Using KEGG pathways for genomic partitioning

Stefan McKinnon Edwards

Centre for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University (DK).

Problem

Given a complex trait on 4,497 bulls, each genotyped to 637,951 markers, the usual approach for estimating marker effects is to assume all markers have equal variance. Instead of trying to attribute how much of the observed variance, each marker accounts for, we estimate how much a subset of markers *jointly* are accountable for.

The model

We partitioned the markers into two groups, and modelled the marker effects with the linear mixed model

$$\mathbf{y} = \mu + \mathbf{X}_{\mathcal{S}}\mathbf{b}_{\mathcal{S}} + \mathbf{X}_{\neg\mathcal{S}}\mathbf{b}_{\neg\mathcal{S}} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of phenotypic variances, μ an intercept, $\mathbf{X}$ a design matrix linking bulls to genotypes, $\mathbf{b}$ a vector of marker effects and $\mathbf{e}$ the residual. The subset $\mathcal{S}$ indicates that the variable is for the markers in the group and $\neg\mathcal{S}$ for markers not in the group. For computational reasons, we considered the equivalent model

$$\mathbf{y} = \mu + \mathbf{g}_{\mathcal{S}} + \mathbf{g}_{\neg\mathcal{S}} + \mathbf{e}$$

where $\mathbf{g}$ is a vector of genetic values. For these values, we assume

$$\begin{pmatrix} \mathbf{g}_{\mathcal{S}} \\ \mathbf{g}_{\neg\mathcal{S}} \\ \mathbf{e} \end{pmatrix} \sim N\left[\mathbf{0}, \begin{pmatrix} \mathbf{I}\sigma^2_e & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\mathcal{S}}\sigma^2_{\mathcal{S}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_{\neg\mathcal{S}}\sigma^2_{\neg\mathcal{S}} \end{pmatrix}\right]$$

i.e. that all genetic values are expected zero with variances σ^2, that all variances are uncorrelated and that the markers can assume one of two variances.

Evaluating a group of markers

We have used KEGG pathways, to select which markers to include in the group. Using an average-information restricted maximum likelihood algorithm ("AI-REML"), for each group, the three variance components (the sigmas) are estimated. With these values estimated, we look at the proportion of variance, that the first group explains. That is, we calculate the ratio

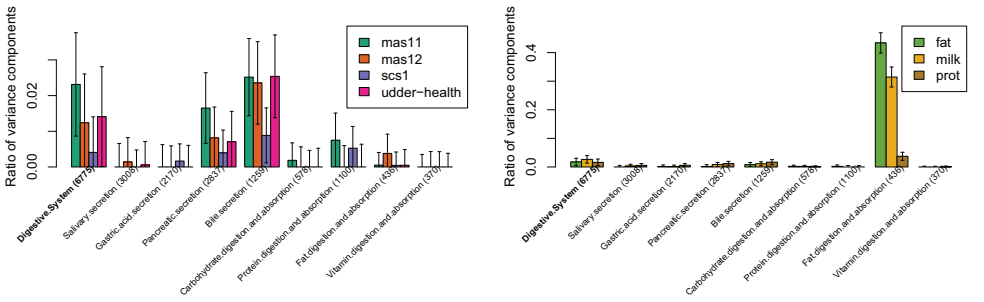
$$\varphi = \frac{\sigma^2_{\mathcal{S}}}{\sigma^2_{\mathcal{S}} + \sigma^2_{\neg\mathcal{S}}}$$

for each KEGG pathway of interest.

Data

For these analysis, we have used both udder health traits and production traits. The health traits are measurements of mastitis ('mas11', 'mas12'), somatic cell count ('scs1') and a combined trait called 'udder-health'. Production traits include fat yield, milk yield and protein yield. As the genotyped animals are bulls, the traits are derieved from the bulls' daugthers.

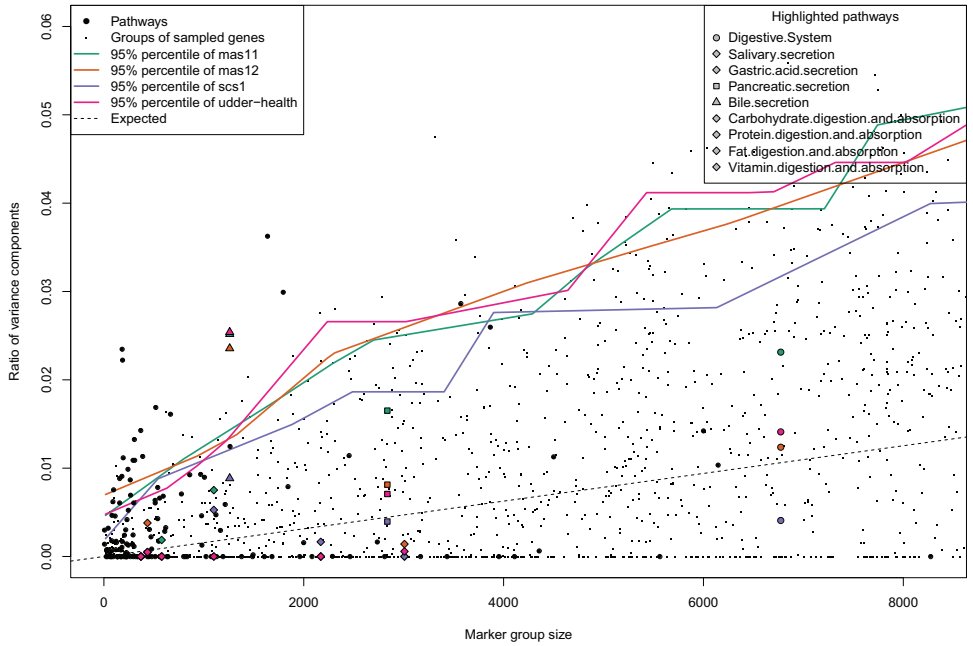
Ratio of variances for a KEGG pathway



Proportion of the variance that the markers associated to the pathways can explain for health traits (left) and production traits (right). Whiskers show standard deviation of the ratio. Number in parenthesis is number of markers in group.

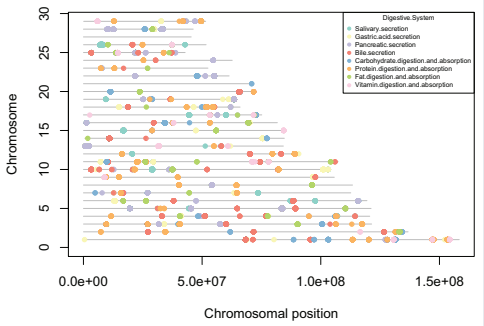
Evaulating significance

But how do we evaluate whether the results are not found by chance? *Sampling!*
We can empircally determine whether the proportion of variance a pathway explains is due to chance, by sampling markers in the same manner, as the markers are selected for the pathways, and computing the ratio for each group of randomly selected genes.



Plot showing how the ratio increases as the number of markers in a group increases, together with KEGG pathways. For the expected ratio, it is assumed that all markers contribute equally to the observed variance.

Correlation



Plot of the genomic positions of markers associated to the KEGG pathways.

Conclusion

- Markers do *not* contribute equally to the observed variance.
- More markers will by chance explain more variance.
- The method is not restricted to KEGG pathways.
- A pathway with a ratio significant different from zero, may still be unsignificant.

AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS