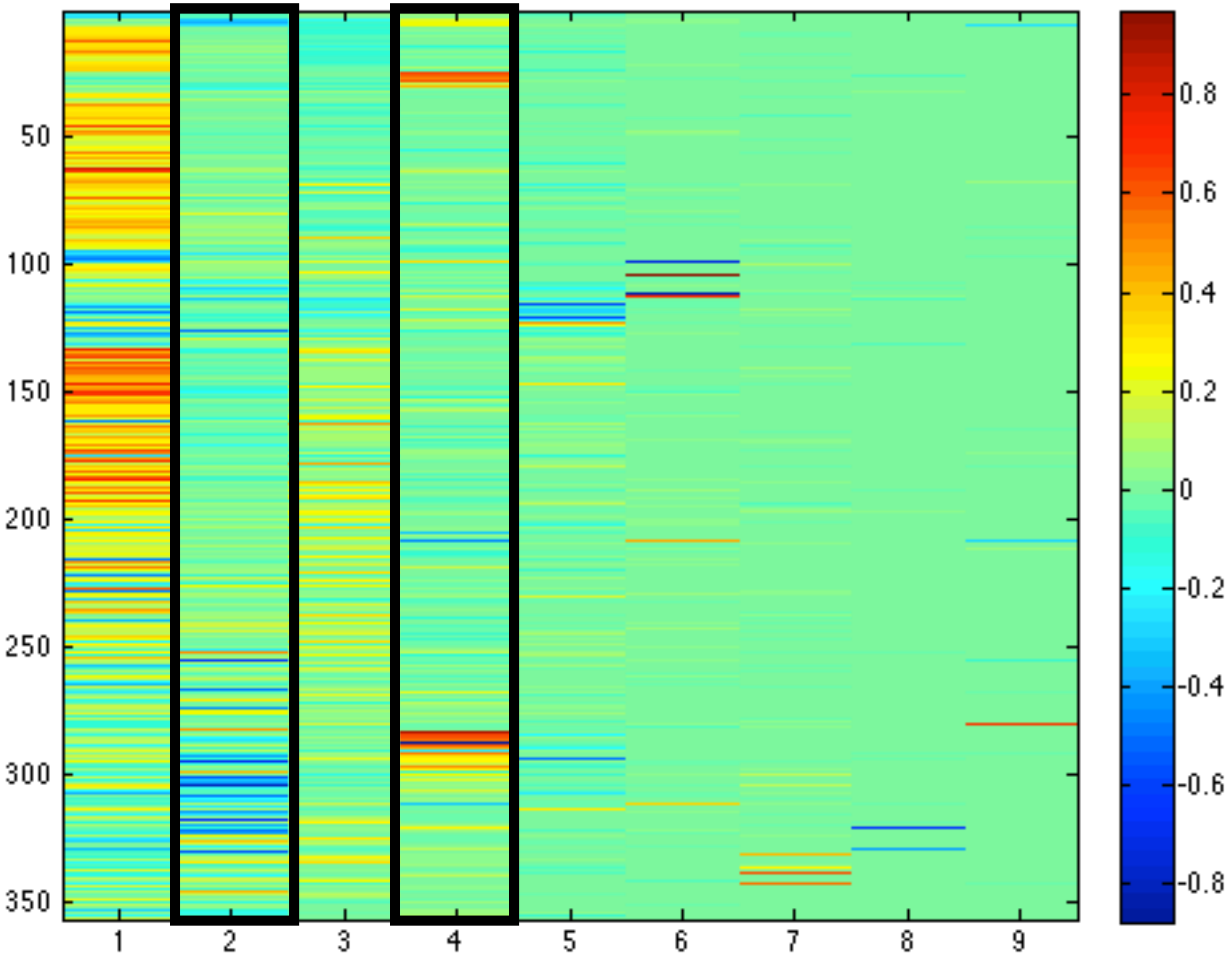
Case study: *Drosophila* gene expression

We estimate that the genetic covariation in expression could be explained by 9 factors

Factor 1 is dense but the remainder are very sparse.

Λ

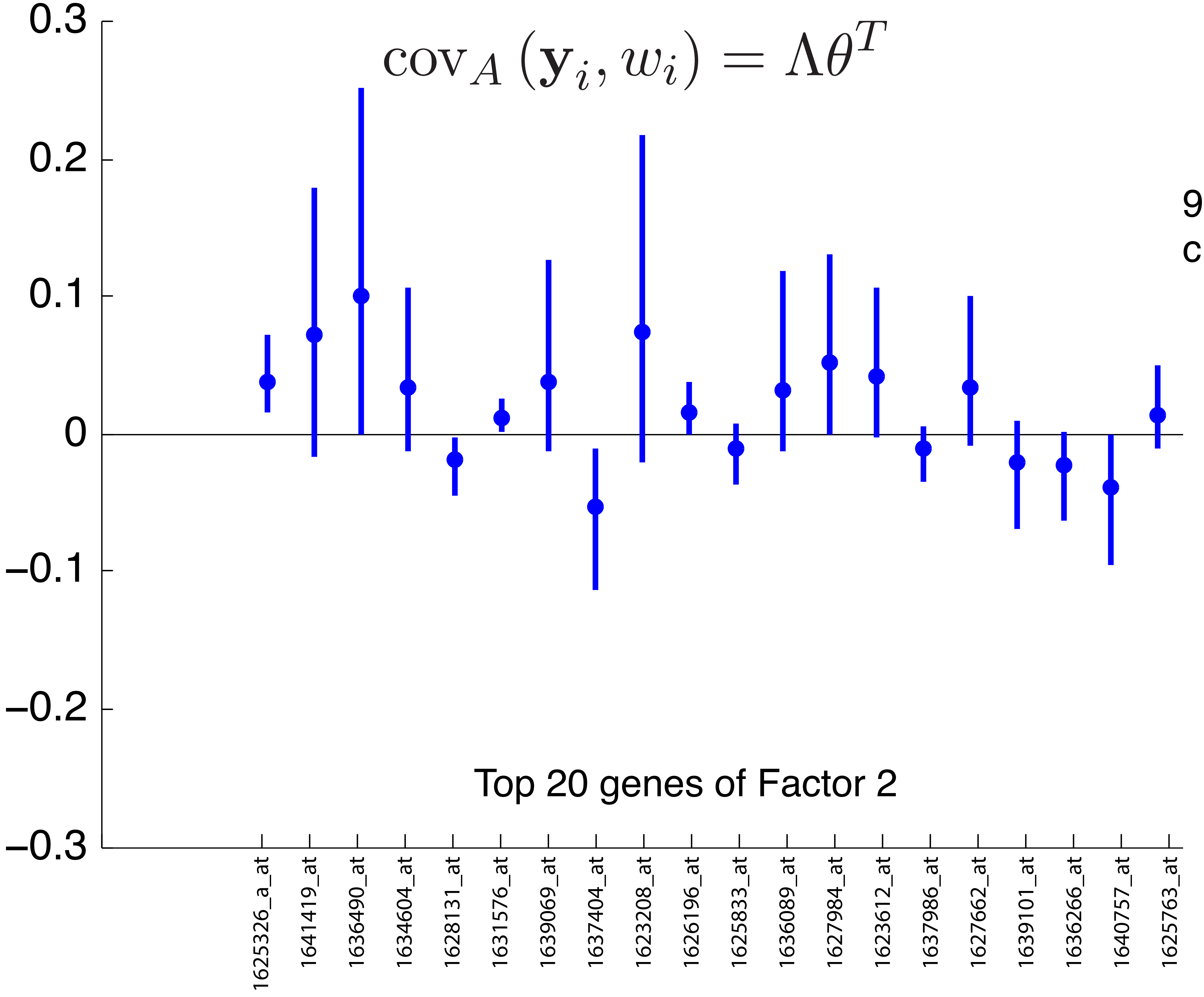
Loadings matrix



Genes related to defense and immune responses

Case study: *Drosophila* gene expression

We can measure genetic covariances with Starvation Resistance



$$\mathrm{cov}_A\left(\mathbf{y}_i, w_i\right) = \Lambda\theta^T$$

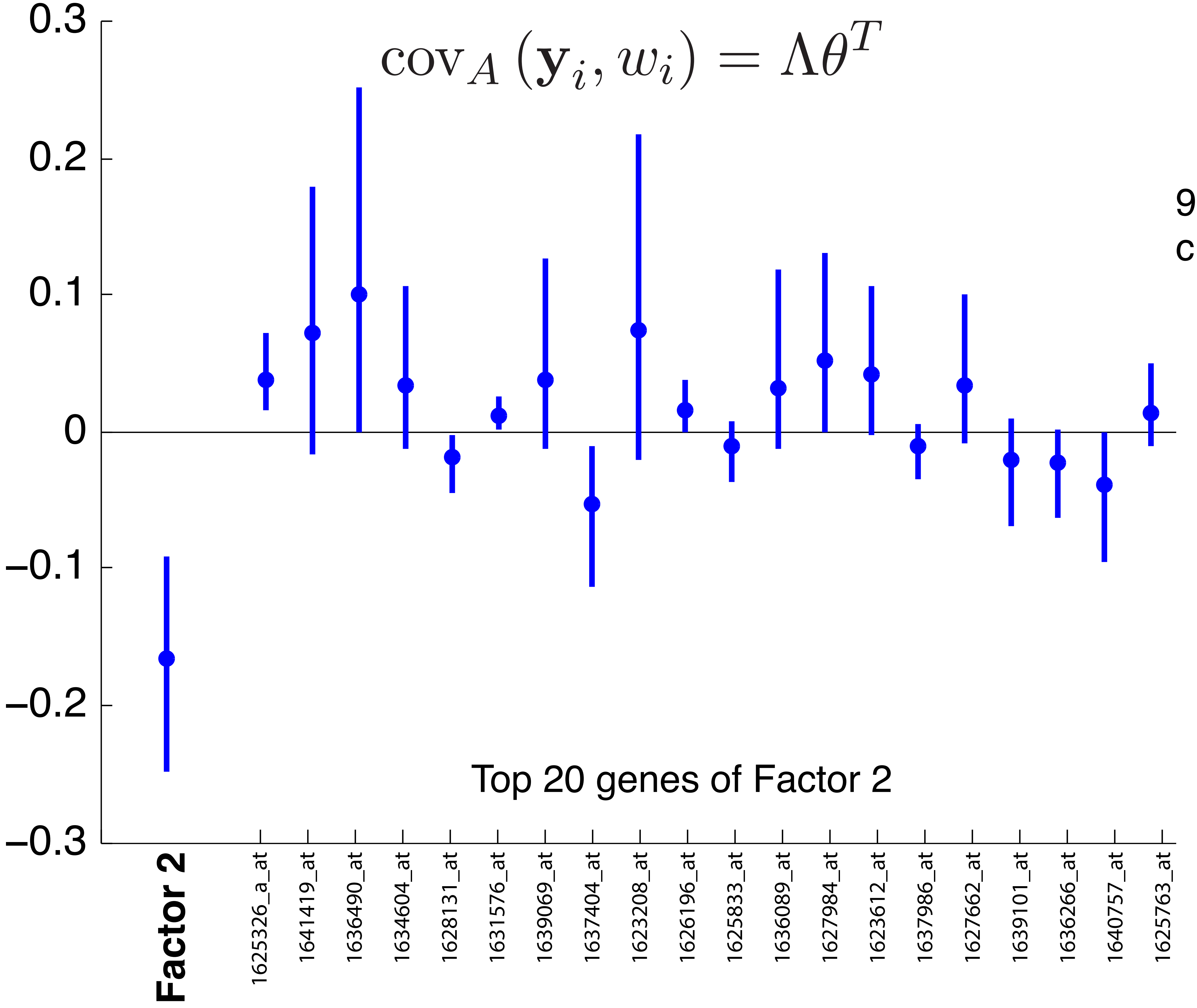
Top 20 genes of Factor 2

95% Posterior
credible intervals

Case study: *Drosophila* gene expression

We can measure genetic covariances with Starvation Resistance

But have more power to identify covariances with underlying traits

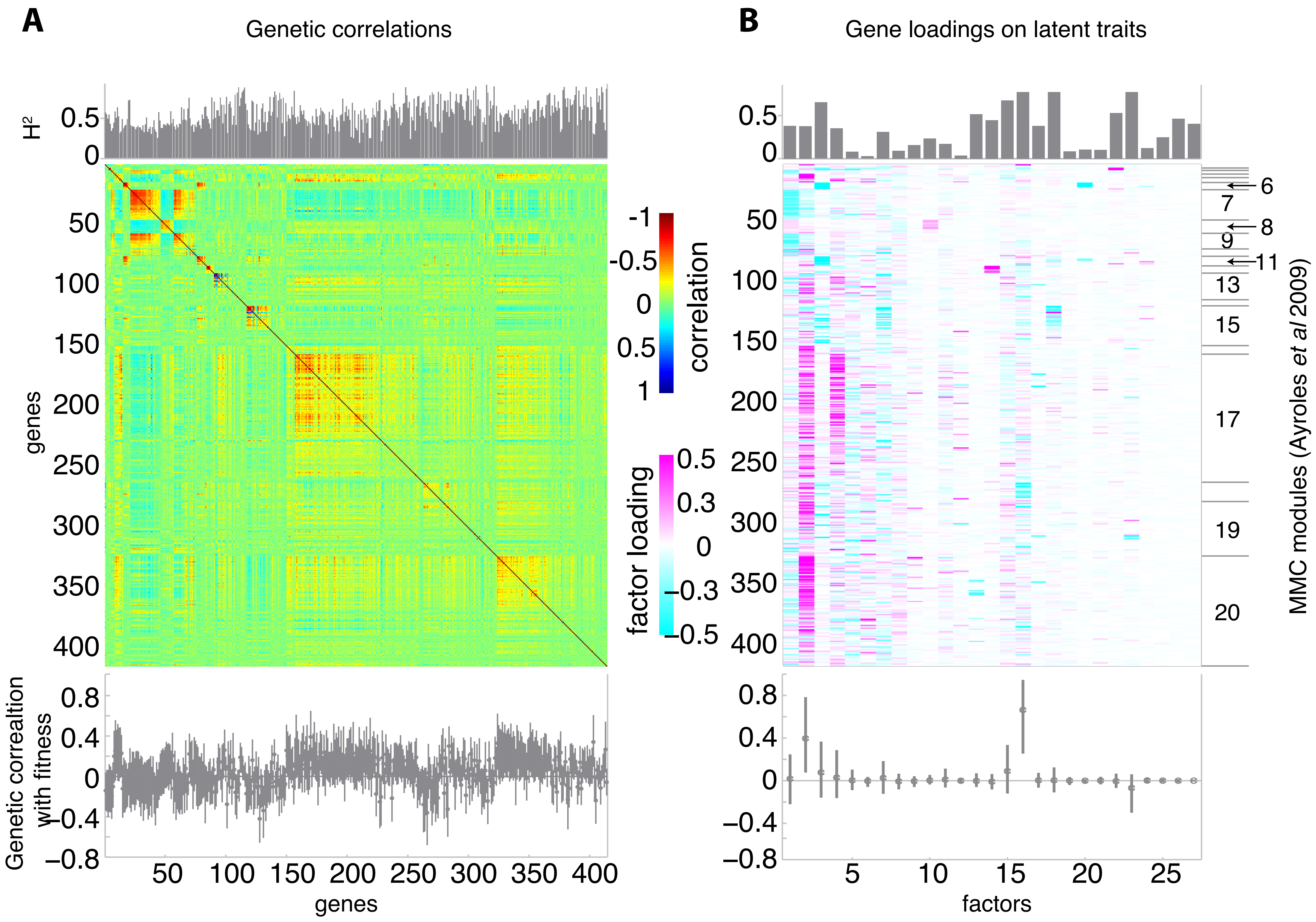


$$\mathrm{cov}_A\left(\mathbf{y}_i, w_i\right) = \Lambda\theta^T$$

95% Posterior credible intervals

Top 20 genes of Factor 2

Drosophila results



Software

Software:

`http://www.stat.duke.edu/~sayan/bfgr/index.shtml`

Extensions and open problems

(1) Simultaneous inference of **G** and kinship matrix.

Extensions and open problems

(1) Simultaneous inference of **G** and kinship matrix.

(2) Local heritability.

Extensions and open problems

(1) Simultaneous inference of **G** and kinship matrix.

(2) Local heritability.

(3) Incorporation with GWAS.

Extensions and open problems

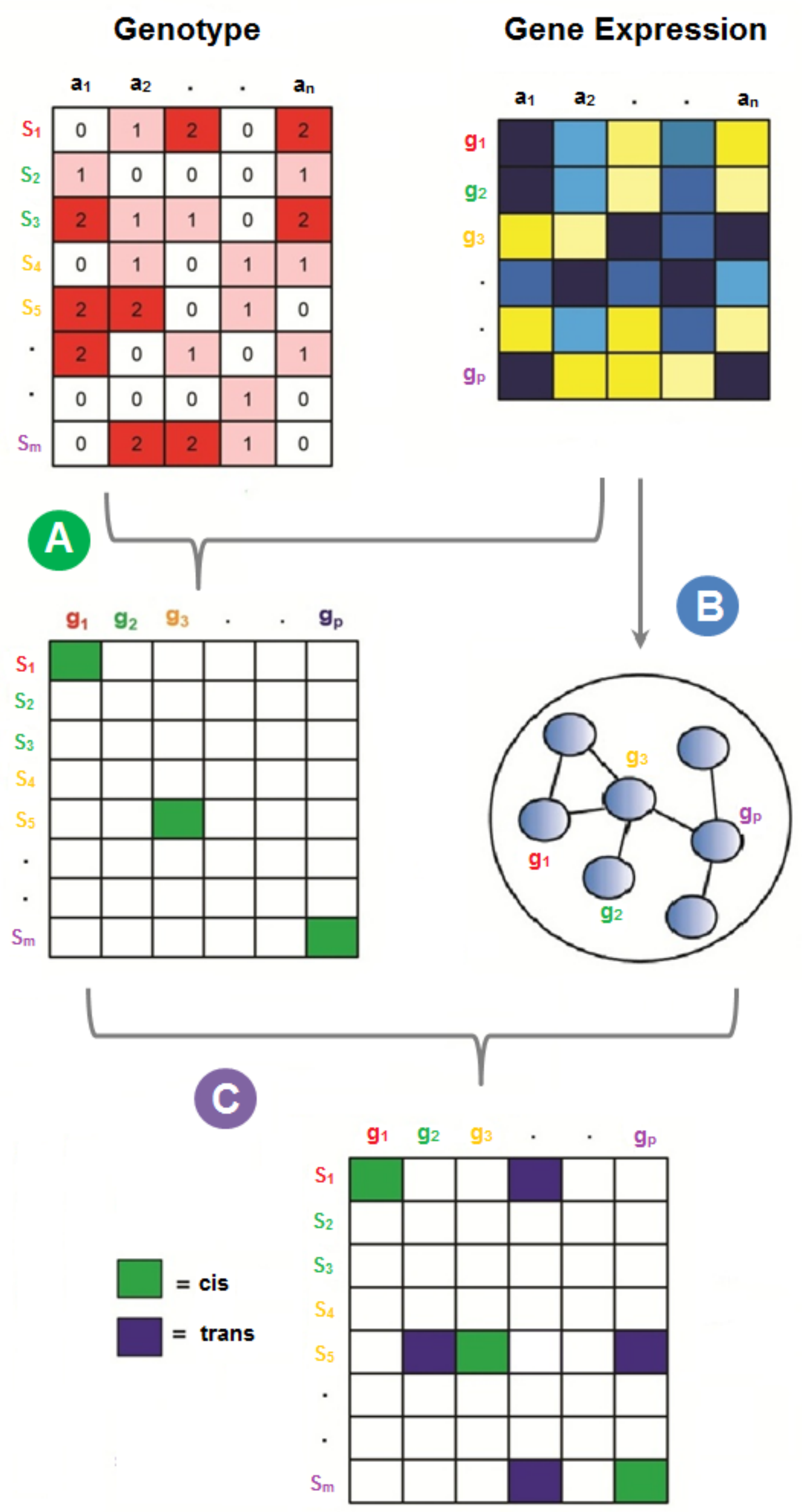
(1) Simultaneous inference of **G** and kinship matrix.

(2) Local heritability.

(3) Incorporation with GWAS.

(4) Discrete traits and time varying traits.

Network-based, Large-scale Identification oF disTal eQTL (NetLIFT)



Genotype

	a_1	a_2	.	.	a_n
S_1	0	1	2	0	2
S_2	1	0	0	0	1
S_3	2	1	1	0	2
S_4	0	1	0	1	1
S_5	2	2	0	1	0
.	2	0	1	0	1
.	0	0	0	1	0
S_m	0	2	2	1	0

Gene Expression

	a_1	a_2	.	.	a_n
g_1					
g_2					
g_3					
.					
.					
g_p					

A

	g_1	g_2	g_3	.	.	g_p
S_1	🟩					
S_2						
S_3						
S_4						
S_5			🟩			
.						
.						
S_m						🟩

B

C

	g_1	g_2	g_3	.	.	g_p
S_1	🟩			🟪		
S_2						
S_3						
S_4						
S_5		🟪	🟩			🟪
.						
.						
S_m				🟪		🟩

🟩 = cis

🟪 = trans

Objective

Dissect genetic and molecular mechanism underlying complex (disease) traits.

Objective

Dissect genetic and molecular mechanism underlying complex (disease) traits.
Standard approaches:

(1) Genome wide association studies (GWAS): Correlations between genetic variants and trait variation.

Objective

Dissect genetic and molecular mechanism underlying complex (disease) traits.
Standard approaches:

(1) Genome wide association studies (GWAS): Correlations between genetic variants and trait variation.

(2) Gene expression studies: correlations between gene expression and trait variation.

Objective

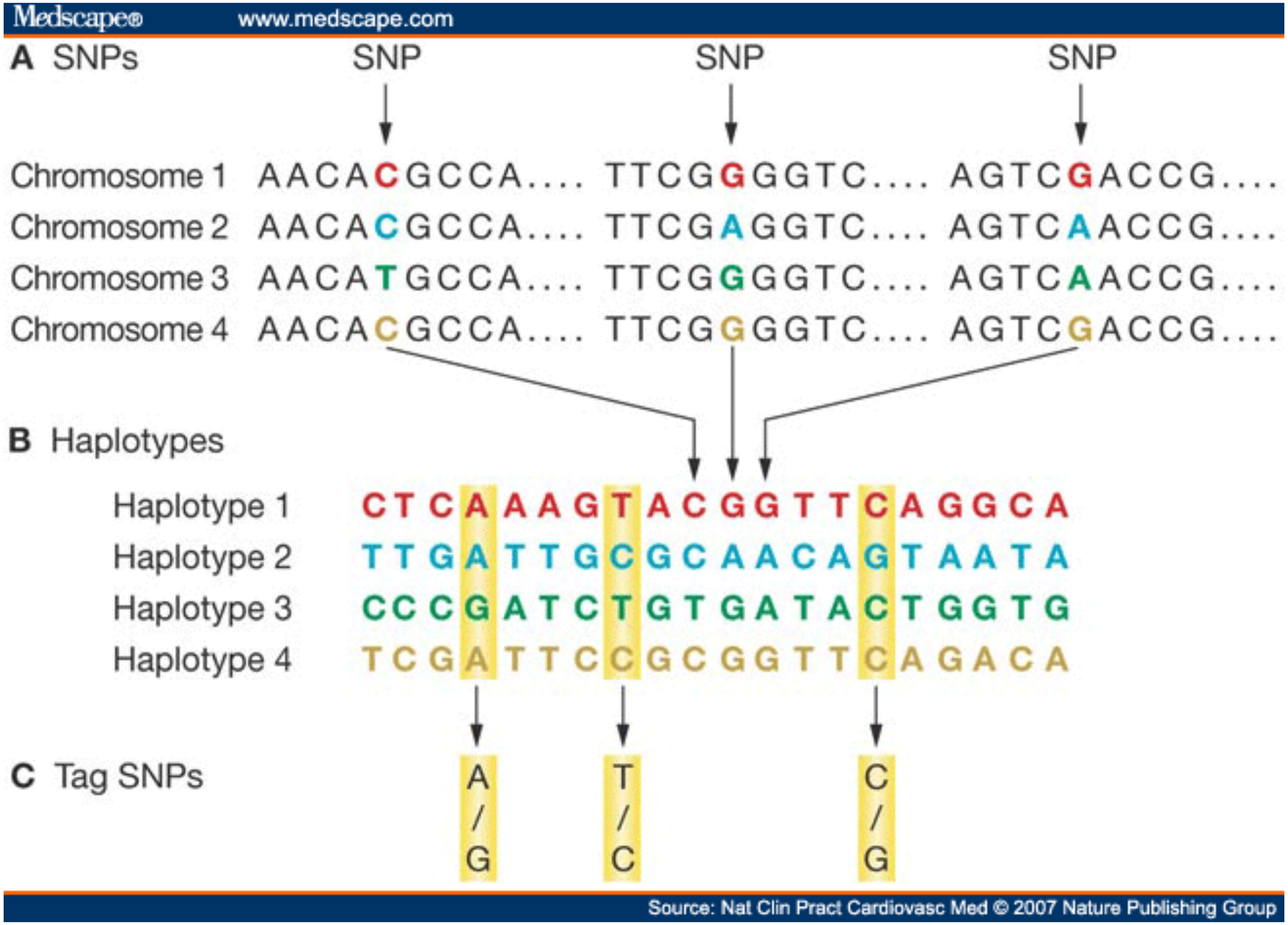
Dissect genetic and molecular mechanism underlying complex (disease) traits.
Standard approaches:

(1) Genome wide association studies (GWAS): Correlations between genetic variants and trait variation.

(2) Gene expression studies: correlations between gene expression and trait variation.

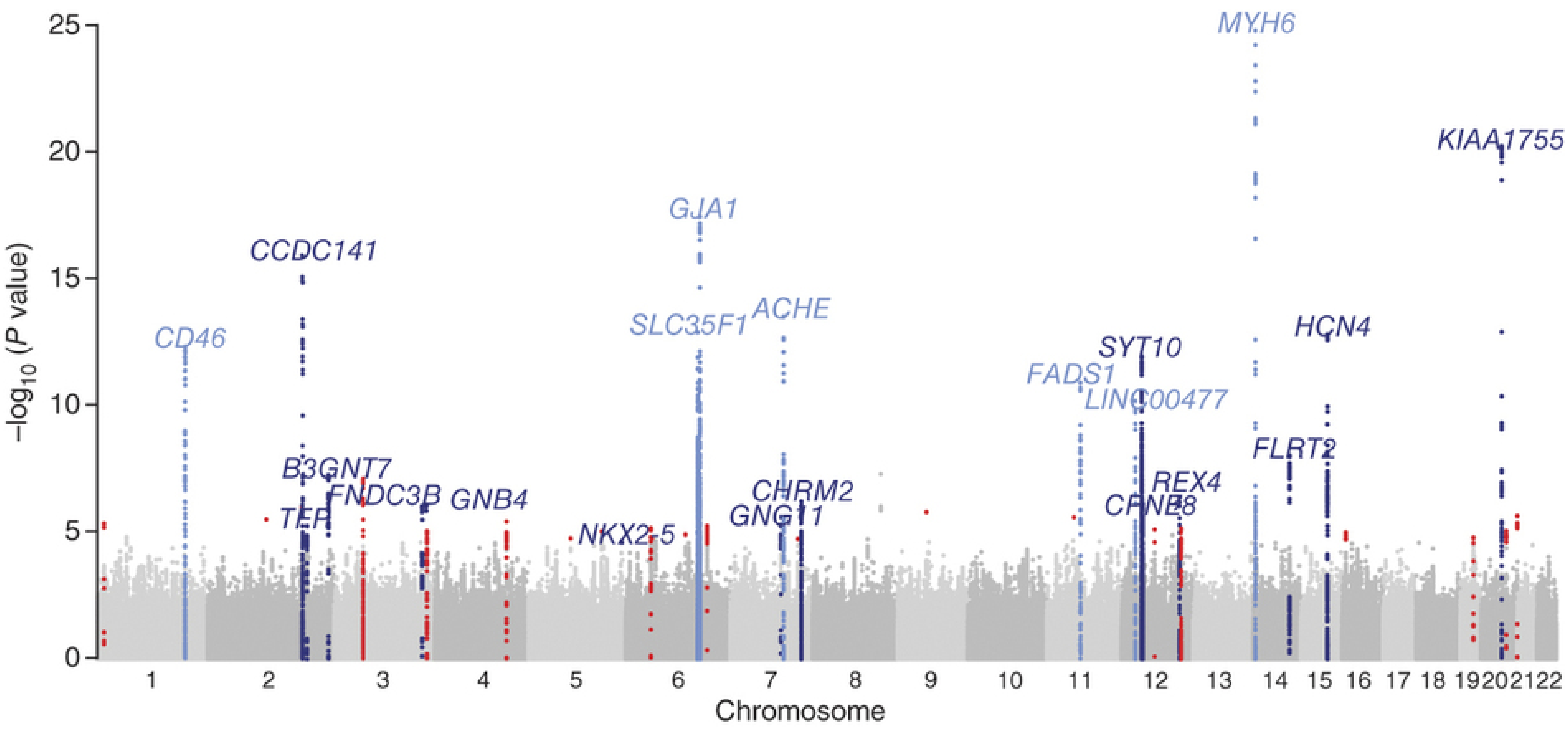
Integration of both approaches for complementary evidence.

Single nucleotide polymorphisms and haplotypes


Medscape®
www.medscape.com
A SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCAACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCAACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
B Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
C Tag SNPs
A/G
T/C
C/G

Source: Nat Clin Pract Cardiovasc Med © 2007 Nature Publishing Group

Genome wide association studies



den Hoed et al 2013.

Challenges

(1) Find single variants, independently contributing to disease.

Challenges

(1) Find single variants, independently contributing to disease.

(2) Issues with population structure, control for LD, etc...

Challenges

(1) Find single variants, independently contributing to disease.

(2) Issues with population structure, control for LD, etc...

(3) Genetic variations have been identified for a wide variety of common complex diseases (GWAS catalog).

Challenges

(1) Find single variants, independently contributing to disease.

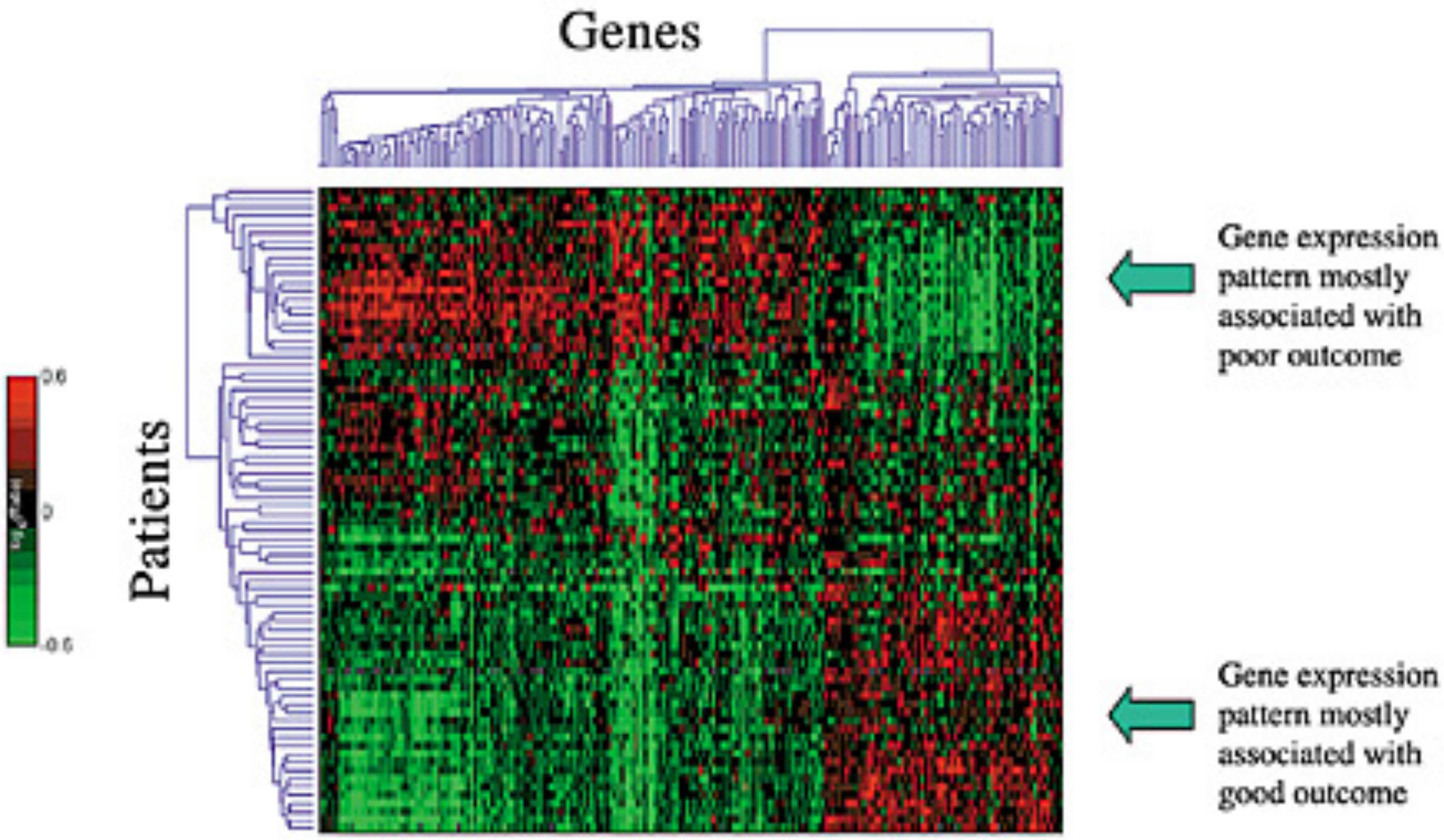
(2) Issues with population structure, control for LD, etc...

(3) Genetic variations have been identified for a wide variety of common complex diseases (GWAS catalog).

(4) Missing heritability: genetic variation explains 5% of height variation.

(5) Very weak predictive power.

Gene expression based studies



Gene expression pattern mostly associated with poor outcome

Gene expression pattern mostly associated with good outcome

Challenges

(1) Signatures or gene lists predictive of disease.

Challenges

(1) Signatures or gene lists predictive of disease.

(2) Sensitive to many environmental factors.

Challenges

(1) Signatures or gene lists predictive of disease.

(2) Sensitive to many environmental factors.

(3) Is a complex trait itself.

Challenges

(1) Signatures or gene lists predictive of disease.

(2) Sensitive to many environmental factors.

(3) Is a complex trait itself.

(4) Causal versus reactive.

Challenges

(1) Signatures or gene lists predictive of disease.

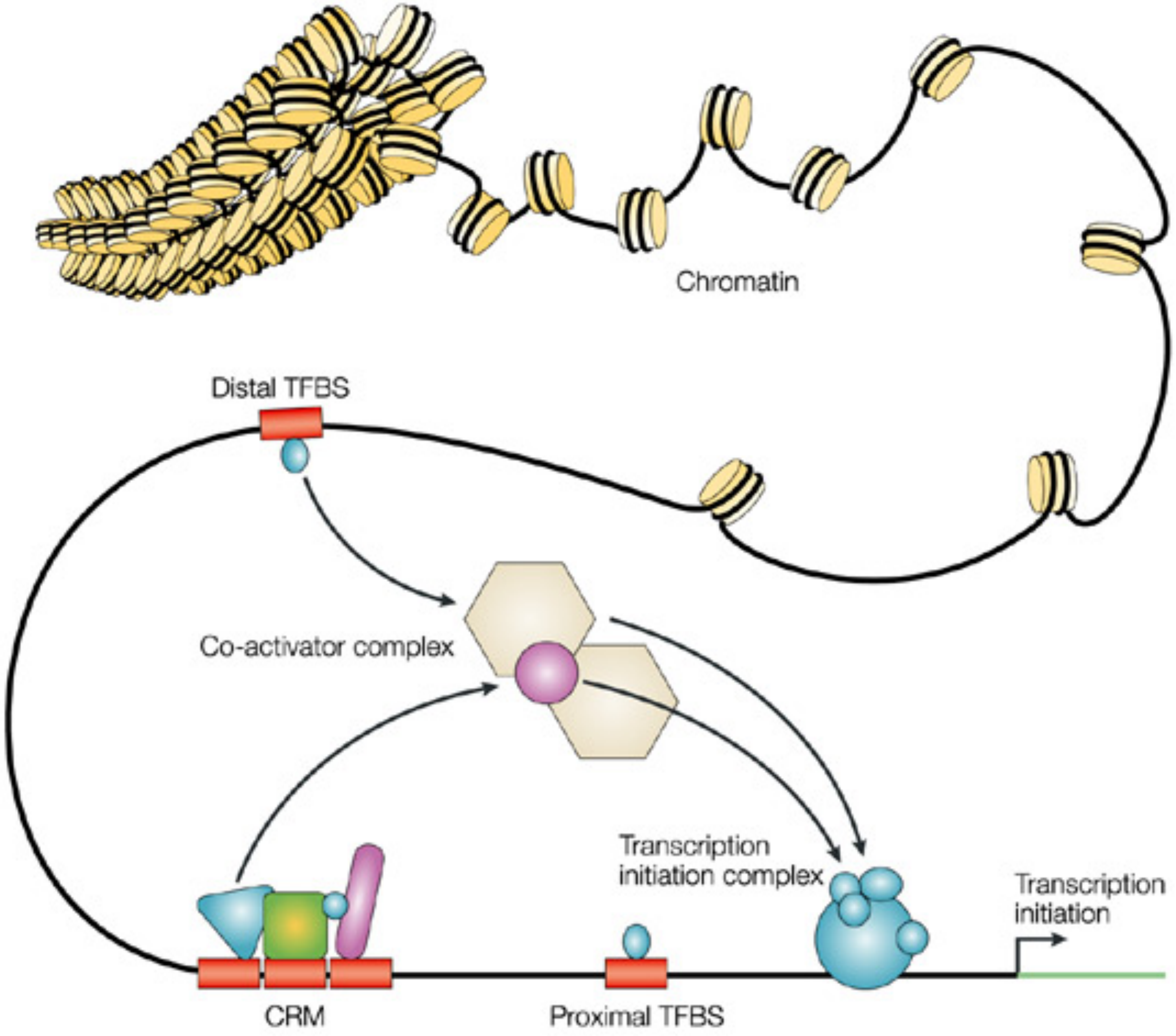
(2) Sensitive to many environmental factors.

(3) Is a complex trait itself.

(4) Causal versus reactive.

(5) Can we find evidence that expression variation predictive of trait variation is genetic.

Transcriptional regulation



Nature Reviews | Genetics

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals:

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.

(2) This association is robust across eQTL thresholds.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.

(2) This association is robust across eQTL thresholds.

(3) Can help with causal versus reactive.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.

(2) This association is robust across eQTL thresholds.

(3) Can help with causal versus reactive.

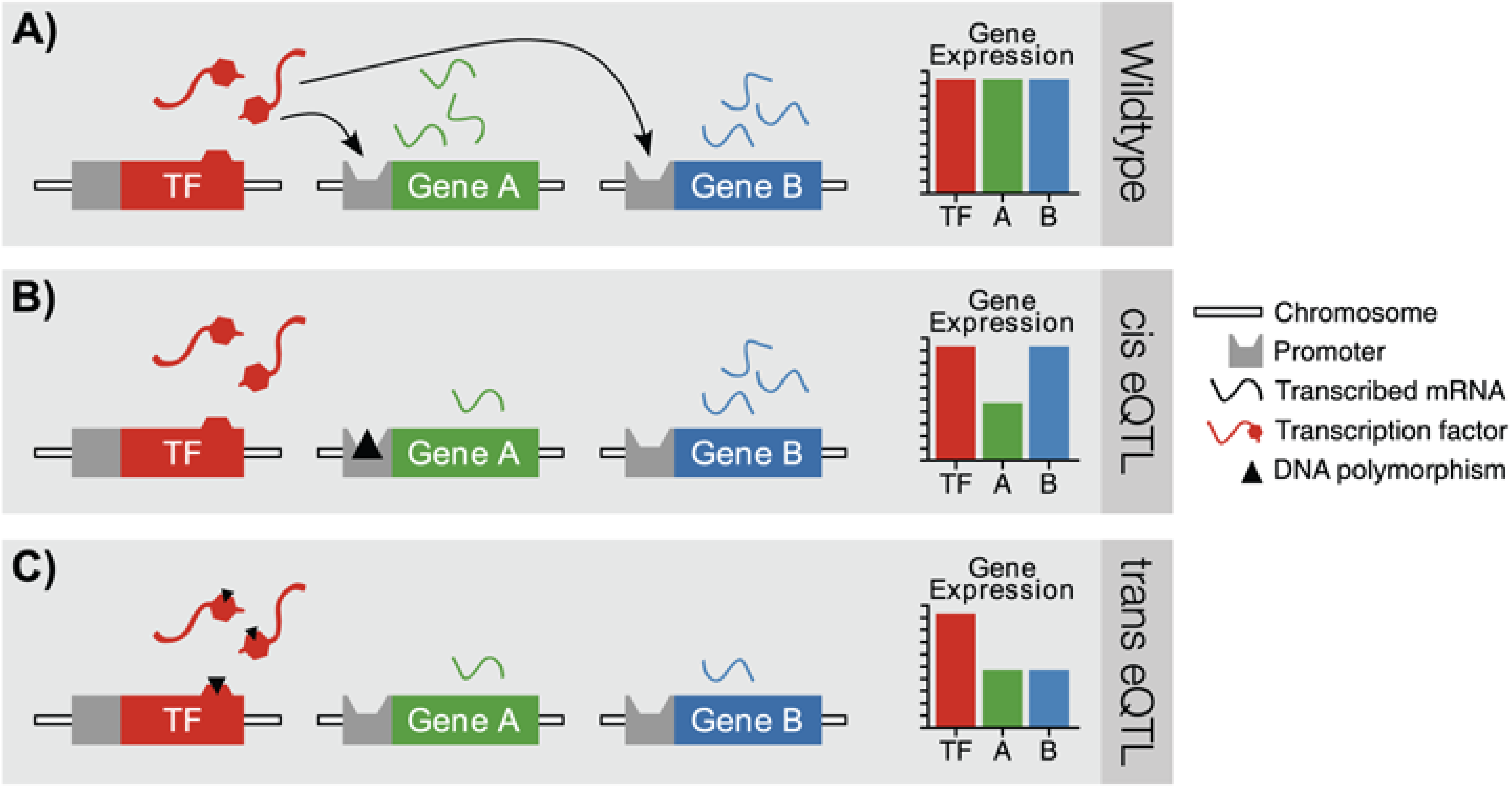
(4) Need expression data and SNP data from same individuals.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.
(2) This association is robust across eQTL thresholds.
(3) Can help with causal versus reactive.
(4) Need expression data and SNP data from same individuals.
(5) Missing heritability still a problem.

cis and trans eQTL



Wolen and Miles 2012.

Mapping cis vs. trans

eQTL meta-study in 5,311 individuals with replication in 2,775 individuals of non-transformed peripheral blood samples by H-J Westra et al 2013

Mapping cis vs. trans

eQTL meta-study in 5,311 individuals with replication in 2,775 individuals of non-transformed peripheral blood samples by H-J Westra et al 2013

397,310 significant unique cis-eQTL SNPs at FDR<.05

Mapping cis vs. trans

eQTL meta-study in 5,311 individuals with replication in 2,775 individuals of non-transformed peripheral blood samples by H-J Westra et al 2013

397,310 significant unique cis-eQTL SNPs at FDR<.05
346 significant unique trans-eQTL SNPs at FDR<.05

Why is the distal signal weak

(1) Testing burden: Number of distal SNPs $\gg$ number of proximal SNPs.

Why is the distal signal weak

(1) Testing burden: Number of distal SNPs $\gg$ number of proximal SNPs.

(2) It is really weaker.

Why is the distal signal weak

(1) Testing burden: Number of distal SNPs $\gg$ number of proximal SNPs.

(2) It is really weaker.

Fisher's infinitesimal (polygenic) model suggested very large number of mutations of infinitesimal effect.
The effect-size distribution of adaptive substitutions is approximately exponential.

A model

Given paired gene expression and SNP data for n individuals: $(X_i, S_i)_{i=1}^{n}$ with $X_i \in \mathbb{R}^{30k}$ and $S_i \in \{0, 1, 2\}^{500k}$

A model

Given paired gene expression and SNP data for n individuals: $(X_i, S_i)_{i=1}^n$ with $X_i \in \mathbb{R}^{30k}$ and $S_i \in \{0, 1, 2\}^{500k}$

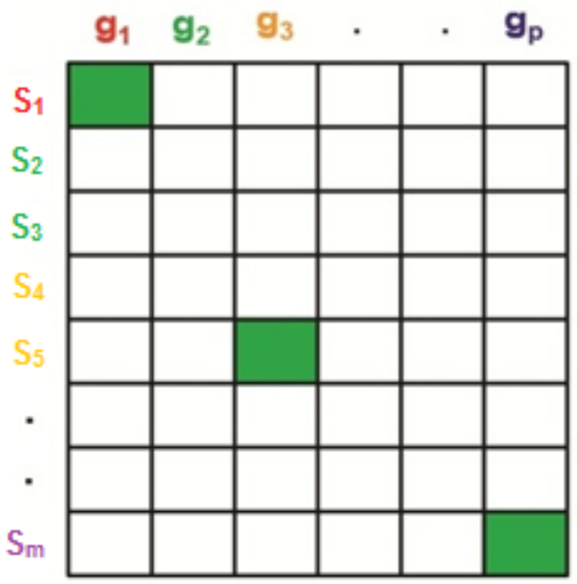
Assume the j-th SNP S^j is distal to the k-th gene X^k

$$e(X^k \mid S^j) = e(X^k \mid X^j) + e(X^j \mid S^j),$$

where X^j is the gene proximal to SNP j.

A strategy

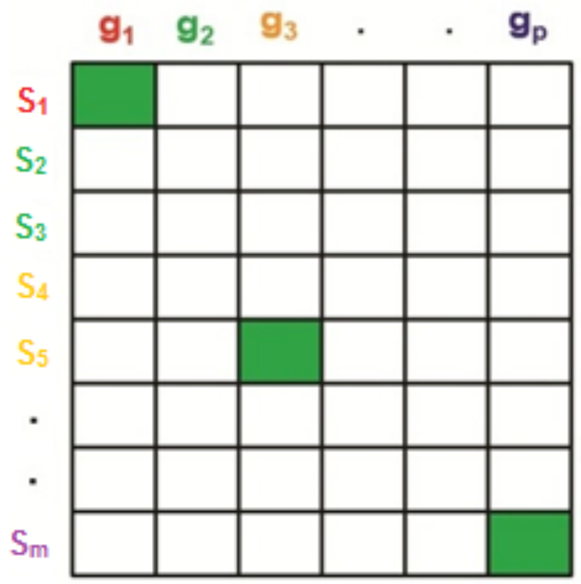
(1) Compute evidence for proximal effects.



	g_1	g_2	g_3	.	.	g_p
S_1	🟩					
S_2						
S_3						
S_4						
S_5			🟩			
.						
.						
S_m						🟩

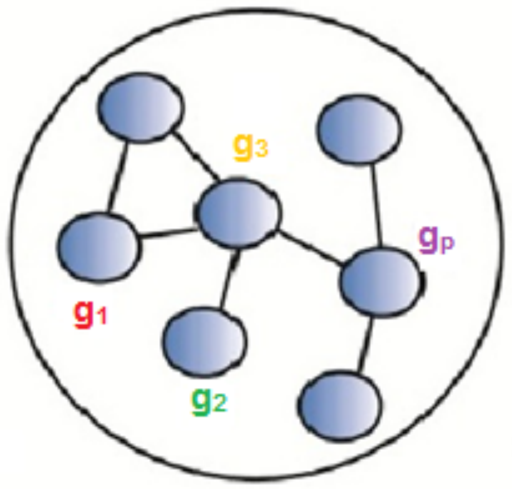
A strategy

(1) Compute evidence for proximal effects.



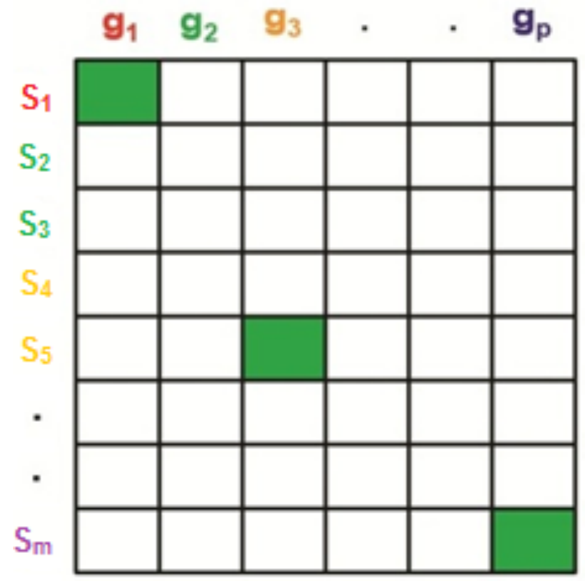
	g_1	g_2	g_3	.	.	g_p
S_1	🟩					
S_2						
S_3						
S_4						
S_5			🟩			
.						
.						
S_m						🟩

(2) Compute evidence for direct gene by gene expression effects – infer a gene network.



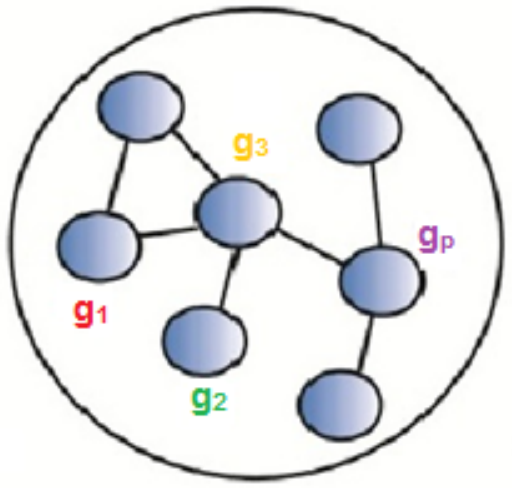
A strategy

(1) Compute evidence for proximal effects.



	g_1	g_2	g_3	.	.	g_p
s_1	🟩					
s_2						
s_3						
s_4						
s_5			🟩			
.						
.						
s_m						🟩

(2) Compute evidence for direct gene by gene expression effects – infer a gene network.



(3) Test for associations between SNPs with proximal effects and genes local to the proximal gene on the gene network.

Step 1. Infer local eQTLs

For each gene $j = 1, ..., p$ and a specified window size assign local SNPs and fit:

$$X^j = \beta_0 + \beta S^1, \quad \cdots \quad X^j = \beta_0 + \beta S^m.$$

Step 1. Infer local eQTLs

For each gene $j = 1, ..., p$ and a specified window size assign local SNPs and fit:

$$X^j = \beta_0 + \beta S^1, \quad \cdots \quad X^j = \beta_0 + \beta S^m.$$

(1) Use FDR q-value of .05 for significant local significance.

Step 1. Infer local eQTLs

For each gene $j = 1, ..., p$ and a specified window size assign local SNPs and fit:

$$X^j = \beta_0 + \beta S^1, \quad \cdots \quad X^j = \beta_0 + \beta S^m.$$

(1) Use FDR q-value of .05 for significant local significance.

(2) If there are multiple locally associated SNPs select the variant with the largest effect size.

Step 2. Infer the gene network

Infer gene network using Sparse PArtial Correlation Estimation (SPACE), Peng et al 2009.

Step 2. Infer the gene network

Infer gene network using Sparse PArtial Correlation Estimation (SPACE), Peng et al 2009.

We want to infer

$$\rho^{ij} = \mathrm{Corr}(X^i, X^j \mid X^{1,\ldots,p\setminus ij})$$

Step 2. Infer the gene network

Infer gene network using Sparse PArtial Correlation Estimation (SPACE), Peng et al 2009.

We want to infer

$$\rho^{ij} = \mathrm{Corr}(X^i, X^j \mid X^{1,\ldots,p\setminus ij})$$

Regularized loss:

$$\hat{\rho}^{ij} = \arg\min_{\rho^{ij}} \left[\frac{1}{2} \sum_{i=1}^{p} \left\| \mathbf{x}^i - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\omega^{jj}}{\omega^{i}\boldsymbol{i}}} \mathbf{x}^i \right\|^2 + \lambda \sum_{1 \leq i < j \leq p} |\rho^{ij}| \right],$$

$\mathbf{x}^i$ is the vector of gene expression for the *i*-th gene, ω^{ii} is the precision of the *i*-th gene, λ set by BIC.

Step 2. Infer the gene network

If the partial correlation is non-zero there is an edge between genes i and j, $\mathbf{E}_{ij} = 1$ if $\hat{\rho}^{ij} \neq 0$.

Step 3. Infer distal eQTL

1. For each gene *j* with a significantly associated SNP perform distal eQTL testing:

Step 3. Infer distal eQTL

1. For each gene *j* with a significantly associated SNP perform distal eQTL testing:
   1. Define the set $\mathcal{S} = \{$all genes within two steps from $j \in \mathbf{E}\}$
   2. For each gene $k \in \mathcal{S}$ regress against the associated SNP S^j for the *j*-th gene.

Step 3. Infer distal eQTL

1. For each gene j with a significantly associated SNP perform distal eQTL testing:
   i. Define the set $\mathcal{S} = \{$all genes within two steps from $j \in \mathbf{E}\}$
   ii. For each gene $k \in \mathcal{S}$ regress against the associated SNP S^j for the j-th gene.
2. Assess significance using Benjamini-Hochberg correction for FDR.

Results on HapMap data

eQTL analysis was performed for 69 Nigerian individuals with RNA-seq data from lymphoblastoid cell lines and genotype data from HapMap in Pickrell et al. 2010.

Results on HapMap data

eQTL analysis was performed for 69 Nigerian individuals with RNA-seq data from lymphoblastoid cell lines and genotype data from HapMap in Pickrell et al. 2010.

Using a window of 200kb and restricting to 9810 top quartile transcripts

(1) 1842 transcripts with local eQTL, FDR $<$ 0.1.

Results on HapMap data

eQTL analysis was performed for 69 Nigerian individuals with RNA-seq data from lymphoblastoid cell lines and genotype data from HapMap in Pickrell et al. 2010.

Using a window of 200kb and restricting to 9810 top quartile transcripts

(1) 1842 transcripts with local eQTL, FDR < 0.1.

(2) Replicated 541 of the 949 transcripts identified in Pickrell et al., remainder were low expression.

Results on HapMap data

eQTL analysis was performed for 69 Nigerian individuals with RNA-seq data from lymphoblastoid cell lines and genotype data from HapMap in Pickrell et al. 2010.

Using a window of 200kb and restricting to 9810 top quartile transcripts

(1) 1842 transcripts with local eQTL, FDR < 0.1.

(2) Replicated 541 of the 949 transcripts identified in Pickrell et al., remainder were low expression.

(3) 1824 transcripts with distal eQTL, FDR < 0.1.

Results on HapMap data

eQTL analysis was performed for 69 Nigerian individuals with RNA-seq data from lymphoblastoid cell lines and genotype data from HapMap in Pickrell et al. 2010.

Using a window of 200kb and restricting to 9810 top quartile transcripts

(1) 1842 transcripts with local eQTL, FDR < 0.1.

(2) Replicated 541 of the 949 transcripts identified in Pickrell et al., remainder were low expression.

(3) 1824 transcripts with distal eQTL, FDR < 0.1.

(4) Pickrell et al. reported no distal eQTL.

Results on HapMap data

eQTL analysis was performed for 69 Nigerian individuals with RNA-seq data from lymphoblastoid cell lines and genotype data from HapMap in Pickrell et al. 2010.

Using a window of 200kb and restricting to 9810 top quartile transcripts

1. 1842 transcripts with local eQTL, FDR < 0.1.
2. Replicated 541 of the 949 transcripts identified in Pickrell et al., remainder were low expression.
3. 1824 transcripts with distal eQTL, FDR < 0.1.
4. Pickrell et al. reported no distal eQTL.
5. SNPs-vs-all genes finds 5 distal associations.

Results on mouse cross

Paired genotype and liver gene expression data from 156 partially inbred mice analyzed as part of the Collaborative Cross Consortium 2012.

Results on mouse cross

Paired genotype and liver gene expression data from 156 partially inbred mice analyzed as part of the Collaborative Cross Consortium 2012.

6,182 eQTL for 5,733 genes for FDR of 5%, 75% eQTL were within 10cM of associated gene.

Results on mouse cross

Paired genotype and liver gene expression data from 156 partially inbred mice analyzed as part of the Collaborative Cross Consortium 2012.
6,182 eQTL for 5,733 genes for FDR of 5%, 75% eQTL were within 10cM of associated gene.

Using a window of 1Mb window

(1) 5,748 genes with a local eQTL.

Results on mouse cross

Paired genotype and liver gene expression data from 156 partially inbred mice analyzed as part of the Collaborative Cross Consortium 2012.
6,182 eQTL for 5,733 genes for FDR of 5%, 75% eQTL were within 10cM of associated gene.

Using a window of 1Mb window

(1) 5,748 genes with a local eQTL.
(2) 774 with at least one distal eQTL.

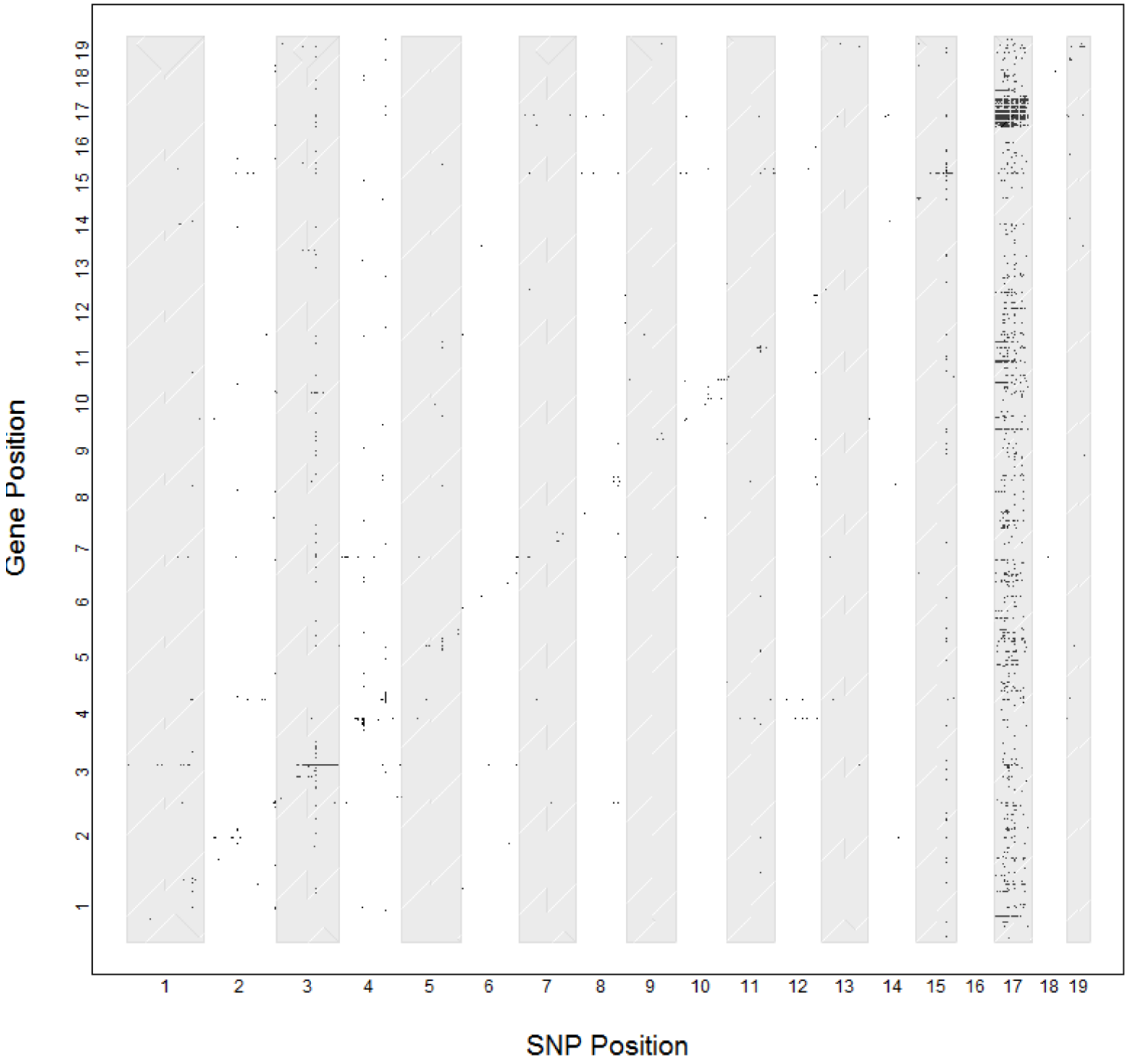
Results on mouse cross

Paired genotype and liver gene expression data from 156 partially inbred mice analyzed as part of the Collaborative Cross Consortium 2012.
6,182 eQTL for 5,733 genes for FDR of 5%, 75% eQTL were within 10cM of associated gene.

Using a window of 1Mb window

(1) 5,748 genes with a local eQTL.
(2) 774 with at least one distal eQTL.
    i. 453 linked to one SNP.
    ii. 87 linked to two SNPs.
    iii. 44 linked to three SNPs.
    iv. 190 linked to four or more SNPs.
    v. 260 multi locus genes linked to a set of 42 hotspot loci on chromosome 17

Results on mouse cross



Extensions and open problems

(1) Bayesian one-step procedure.

Extensions and open problems

(1) Bayesian one-step procedure.

(2) Incorporating other genomic features.

Extensions and open problems

(1) Bayesian one-step procedure.

(2) Incorporating other genomic features.

(3) Explicit use of the effect size distributions.

Extensions and open problems

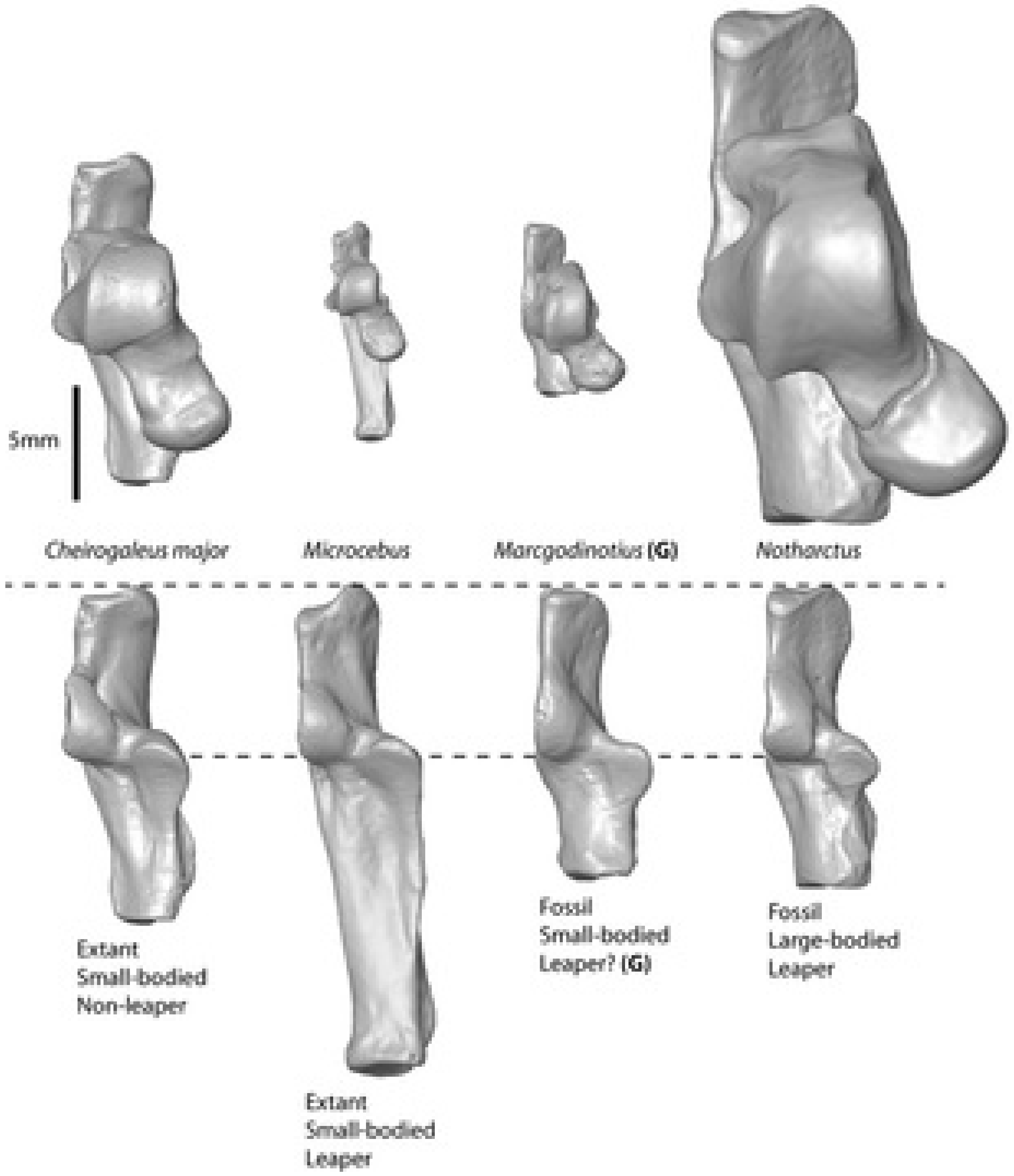
(1) Bayesian one-step procedure.

(2) Incorporating other genomic features.

(3) Explicit use of the effect size distributions.

(4) Replacing FDR with local FDR.

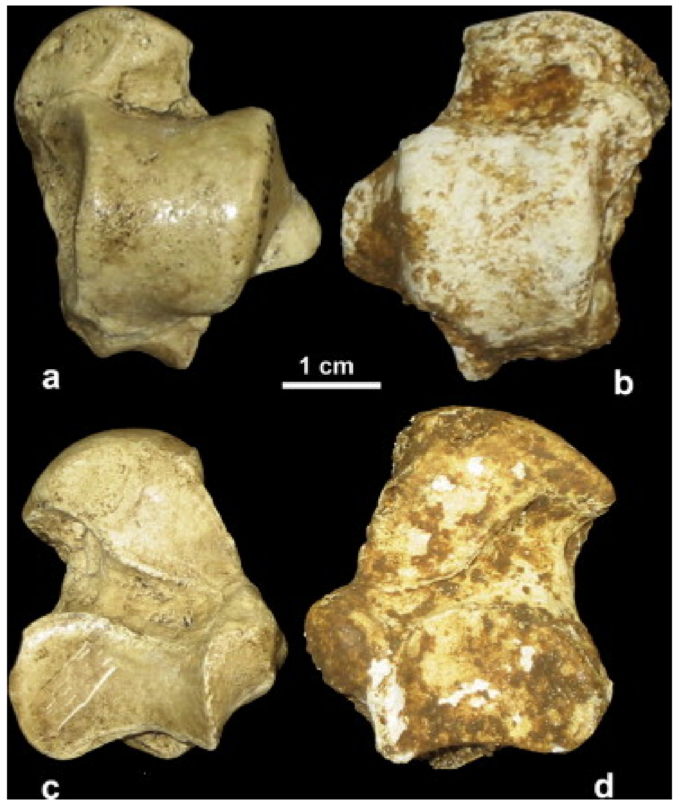
Shapes as traits



From D. Boyer.

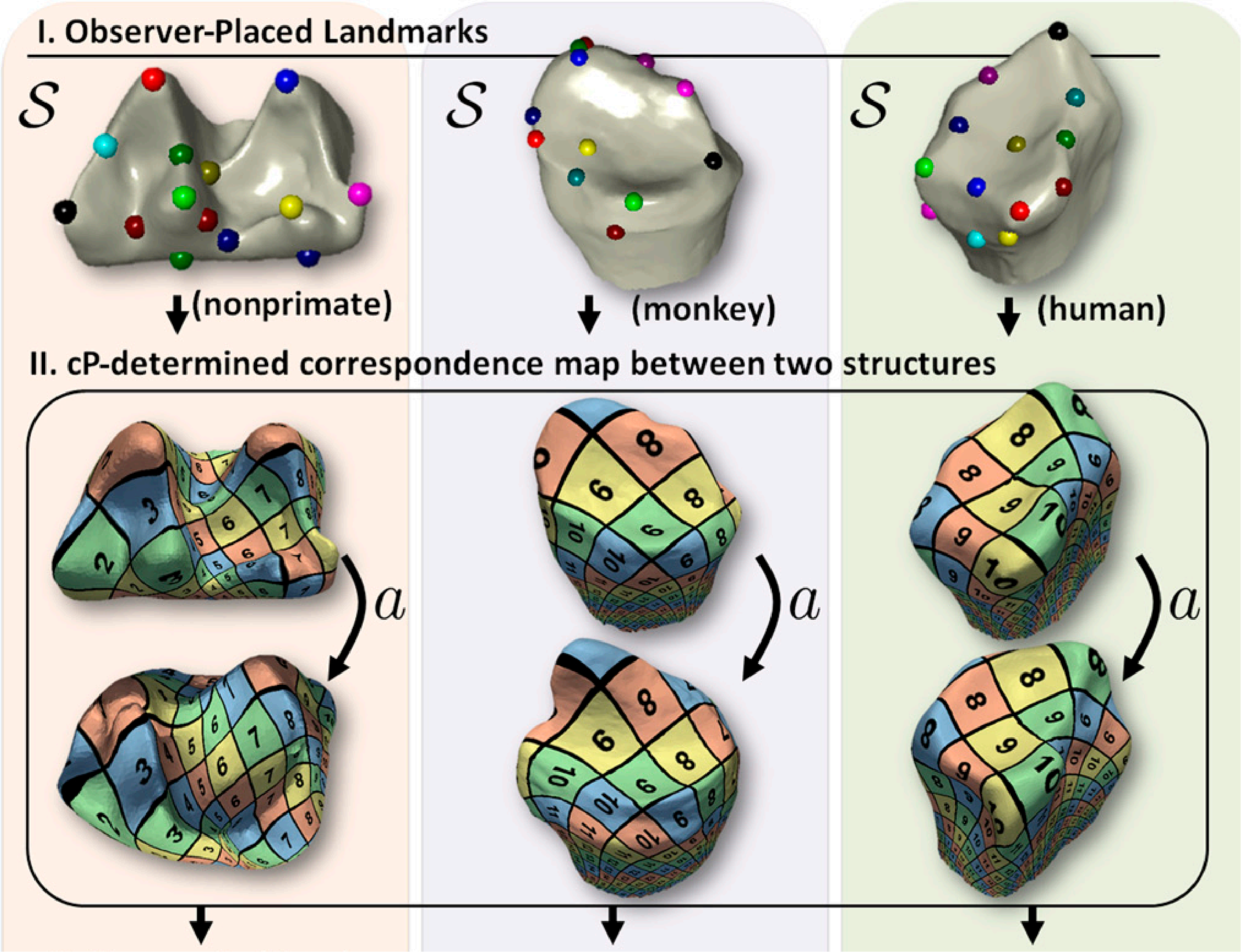
A problem in morphology

Distance between ankle bones across primates for evolutionary analysis.



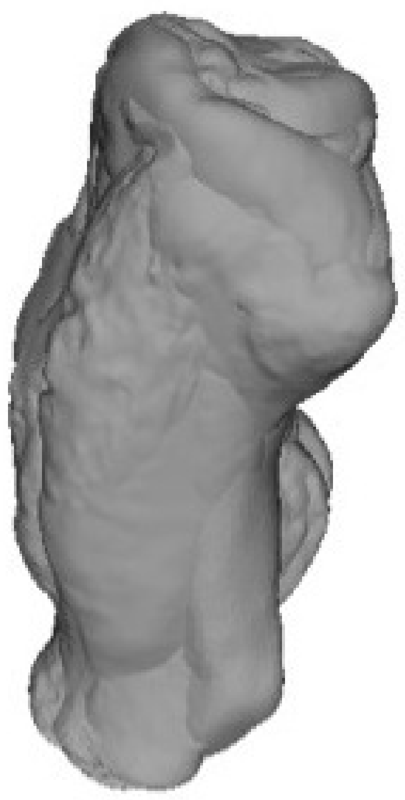
Algorithms to automatically quantify the geometric similarity of anatomical surfaces, Boyer et. al. PNAS 2011.

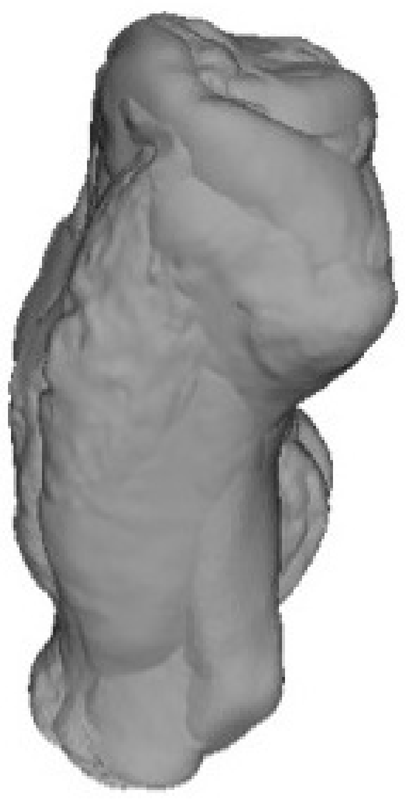
Geometric algorithm


I. Observer-Placed Landmarks
$\mathcal{S}$
$\mathcal{S}$
$\mathcal{S}$
(nonprimate)
(monkey)
(human)
II. cP-determined correspondence map between two structures
a
a
a

Topological methods

What happens when the shapes are not isomorphic ?





Topological methods

Broken claw tips.

Euler characteristic

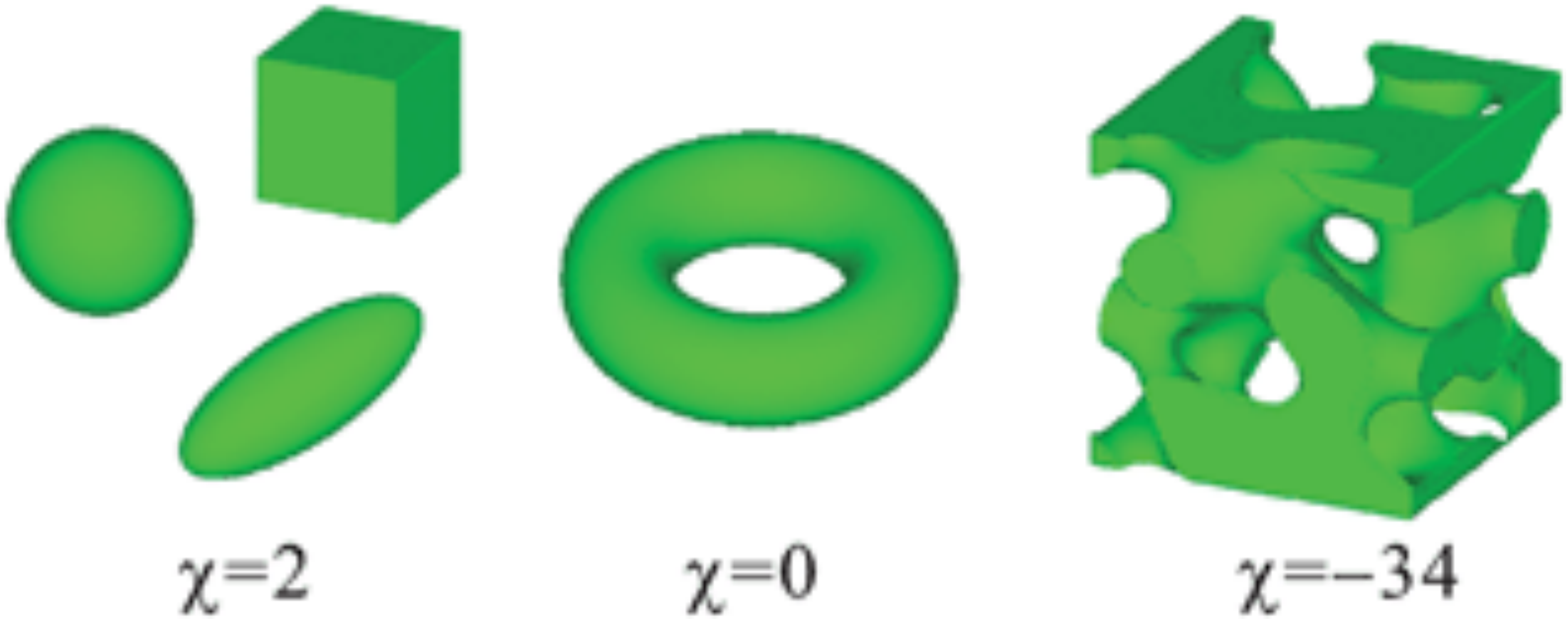
Given a shape M the Euler characteristic is

$$\chi(M) = \sum_{i=0}^{d} (-1)^i \beta_i = \#\text{vertices} - \#\text{edges} + \#\text{faces}.$$

Euler characteristic

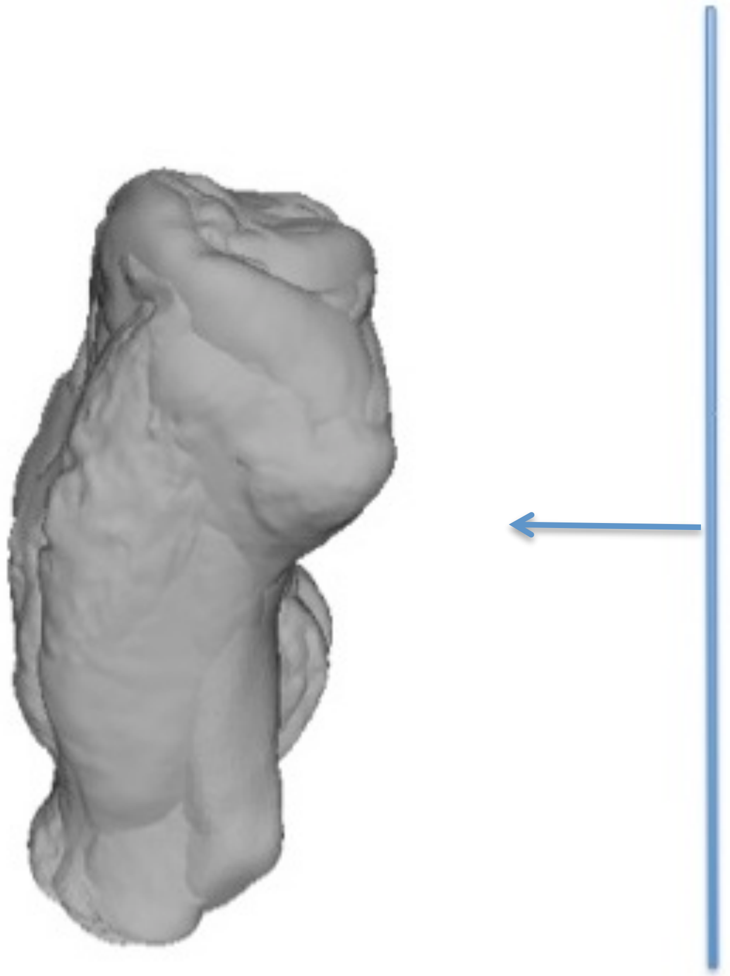
Given a shape M the Euler characteristic is

$$\chi(M) = \sum_{i=0}^{d} (-1)^i \beta_i = \#\text{vertices} - \#\text{edges} + \#\text{faces}.$$



Back to bones

The idea of a height function



Summary statistic

M is simplicial complex in $\mathbb{R}^d$ and $v \in S^{d-1}$ is a unit vector.
$\chi(M, v)$ captures changes in topology of

$$M(v)_r = \{\Delta \in M : x \cdot v \leq r \text{ for all } x \in \Delta\}.$$

Summary statistic

M is simplicial complex in $\mathbb{R}^d$ and $v \in S^{d-1}$ is a unit vector. $\chi(M, v)$ captures changes in topology of

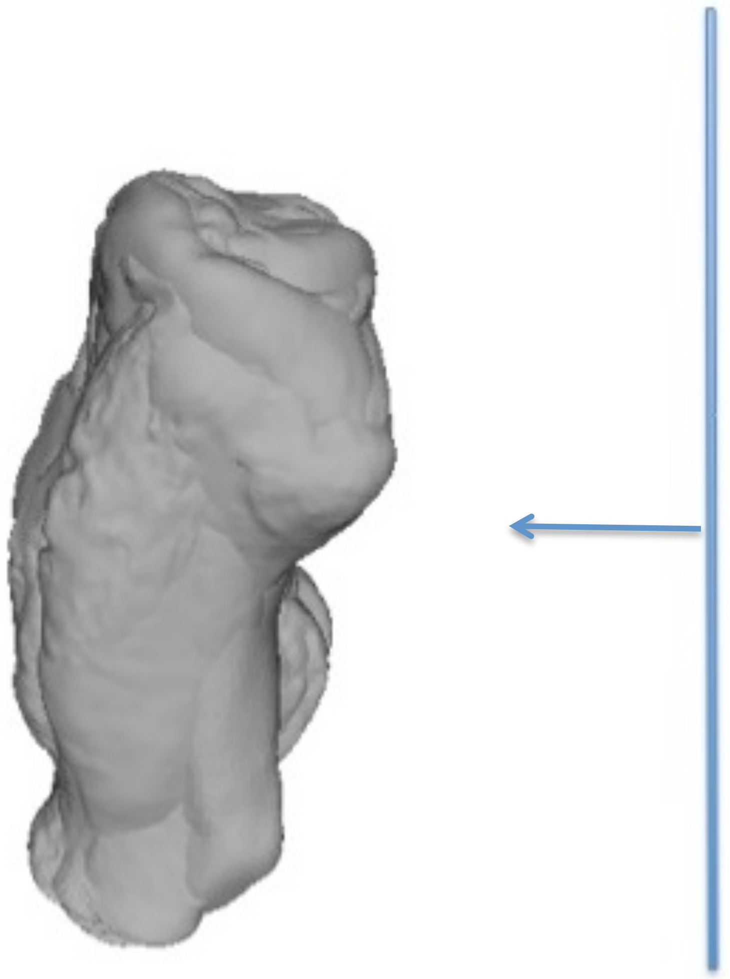
$$M(v)_r = \{\Delta \in M : x \cdot v \leq r \text{ for all } x \in \Delta\}.$$

Definition

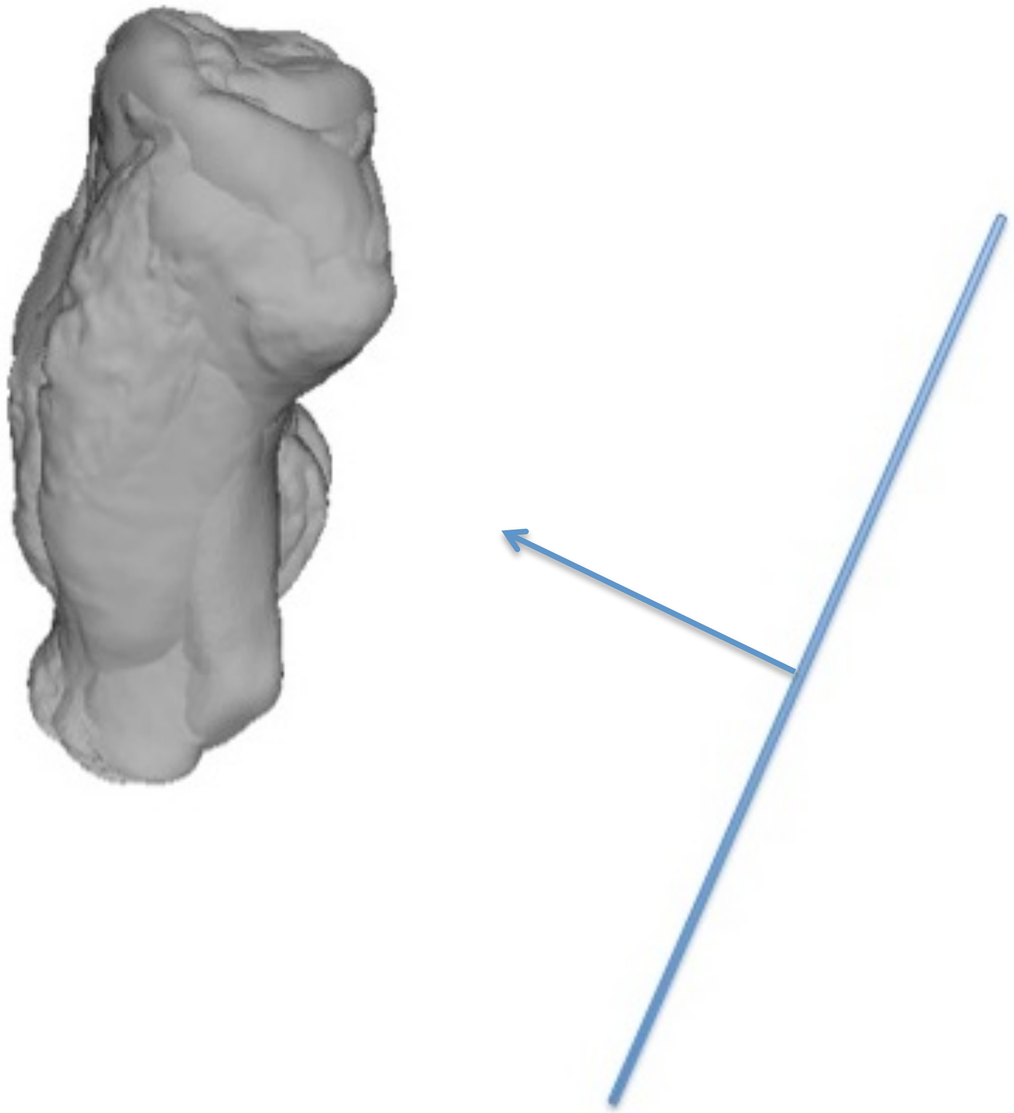
The Euler characteristic transform of $M \in \mathbb{R}^d$ is the function

$$\begin{aligned} \mathrm{ECT}(M) : S^{d-1} &\to L_2(\mathbb{R}) \\ v &\mapsto \chi(M, v). \end{aligned}$$

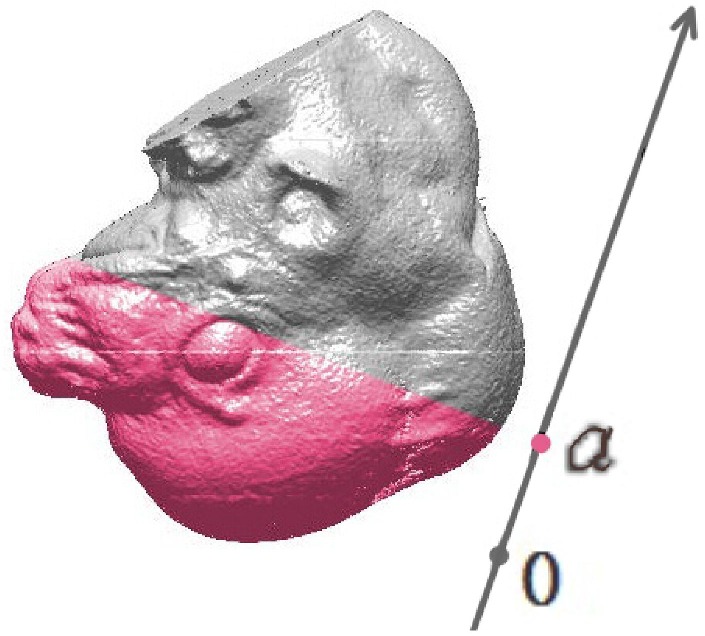
Height function: v_1

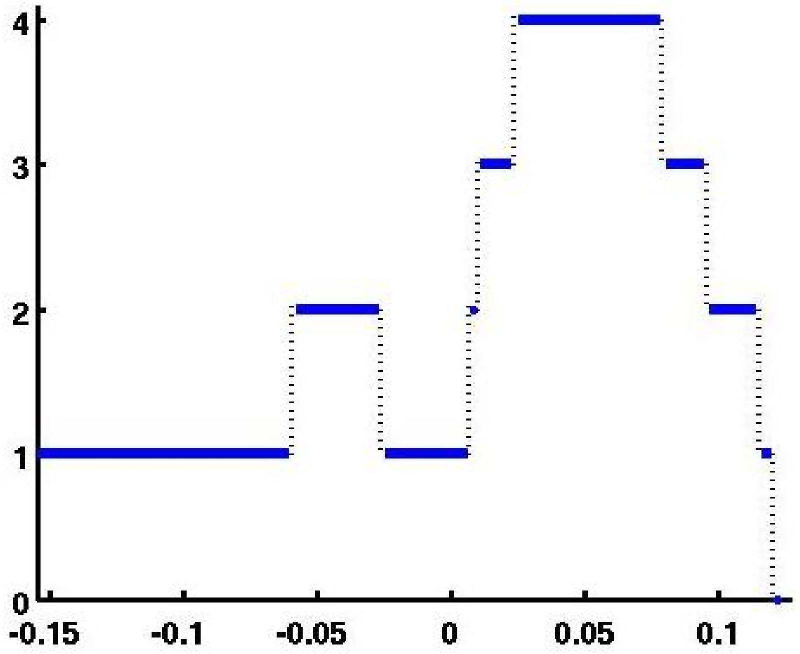


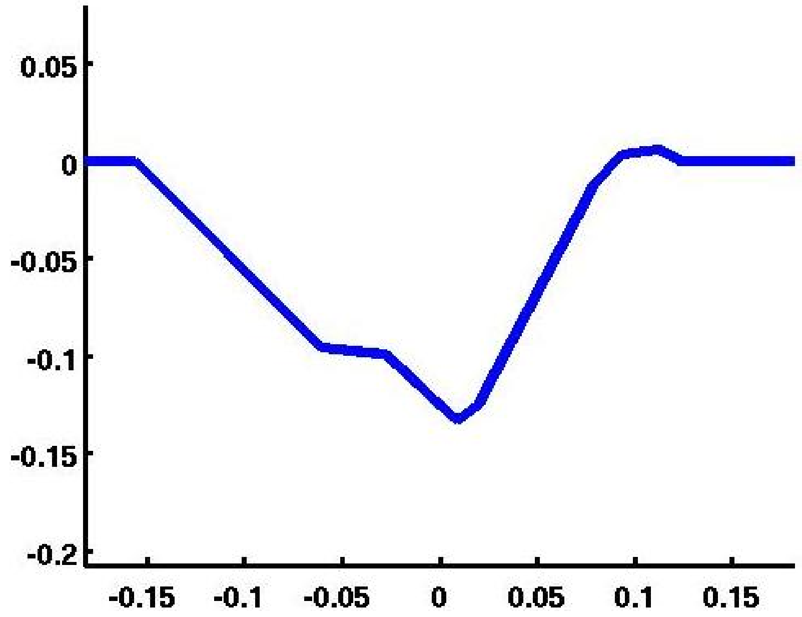
Height function: v_2



Euler characteristic curve







Distances

$\mathcal{M}_d$ is the space of finite simplicial complexes in $\mathbb{R}^d$.

Distances

$\mathcal{M}_d$ is the space of finite simplicial complexes in $\mathbb{R}^d$.

The distance between two surfaces M_1, M_2 is

$$d_{\mathcal{M}_d}(M_1, M_2) := \int_{S^{d-1}} d(\chi(M_1, v), \chi(M_2, v)) dv.$$

Sufficient statistic

Given $X \sim f_\theta \in \mathcal{F}$, a statistic $T = T(X)$ is sufficient if for the parameter θ if for all sets B the probability $\mathbb{P}[X \in B \mid T(X) = t]$ does not depend on θ

$$\mathbb{P}[X \mid T(X) = t, \theta] = \mathbb{P}[X \mid T(X) = t].$$

Sufficient statistic

Given $X \sim f_\theta \in \mathcal{F}$, a statistic $T = T(X)$ is sufficient if for the parameter θ if for all sets B the probability $\mathbb{P}[X \in B \mid T(X) = t]$ does not depend on θ

$$\mathbb{P}[X \mid T(X) = t, \theta] = \mathbb{P}[X \mid T(X) = t].$$

Sufficiency of the ECT

Theorem (Turner-M-Boyer)

The Euler characteristic transform is injective when the domain is $\mathcal{M}_d$ *for* $d = 2, 3$.

Sufficiency of the ECT

Theorem (Turner-M-Boyer)

The Euler characteristic transform is injective when the domain is $\mathcal{M}_d$ for $d = 2, 3$.

Corollary (Turner-M-Boyer)

For a density function $f(x; \theta)$ with supp$(f) \subseteq \mathcal{M}_d$ ($d = 2, 3$) the ECT is a sufficient statistic.

Exponential family and ECT

Denote the Euler characteristic curve for each direction: $f(y) = \chi(M, v)$ Define the integral of $f(y)$ as $F(x) = \int_0^x f(y)dy$.

Exponential family and ECT

Denote the Euler characteristic curve for each direction: $f(y) = \chi(M, v)$ Define the integral of $f(y)$ as $F(x) = \int_0^x f(y)dy$.

This results in K smooth curves $\{F_1, ..., F_K\}$.

Exponential family and ECT

Denote the Euler characteristic curve for each direction: $f(y) = \chi(M, v)$ Define the integral of $f(y)$ as $F(x) = \int_0^x f(y) dy$.

This results in K smooth curves $\{F_1, ..., F_K\}$.

Exponential family model

$$p_\theta(x) = a(\theta)\, h(x)\, \exp\left(- \sum_{k=1}^{K} \langle \theta, F_k(x) \rangle \right).$$

The matrix variate normal

Define $\mathbf{F} = [F_1 F_2 \cdots F_K]$ as a $K \times T$ matrix and

$$p(\mathbf{F} \mid \mathbf{A}, \mathbf{U}, \mathbf{V}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[\mathbf{V}^{-1}(\mathbf{F} - \mathbf{A})^T\mathbf{U}^{-1}(\mathbf{F} - \mathbf{A})]\right)}{(2\pi)^{KT/2}|\mathbf{V}|^{L/2}|\mathbf{U}|^{K/2}},$$

A models mean

U models covariance between curves

V models covariance between points in a curve.

The matrix variate normal

Define $\mathbf{F} = [F_1 F_2 \cdots F_K]$ as a $K \times T$ matrix and

$$p(\mathbf{F} \mid \mathbf{A}, \mathbf{U}, \mathbf{V}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[\mathbf{V}^{-1}(\mathbf{F} - \mathbf{A})^T\mathbf{U}^{-1}(\mathbf{F} - \mathbf{A})]\right)}{(2\pi)^{KT/2}|\mathbf{V}|^{L/2}|\mathbf{U}|^{K/2}},$$

A models mean
U models covariance between curves
V models covariance between points in a curve.

The given n meshes $(M_1, ..., M_n)$ we can define a likelihood model

$$\mathrm{Lik}(M_1, ..., M_n \mid \mathbf{A}, \mathbf{U}, \mathbf{V}) = \prod_{i=1}^{n} p(\mathbf{F}(M_i) \mid \mathbf{A}, \mathbf{U}, \mathbf{V}), \qquad (4)$$

Picture of heel bone

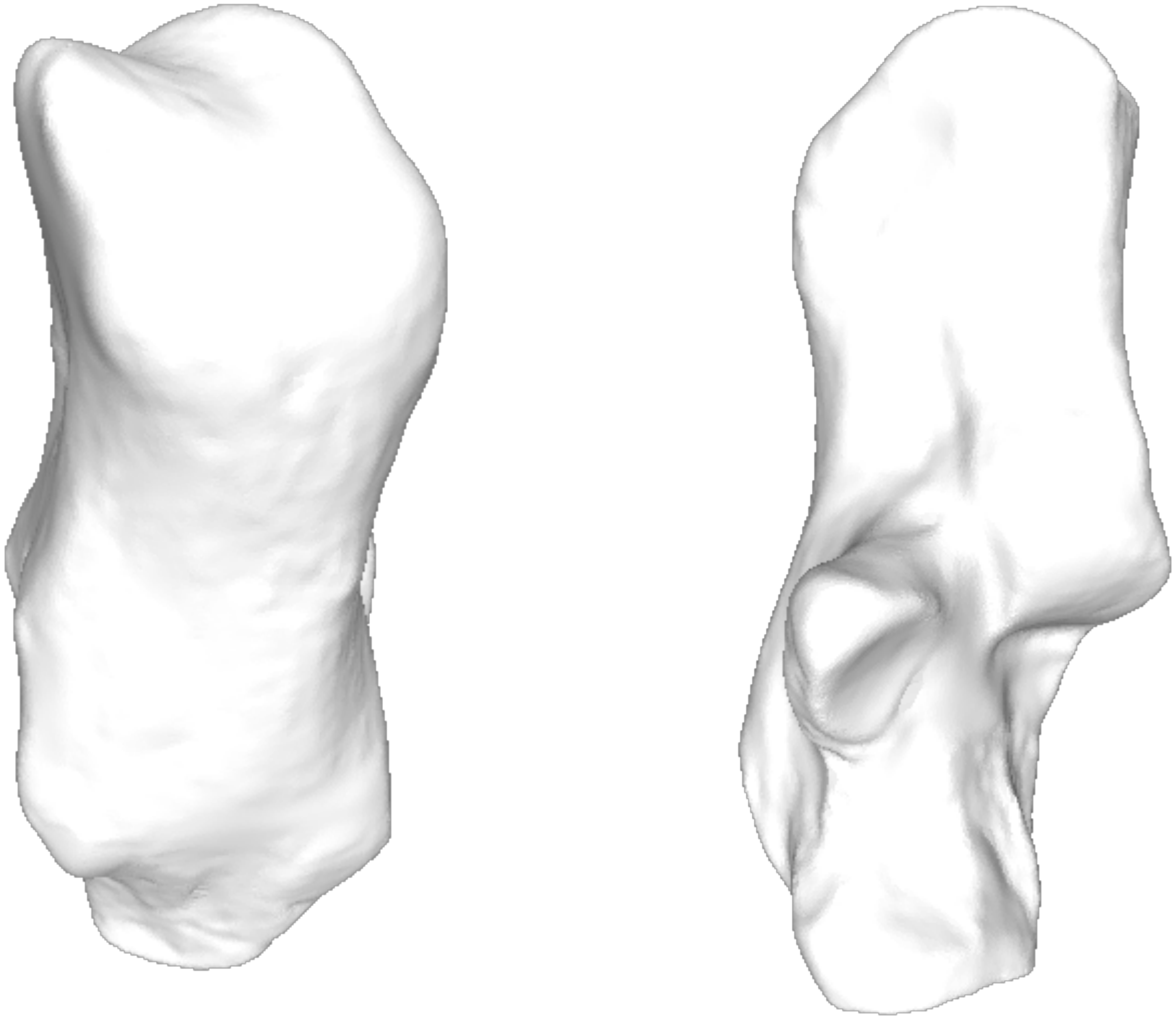
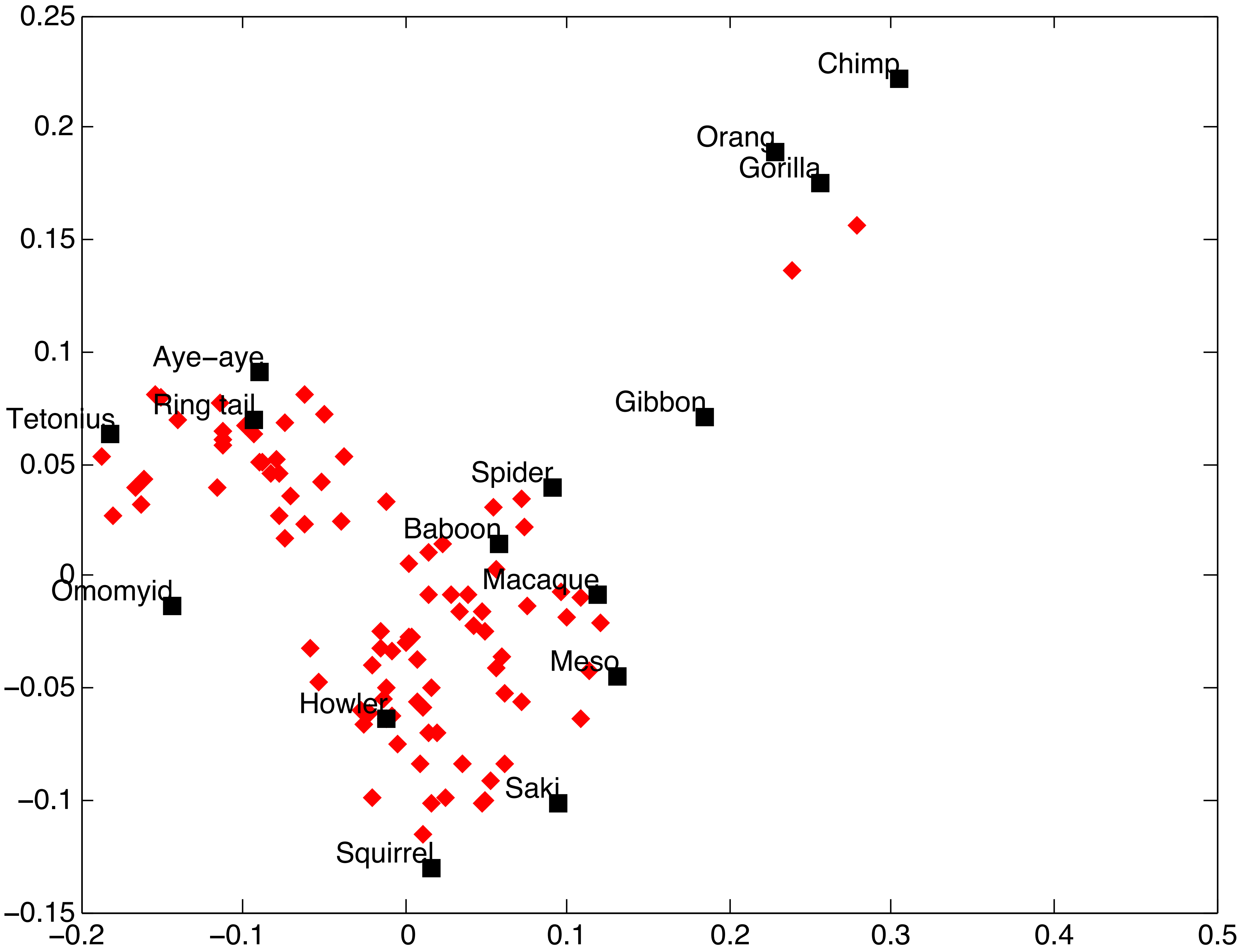


Figure : Images of a calcaneus from two different angles.

106 primates



Primate calcanei

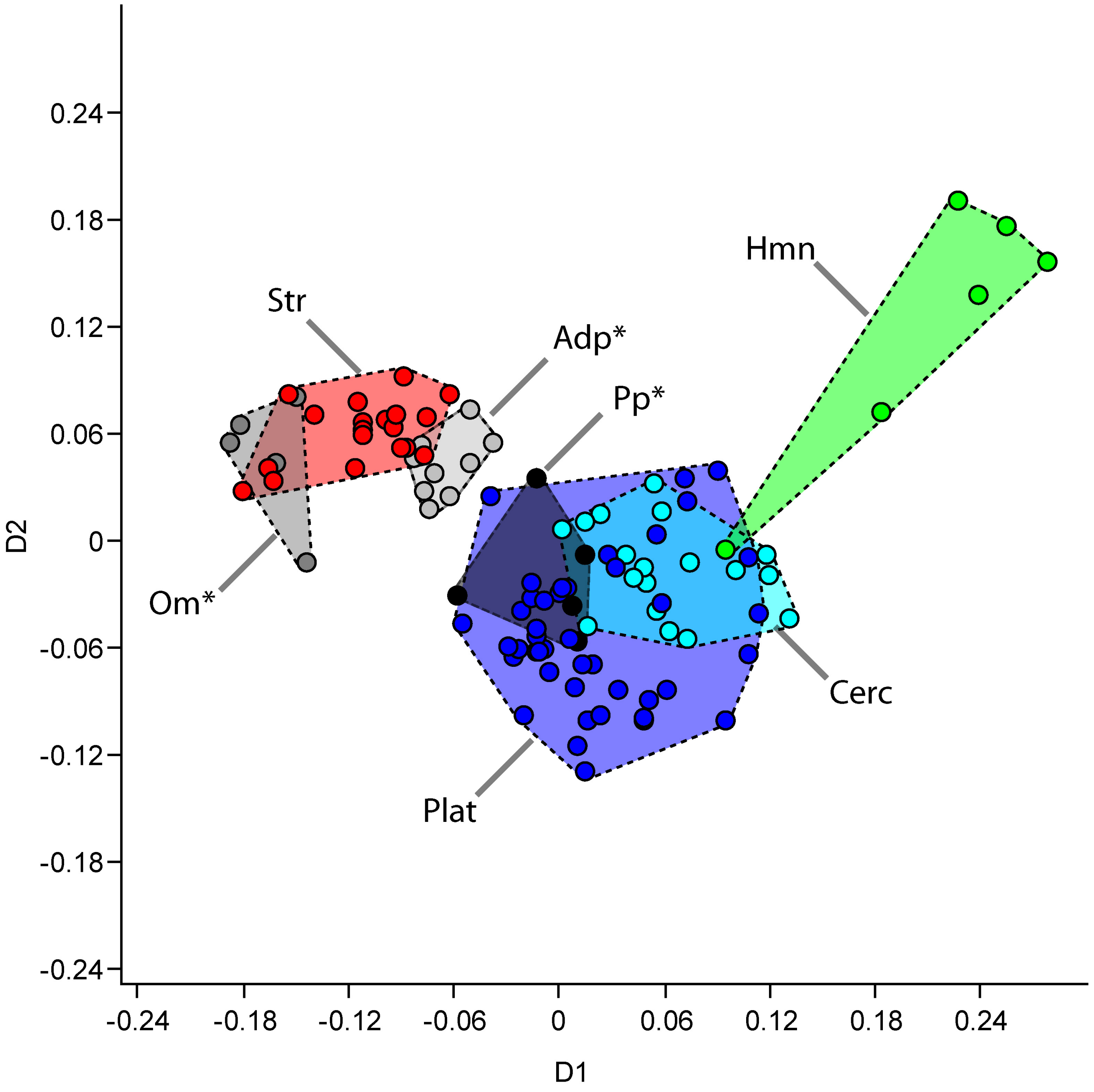
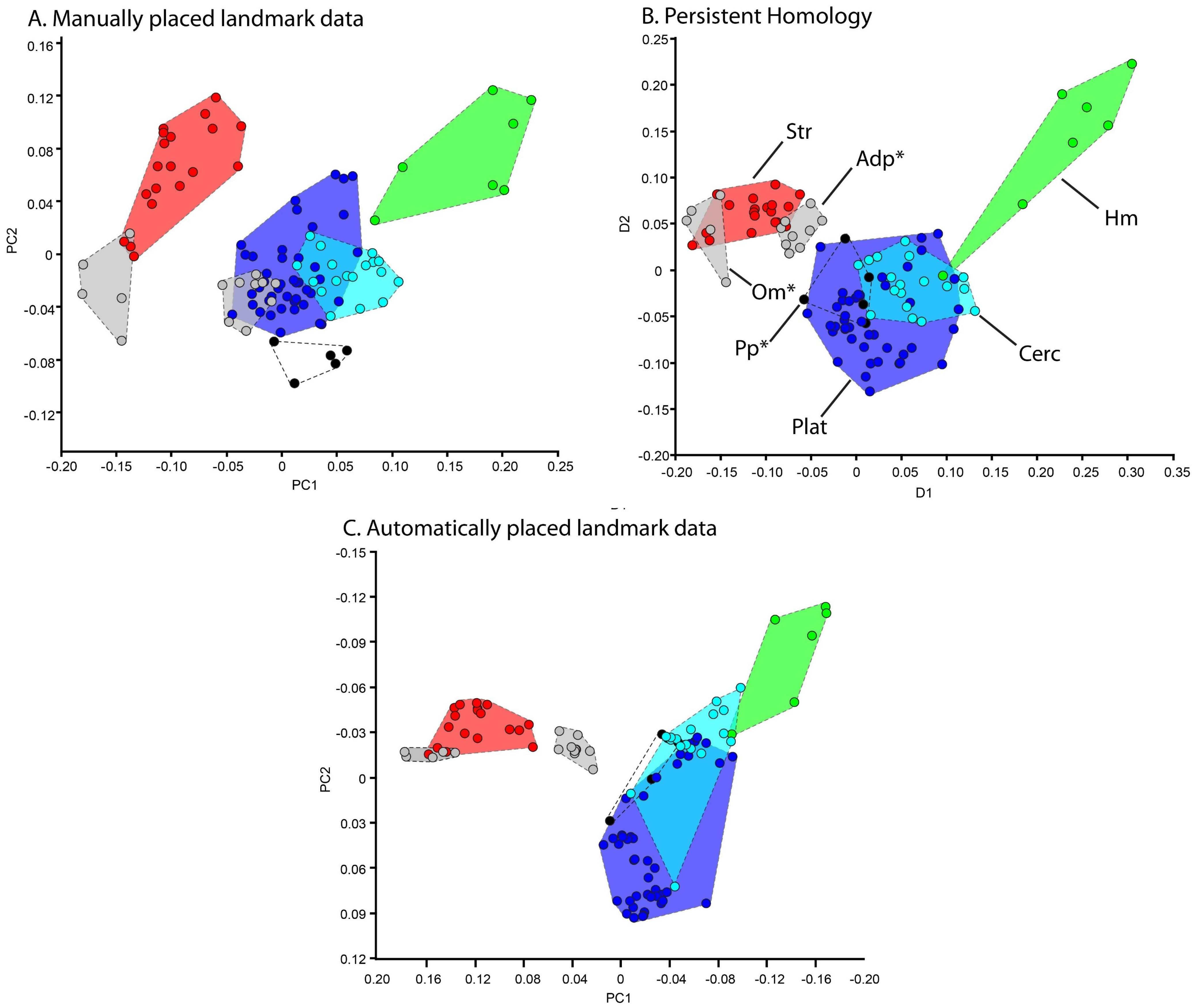


Figure : Phenetic clustering of phylogenetic groups of primate calcanei ($n = 106$). 67 genera are represented. Asterisks indicate groups of extinct taxa. Abbreviations: Str, Strepsirrhines; Plat, platyrrhines; Cerc, Cercopithecoids; Om, Omomyiforms; Adp, Adapiforms; Pp, parapithecids; Hmn, Hominoids. Note that more primitive prosimian taxa cluster separately from simians (Om, Adp, Str.). Also note that monkeys (Plat, Cerc, Pp) cluster mainly separately from apes (Hmn).

Comment from Doug

”In at least one way the method matched shapes with family groups better than any of the other previous methods... it linked a Hylobates specimen with the the other ape specimens (pan, gorilla, pongo, and oreopithecus). Previous both hylobatids (which ARE apes) always ended up closest to some Alouatta specimens.”

Comparing methods

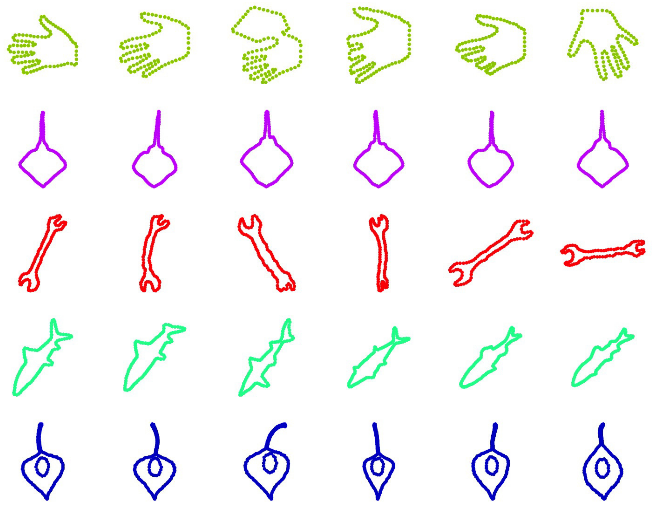


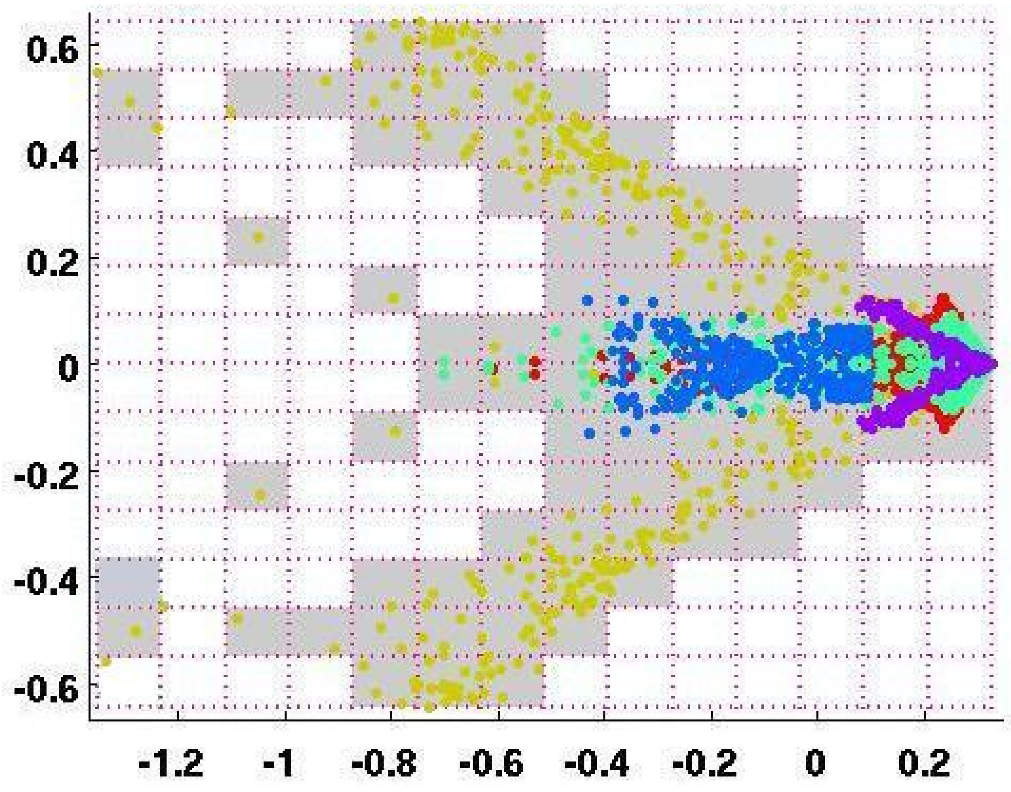
A. Manually placed landmark data

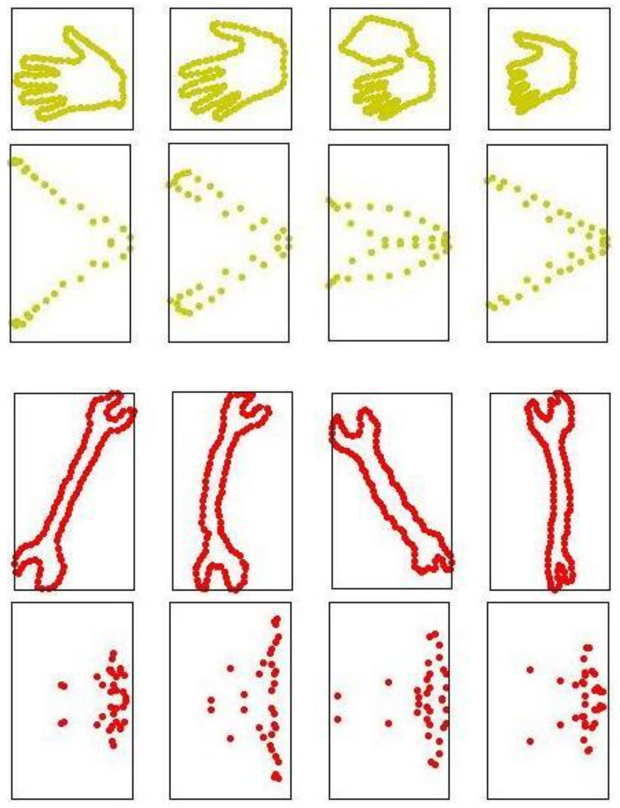
B. Persistent Homology

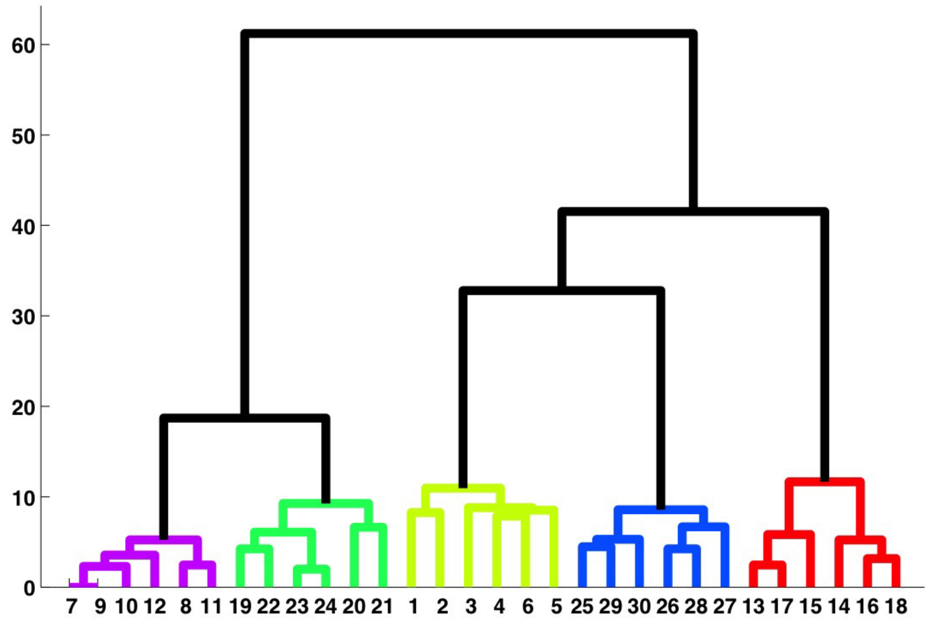
C. Automatically placed landmark data

A shape library

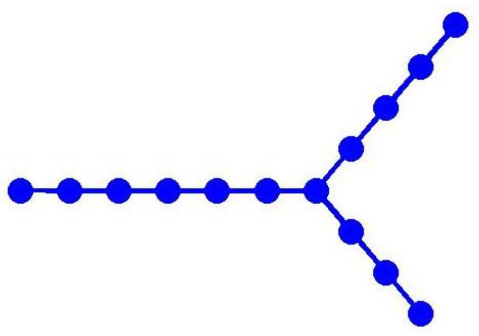


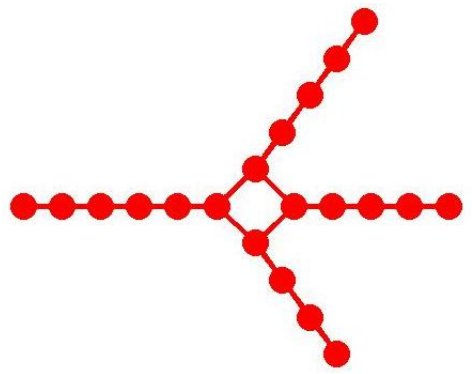


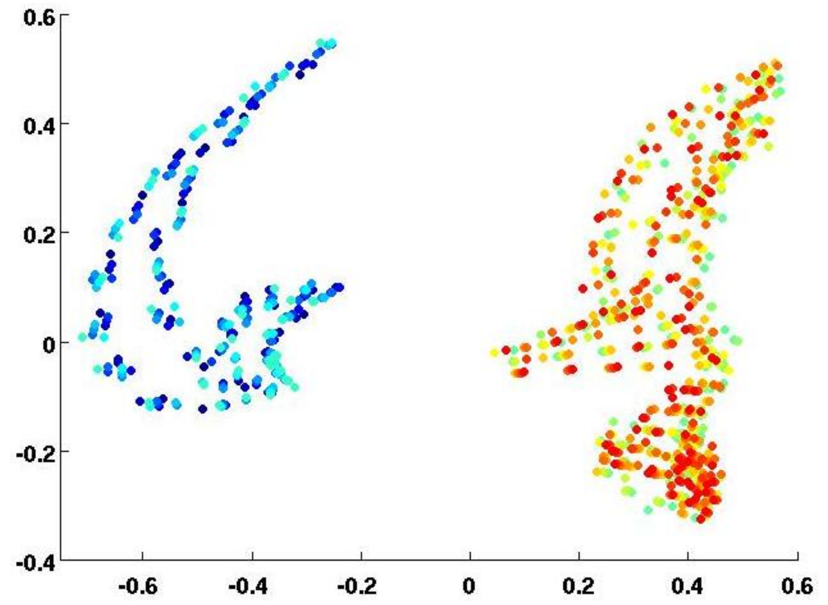




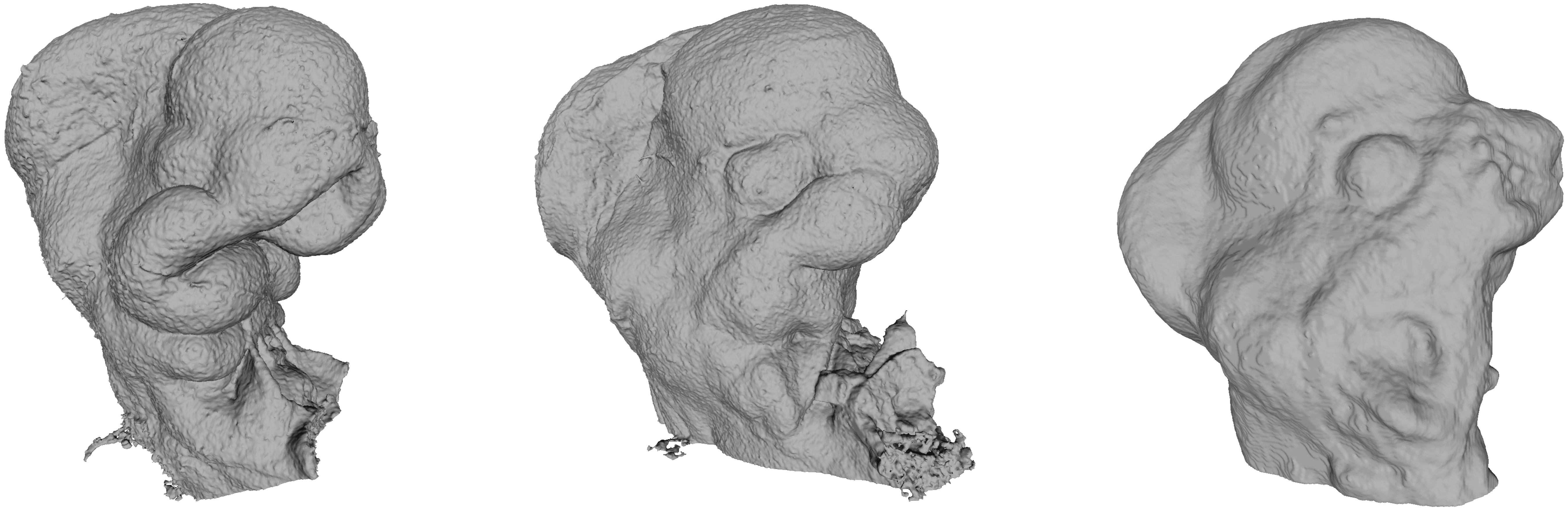
Can you hear the shape of a network ?



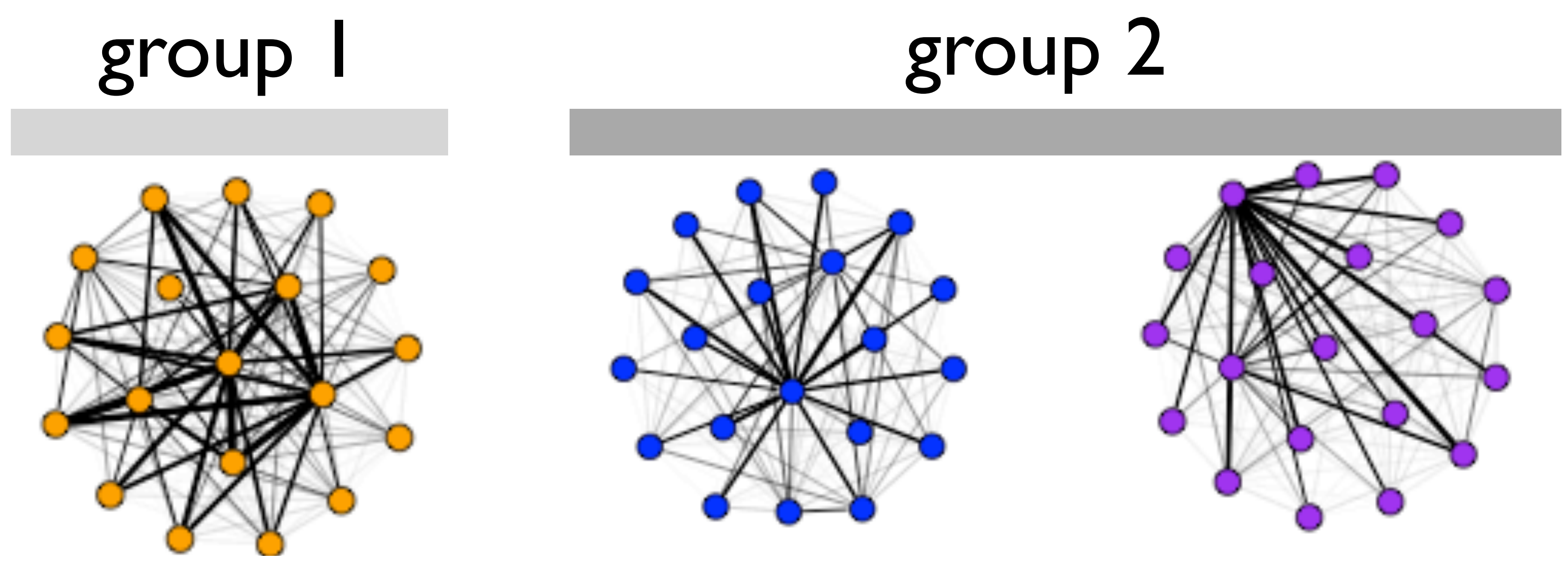


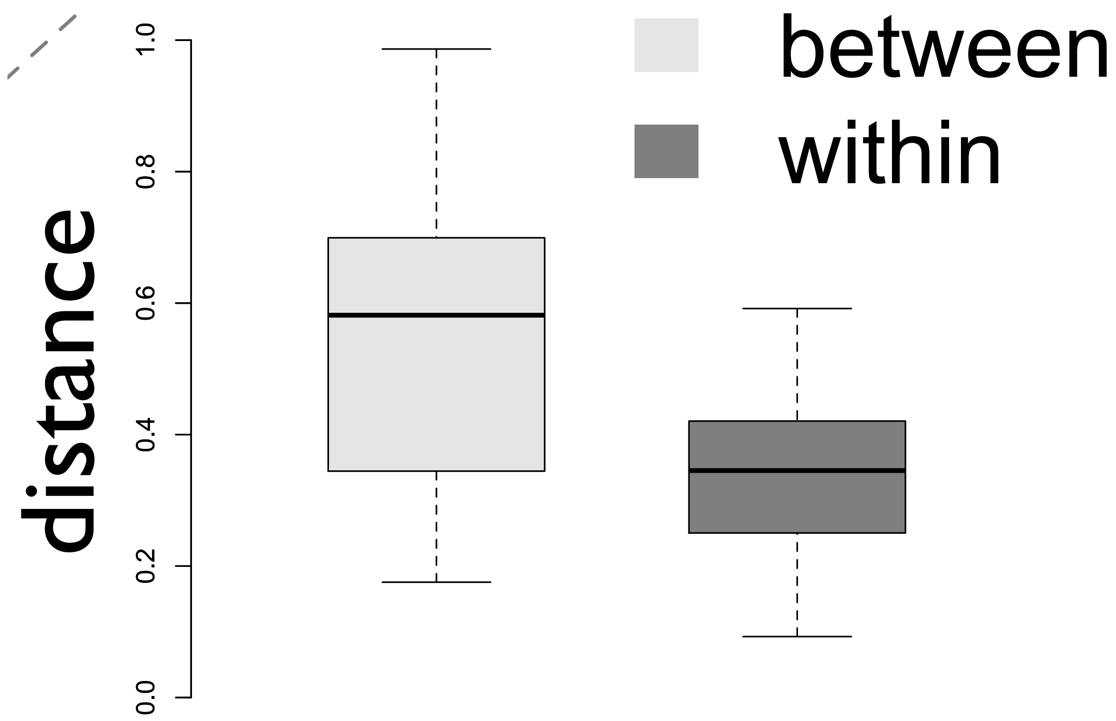


Association studies of shape phenotypes



Variation in baboon microbiome networks





Open problems

(1) Other transforms.

Open problems

(1) Other transforms.

(2) Sampling theory for surfaces.

Open problems

(1) Other transforms.

(2) Sampling theory for surfaces.

(3) Localized transforms.

Open problems

(1) Other transforms.

(2) Sampling theory for surfaces.

(3) Localized transforms.

(4) Statistical and quantitative genetics of shape traits.

Funding

- Center for Systems Biology at Duke
- NSF DMS and CCF
- DARPA
- AFOSR
- NIH