Influence of Family Structure on Variance Decomposition

***S.M. Edwards*, P.M. Sarup, and P. Sørensen.**

*Center of Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University

ABSTRACT: Partitioning genetic variance by sets of randomly sampled genes for complex traits in *D. melanogaster* and *B. taurus*, has revealed that family structure can affect variance decomposition. In fruit flies, we found that a high likelihood ratio is correlated with a high proportion of explained genetic variance. However, in Holstein cattle, a group of genes that explained close to none of the genetic variance could also have a high likelihood ratio. This is still a good separation of signal and noise, but instead of capturing the genetic signal in the marker set being tested, we are instead capturing pure noise. Therefore it is necessary to use both criteria, high likelihood ratio in favor of a more complex genetic model and proportion of genetic variance explained, to identify biologically important gene groups.

Key words: linear mixed models; genetics; Holstein dairy; cattle; fruit flies.

INTRODUCTION

The traditional statistical modelling that simultaneously fit all genetic variants does so in a “Black box” approach that models each genetic variant both equally and independently. A disadvantage of this is that the modelling does not provide any insight into the biological mechanisms underlying the trait.

We aimed to circumvent these problems by integrating external information, such as KEGG pathways (Kanehisa and Goto (2000); Kanehisa et al (2012)), into variance decomposition of complex traits, whereby we estimated the joint contribution from genetic variants found in the genes linked to biological pathways.

In the transition from health and production traits in Danish Holstein dairy cattle (Su et al. (2012)) to other complex traits in the common fruit fly *Drosophila melanogaster* (MacKay et al. (2012)), we have found that differences in family structures give rise to some interesting patterns in the variance decomposition. The Holstein dairy cattle population is characterised the intensive use of a limited number of sires, whereas the fly data used here consists of multiple ‘lines’, where each line is almost entirely inbred ($F = 0.986$, MacKay et al. (2012)) and highly homozygous; this allows for repetitive sampling of phenotypes from a highly homogenous population (i.e. each line).

Here we present some results from partitioning the phenotypic variance by sets of randomly sampled genes in both startle response in *D. melanogaster* and mastitis in *Bos taurus*. The results found here are not unique to these two traits in neither fruit flies nor dairy cattle, respectively, but it is rather an effect of differences in family structure as other traits (e.g. starvation or production traits, respectively) in these two organisms yielded similar results.

Materials and Methods

Startle Response. SNP data are called from raw sequence data from MacKay et al. (2012), and kept if the coverage was greater than 2X but less than 30X, if the minor allele frequency (MAF ≥ 0.025) was present in at least four lines and if they were called in at least 60 lines resulting in a total of 2,476,804 SNPs distributed on five chromosome arms (2L, 2R, 3L, 3R and X). Missing genotypes were imputed using Beagle Version 3.3.1.

Records of startle response are obtained by placing flies in a test tube, knocking the flies down by tapping the tube and then measuring how many seconds the flies were active, up to 45 seconds (Yamamoto et al. (2009)). The data used here consists of 167 lines each with, on average, 80 observations with values from 1 s to 45 s, mean at 29 s.

Mastitis. Genotype data was retrieved from imputed HD chip data on 4,497 Danish Holstein bulls (Su et al. (2012)), covering a total of 637,951 markers, i.e. SNPs, of which 621,217 have a minimal allele frequency above 0.01. The HD chip is mapped to the UMD3.1 Bovine Genome assembly (Zimin et al. (2009)); gene mapping for this assembly was downloaded September 1st 2011 from the website (ftp://ftp.cbcb.umd.edu/pub/data/assembly/-Bos_taurus/Bos_taurus_UMD_3.1/annotation/UMD3.1-.gff.gz), containing 26,352 genes with an Entrez Gene ID.

For the bovine data, we have chosen the health trait ‘Mastitis 1.2’ (hereafter simply referenced as ‘Mastitis’), records of treatments for mastitis in first lactation in the period of 10 days before to 305 days after calving. The mastitis records are estimated breeding values, mean 96.25, std.dev. 9.71.

Model. Using DMU (Madsen and Jensen (2012)), we estimated variance components for the following models:

$$M_p: \quad y = \mu + g_S + g_{\neg S} + e$$

$$M_0: \quad y = \mu + g + e$$

where M_p is the partitioned model, M_0 is the reduced model, y the vector of observations, μ mean, genetic effects $g_i \sim \mathrm{N}(0, \mathbf{G}_i\sigma_i^2)$ and residuals $e \sim \mathrm{N}(0, \mathbf{I}\sigma^2)$. Using groups of randomly sampled genes to partition the markers into two sets, S and $\neg S$, $\mathbf{G}_i$ is constructed as $\frac{\mathbf{ZZ}'}{v_p}$, where $\mathbf{Z}$ is the scaled and centred marker incidence matrix for the subset of markers in set S and $\neg S$, respectively, or all makers for M_0. $v_p = 2\sum_i p_i(1-p_i)$ for mastitis or just number of markers for startle response. For all results, only sets where the AI-REML algorithm had converged are utilised.

The likelihood ratio (LR) is defined as twice the difference between the log-likelihood of model M_p and M_0. Proportion of explained genetic variance by a set S is defined as:

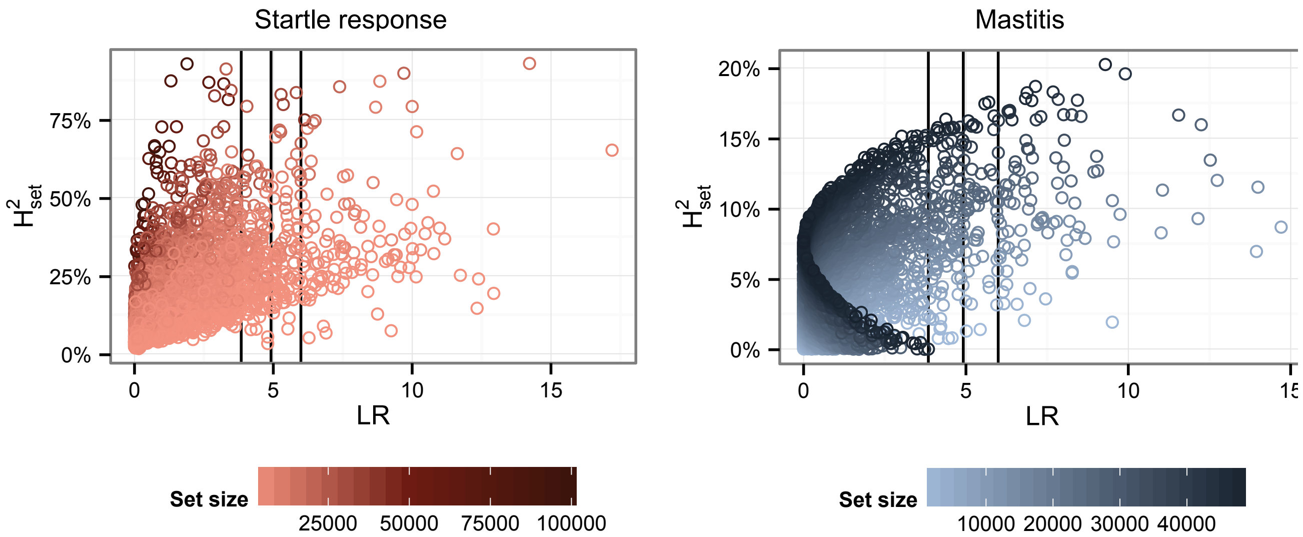

Figure 1: Different patterns of the relationship between likelihood ratios (x-axis) and proportion of explained genetic variance (y-axis), in startle response in fruit flies and mastitis in dairy cattle. Each point corresponds to a set of randomly sampled genes and colour coded by the number of markers in the set; note that startle response has a darker scale than mastitis to emphasise the larger sets. Vertical black lines corresponds to 95th percentile of the theoretical χ^2 distributions; left to right: χ^2_1, $(\chi^2_1 + \chi^2_2)/2$ and χ^2_2.

$$\mathrm{H}^2_{set} = \frac{\mathrm{Var}(g_S)}{\mathrm{Var}(g_S)+\mathrm{Var}(g_{\neg S})}$$

where $\mathrm{Var}(g_S)$ and $\mathrm{Var}(g_{\neg S})$ are the variance components σ^2_S and $\sigma^2_{\neg S}$ in mastitis, but scaled by the mean of the diagonals of $\mathbf{G}_S$ and $\mathbf{G}_{\neg S}$ in startle response.

Mapping genes to markers. Genetic markers were associated with a gene, if the chromosomal position of the marker was between the start and stop chromosomal position of the gene. This resulted in 88 % of markers mapped to genes in fruit flies and 39 % in dairy cattle.

The sets are composed of randomly sampled genes, resulting in sets with up to 100,000 and 50,000 markers for startle response and mastitis, respectively.

The complete set of results and a report with all figures can be found on fig**share** at http://goo.gl/FjVm91

Results and Discussion

In this study, we have applied a linear mixed modelling approach for examining the joint contribution to a complex trait from the genetic markers associated to a set of randomly sampled genes. Although it seems as a rather brute-force approach for gaining biological insight, these results reveal how much of the genetic variance can be explained jointly “by chance” by a *dispersed set* of genetic markers, and can therefore be used in collaboration with analyses on e.g. biological pathways or genomic features, to indicate the significance of these.

Fitting sets of randomly sampled genes, we found 28 % and 86 %, respectively, were converged for startle response and mastitis, respectively. The high level of non-converged sets in startle response might be attributed to these having a $\mathrm{Var}(g_S)$ close to zero, giving the AI-REML algorithm problems converging on the border of the parameter space. In these cases, $\mathrm{Var}(g_{\neg S})$ usually captures the genetic variance.

Even though the LR distributions seem alike (we will return to those) between the two traits, there does not seem to be a correlation between set size and LR for startle response, while there could be a slight increase in LR for mastitis for larger set sizes. In startle response, a set of random genes explains near 100 % of the genetic variance, while the maximal proportion of explained genetic variance by a set of random genes in mastitis is approx. 20 %. In both traits the maximal proportion of explained genetic variance increases for an increasing set size.

Comparing the relationship of H^2_{set} and LR, as done in Figure 1, reveals some interesting patterns. For startle response, the scatter plot resembles a slanted volcano plot, where a higher LR has a higher H^2_{set}. The larger sets are restricted to the lower values of LR, but can still account for a high proportion of explained genetic variance. For mastitis, there are hyperbolae for each group of set sizes and it would seem that for a certain set size is a ‘default’ proportion of explained genetic variance at LR = 0, and for increasing values of LR, H^2_{set} either increases or decreases. However, the decreasing ‘branch’ of the hyperbola does not reach higher LR than the 95th percentile of χ^2_1.

In these models, to achieve a high LR, it is crucial to describe the genetic relationship based on the causal genetic markers for the trait, or markers in high LD with these. In Holstein cattle, even a small proportion of the available markers will describe the genetic relationship, due to large LD blocks, whereas a similar proportion of genetic markers in fruit flies will have difficulty describing the genetic relationship, as there is more divergence between the lines.

A high LR can also be interpreted as the ability for the genetic markers in set S to separate the genetic signal (governing the trait) from the background noise of the genome. This is evident in mastitis, where there are sets that account for almost none of the genetic variance, but still retain a high LR. This is still a good separation of signal and noise, but instead of capturing the genetic signal in the

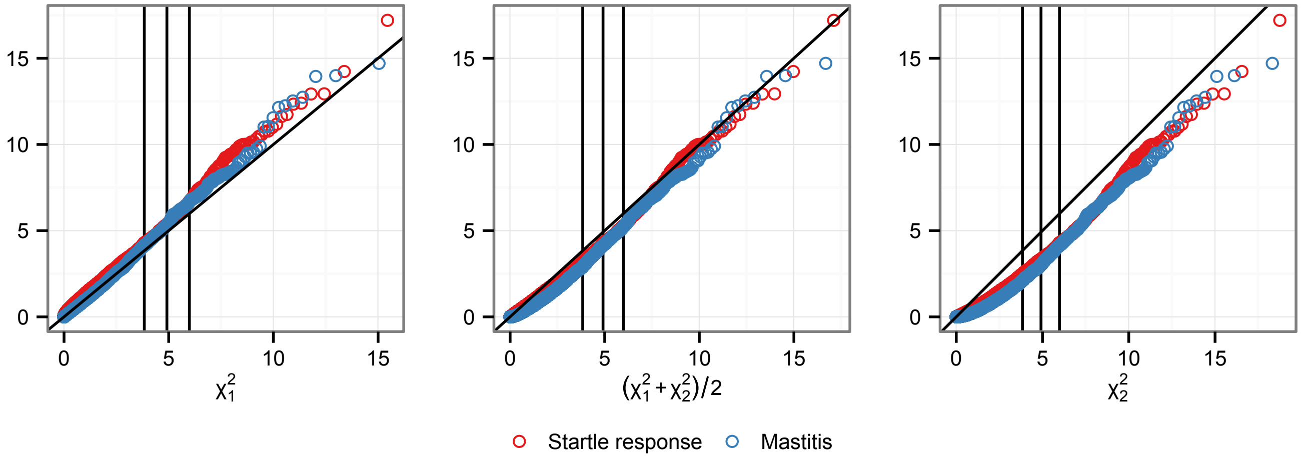

Figure 2: Likelihood ratio distributions appear similar to theoretical chi-squared distributions with 1 or 2 degrees of freedom or mixture of the two. Vertical black lines are as in figure 1.

set being tested, we are instead capturing pure noise. We do not experience this in startle response, as a set of genetic markers that contribute with nothing but noise, are unlikely to be able to describe the genetic relationship in fruit flies. In mastitis in dairy cattle, they can.

The empirical likelihood ratio distributions are displayed in Figure 2; although they appear similar to the theoretical chi-squared distributions, Kolmogorov-Smirnov tests reveal that they deviate significantly ($p \ll 0.01$), with the exception of mastitis not being significantly different from χ_1^2. The ratio of observed LR above the 95th percentile of the theoretical χ_1^2 distribution is 6.1 % and 5.8 % for startle response and mastitis, respectively, while for 2 degrees of freedom or mixture of 1 and 2 degrees of freedom, the proportion is 3.4 % or below. It is important to note that the test procedure used in this study is different from a standard likelihood ratio test used to test whether the variance component for a random effect is different from zero. In this context it has been shown that the empirical distribution of the likelihood ratios resembles a mixture of χ^2 distributions with 1 or 2 degrees of freedom (Self and Liang (1987)). However the results here indicate that a likelihood ratio test based on testing random gene groups follow a mixture of χ_1^2 and χ_2^2.

Conclusion

In this study we have compared variance decomposition in two species with different family structures. It is clear that the differences in family structure have an impact on the variance decomposition and the partitioning of genetic signal and noise.

We also found that using a χ^2 test for a likelihood ratio testing is an adequate choice for likelihood ratio tests; however in the Holstein cattle it is also necessary to include the estimated proportion of explained genetic variance to indicate the significance of a set of genetic markers selected by a a group of genes.

Acknowledgments

The work was in part funded by Quantomics a collaborative project under the 7th Framework Programme (contract no. KBBE-2A-222664) and the Danish Strategic Research Council (GenSAP: Centre for Genomic Selection in Animals and Plants, contract no. 12-132452).

Literature Cited

Kanehisa, M., Goto, S. (2000). *Nucleic Acids Res.* 28:27–30

Kanehisa, M., Goto, S., Sato, Y., et al. (2012). *Nucleic Acids Res.* 40(Database issue):109–14

Madsen, P., and Jensen, J. (2012). DMU. Version 6, release 5.1.

Mackay, T. F. C., Richards, S., Stone, E. A., et al. (2012). *Nature* 482:173–178

Self, S. G. and Liang, K. Y. (1987). *J. Amer. Statist. Assoc.* 82:605-610

Su, G., Brøndum, R. F., Ma, P., et al. (2012). *J. Dairy Sci.* 95:4657–65

Yamamoto, A., Anholt, R. R. H. and Mackay, T. F. C. (2009). *Genet. Res.* 91:373–382

Zimin, A. V., Delcher, A. L., Florea, L., et al. (2009). *Genome Biol.* 10:R42