Bootstrap test of ordered RIG for multiple testing in genomics of Quantitative Trait Loci in yeasts

E. Mirkes, T. Walsh, E.J. Louis, and A.N. Gorban

University of Leicester, UK

Plan

- Multiple testing of correlations: three approaches:
  - Bonferroni corrections
  - False discovery rate
  - Bootstrap of correlations
- Quantitative Trait Loci (QTL) in heat selected yeast – experiment description
- Results of data analysis:
  - Do the selected genomes differ from the unselected ones?
  - Do the QTL interact in heat selected genomes?
  - Are the statistically significant correlations large?
- Conclusion

Multiple testing of correlations: The applied problem

- A sample of *N* objects is studied;
- We measure *n* attributes $X_1,\ldots,X_n$ of these objects;
- R_{ij} is a measure of dependence between X_i and X_j and $\hat{R}_{ij}$ is its sample estimate, for example the relative information gain;
- **Which dependencies are significant**?

(There are *n*(*n*-1) ordered pairs of attributes and some of $\hat{R}_{ij}$ may be large by chance.)

P-hunting

- Study statistics of $\hat{R}_{ij}$ for each pair (X_i,X_j).
- For a given sufficiently small *p*-value p_0 find the borders r_{ij} such that

$$\mathrm{P}\left(\hat{R}_{ij} > r_{ij}\right) < p_0.$$

- Find pairs (X_i,X_j) with $\hat{R}_{ij} > r_{ij}$.
- Call all connections between (X_i,X_j) significant with *p*-value p_0 (or the significance level 1- p_0).

P-hunting

- Study statistics of $\hat{R}_{ij}$ for each pair (X_i,X_j).
- For a given sufficiently small *p*-value p_0 find the borders r_{ij} such that

$$\mathrm{P}\left(\hat{R}_{ij} > r_{ij}\right) < p_0.$$

- Find pairs (X_i,X_j) with $\hat{R}_{ij} > r_{ij}$.
- Call all connections between (X_i,X_j) significant with *p*-value p_0 (or the significance level 1- p_0).

Nowadays everybody knows
that *p*-hunting is incorrect

Bonferroni correction – uniform estimate in *i,j*

- Notice that

$$\mathrm{P}\left(\forall i,j\ \hat{R}_{ij} < r_{ij}\right) \geq 1 - \sum_{i,j} \mathrm{P}\left(\hat{R}_{ij} > r_{ij}\right)$$

- Select a sufficiently small *p*-value p_0.
- Study statistics of $\hat{R}_{ij}$ for each pair (X_i,X_j).
- Find the borders r_{ij} such that

$$\mathrm{P}\left(\hat{R}_{ij} > r_{ij}\right) < {}^{p_0}/_{number\ of\ pairs}$$

- Find pairs (X_i,X_j) with $\hat{R}_{ij} > r_{ij}$.
- Call these dependencies significant with *p*-value p_0 (or the significance level 1- p_0).

Bonferroni correction – uniform estimate in *i,j*

- Notice that

$$\mathrm{P}\left(\forall i,j\ \hat{R}_{ij} < r_{ij}\right) \geq 1 - \sum_{i,j} \mathrm{P}\left(\hat{R}_{ij} > r_{ij}\right)$$

- Select a sufficiently small *p*-value p_0.
- Study statistics of $\hat{R}_{ij}$ for each pair (X_i,X_j).
- Find the borders r_{ij} such that

$$\mathrm{P}\left(\hat{R}_{ij} > r_{ij}\right) \leq p_0 / number\ of\ pairs$$

- Find pairs (X_i,X_j) with $\hat{R}_{ij} > r_{ij}$.
- Call these dependencies significant with *p*-value p_0 (or the significance level $1-p_0$).

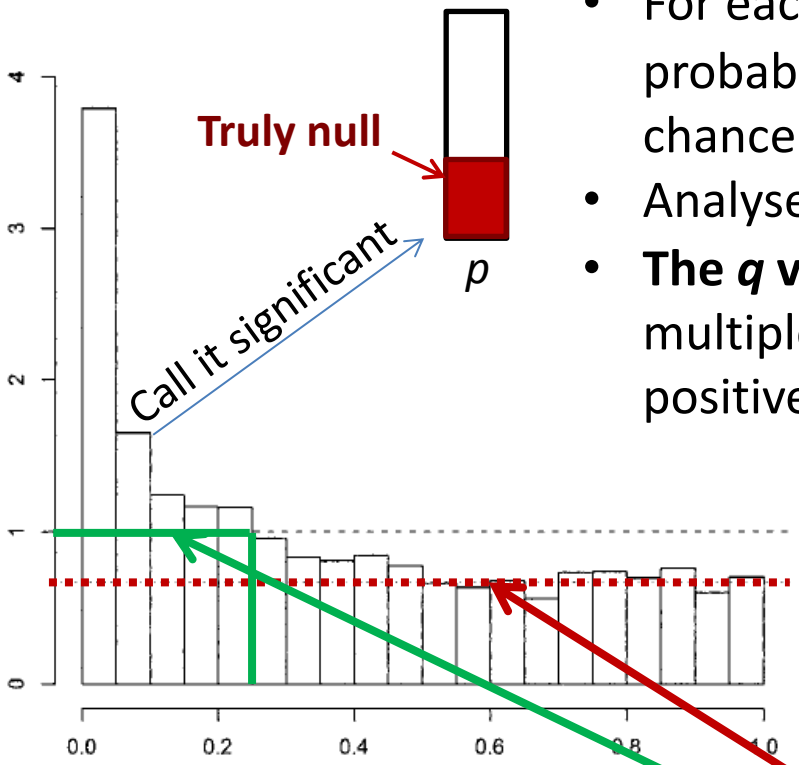
May be too conservative – the denominator, the number of pairs, is too big, and we do not always need such a uniform estimate

False discovery rate model

Statistical significance for genomewide studies

John D. Storey*† and Robert Tibshirani‡

9440–9445 | PNAS | August 5, 2003 | vol. 100 | no. 16



A histogram of *p* values.

- For each observed $\hat{R}_{ij}$ find the *p*-value p_{ij}, that is the probability that such (or larger) value is observed by chance for independent X_i, X_j.
- Analyse the distribution of the *p*-values p_{ij}:
- **The *q* value** of a particular feature (correlation) in multiple testing is the expected proportion of false positives incurred when calling that feature significant.

False discovery rate≈
≈ pFDR$_i$=P(feature *i* is truly null|feature *i* is significant)

False positive rate=
=p_i=Pr(feature *i* is significant|feature *i* is truly null)

P(feature *i* is truly null)(*p*)=constant

Pr(feature *i* is significant)(*p*)=1 if $p \le p_0$ and =0 if $p > p_0$

From multiple testing to simple testing by order statistics 1. Maximal correlation

- A sample of *N* objects is studied;
- We measure *n* attributes $X_1,\ldots,X_n$;
- We know individual distribution functions for X_i ;
- R_{ij} is a measure of dependence between X_i and X_j and $\hat{R}_{ij}$ is its sample estimate;
- Assume that we know the distribution

  $$\mathbf{P_{max}(r)=P(max_{ij}\,\hat{R}_{ij}>r)}$$

  for independent $X_1,\ldots,X_n$ in a sample of *N* objects
- Let $P_{max}(r_0)<p_0$;
- We observe $\mathbf{max_{ij}\hat{R}_{ij}>r_0}$ → the probability of this observation by chance is less then p_0.

From multiple testing to simple testing by order statistics 2. General procedure

- Order the first *k* sample estimates:
$$\hat{R}_{ij}^{1} \geq \hat{R}_{ij}^{2} \geq \cdots \geq \hat{R}_{ij}^{k};$$
- Assume that we know the distribution
$$\mathbf{P}_l(r) = \mathbf{P}(\hat{R}_{ij}^{l} > r)$$
for *l*=1,…,*k* and independent $X_1,\ldots,X_n$ in a sample of *N* randomly chosen objects;
- Select *p*-value p_0. Let $P_l(r_l)<p_0$ for *l*=1,…,*k*.
- If we observe $\hat{R}_{ij}^{l} > r_l$ ***for all*** *l*=1,…,*k* then these correlations are significant with the level 1-p_0 .

Where can we take the disrtibutions $\mathbf{P}_l(r) = \mathbf{P}(\hat{R}_{ij}^l > r)$?

- The number of attributes X_i is significantly smaller than the number of pairs (X_i,X_j);
- Therefore, we can often rely on the observed individual distributions of X_i (even with Bonferroni corrections) much more than on the empirical correlations;
- Bootstrap: generate sufficiently many samples of *N* objects with independent and properly distributed X_i and estimate $\mathbf{P}(\hat{R}_{ij}^l > r)$…

In the case study below

- The number of attributes is *n*=16;
- The sample size is *N*=896;
- The correlation measure is

$$\text{RIG}(X|Y) = \frac{Entropy(X) - Entropy(X|Y)}{Entropy(X)};$$

- The number of pairs is 240 (RIG is non-symmetric).

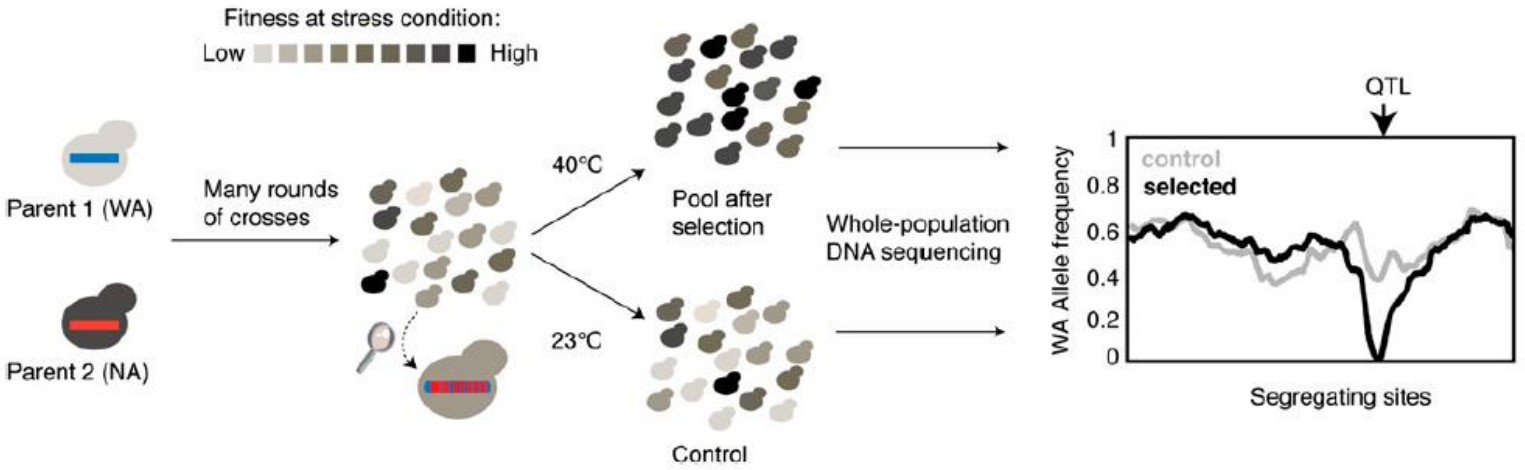
We used genomes for the 960 heat selected individuals and of 172 unselected individuals

Revealing the genetic structure of a trait by sequencing a population under selection

Leopold Parts,[1,6] Francisco A. Cubillos,[2] Jonas Warringer,[3,4] Kanika Jain,[2] Francisco Salinas,[2] Suzannah J. Bumpstead,[1] Mikael Molin,[3] Amin Zia,[5] Jared T. Simpson,[1] Michael A. Quail,[1] Alan Moses,[5] Edward J. Louis,[2] Richard Durbin,[1] and Gianni Liti[2,6]

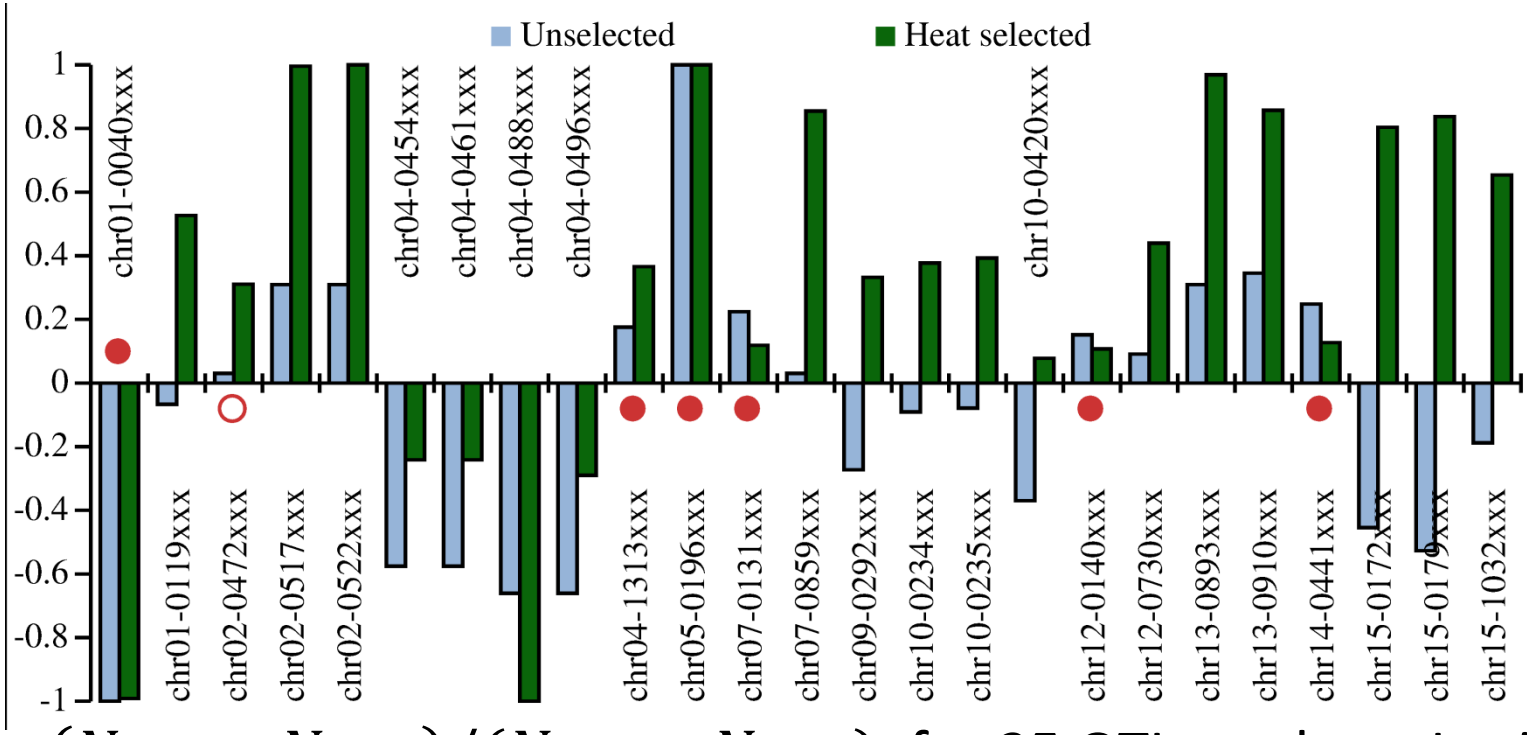
21:1131–1138 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

Genome Research
www.genome.org



(i) Crossing different strains, (ii) growing the pool in a restrictive condition, (iii) sequencing total DNA from the pool.

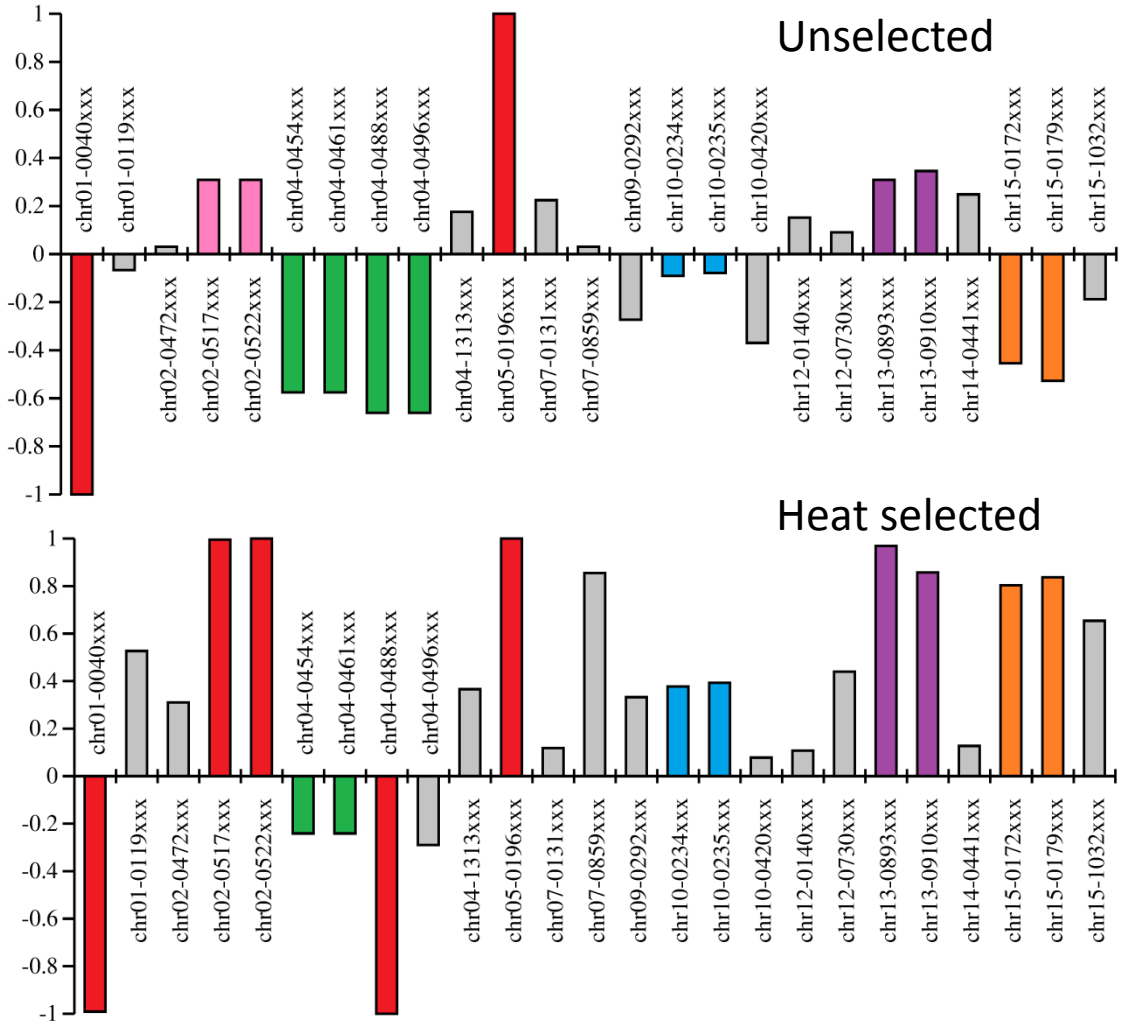
NA/WA alleles for QTL markers in unselected and heat selected samples



$(N_{\mathrm{NA}} - N_{\mathrm{WA}})/(N_{\mathrm{NA}} + N_{\mathrm{WA}})$ for 25 QTL markers. Loci with (*Bonferroni corrected*) $p>0.1$ are marked by solid circle and with $0.1 \geq p \geq 0.01$ by circle. All other $p<0.01$.

(Here, *p*-value is the probability to observe the same or larger difference between distributions of alleles in a locus for selected and unselected samples if the probability distributions are the same.)

Distribution of $(N_{NA} - N_{WA})/(N_{NA} + N_{WA})$ for unselected and heat selected pools

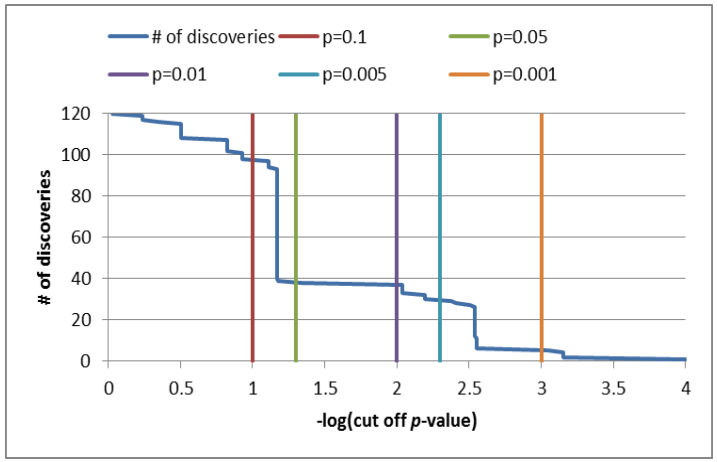


Red - one parent allele (constant loci) and almost constant loci (the fraction of one of the alleles is greater than 99%).

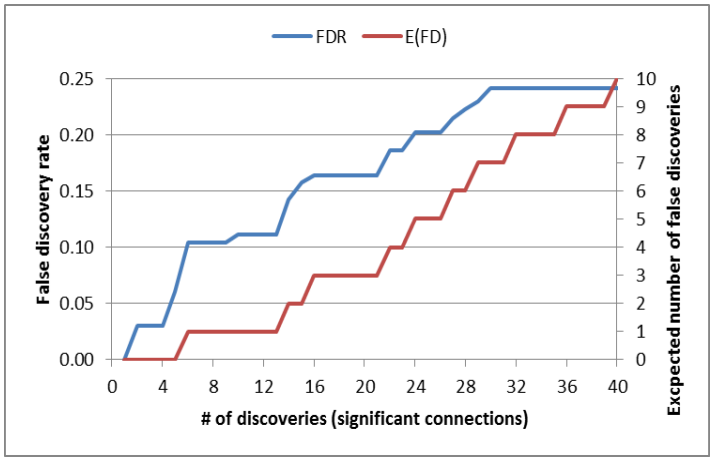
Magenta, Green, Blue, Violet, and Brown - different groups of linked loci.

Grey - all other loci (not linked).

The number of significant correlations and estimated number of false discoveries



The number of significant connections with respect to *p*-value for BToRIG.



The false discovery rate and expected number of false discoveries as a function of the number of discoveries.

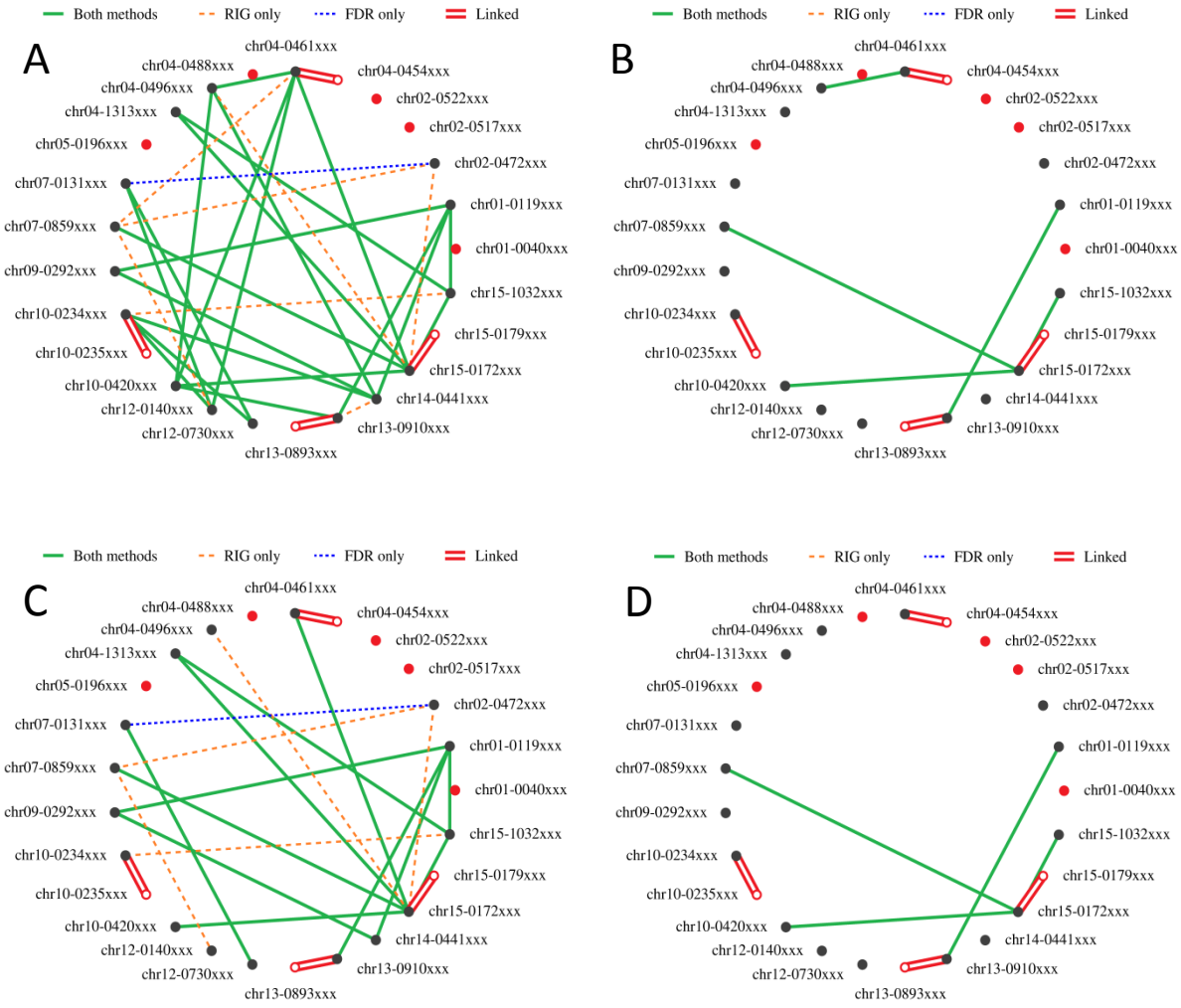
Which links are important for the heat tolerance?

Significant does not mean large.

What else should we take into account?

- The ***level*** of correlations (if RIG<0.01 then it is difficult to find a solid reason to consider this connection as important when the number of candidates is ~20);
- The ***novelty***: there should be a difference between the links in the heat selected and unselected samples
- The novelty may be *measured* by RIG ratio $RIG_{selected}/RIG_{unselected}$ or by relative entropy H(selected|unselected)

Significantly dependent loci for heat selected pools



A) Significant connections.

B) Significant connections with $RIG \geq 0.01$.

C) Significant connections with $RIG_{selected}/RIG_{unselected} \geq 2$ or H(selected|unselected)>0.5.

D) Significant connections with RIG ≥ 0.01, and $RIG_{selected}/RIG_{unselected} \geq 2$ or H(selected|unselected)>0.5.

Red solid circles – the constant loci.

Red circles with white centres - the loci in linkage disequilibrium with other loci (*doubled red line*).

Solid green lines connect loci defined as significantly dependent by DFR and BToRIG.

Conclusion

- Bootstrap test of ordered correlation measures is efficient when the number of pairs is much larger than the number of attributes;
- It works and the results in the case study are (surprisingly) similar to the False Discovery Rate approach which has very different backgrounds.
- Multiple testing of significance of associations after selection should be supplemented by the evaluation of importance (size) and novelty.

- For the heat tolerance of yeasts, a statistical analysis of entropy and information gain in genotypes of a selected population can reveal further interactions than previously seen.
- Various non-random associations were found across the genome both within chromosomes and between chromosomes.

E-print

EM Mirkes, T Walsh, EJ Louis, AN Gorban, Long and short range multi-locus QTL interactions in a complex trait of yeast, **arXiv:1503.05869 [q-bio.GN]**