Systems genetics with graphical Markov models

Robert Castelo

robert.castelo@upf.edu

@robertclab

Dept. of Experimental and Health Sciences (DCEXS)
Universitat Pompeu Fabra (UPF)

upf.

Barcelona

Machine Learning for Personalized Medicine
Satellite Symposium of the ESHG Conference
Barcelona, May 19th, 2016

upf.

DCEXS/UPF is located at the Barcelona Biomedical Research Park (PRBB)

Joint work with

Inma Tur
Kernel Analytics, Barcelona

Alberto Roverato
University of Bologna

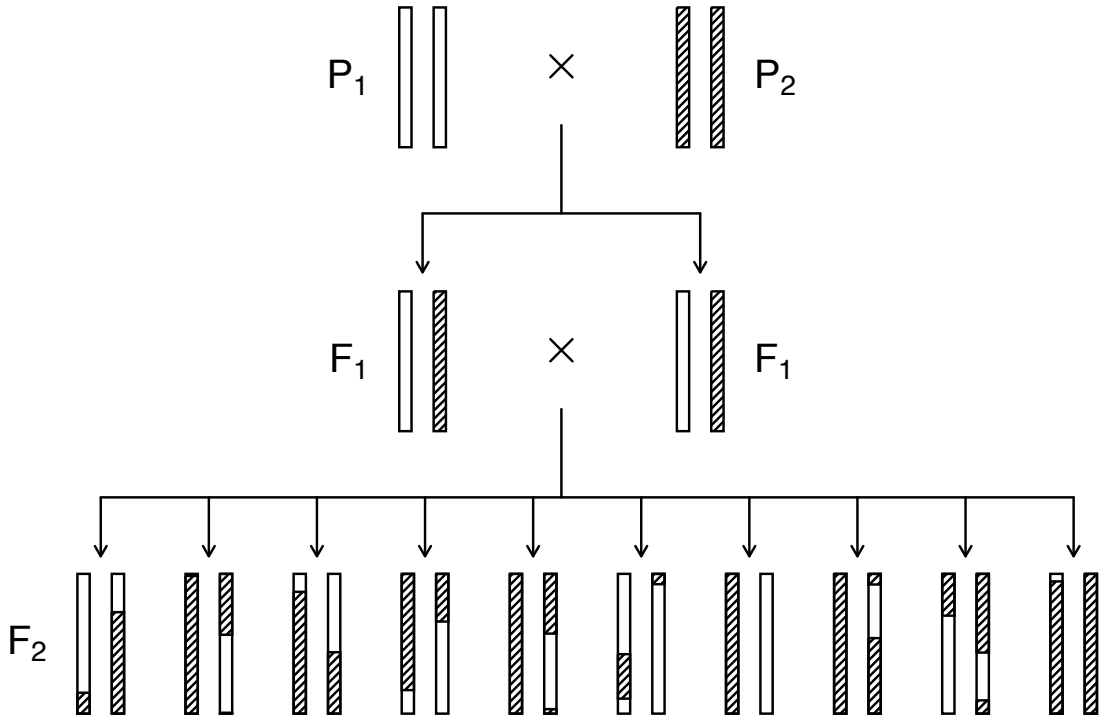
I. Tur, A. Roverato and R. Castelo. Mapping eQTL networks with mixed graphical Markov models. *Genetics*, 198(4):1377-1383, 2014. `http://arxiv.org/abs/1402.4547`

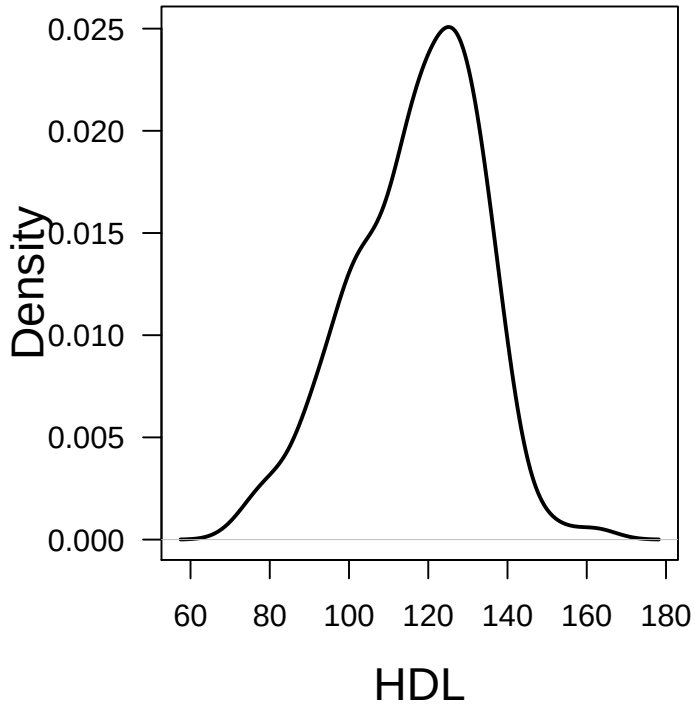
upf.

Motivation - Quantitative genetics

Primary goal: finding the genetic basis of complex (quantitative) higher-order phenotypes (traits).

Intercross (Fig. by Karl Broman in "Introduction to QTL mapping in model organisms")





Leduc *et al.* Using bioinformatics and systems genetics to dissect HDL-cholesterol genetics in an MRL/MpJ x SM/J intercross. *Journal of Lipid Research*, 53:1163-1175, 2012.

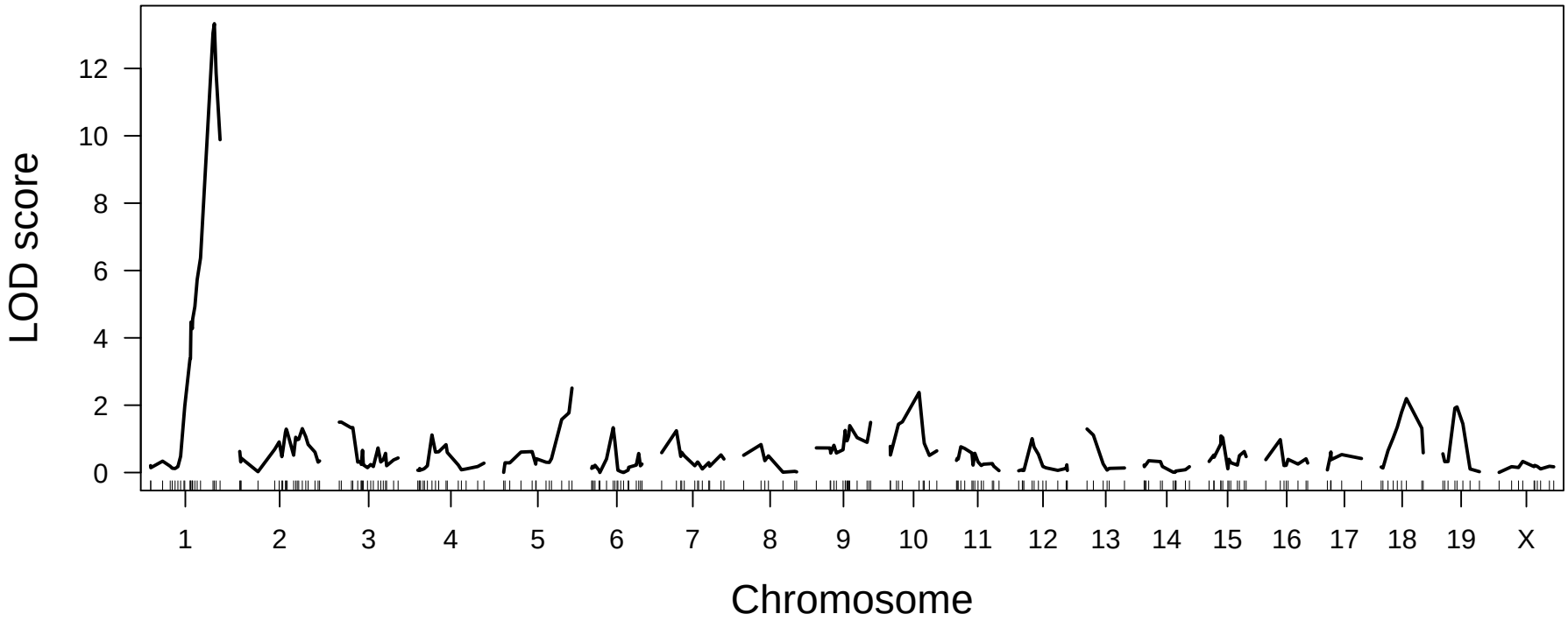
upf.

Motivation - Quantitative genetics

Find DNA sites along the genome associated to the phenotype, known as *quantitative trait loci* (QTLs). Simplest approach: regress phenotype on each marker (Soller, 1976), calculating the so-called logarithm of odds (LOD) score.

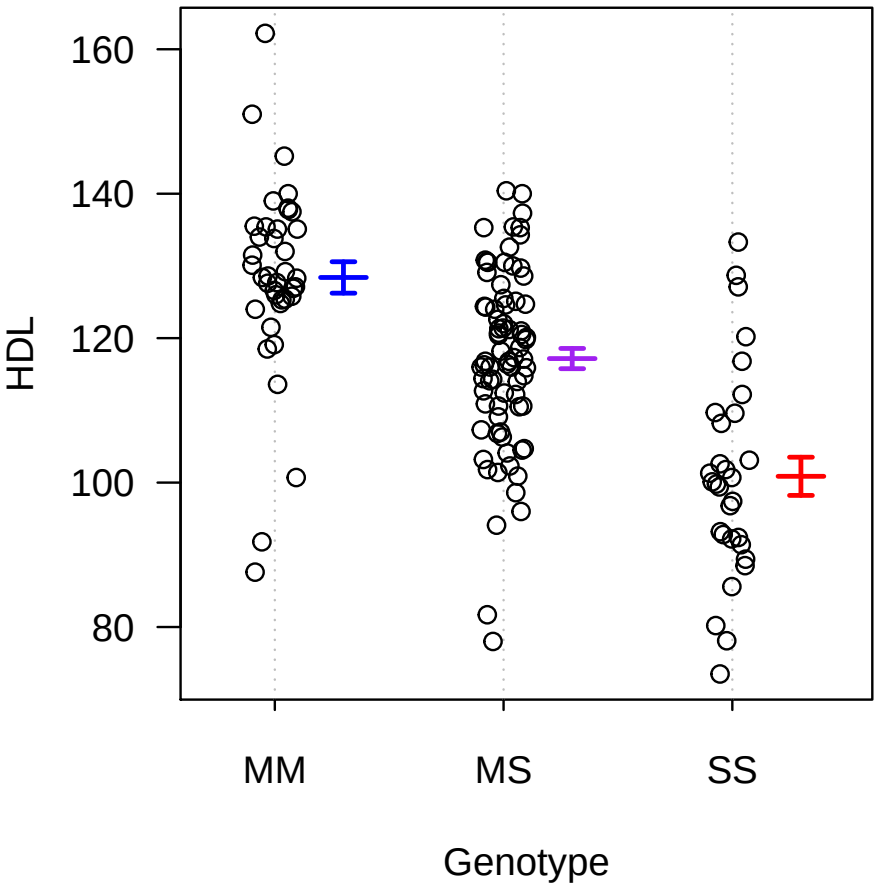
$$H_0 : y_i \sim \mathcal{N}(\mu_0, \sigma_0^2) \qquad H_1 : y_i | g_i \sim \mathcal{N}(\mu_{g_i}, \sigma_1^2)\,.$$

$$\text{LOD} = \log_{10} \frac{\mathcal{L}_1}{\mathcal{L}_0} = \frac{n}{2} \log_{10} \frac{\text{RSS}_0}{\text{RSS}_1}\,.$$



Motivation - Quantitative genetics

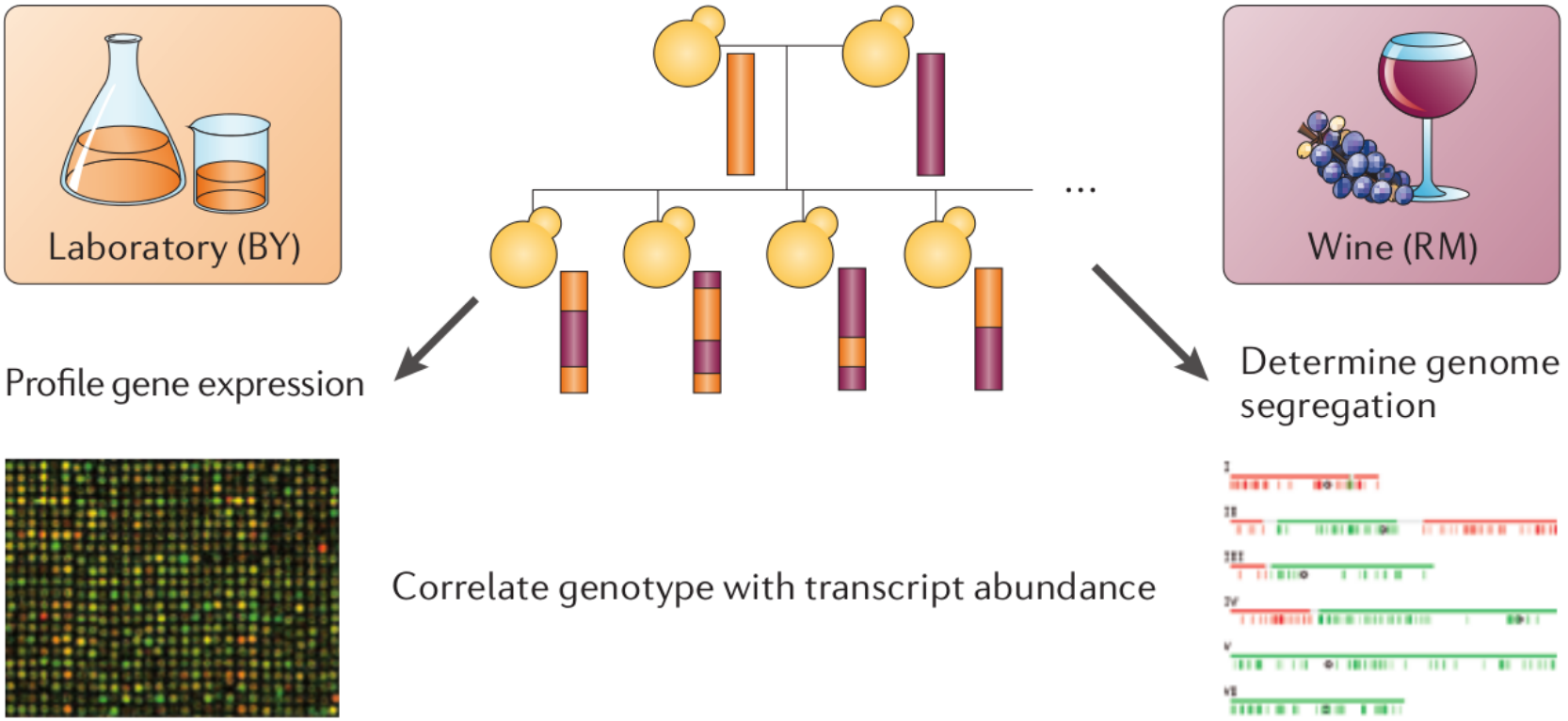
Estimate the effect size of found QTLs using, for instance, the percentage of variance explained by the QTL.



$$\eta^2 = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{(n-1)\cdot s_Y^2} = 0.346\,.$$

About 35% of the variability in HDL levels is explained by this QTL.

Motivation - Quantitative genetics on genomics data



Yeast BY x RM cross (Fig. by Rockman and Kruglyak, 2006). The resulting data published by Brem and Kruglyak (2005) consists of $\sim 6,000$ genes and $\sim 3,000$ genotype markers.

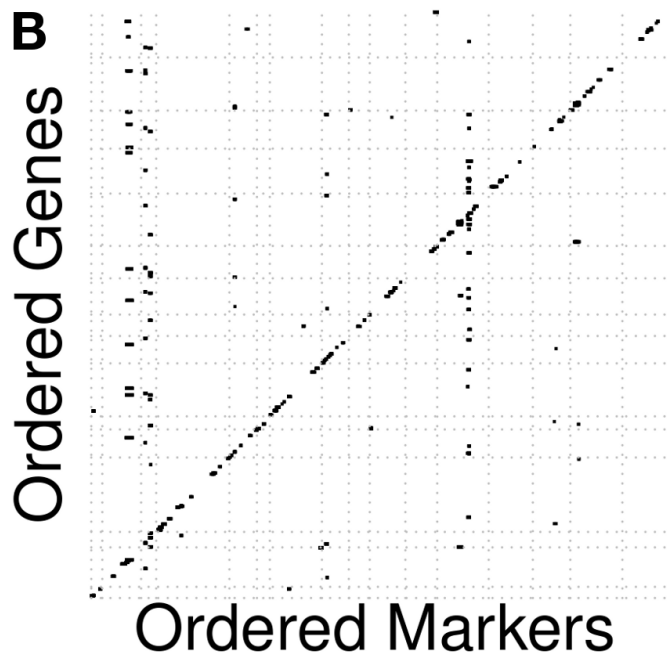
DNA sites along the genome associated to gene expression are called *expression QTLs* (eQTLs).

Motivation - Quantitative genetics on genomics data

Straightforward approach: apply classical QTL analysis methods independently on each gene expression profile (Soller, 1976):

$$\left.\begin{array}{rl} H_0 : y \sim & \mathcal{N}(\mu_0, \sigma_0^2) \\ H_1 : y|g \sim & \mathcal{N}(\mu_g, \sigma_1^2) \end{array}\right\} \quad \text{LOD} = \log_{10} \frac{\mathcal{L}_1}{\mathcal{L}_0} = \frac{n}{2} \log_{10} \frac{\text{RSS}_0}{\text{RSS}_1} .$$

Plot location of genome-wide significant eQTLs with respect to both, eQTL and gene genomic position (*dot plot*).



Motivation - Quantitative genetics on genomics data

- Let Γ denote the an index set for all genes with $p_\Gamma = |\Gamma|$ (thousands).
- Let n denote the number of profiled individuals (tens, hundreds).
- Let $Y = \{y_{ij}\}_{p_\Gamma \times n}$ denote the matrix of gene expression values with $p_\Gamma \gg n$:

Y	1	2	$\dots$	n
g_1	y_{11}	y_{12}	$\cdots$	y_{2n}
g_2	y_{21}	y_{22}	$\cdots$	y_{2n}
g_3	y_{31}	y_{32}	$\cdots$	y_{3n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
g_{p_Γ}	$y_{p_\Gamma 1}$	$y_{p_\Gamma 2}$	$\cdots$	$y_{p_\Gamma n}$

- Gene expression is a high-dimensional multivariate trait.

upf.

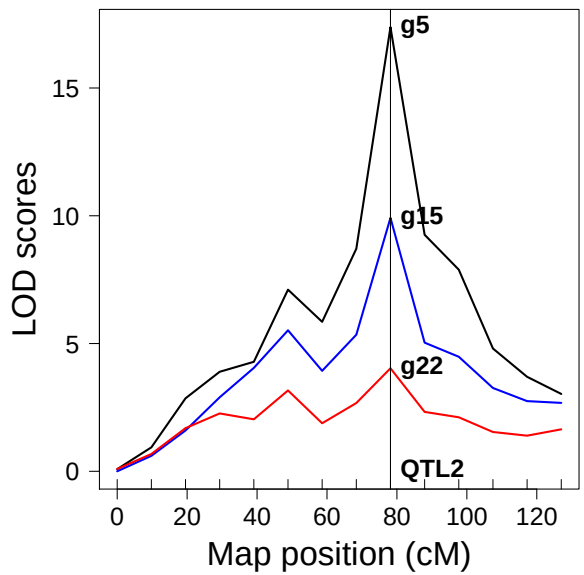
Motivation - Quantitative genetics on genomics data

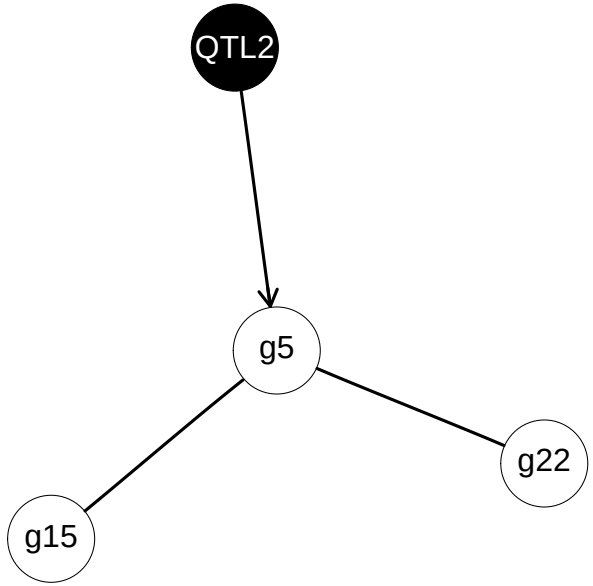
- Gene expression measurements by high-througthput instruments are the result of multiple types of **effects**:
  - **Genetic**: DNA polymorphisms affecting transcription initiation and RNA processing.
  - **Molecular**: RNA-binding events affecting post-transcriptional regulation (e.g., RNA degradation).
  - **Environmental**: response of the cell to external stimuli.
  - **Technical**: sample preparation protocols or laboratory conditions create sample-specific biases affecting most of the genes.
- All these effects render expression measurements in Y highly-correlated, thereby complicating the distinction between **direct** and **indirect** effects.

upf.

Motivation - Quantitative genetics on genomics data

Think of genes and eQTLs as forming a network, which we shall call an *eQTL network*.

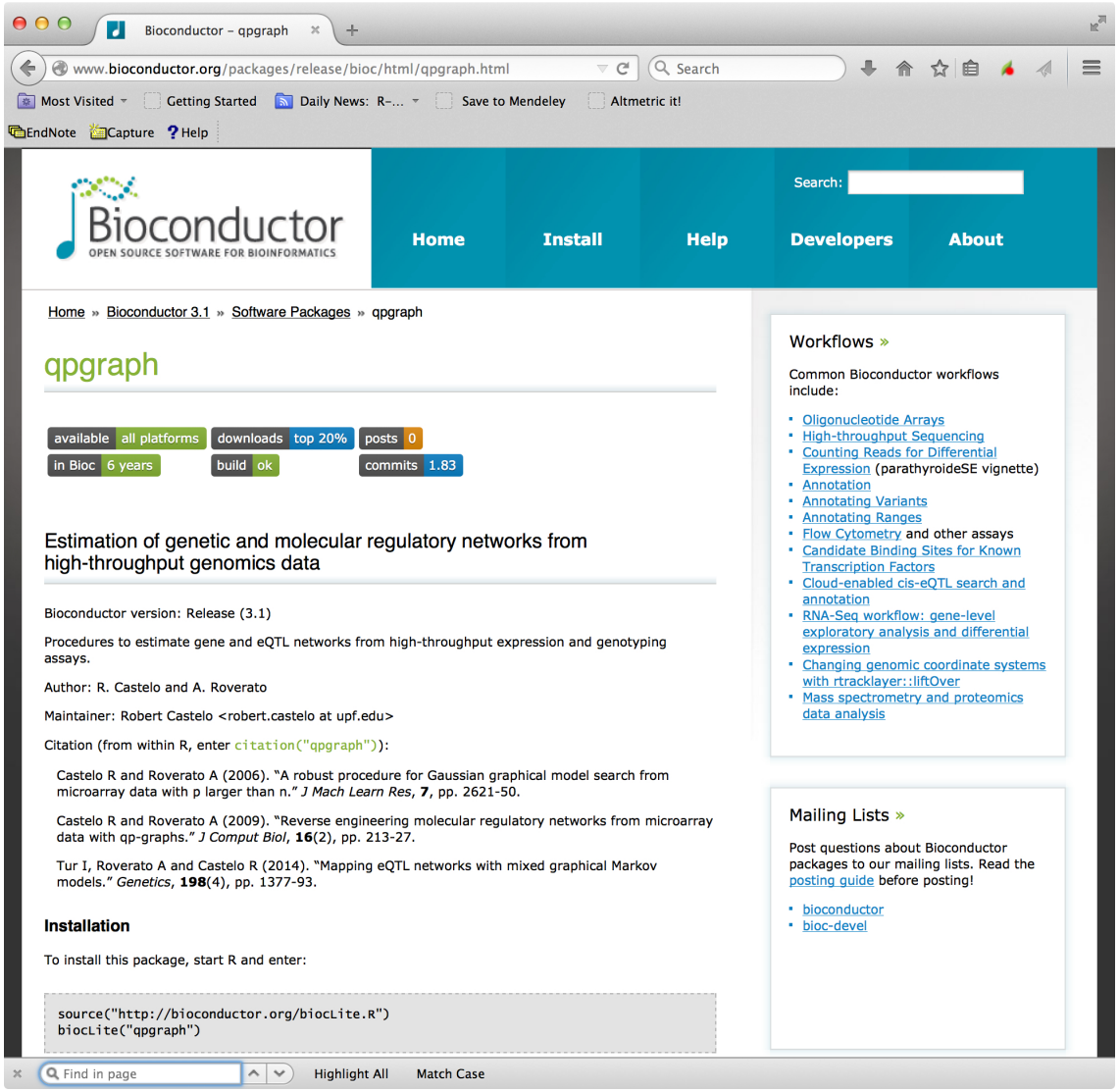




Assume that gene expression forms a p_Γ-multivariate sample following a conditional Gaussian distribution given the joint probability of all eQTLs

$\Longrightarrow$ mixed Graphical Markov model (Lauritzen and Wermuth, 1989)

Software availability: the R/Bioconductor package qpgraph



Available at `http://bioconductor.org/packages/qpgraph`

Outline

upf.

Outline

upf.

Overview of GMMs - undirected Gaussian GMMs

Let X_V be continuous r.v.'s and $G = (V, E)$ an undirected labeled graph:

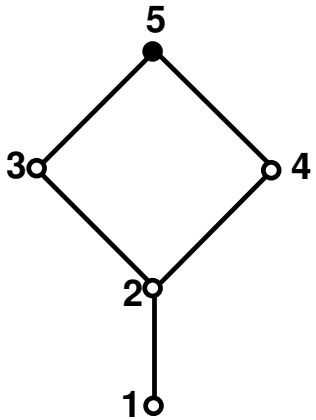
- $V = \{1, ..., p\}$ are the vertices of G
- $X_V \sim P(X_V) \equiv \mathcal{N}(\mu, \Sigma)$
- μ is the p-dimensional mean vector
- $\Sigma = \{\sigma_{ij}\}_{p \times p}$ is the covariance matrix
- $\Sigma^{-1} = \{\kappa_{ij}\}_{p \times p}$ is the concentration matrix
- Note that Pearson and partial correlation coefficients follow from scaling covariance (Σ) and concentration (Σ^{-1}) matrices, respectively:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \qquad \rho_{ij.R} = \frac{-\kappa_{ij}}{\sqrt{\kappa_{ii}\kappa_{jj}}}, R = V \backslash \{i, j\}\,.$$

upf.

Overview of GMMs - undirected Gaussian GMMs

- Let $G = (V, E)$ be an undirected graph with $V = \{1, \ldots, p\}$, a Gaussian graphical model can be described as follows:



$$\Sigma^{-1} = \begin{pmatrix} \kappa_{11} & \kappa_{12} & 0 & 0 & 0 \\ \kappa_{21} & \kappa_{22} & \kappa_{23} & \kappa_{24} & 0 \\ 0 & \kappa_{32} & \kappa_{33} & 0 & \kappa_{35} \\ 0 & \kappa_{42} & 0 & \kappa_{44} & \kappa_{45} \\ 0 & 0 & \kappa_{53} & \kappa_{54} & \kappa_{55} \end{pmatrix}$$

- A probability distribution $P(X_V)$ is undirected Markov w.r.t. G if

$$(i,j) \notin E \;\; \Rightarrow \;\; \kappa_{ij} = 0 \;\; \Leftrightarrow \;\; X_i \perp\!\!\!\perp X_j | X_V \backslash \{X_i, X_j\}$$

- These models are also known as covariance selection models (Dempster, 1972) or concentration graph models (Cox and Wermuth, 1996).
- Two vertices i and j are **separated** in G by a subset $S \subset V \backslash \{i,j\}$ iff every path between i and j intersects S, denoted hereafter by $i \perp_G j | S$.
- Global Markov property (Hammersley and Clifford, 1971):

$$i \perp_G j | S \Rightarrow X_i \perp\!\!\!\perp X_j | X_S \, .$$

upf.

Overview of GMMs - undirected Gaussian GMMs

Consider simulating an undirected Gaussian GMM by simulating a covariance matrix Σ such that

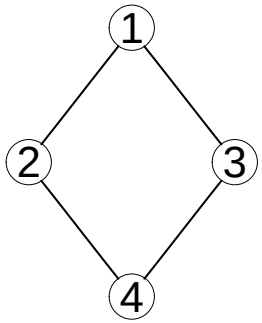
1. Σ is positive definite ($\Sigma \in S^+$),
2. the off-diagonal cells of the scaled Σ corresponding to the present edges in G match a given marginal correlation ρ,
3. the zero pattern of Σ^{-1} matches the missing edges in G.

This is not straightforward since setting directly off-diagonal cells to zero in some initial $\Gamma \in S^+$ will **not** typically lead to a positive definite matrix.

upf.

Overview of GMMs - undirected Gaussian GMMs

Let Γ^G be an *incomplete matrix* with elements $\{\gamma_{ij}\}$ for $i = j$ or $(i, j) \in G$.



$$\Gamma^G = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & * \\ \gamma_{21} & \gamma_{22} & * & \gamma_{24} \\ \gamma_{31} & * & \gamma_{33} & \gamma_{34} \\ * & \gamma_{42} & \gamma_{43} & \gamma_{44} \end{pmatrix}$$

Γ is a *positive completion* of Γ^G if $\Gamma \in S^+$ and $\{\Gamma^{-1}\}_{ij} = 0$ for $i \neq j$, $(i, j) \notin G$.

Draw Γ^G from a Wishart distribution $W_p(\Lambda, p)$; $\Lambda = \Delta R \Delta$, $\Delta = \text{diag}(\{\sqrt{1/p}\}_p)$ and $R = \{R_{ij}\}_{p \times p}$ where $R_{ij} = 1$ for $i = j$ and $R_{ij} = \rho$ for $i \neq j$.

It is required that $\Lambda \in S^+$ and this happens if and only if $-1/(p-1) < \rho < 1$.

Finally, to obtain $\Sigma \equiv \Gamma$ from Γ^G, `qpgraph` uses the regression algorithm by Hastie, Tibshirani and Friedman (2009, pg. 634) as matrix completion algorithm.

upf.

Overview of GMMs - mixed GMMs

- Let Δ denote the set of vertices indexing discrete r.v.'s $I_\delta, \delta \in \Delta$.
- Let Γ denote the set of vertices indexing continuous r.v.'s $Y_\gamma, \gamma \in \Gamma$.
- Let $G = (V, E)$ be a graph with marked vertices $V = \Delta \cup \Gamma$, where $p_\Delta = |\Delta|$, $p_\Gamma = |\Gamma|$, $p = p_\Delta + p_\Gamma$, and E be the edge set.
- Vertices in V index the r.v.'s $X = (I, Y)$, where Y correspond to genes, I to markers or eQTLs, and the joint sample space of X is denoted by,

$$x = (i, y) = \{(i_\delta)_{\delta \in \Delta}, (y_\gamma)_{\gamma \in \Gamma}\} ,$$

  where i_δ denote discrete genotype alleles with $i \in \mathcal{I}$, and y_γ denote continuous expression values.

- Assume $y \sim \mathcal{N}_{|\Gamma|}(\mu(i), \Sigma(i))$ with moment parameters $(p(i), \mu(i), \Sigma(i))$,

$$f(x) = f(i, y) = p(i)|2\pi\Sigma(i)|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(y - \mu(i))^T \Sigma(i)^{-1}(y - \mu(i))\right\} .$$

upf.

Overview of GMMs - mixed GMMs

- $p(i)$ is the probability that $I = i$, and $\mu(i)$ and $\Sigma(i)$ are the conditional mean and conditional covariance matrix of Y.
- If the covariance matrix is constant across $i \in \mathcal{I}$, i.e., $\Sigma(i) \equiv \Sigma$, then the model is *homogeneous*. Otherwise, the model is said to be *heterogeneous*.
- We can write the logarithm of the density in terms of the canonical parameters $(g(i), h(i), K(i))$:

$$\log f(i, y) = g(i) + h(i)^T y - \frac{1}{2} y^T K(i) y \,,$$

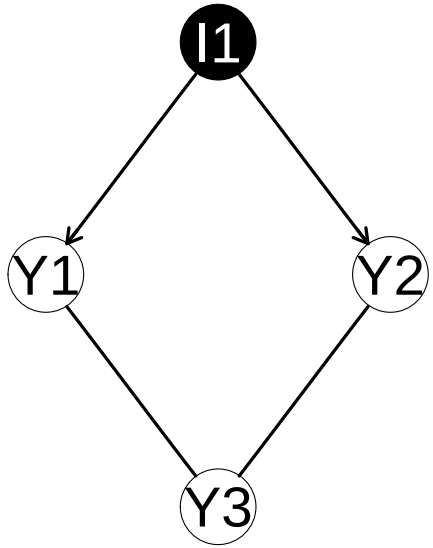
where

$$\begin{aligned} g(i) &= \log(p(i)) - \frac{1}{2}\log|\Sigma(i)| - \frac{1}{2}\mu(i)^T\Sigma(i)^{-1}\mu(i) - \frac{|\Gamma|}{2}\log(2\pi)\,, \\ h(i) &= \Sigma(i)^{-1}\mu(i)\,, \\ K(i) &= \Sigma(i)^{-1}\,. \end{aligned}$$

Overview of GMMs - mixed GMMs

Simplifying assumptions (in the context of genetical genomics data):

1. Discrete genotypes affect gene expression and not the other way around.
2. Joint distribution of X is a conditional Gaussian distribution $X_V \sim \mathcal{N}_{p_Y}(\mu(i), \Sigma(i))$ with $i \in \mathcal{I}$.
3. Genotype alleles affect only mean expression levels of genes and **not** the correlations between them, i.e., $\Sigma(i) \equiv \Sigma$ is *constant* throughout $i \in \mathcal{I}$.
4. Discrete r.v.'s are simulated as being marginally independent between them.
5. Every continuous r.v. cannot depend on more than one discrete r.v.



upf.

Overview of GMMs - mixed GMMs

- Given a suitable covariance matrix Σ, under $\Sigma(i) \equiv \Sigma$, we can calculate conditional mean vectors $\mu(i)$ as function of the canonical parameters $h(i)$,

$$\mu(i) = \Sigma \cdot h(i) \, .$$

- Simulate $h(i)$ assuming genotypes with two possible alleles and independent eQTLs given an additive effect $a_{\delta\gamma} = \mu_\gamma(1) - \mu_\gamma(2)$ of an eQTL I_δ on a gene Y_γ.

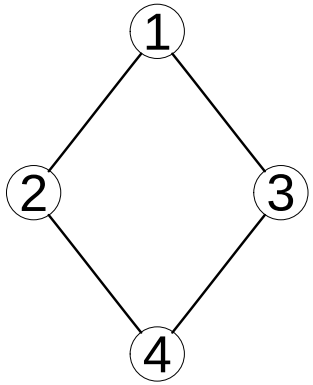
- Full details in Tur, Roverato and Castelo. Mapping eQTL networks with mixed graphical Markov models. *Genetics*, 198(4):1377-1383, 2014.

upf.

Overview of GMMS - simulation using qpgraph

Gaussian GMMs

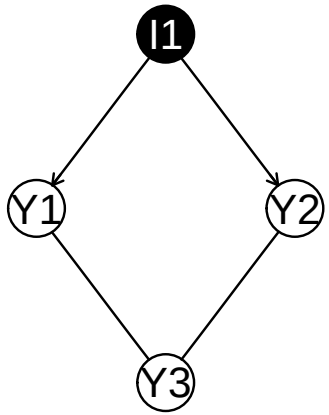
$$X_V \sim \mathcal{N}_p(\mu, \Sigma)$$




```
> library(qpgraph)
> set.seed(12345)
> gmm <- rUGgmm(dRegularGraphParam())
> round(solve(gmm$sigma), digits=1)

     1    2    3    4
1  9.5 -3.4 -7.2  0.0
2 -3.4  5.9  0.0 -2.3
3 -7.2  0.0  8.2  0.9
4  0.0 -2.3  0.9  2.3

> plot(gmm)
```

Homogeneous Mixed GMMs

$$X_V \sim \mathcal{N}_p(\mu(i), \Sigma(i)) \text{ with } \Sigma(i) \equiv \Sigma$$



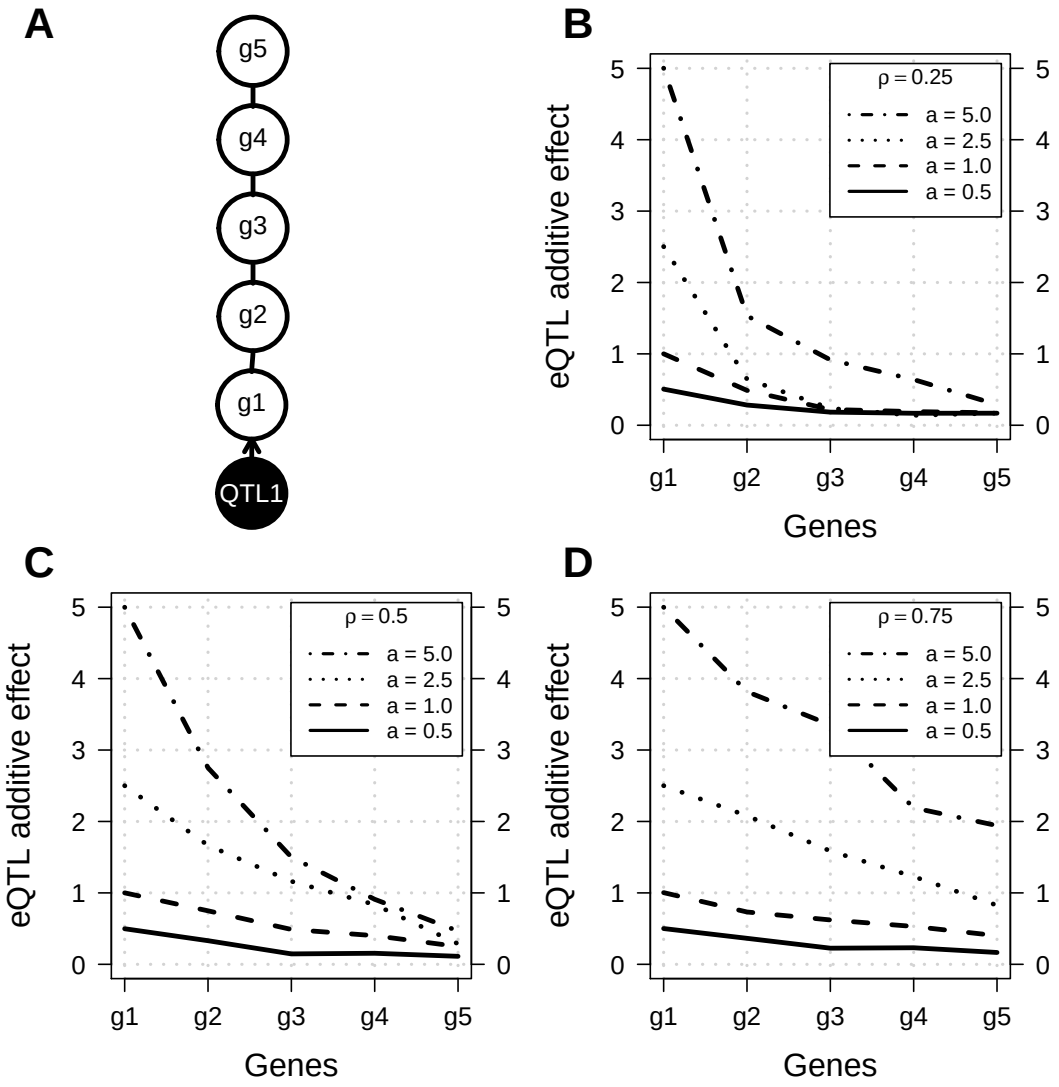

```
> library(qpgraph)
> set.seed(12345)
> gmm <- rHMgmm(dRegularMarkedGraphParam())
> round(solve(gmm$sigma), digits=1)

     Y1   Y2   Y3
Y1 11.0  0.0 -7.2
Y2  0.0  1.2 -1.6
Y3 -7.2 -1.6  8.2

> gmm$mean()

         Y1        Y2        Y3
1 0.4720734 0.9669291 0.7242007
2 1.4720734 1.9669291 1.7934027

> plot(gmm)
```

upf.

Outline

upf.

Propagation of eQTL (genetic) additive effects



eQTL additive effects propagate proportionally to marginal correlations ρ between genes.

Outline

upf.

Conditional independence in mixed GMMs

- Classical ($p \gg n$) approach: use conditional independence to distinguish direct from indirect eQTL associations,

$$X_\delta \perp\!\!\!\perp X_\gamma | X_{V\setminus\{\delta,\gamma\}}\,, \quad \delta \in \Delta, \gamma \in \Gamma,$$

- and direct from indirect gene-gene associations,

$$X_\gamma \perp\!\!\!\perp X_\zeta | X_{V\setminus\{\gamma,\zeta\}} \quad \gamma, \zeta \in \Gamma.$$

- For $\Sigma \equiv \Sigma(i)$, the log-likelihood ratio statistics are (Lauritzen, 1996):

$$D_{\delta\gamma.\,V\setminus\{\delta,\gamma\}} = -2\ln\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right) = -2\ln\left(\frac{|ssd_\Gamma||ssd_{\Gamma^*}(\Delta^*)|}{|ssd_{\Gamma^*}||ssd_\Gamma(\Delta^*)|}\right)^{n/2},$$

$$D_{\gamma\zeta.\,V\setminus\{\gamma,\zeta\}} = -2\ln\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right) = -2\ln\left(\frac{|ssd_\Gamma||ssd_{\Gamma\setminus\{\gamma,\zeta\}}|}{|ssd_{\Gamma\setminus\{\gamma\}}||ssd_{\Gamma\setminus\{\zeta\}}|}\right)^{n/2},$$

respectively, where $\Gamma^* = \Gamma\setminus\{\gamma\}$ and $\Delta^* = \Delta\setminus\{\delta\}$.

upf.

Conditional independence in mixed GMMs

- The likelihood function $\mathcal{L}_1$ for the homogeneous, saturated model attains its maximum if and only if $n \geq |\Gamma| + |\mathcal{I}|$. Unfortunately, since $p \gg n$, we cannot directly test for full-order conditional independence.

- However, MLEs exist for limited-order conditional independences given subsets of genes Q such that $|Q| < (n-2)$.

- Let X_α and X_γ, with $\gamma \in \Gamma$ and let $Q \subset \Gamma$. If Q separates α from γ in the underlying G we can find this out by testing whether $X_\alpha \perp\!\!\!\perp X_\gamma | X_Q$.

- Assume $V = \{\alpha, \gamma, Q\}$. Saturated and constrained models differ in one single edge. This makes them decomposable and collapsible onto $X_{V\setminus\{\gamma\}}$:

$$f_V = f_{\gamma|V\setminus\{\gamma\}} \cdot f_{V\setminus\{\gamma\}}\,,$$

leading to $\mathcal{L}_0 = \mathcal{L}^0_{\gamma|V\setminus\{\gamma\}} \cdot \mathcal{L}^0_{V\setminus\{\gamma\}}$ and $\mathcal{L}_1 = \mathcal{L}^1_{\gamma|V\setminus\{\gamma\}} \cdot \mathcal{L}^1_{V\setminus\{\gamma\}}$.

upf.

Conditional independence in mixed GMMs

- Since $\mathcal{L}^0_{V\setminus\{\gamma\}} = \mathcal{L}^1_{V\setminus\{\gamma\}}$, we can calculate the pure continuse case as,

$$D_{\gamma\zeta.Q} = -2\ln\left(\frac{\mathcal{L}^0_{\gamma|\,V\setminus\{\gamma\}}}{\mathcal{L}^1_{\gamma|\,V\setminus\{\gamma\}}}\right) = -2\ln\left(\frac{\hat{\sigma}^0_{\gamma|\,V\setminus\{\gamma\}}}{\hat{\sigma}^1_{\gamma|\,V\setminus\{\gamma\}}}\right)^{-n/2},$$

  where $\hat{\sigma}^0_{\gamma|\,V\setminus\{\gamma\}} = \mathrm{RSS}_0$ and $\hat{\sigma}^1_{\gamma|\,V\setminus\{\gamma\}} = \mathrm{RSS}_1$, and therefore,

$$D_{\gamma\zeta.Q} = -2\ln\left(\frac{\mathrm{RSS}_1}{\mathrm{RSS}_0}\right)^{n/2} = -2\ln(\Lambda_{\gamma\zeta.Q})^{n/2},$$

  which follows asymptotically a χ^2_{df} with $df = 1$.

- Analogously, the mixed case can be written as,

$$D_{\delta\gamma.Q} = -2\ln\left(\frac{\mathrm{RSS}_1}{\mathrm{RSS}_0}\right)^{n/2} = -2\ln(\Lambda_{\delta\gamma.Q})^{n/2},$$

  which follows asymptotically a χ^2_{df} with $df = |\mathcal{I}_{\Delta^*}|(|\mathcal{I}_\delta| - 1)$.

upf.

Conditional independence in mixed GMMs

From the relationship between χ^2_k and gamma $\Gamma(k/2, 2)$ distributions (Rao, 1973; Lauritzen, 1996) it can be shown that,

$$\begin{aligned}
\Lambda_{\gamma\zeta.Q} &\sim B\left(\frac{n - |\Gamma| - |\mathcal{I}| + 1}{2}, \frac{1}{2}\right) \\
\Lambda_{\delta\gamma.Q} &\sim B\left(\frac{n - |\Gamma| - |\mathcal{I}| + 1}{2}, \frac{|\mathcal{I}_{\Delta^*}|(|\mathcal{I}_\delta| - 1)}{2}\right),
\end{aligned}$$

exactly. Likewise, using the relationship between the beta and F distributions (Rao, 1973) we can also calculate the F-statistics

$$\begin{aligned}
F_{\gamma\zeta.Q} &= \frac{1}{n - |\Gamma| - |\mathcal{I}| + 1} \cdot \frac{\Lambda_{\gamma\zeta.Q}}{1 - \Lambda_{\gamma\zeta.Q}}, \\
F_{\delta\gamma.Q} &= \frac{|\mathcal{I}_{\Delta^*}|(|\mathcal{I}_\delta| - 1)}{n - |\Gamma| - |\mathcal{I}| + 1} \cdot \frac{\Lambda_{\delta\gamma.Q}}{1 - \Lambda_{\delta\gamma.Q}},
\end{aligned}$$

which, again in terms of mixed GMM parameters, follow exactly

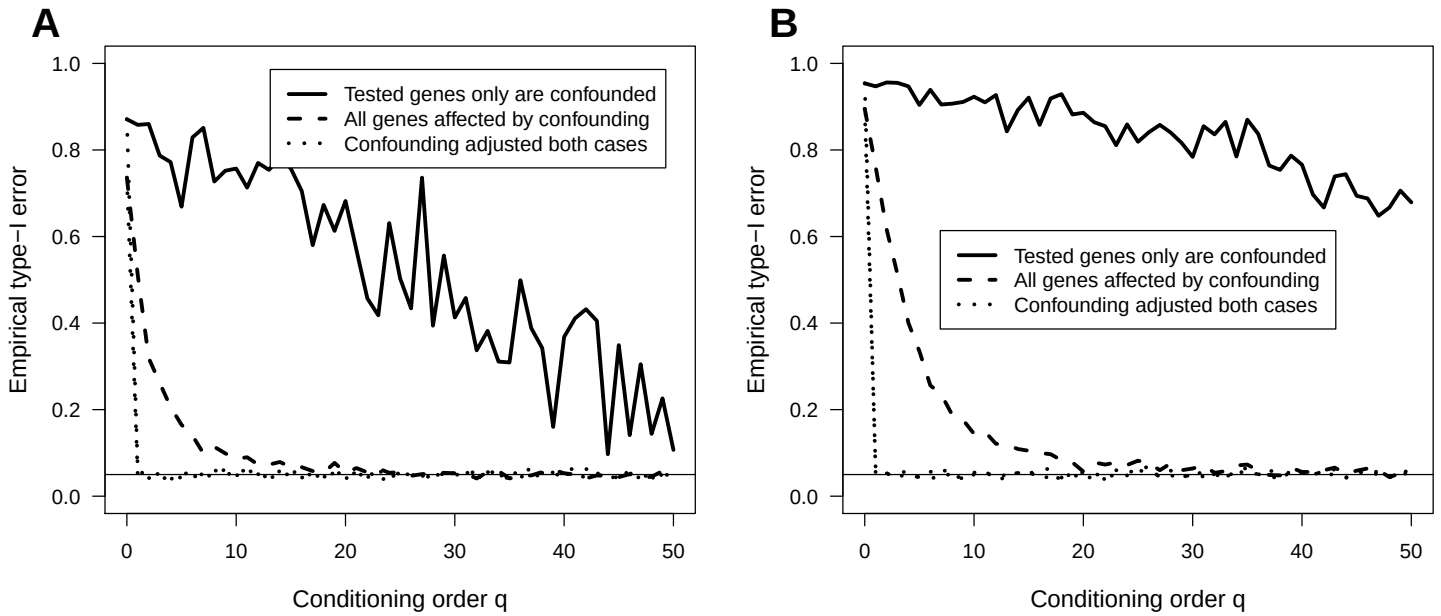
$$\begin{aligned}
F_{\gamma\zeta.Q} &\sim F(1, n - |\Gamma| - |\mathcal{I}| + 1), \\
F_{\delta\gamma.Q} &\sim F(|\mathcal{I}_{\Delta^*}|(|\mathcal{I}_\delta| - 1), n - |\Gamma| - |\mathcal{I}| + 1).
\end{aligned}$$

upf.

Conditional independence in mixed GMMs

- Confounding effects in expression data affecting all genes can be implicitly adjusted by conditoning on higher-order associations.

- Simulate an eQTL network with 100 disconnected genes, where one of them has an one eQTL with $a = 2.5$. Include a continuous confounding factor either affecting all genes or affecting only the two genes, or the gene and the marker, being tested, with $\rho = 0.5$. Sample data sets with $n = 100$.



Outline

upf.

q-order correlation graphs

- We would like to use full-order conditional independence to estimate the direct association between two genes, or a genotype marker and a gene, adjusting for every other gene and intervining factor.

- We cannot use directly full-order conditional indpendence because in our data $p \gg n$, and moreover, p is of very high-dimension.

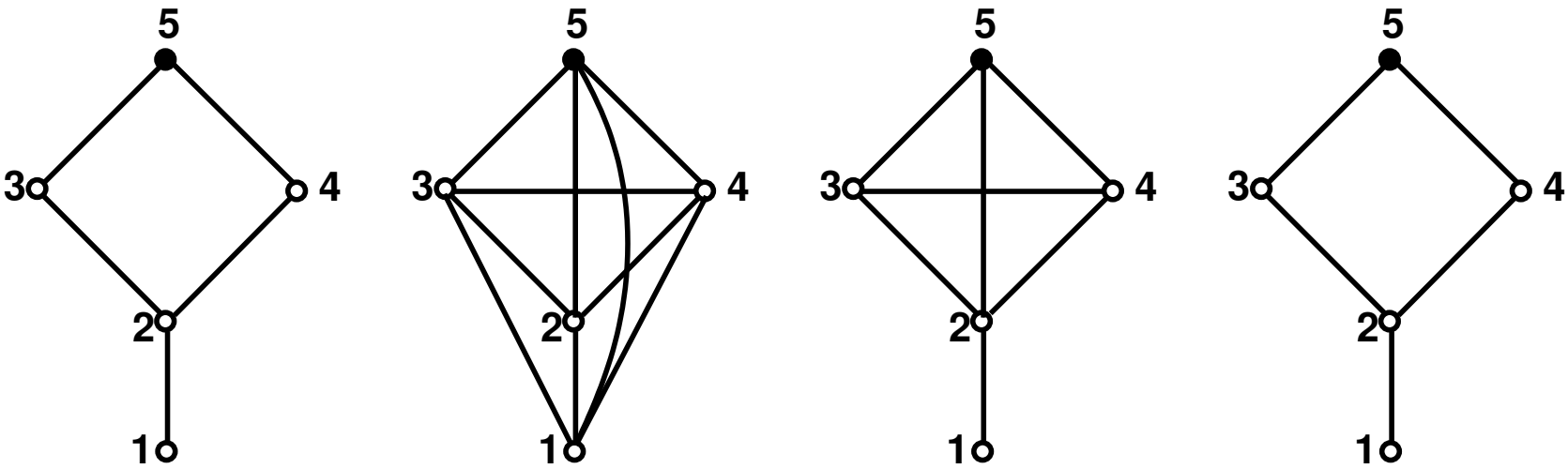
- Observation: the underlying molecular and functional relationships are **sparse**, that is, the fraction of interactions present in a specific cellular state under study is much smaller than the total number of possible interactions.

upf.

q-order correlation graphs

- If the underlying G is **sparse**, we can expect to explain many of the indirect associations by conditioning on subsets Q with $|Q| = q$ and $q < (n-2)$.

- The mathematical object that results from testing q-order correlations is called a q-order correlation graph, or qp-graph (Castelo and Roverato, 2006), and is denoted by $G^{(q)} = (V, E^{(q)})$.



G $G^{(0)}$ $G^{(1)}$ $G^{(2)}$

q-order correlation graphs

- To estimate $G^{(q)}$ we use a quantity called the *non-rejection rate* (NRR).
- Let $\mathcal{Q}_{ij}^q = \{Q \subseteq V \backslash \{i,j\} : |Q| = q\}$ and let T_{ij}^q be a binary r.v. associated to the pair of vertices (i,j) that takes values from the following three-step procedure:
  1. A subset Q is sampled from $\mathcal{Q}_{ij}^q$ uniformly at random.
  2. Test the null hypothesis of conditional independence $H_0 : X_i \perp\!\!\!\perp X_j | X_Q$.
  3. If H_0 is rejected then T_{ij}^q takes value 0, otherwise takes value 1.
- T_{ij}^q follows a Bernoulli distribution and the NRR, denoted as ν_{ij}^q, is defined as its expectancy

$$\nu_{ij}^q := \mathrm{E}[T_{ij}^q] = \Pr(T_{ij}^q = 1)\,.$$

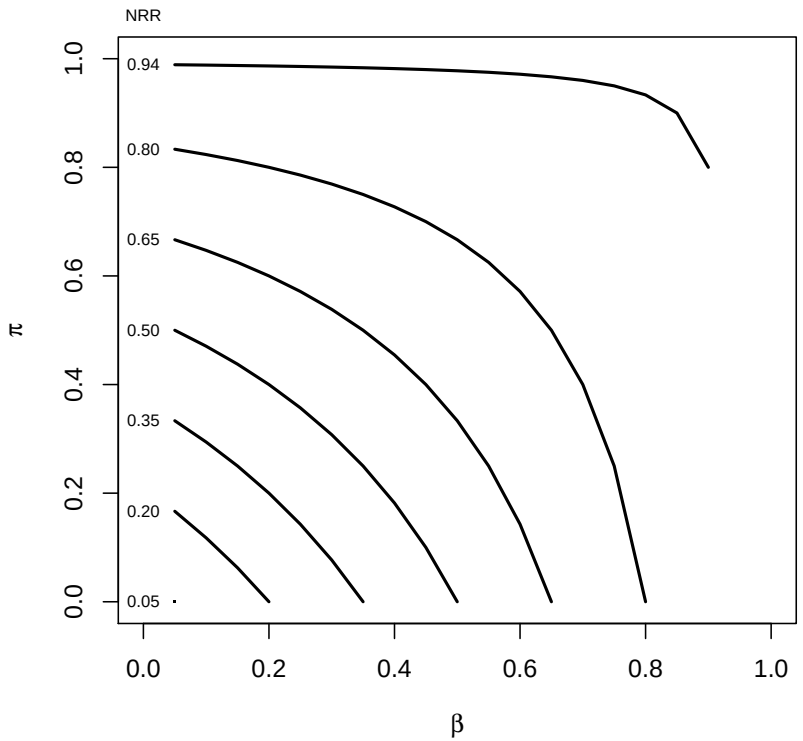
upf.

q-order correlation graphs

It can be shown (Castelo and Roverato, 2006) that the theoretical NRR is,

$$\nu_{ij}^{q} = \beta_{ij}(1 - \pi_{ij}^{q}) + (1 - \alpha)\pi_{ij}^{q} \,,$$

where π_{ij}^{q} is the fraction of vertex subsets of size q separating vertices i and j in G, α is the significance level of the tests and β_{ij} is the average value of the type-II error throughout the tests between vertices i and j.



q-order correlation graphs

- An estimate $\hat{\nu}_{ij}^{q}$ of the NRR can be obtained by testing $X_i \perp\!\!\!\perp X_j | X_Q$ for every $Q \in \mathcal{Q}_{ij}^{q}$.

- However, since $|\mathcal{Q}_{ij}^{q}|$ can be prohibitively large, we use a limited number of subsets $Q \in \mathcal{Q}_{ij}^{q}$, such as one-hundred, sampled uniformly at random.

- We can also explicitly adjust for confounding factors and other covariates $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ by sampling from

$$\mathcal{Q}_{ij.\mathcal{C}}^{q} = \{Q \subseteq \{V \backslash \{i,j\}\} \cup \mathcal{C} : \mathcal{C} \subseteq Q \text{ and } |Q| = q\}.$$

- A qp-graph estimate $\hat{G}_{\epsilon}^{(q)}$ can be obtained by selecting edges (i,j) that meet a maximum cutoff value ϵ:

$$\hat{G}_{\epsilon}^{(q)} := \{(V, E^{(q)}) : (i,j) \in E^{(q)} \Leftrightarrow \hat{\nu}_{ij}^{q} < \epsilon\}.$$

upf.

Outline

upf.

A three-step estimation strategy for eQTL networks

We propose to use conditional independence and q-order correlation graphs to estimate eQTL networks in a strategy consisting of three steps:

1. Estimate the qp-graph $G^{(0)}$ under some standard framework such as the null hypothesis of no-eQTL at each marker (correcting p-values by multiple testing), or under the global null hypothesis of no-eQTL anywhere in the genome (calculating p-values by permutation).

2. Estimate a qp-graph $G^{(q)} \subseteq G^{(0)}$ for one or more q values and restrict edges in $G^{(0)}$ to those also present in $G^{(q)}$.

3. Among eQTLs in $G^{(q)} \subseteq G^{(0)}$ that are in the same chrosomosome and target a common gene, perform a forward-selection strategy at some significance level α, to discard redundant associations tagging the same causal eQTL.

upf.

A three-step estimation strategy - data simulation

- We will illustrate this three-step estimation strategy with simulated data.
- Simulate genetic map with 9 chromosomes, 10 markers per chromosome.

```
> detach("package:qpgraph") ## remove qpgraph from R's search path
> library(GenomeInfoDb)     ## to enable a correct overlaading of
> library(qtl)              ## the R/qtl function sim.cross() by
> library(qpgraph)          ## the qpgraph package
> map <- sim.map(len=rep(100, times=9),
+                n.mar=rep(10, times=9),
+                anchor.tel=FALSE,
+                eq.spacing=TRUE,
+                include.x=FALSE)
```

- Simulate eQTL network with 50 genes, 25 have local eQTLs and 5 eQTL hotspots *trans*-acting (distant) on 5 other genes. Each gene is also connected to other two genes.

```
> set.seed(12345)
> sim.eqtl <- reQTLcross(eQTLcrossParam(map=map, genes=50, cis=0.5, trans=rep(5, 5)),
+                        a=2, rho=0.5)
```

- Simulate data from this eQTL network model.

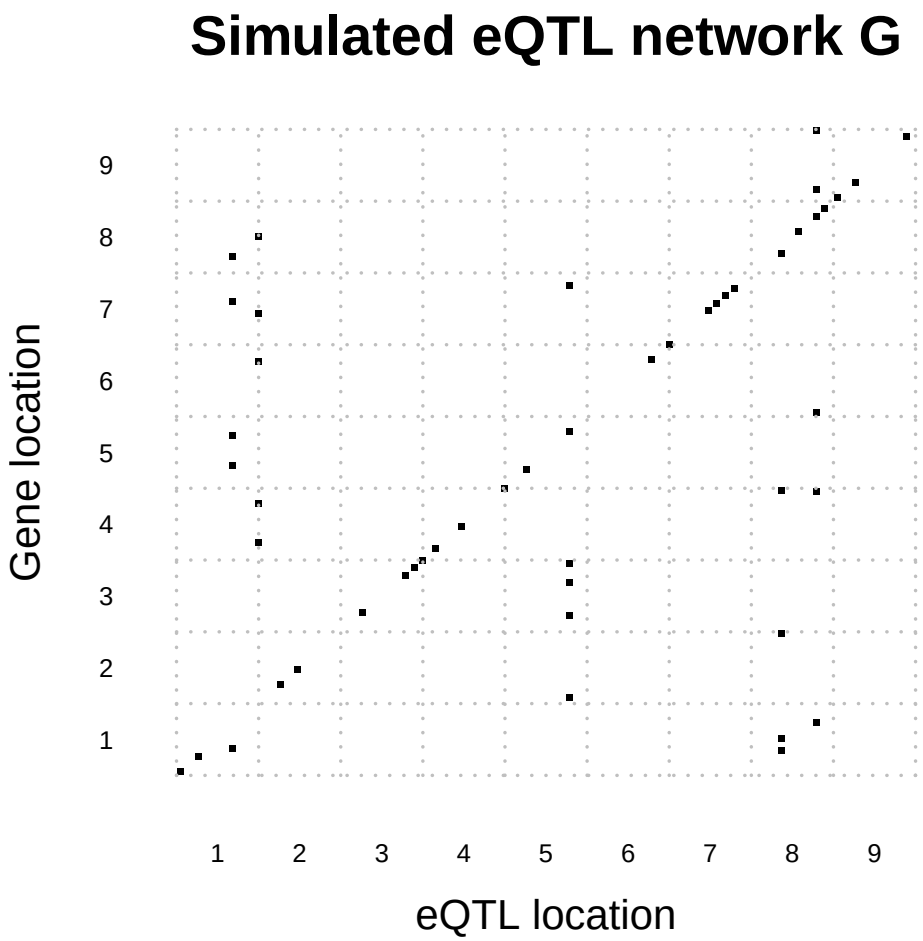
```
> set.seed(12345)
> cross <- sim.cross(map, sim.eqtl, n.ind=100)
```

upf.

A three-step estimation strategy - data simulation

Display the dot plot of the simulated eQTL associations.

```
> plot(sim.eqtl, main="Simulated eQTL network G", cex.lab=1.5, cex.main=2)
```



A three-step estimation strategy - parameter setup

- Pull the gene annotation from the simulated eQTL network object.

```
> annot <- data.frame(chr=as.character(sim.eqtl$genes[, "chr"]),
+                     start=sim.eqtl$genes[, "location"],
+                     end=sim.eqtl$genes[, "location"],
+                     strand=rep("+", nrow(sim.eqtl$genes)),
+                     row.names=rownames(sim.eqtl$genes),
+                     stringsAsFactors=FALSE)
```

- Translate the simulated cM positions to physical positions using a fixed rate of 5 Kb/cM.

```
> pMap <- lapply(map, function(x) x * 5)
> class(pMap) <- "map"
> annot$start <- floor(annot$start * 5)
> annot$end <- floor(annot$end * 5)
```

- Create a *Seqinfo* object of the simulated genome describing its chromosome names and lengths using the 5 Kb/cM rate.

```
> genome <- Seqinfo(seqnames=names(map), seqlengths=rep(100 * 5, nchr(pMap)),
+                   NA, "simulatedGenome")
```

- Create a parameter object of class *eQTLnetworkEstimationParam*.

```
> param <- eQTLnetworkEstimationParam(cross, physicalMap=pMap,
+                                     geneAnnotation=annot, genome=genome)
```

upf.

A three-step estimation strategy - first step

- Calculate all marginal associations between markers and genes.

```
> eqtlnet.q0 <- eQTLnetworkEstimate(param, ~ marker + gene, verbose=FALSE)
> eqtlnet.q0
```

```
eQTLnetwork object:
  Genome: simulatedGenome
  Input size: 90 markers 50 genes
  Model formula: ~marker + gene
```

- Obtain a first estimate $G^{(0)}$ of the eQTL network by selecting associations at FDR < 0.05.

```
> eqtlnet.q0.fdr <- eQTLnetworkEstimate(param, estimate=eqtlnet.q0,
+                                       p.value=0.05, method="fdr")
> eqtlnet.q0.fdr
```

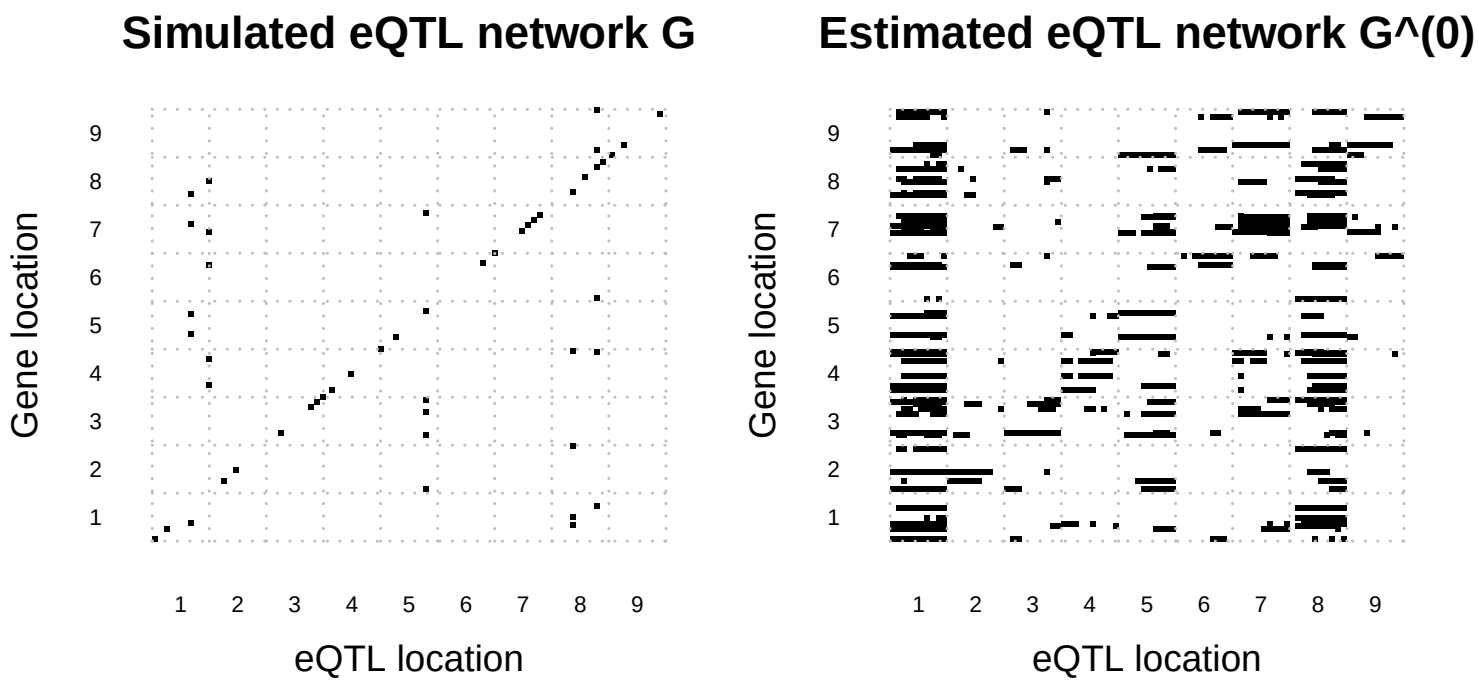
```
eQTLnetwork object:
  Genome: simulatedGenome
  Input size: 90 markers 50 genes
  Model formula: ~marker + gene (q = 0,)
  G^(0,): 140 vertices and 1996 edges corresponding to
          1015 eQTL and 981 gene-gene associations meeting
          a fdr-adjusted p-value < 0.05
          and involving 50 genes and 87 eQTLs
```

upf.

A three-step estimation strategy - first step

$G^{(0)}$ contains all marginal associations with FDR < 0.05.

```
> par(mfrow=c(1, 2))
> plot(sim.eqtl, main="Simulated eQTL network G", cex.lab=1.5, cex.main=1.8)
> plot(eqtlnet.q0.fdr, main="Estimated eQTL network G^(0)", cex.lab=1.5, cex.main=1.8)
```



A three-step estimation strategy - second step

- Calculate NRR values ν_{ij}^{q} with $q = 3$ between markers and genes.

```
> eqtlnet.q0.fdr.nrr <- eQTLnetworkEstimate(param, ~ marker + gene | gene(q=3),
+                                           estimate=eqtlnet.q0.fdr, verbose=FALSE)
> eqtlnet.q0.fdr.nrr
eQTLnetwork object:
  Genome: simulatedGenome
  Input size: 90 markers 50 genes
  Model formula: ~marker + gene | gene (q = 0,3)
  G^(0,3): 140 vertices and 1996 edges corresponding to
           1015 eQTL and 981 gene-gene associations meeting
           a fdr-adjusted p-value < 0.05
           and involving 50 genes and 87 eQTLs
```

- Obtain a second estimate $G^{(q)}$ of the eQTL network by selecting associations at FDR < 0.05 and with NRR value $\nu_{ij}^{q} < 0.1$.

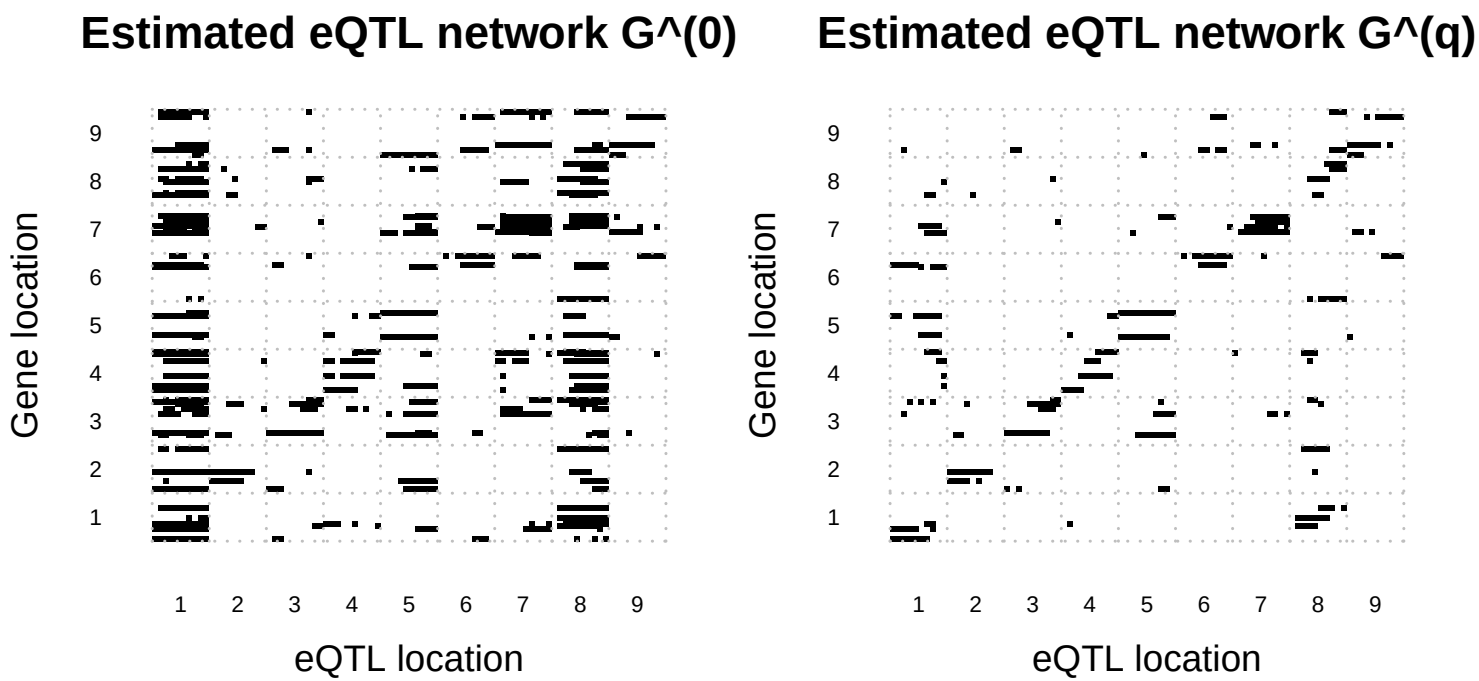
```
> eqtlnet.q0.fdr.nrr <- eQTLnetworkEstimate(param, estimate=eqtlnet.q0.fdr.nrr,
+                                           epsilon=0.1)
> eqtlnet.q0.fdr.nrr
eQTLnetwork object:
  Genome: simulatedGenome
  Input size: 90 markers 50 genes
  Model formula: ~marker + gene | gene (q = 0,3)
  G^(0,3): 140 vertices and 440 edges corresponding to
           293 eQTL and 147 gene-gene associations meeting
           a fdr-adjusted p-value < 0.05,
           a non-rejection rate epsilon < 0.10
           and involving 50 genes and 85 eQTLs
```

upf.

A three-step estimation strategy - second step

$G^{(q)} \subseteq G^{(0)}$ has lost most of the vertical bands in $G^{(0)}$.

```
> par(mfrow=c(1, 2))
> plot(eqtlnet.q0.fdr, main="Estimated eQTL network G^(0)", cex.lab=1.5, cex.main=1.8)
> plot(eqtlnet.q0.fdr.nrr, main="Estimated eQTL network G^(q)", cex.lab=1.5, cex.main=1.8)
```



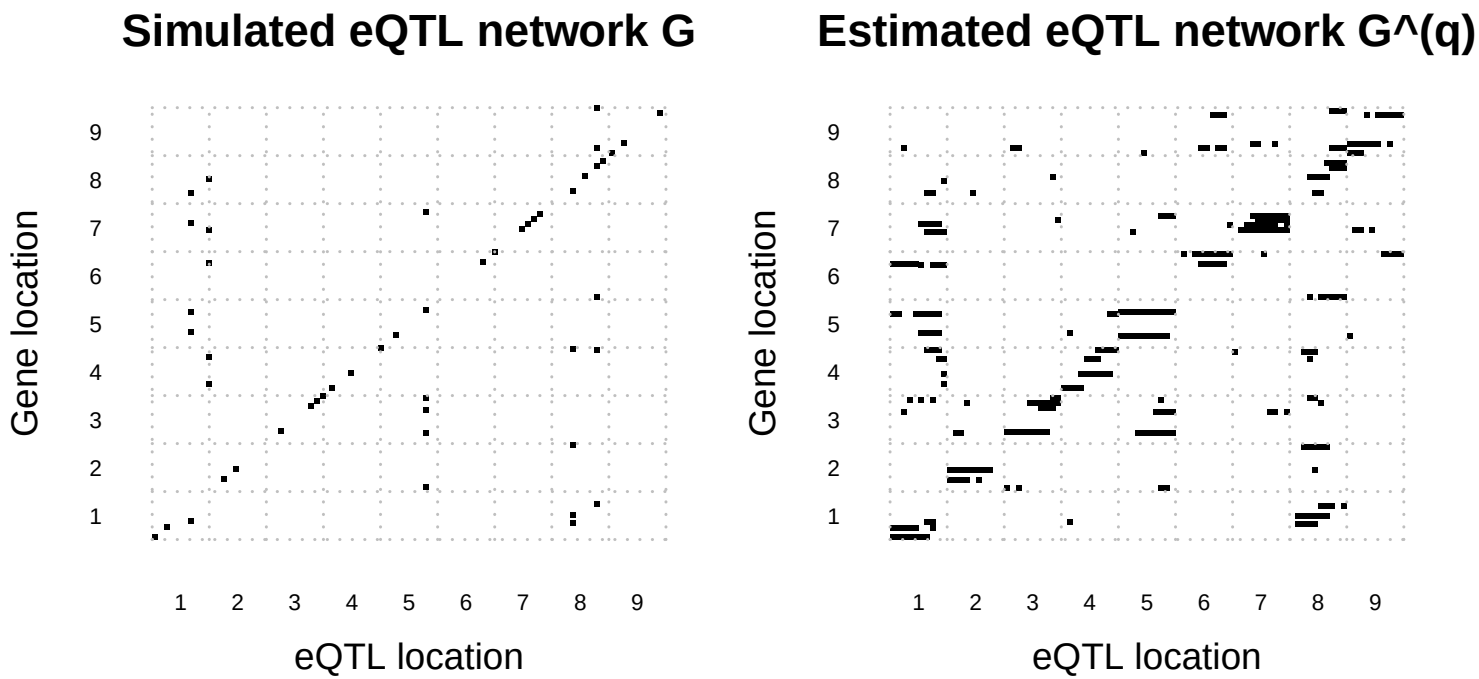
A three-step estimation strategy - third step

- Examine the median number of eQTLs per gene.

```
> eqtls <- alleQTL(eqtlnet.q0.fdr.nrr)
> median(sapply(split(eqtls$QTL, eqtls$gene), length))
```

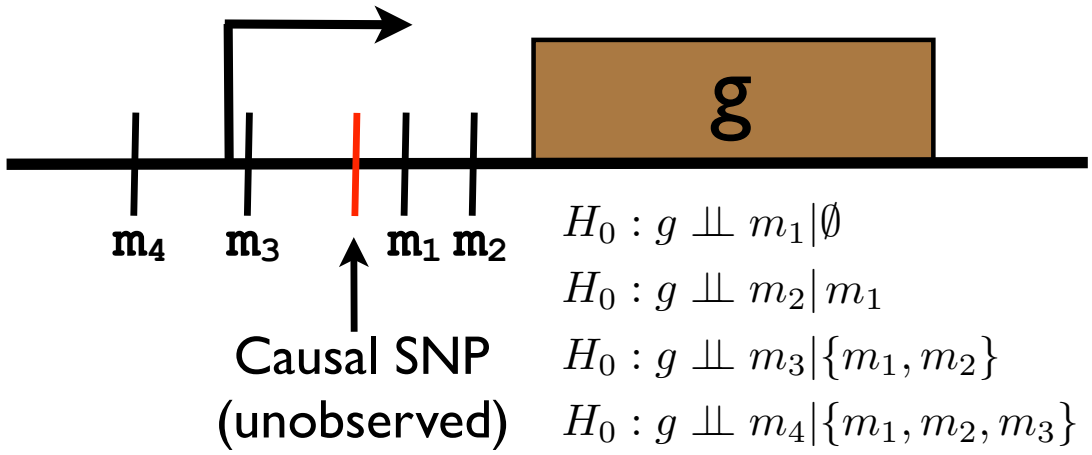
```
[1] 6
```

- Note that while we have simulated at most one eQTL per gene, we have currently estimated a median of 6 eQTLs per gene.



A three-step estimation strategy - third step

- Perform a forward selection procedure at a nominal significance level $\alpha < 0.05$ to remove redundant associations tagging the same causal eQTL.



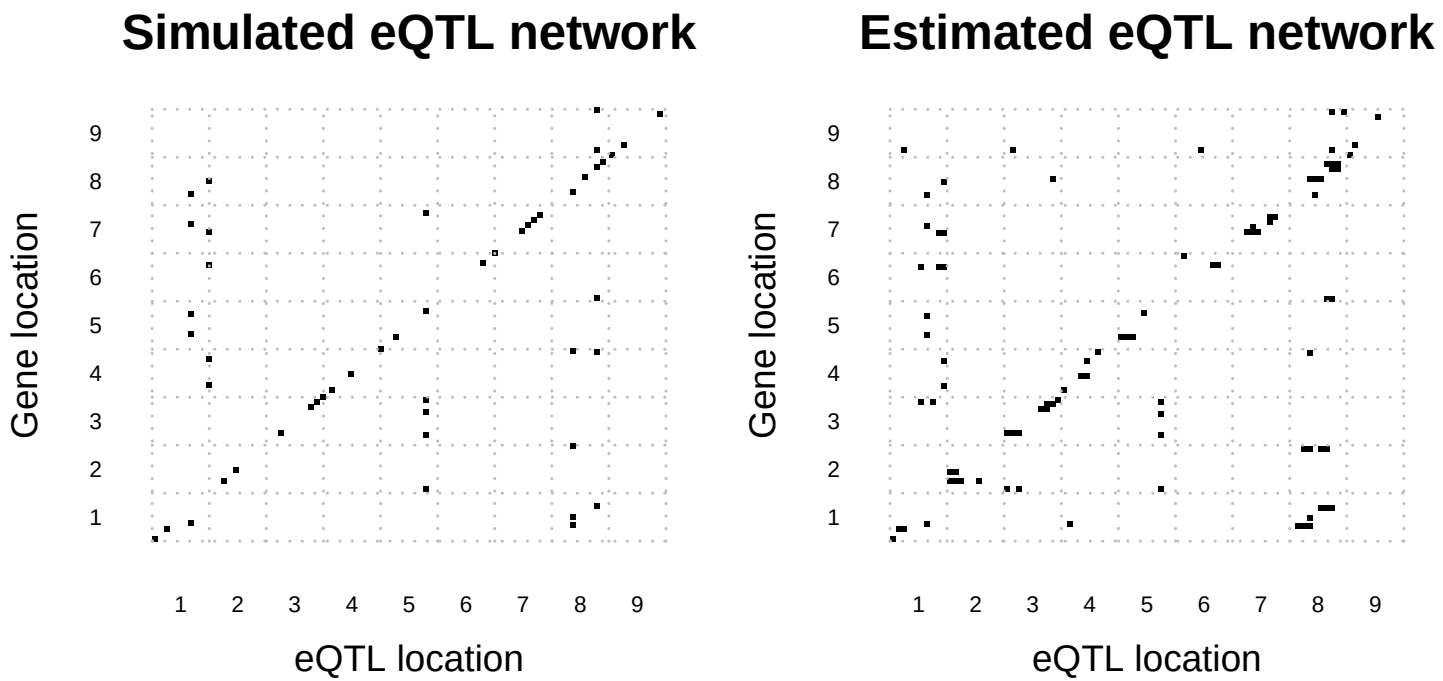

```
> eqtlnet.q0.fdr.nrr.sel <- eQTLnetworkEstimate(param, estimate=eqtlnet.q0.fdr.nrr,
+                                               alpha=0.05)
> eqtlnet.q0.fdr.nrr.sel

eQTLnetwork object:
  Genome: simulatedGenome
  Input size: 90 markers 50 genes
  Model formula: ~marker + gene | gene (q = 0,3)
  G^(0,3,*): 140 vertices and 238 edges corresponding to
             91 eQTL and 147 gene-gene associations meeting
             a fdr-adjusted p-value < 0.05,
             a non-rejection rate epsilon < 0.10,
             a forward eQTL selection significance level alpha < 0.05
             and involving 50 genes and 50 eQTLs
```

upf.

A three-step estimation strategy - third step

Most horizontal bands in $G^{(q)}$ have disappeared.

```
> par(mfrow=c(1, 2))
> plot(sim.eqtl, main="Simulated eQTL network", cex.main=2, cex.lab=1.5)
> plot(eqtlnet.q0.fdr.nrr.sel, main="Estimated eQTL network", cex.main=2, cex.lab=1.5)
```

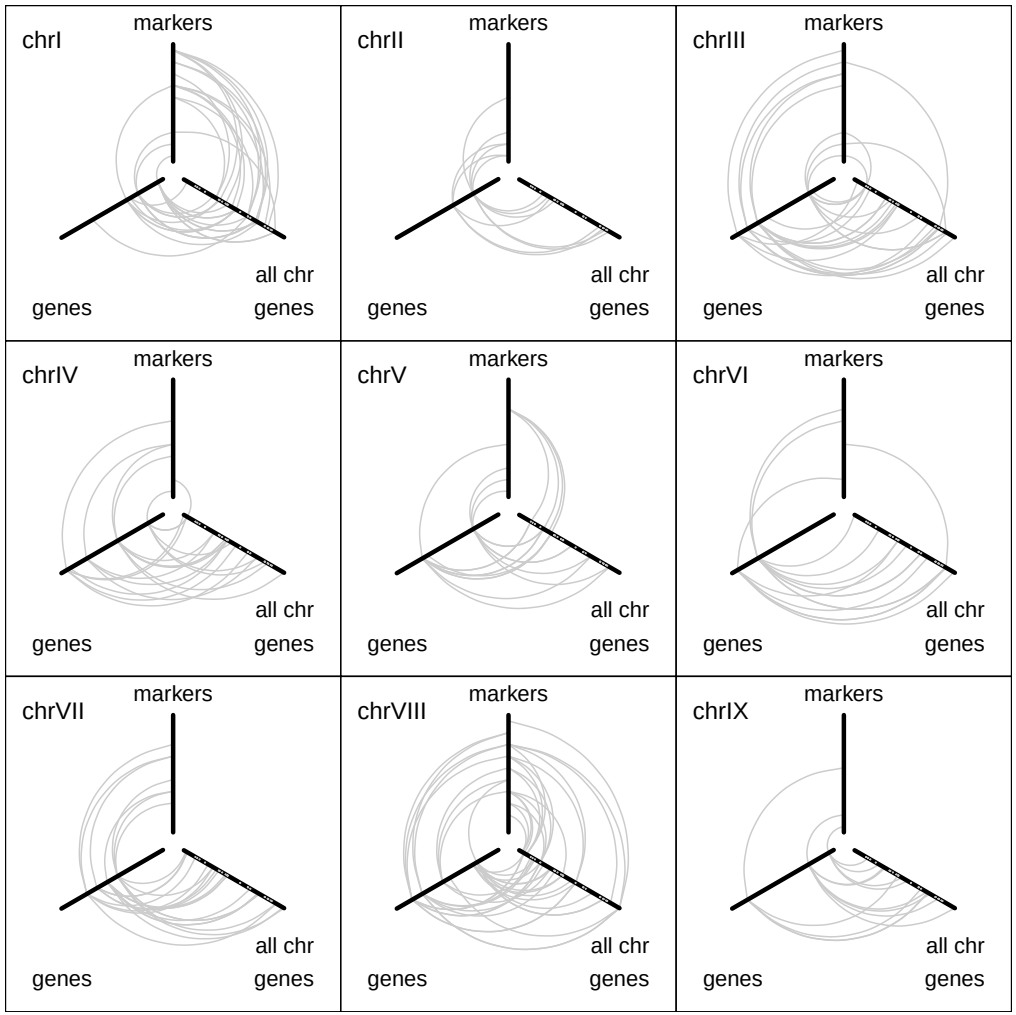


Outline

upf.

Visualization - from dot plot to hive plot

Visualize the gene-gene dimension simultaneously with eQTLs using Hive plots (Krzywinski *et al.*, 2012).



Outline

upf.

Analysis of a yeast cross - parameter setup

- We reanalyzed the yeast data from Brem and Kruglyak (2005), first calculating an estimate $G^{(0)}$ by doing all pairwise marginal tests and selecting edges at FDR $< 1\%$.

- Second, we estimated NRR values ν_{ij}^{q} between every possible pair of marker-gene and gene-gene in $G^{(0)}$, using conditioning subsets restricted to the genes and $q = \{25, 50, 75, 100\}$. The resulting estimates $\nu_{ij}^{q_k}$, $q_k \in q$, were averaged $\nu_{ij}^{\bar{q}} = \frac{1}{|q|} \sum_{q_k} \nu_{ij}^{q_k}$, to account for the uncertainty in the choice of q (Castelo and Roverato, 2009).

- Considered a conservative cutoff $\epsilon = 0.1$ on $\nu_{ij}^{\bar{q}}$, which selects edges with more than 90% of rejected tests, and obtained $G_{0.1}^{(\bar{q})}$ having $|E_{0.1}^{(\bar{q})}| = 4,110$ edges from which 2,448 were eQTLs and the rest gene-gene associations.

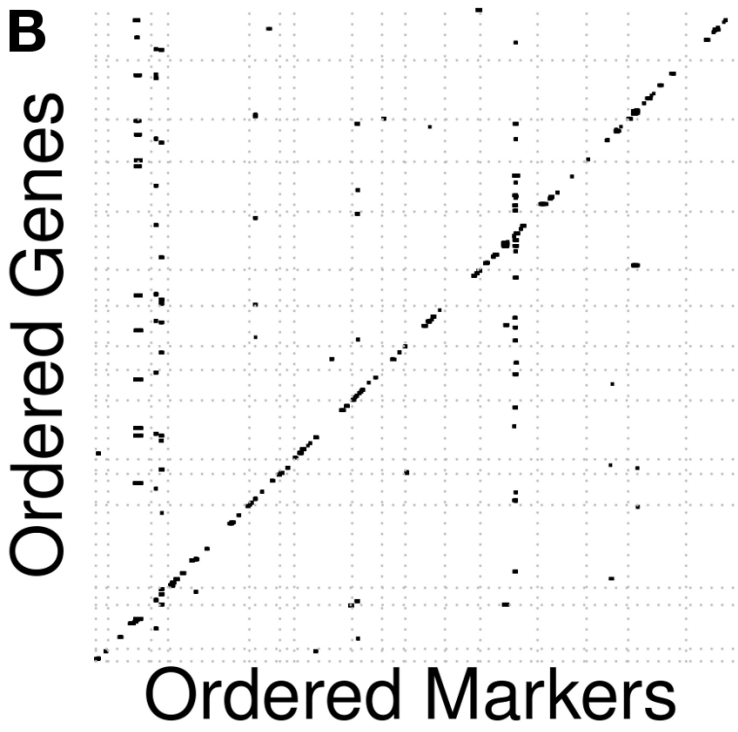
- Redundant eQTL associations were removed by a forward selection procedure with $\alpha = 0.05$.

upf.

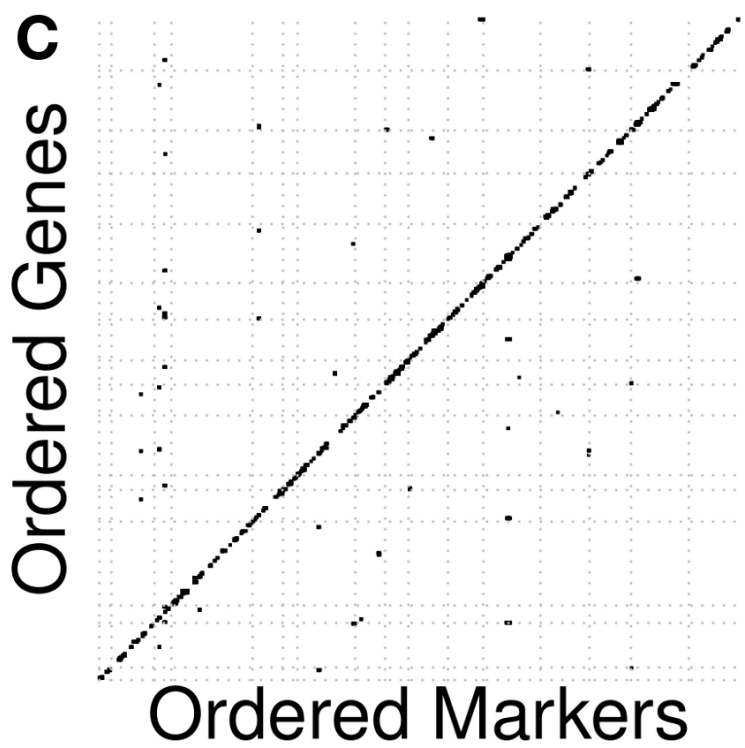
Analysis of a yeast cross - comparative performance

Compare $G_{0.1}^{(\bar{q})}$ with the top 2,448 marker-gene pairs with highest marginal LOD score, in a straightforward single-marker regression approach.

Marginal LOD score



qp-graph $G_{0.1}^{(\bar{q})}$



qpgraph yields a higher enrichment of local eQTLs and fewer vertical bands.

upf.

Analysis of a yeast cross - comparative performance

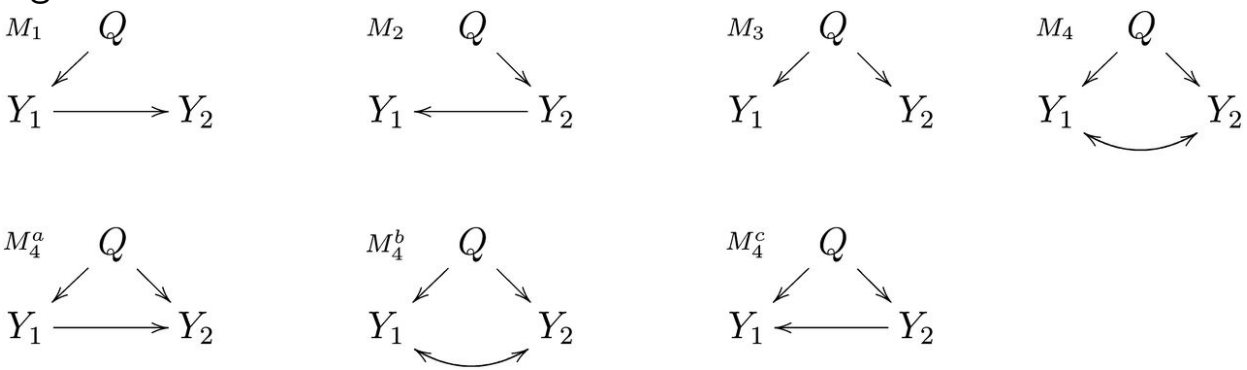
Compare with the causal inference approach of Chaibub Neto *et al.* (2013).

Genetics Society of America
GENETICS
Home Journal Information Subscriptions & Services Collections Y

Modeling Causality for Pairs of Phenotypes in System Genetics

Elias Chaibub Neto*, Aimee T. Broman†, Mark P. Keller†, Alan D. Attie†, Bin Zhang*, Jun Zhu* and Brian S. Yandell*,‡,1

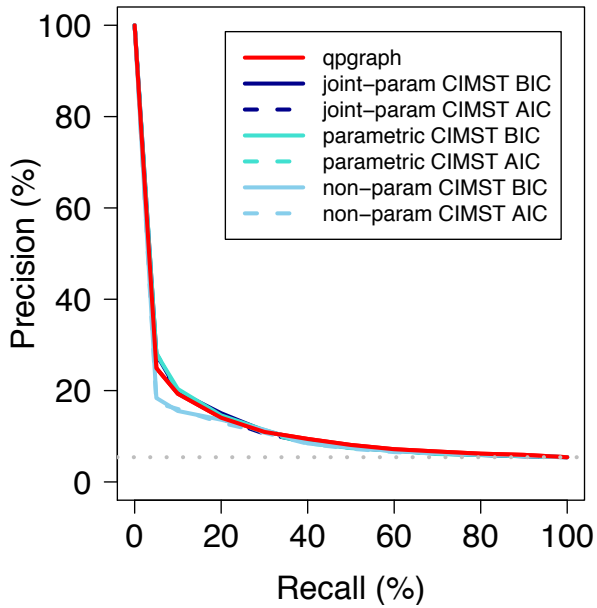
Fig. 1



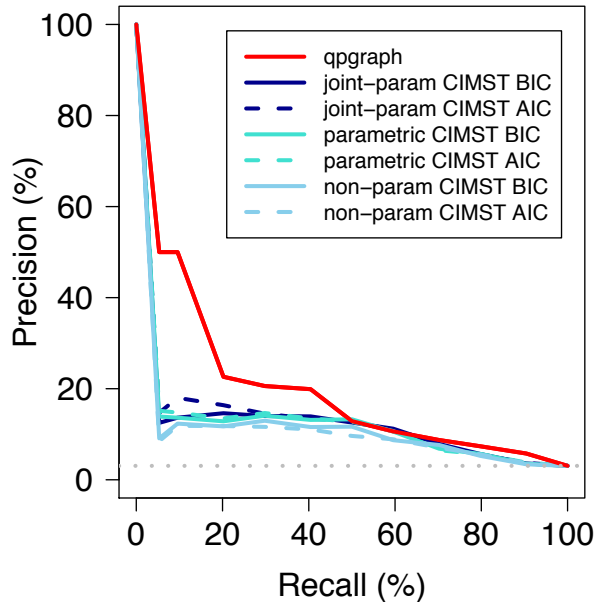
Analysis of a yeast cross - comparative performance

Precision-recall curves against a bronze standard formed by KO genes and their putative targets derived from differential expression (left) and restricted to curated transcriptional regulatory relationships on Yeastract (right).

Hughes *et al.* (2000)



Hughes *et al.* (2000) ∩ Yeastract

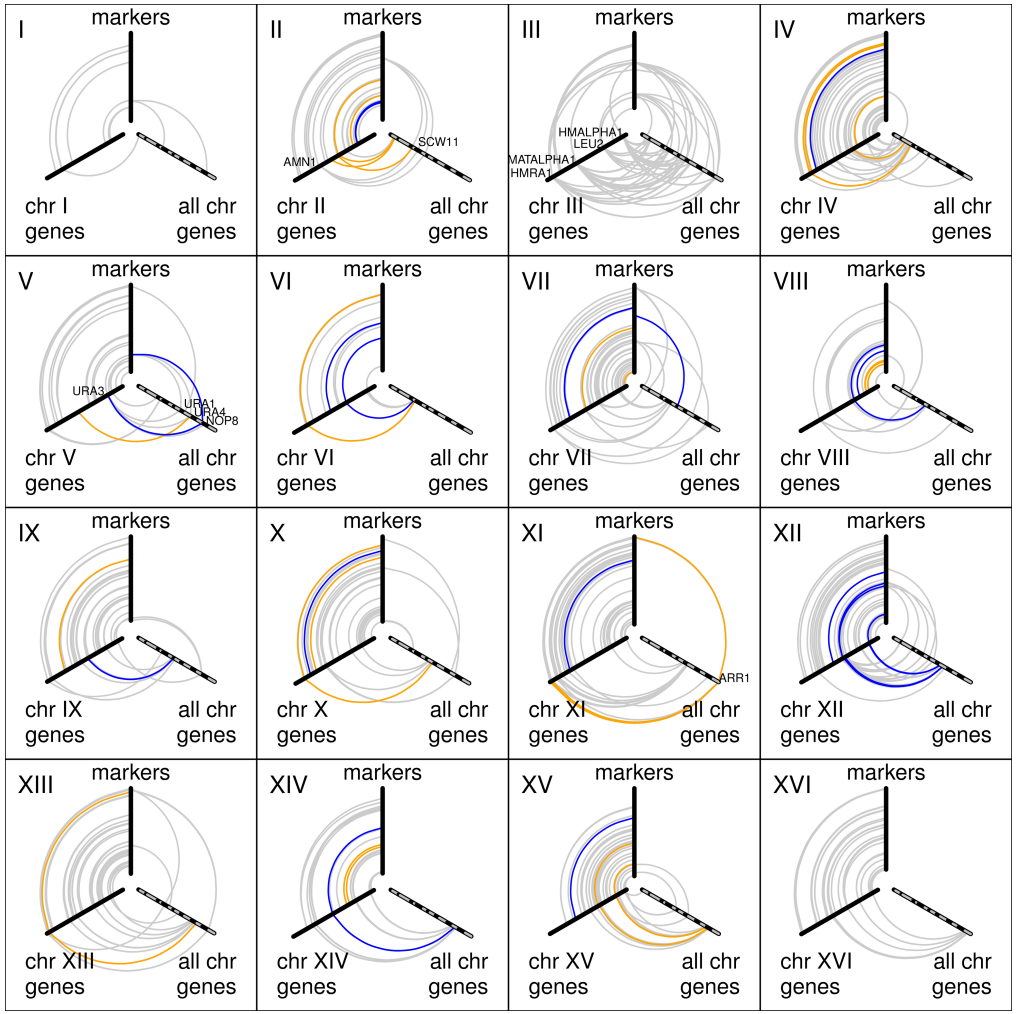


qpgraph performs similarly in identifying differential expression KO associations, but it improves in identifying direct regulatory associations.

upf.

Genetic control of gene expression across chromosomes

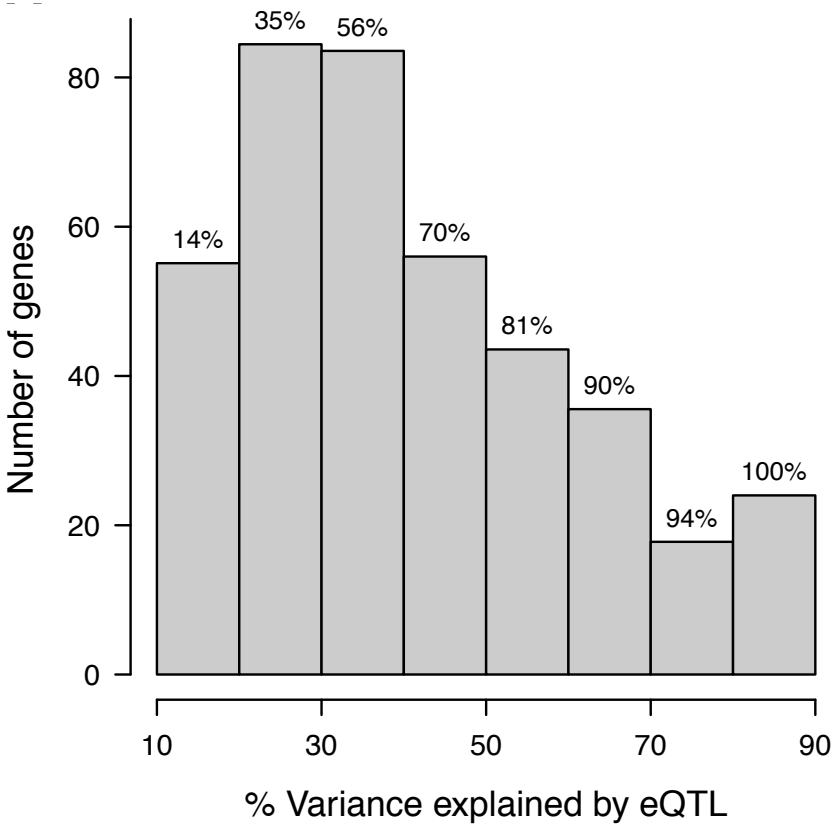
Display of the differential genetic control of gene expression across chromosomes by means of Hive plots (Krzywinski *et al.*, 2012).

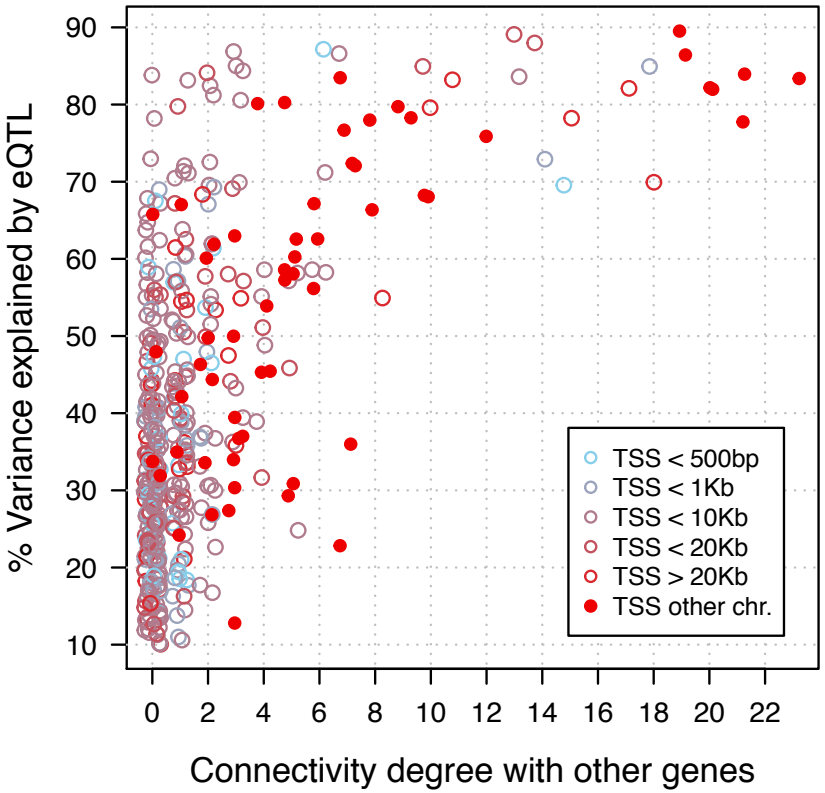


upf.

Analysis of a yeast cross - magnitude of effects

Estimation of the percentage of variance in gene expression explained by eQTLs.



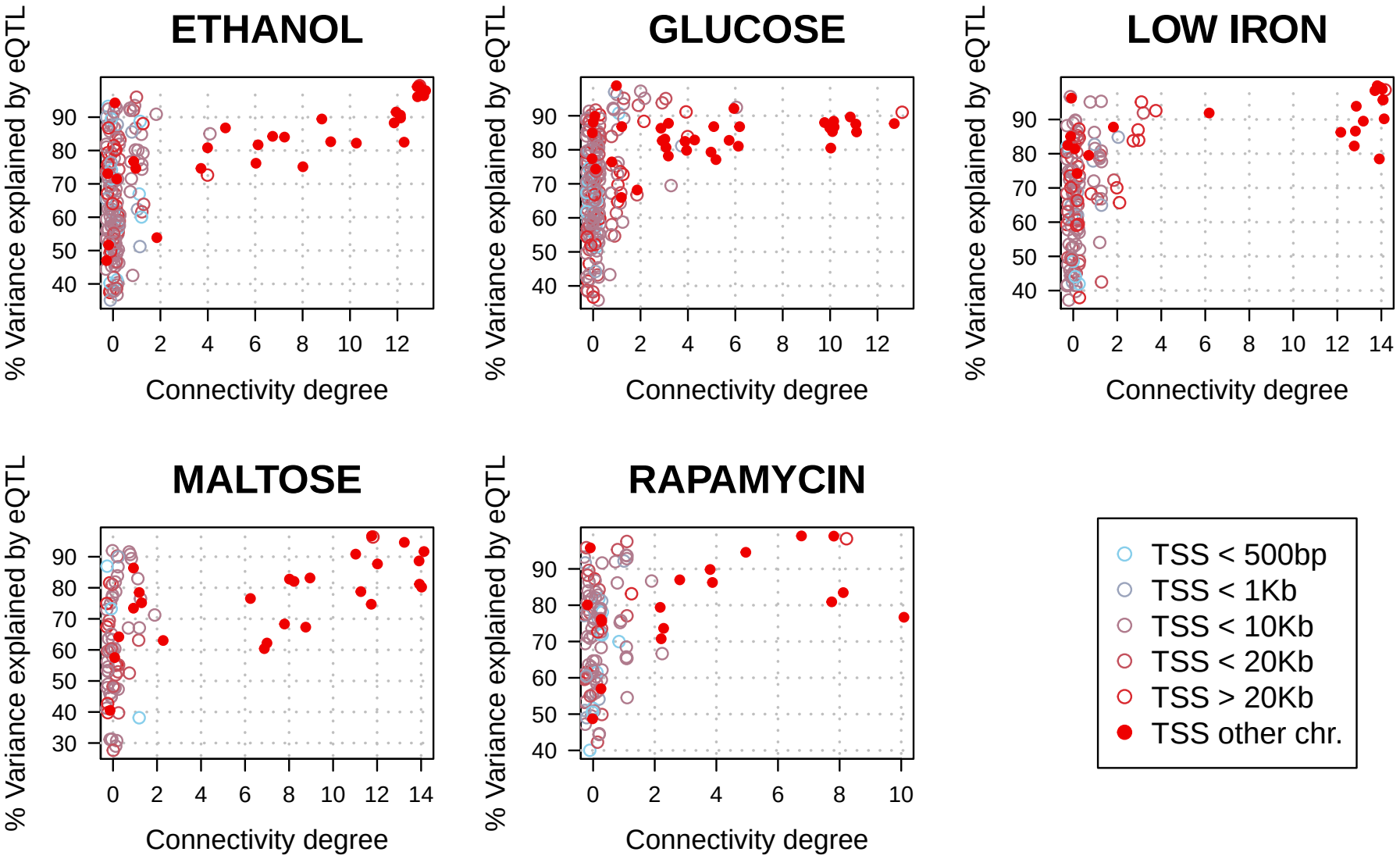


eQTLs explain most of the expression variablity of network hub genes.

upf.

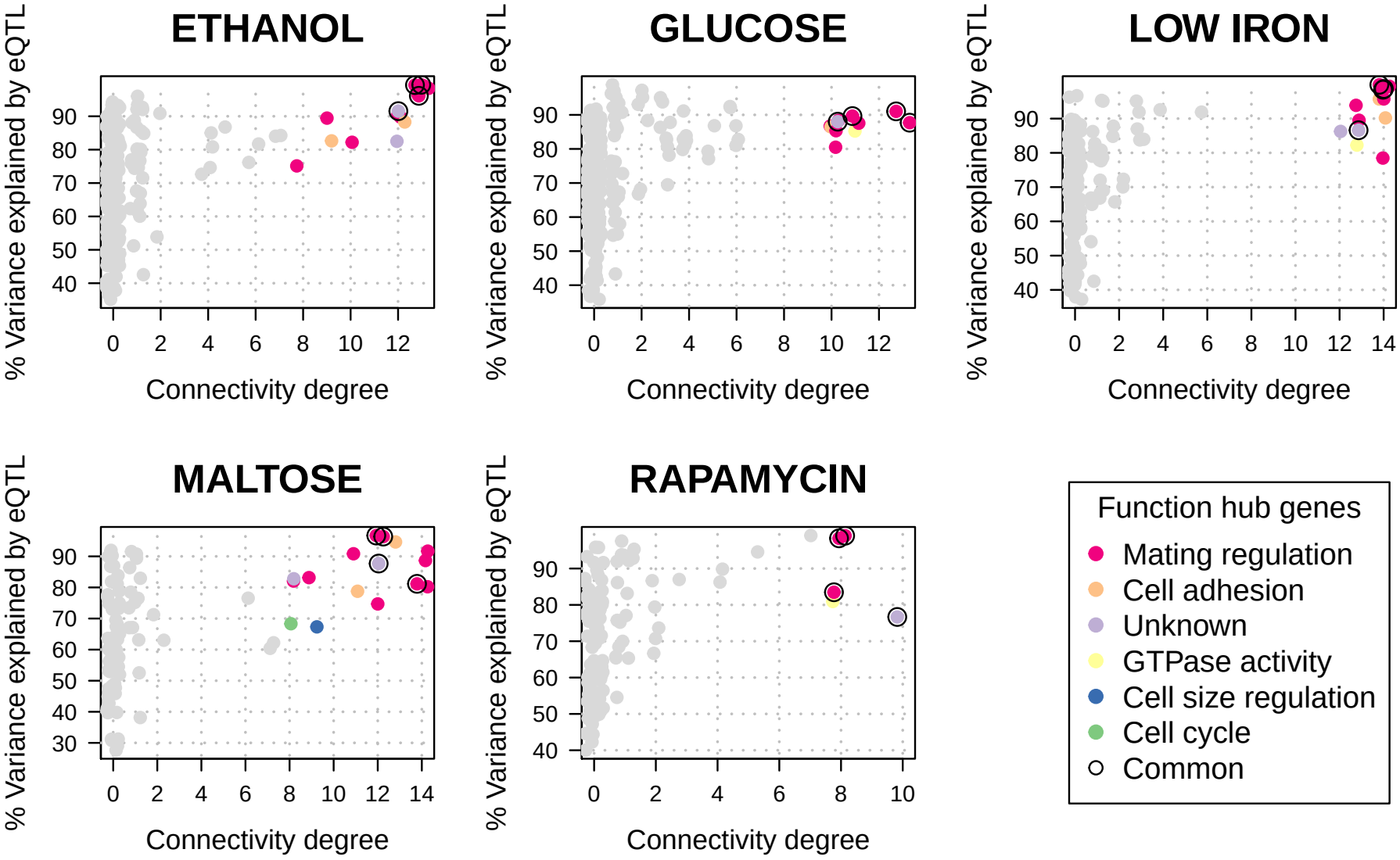
Analysis of a yeast cross - magnitude of effects

Independent data from Gagneur *et al.* (2013) show the same pattern.



Analysis of a yeast cross - magnitude of effects

Most hub genes with more than 7 connections are involved in mating regulation.



Outline

upf.

Concluding remarks

Limited-order correlation graphs, or qp-graphs, use conditional independence on marginal distributions to robustly infer eQTL and gene-gene associations.

Mixed GMMs allow one to embrace the complexity of a high-dimensional multivariate trait, to study the genetic control of gene **networks**.

By simulation, we showed that eQTL additive effects propagate throughout the network proportionally to the marginal correlation between genes.

There are other ways to use mixed GMMs in the $p \gg n$ setting, such as penalized likelihood group-lasso norm approaches (Lee and Hastie, 2014).

upf.

Bibliography and Acknowledgements

Bibliography (available at `http://functionalgenomics.upf.edu`):

- Castelo R and Roverato A. A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research*, 7:2621-2650, 2006.
- Castelo R and Roverato A. Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *Journal of Computational Biology*, 16:213-227, 2009.
- Tur I, Roverato A and Castelo R. Simulation of molecular regulatory networks with graphical models. Slides of a talk at the userR! 2013 conference. http://dx.doi.org/10.6084/m9.figshare.745372
- Tur I, Roverato A and Castelo R. Mapping eQTL networks with mixed graphical Markov models. *Genetics*, 198(4):1377-1383, 2014.

Data: Julien Gagneur for the genotype and expression data from Gagneur *et al.* PLOS Genet. (2013).

Funding:

- Spanish MINECO project grants [TIN2011-22826, TIN2015-71079-P]
- Catalan research group grant [2014-SGR-1121]

Software: The `qpgraph` package is available at http://www.bioconductor.org.

Follow news and bugfixes about `qpgraph` in `@robertclab`.

upf.