Gene Based Association Approach Identify Genes Across Stress Traits in Fruit Flies

Palle Jensen[*,†], Stefan McKinnon Edwards[*], Pernille Merete Sarup[*], and Peter Sørensen[*]

[*]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University
[†]Section of Genetics, Ecology and Evolution, Department of Bioscience, Aarhus University

ABSTRACT: Identification of genes explaining variation in quantitative traits or genetic risk factors of human diseases requires both good phenotypic- and genotypic data, but also efficient statistical methods. Genome-wide association studies may reveal association between phenotypic variation and variation at nucleotide level, thus potentially identify genetic variants. However, testing million of polymorphic nucleotide positions requires conservative correction for multiple testing which lowers the probability of finding genes with small to moderate effects. To alleviate this, we apply a gene based association approach grouping variants accordingly to gene position, thus lowering the number of statistical tests performed and increasing the probability of identifying genes with small to moderate effects. Using this approach we identify numerous genes associated with different types of stresses in *Drosophila melanogaster*, but also identify common genes that affects the stress traits.

Key words: *Drosophila melanogaster*; environmental stress; genetic variants.

INTRODUCTION

Identification of genetic variants and genes explaining quantitative traits are central topics in modern biology ranging from evolutionary genetics to animal and plant breeding and human health.

Genome-wide association studies (GWAS) utilize genetic variation at nucleotide level (single nucleotide polymorphism, SNPs) to associate phenotypic variability with genetic variation assuming SNPs being in linkage disequilibrium with the causal variants. GWAS rarely identify all causal genes, and only identify genetic variants explaining a small proportion of the total genetic variance (Witte (2010)). Variants associated with phenotypes does not seem be randomly distributed across genomes, but are enriched for genes and biological pathways (Allen et al. (2010)). Grouping SNPs by their physical association to a gene will likely increase the probability of finding association. Firstly, performing millions of independent tests require subsequent adjustment because of multiple testing. Reducing the number of tests performed by limiting to the number of genes allows less conservative adjustment. Secondly, genome-wide significant SNPs only capture variants with large effects. Aggregating small effects of a number SNPs located within genes may increase their signal and thus increase the likelihood of detection.

The objective of this study was to apply a gene based association approach to identify genes important for different types of stress in fruit flies. We apply our method to a public available genetic resource, the ***D****rosophila melanogaster* **G**enetic **R**eference **P**anel (DGRP) (Mackay et al. (2012)).

MATERIALS AND METHODS

Data. The phenotypic- and genomic data applied originate from a public available reference population, the ***D****rosophila melanogaster* **G**enetic **R**eference **P**anel (DGRP) (Mackay et al. (2012)). The population was originally caught in Raleigh, North Carolina, USA and consists of 200 fully inbred (F≈1), independent lines, obtained using 20 generations of full-sib mating.

Initially SNPs were called from raw sequence data (as described in Mackay et al. (2012)) and included with coverage greater than 2X but less than 30X for which the minor allele frequency was present in at least 4 lines and if the SNP was called in minimum 60 lines. We imputed missing genotypes using Beagle Version 3.3.1 (Browning and Browning (2009)) resulting in a total of 2.5 million SNPs distributed among 140 million base pairs across the two autosomes (2L, 2R, 3L and 3R) and one sex chromosome, X.

We used three phenotypes in our analysis; chill coma recovery, starvation resistance and startle response (Mackay et al. (2012)).

Single-variant statistical analysis. Genome-wide association of single variants were conducted using linear mixed models in R-software (R Core Team (2013)) with lme4-package (Bates et al. (2013)); $y = \mu + L + R + S + SNP + \varepsilon$ with a Gaussian approximation of the traits. Phenotypic variances were partitioned between the DGRP lines (L, random), sex differences (S, random), replication (R, random) and a random error term, ε. Genotype effects (SNP, fixed) were assessed by comparing the model to a null model neglecting the fixed effect using a likelihood ratio test; $\Lambda = 2(l_N - l_G)$, where l_N and l_G were the log-likelihood for the null model and the alternative model including the SNP genotype effect. The test statistic, Λ, has an approximate χ^2-distribution with 1 degree of freedom.

Multi-variant statistical analysis. To alleviate the challenges posed by type 2 errors when performing high numbers of correlated tests, we apply a gene-based enrichment test coupled with permutation test to improve the statistical power of identifying true associations. We do this by grouping SNPs according to their physical location within annotated genes (using *Drosophila* annotation from Bioconductor (Carlson (2013))), thus reducing the number of tests from millions to ten of thousands.

We propose a summary statistic to identify genes that can explain the phenotypic variation within the traits. It is based on the χ^2-values from the association of individual genetic variants to the phenotypes using the single-variant

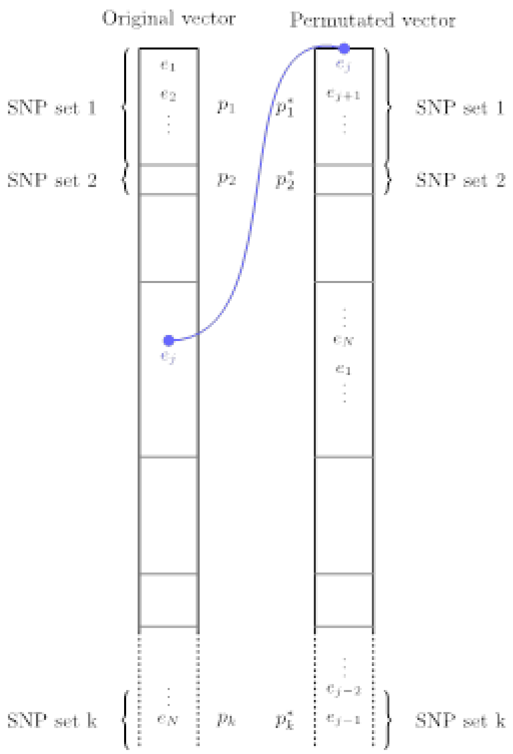

Figure 1. Gene-based permutation enrichment test.

approach described above. By summing the χ^2-values we imitate a genetic model capturing variants with small to moderate effects (Jiang and Gentleman (2007); Newton et al. (2007)). Using a permutation approach the observed summary statistic for a particular SNP set is compared to an empirical distribution for the summary statistics of random samples of SNP sets of same size. Consider a vector of test statistics, one for each SNP tested in the single-variant approach, ordered after their physical position of the SNP on the genome. As a consequence of linkage disequilibrium closely linked SNPs will likely be correlated, which will affect the distribution of the summary statistics, thus to account for this correlation structure we used the following procedure, figure 1. Let the vector of observed test statistics be ordered accordingly to the physical position on the genome of the corresponding SNPs. SNPs are then mapped to genes using the coordinates for the physical location of the genes on the genome. Let the elements in this vector be numbered *1*, *2*, …, *N*. The permutation consists of two steps. 1) Randomly pick an element (e_j) from this vector. Let this j^{th} test statistic be the first element in the permuted vector and the remaining elements ordered e_{j+1}, e_{j+2}, …, e_N, e_1, e_2, …, e_{j-1} accordingly to their original numbering. Thus, all elements from the original vector are now shifted to a new position starting with e_j. The mapping of genes is however kept fixed as to the original maping. 2) A summary statistic is computed for each SNP set based on the original SNP set position in the original vector of test statistics (figure 1). Hereby the link between SNPs and genes are broken while retaining the correlation structure among test statistics. Step 1 and 2 are repeated 10,000 times and from this empirical distribution of summary test statistics for each SNP set a *P*-value can be obtained. This empirical *P*-value corresponds to a one-side test of the proportion of randomly sampled summary statistics that are larger than the observed summary statistic. The arbitrary significance level was set to 0.01.

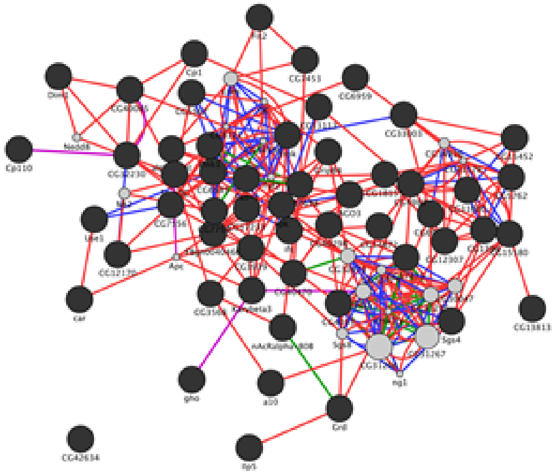

Figure 2. Network of enriched overlapping genes for chill coma recovery and startle response. Black nodes are enriched overlapping genes, and grey nodes are predicted interacting missing genes (see Warde-Farley et al., (2010)). Red edges are co-expressed genes, blue edges are co-localization, pink edges are physical interactions and green edges are shared protein domains.

Genes identified using the method described above was compared across traits. This was done by constructing an incidence matrix with *n* rows corresponding to the number of SNP sets (n = 14641 genes) and *m* columns corresponding to number of traits (m = 3). If a summary statistic for the SNP set was above the significance level the corresponding element in the incidence matrix was set to 1, otherwise zero. The observed overlap was then compared to an empirical distribution of the overlap. For a total of 10,000 times the elements within each column was permutated and the overlap among columns was recorded. The probability of the found overlap was estimated under the null hypothesis of independence of association among traits. We determined the empirical *P*-value of a one-side test as the fraction of all random permutations that was larger or equal to the observed overlap among traits at 5% level.

Population genetic parameters. Variance components were estimated using the AI-REML algorithm implemented in the DMU software (Madsen and Jensen (2012)) by fitting the linear mixed model $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zg} + \boldsymbol{e}$, where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{b}$ is a vector of fixed sex effect ($\boldsymbol{b} \sim N(0, \boldsymbol{I}\sigma_b^2)$) $\boldsymbol{g}$ is a vector of random genetic effects ($\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_g^2)$) and $\boldsymbol{e}$ is a vector of random error terms ($\boldsymbol{e} \sim N(0, \boldsymbol{D}\sigma_e^2)$). **X** and **Z** are design matrices relating records to fixed and random effects. Using best linear unbiased prediction (BLUP) the additive genetic ($\widehat{\boldsymbol{g}}_a$) and genomic values ($\widehat{\boldsymbol{g}}_g$) were predicted as:

$$Var(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{Z} \cdot Var(\boldsymbol{g}) \cdot \boldsymbol{Z}' + Var(\boldsymbol{e})$$

$$\boldsymbol{V} = \boldsymbol{Z} \cdot \boldsymbol{G} \cdot \boldsymbol{Z}' \cdot \sigma_g^2 + \boldsymbol{I} \cdot \sigma_e^2 \cdot \widehat{\boldsymbol{g}}$$

Table 1 Number of associated genes and number of overlapping genes

	Starvation	Startle	Chill coma
Starvation	**753**		
Startle	38 (0.54)	**736**	
Chill coma	34 (0.86)	58 (0.02)	**766**

The significant level of overlap between phenotypes is in parenthesis. Numbers of associated genes within trait are in bold and significant overlaps between traits are highlighted with red.

$$\hat{\boldsymbol{g}} = \boldsymbol{G} \cdot \boldsymbol{Z}' \cdot \sigma_g^2 \cdot \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}})$$

We distinguish between genetic and genomic variance using either the assumption of independence among lines using an identity matrix as correlation structure (**G**=**I**) or a correlation structure computed from genomic data. In the latter case, **G** is computed as in VanRaden (2008).

Pairwise genetic- and genomic correlations were calculated between $\hat{g}_a$ and $\hat{g}_g$. Furthermore, the genetic- and genomic broad sense heritability for each trait was computed as the fraction of additive genetic or genomic variance (scaled by the mean of diagonal elements of **G**) explained of total phenotypic variance.

RESULTS AND DISCUSSION

We used a gene-based approach to identify genes associated with stress traits in fruit flies. Grouping SNPs according to genes lowered the number of statistical tests considerable. Still, at 5% significance level about 730 genes are expected to be associated due to chance; we found 736 to 766 genes significant in the three traits (table 1). In all cases however, we identify more genes than expected. Associated top genes for chill coma recovery include *Dim1* (neurogenesis and mitotic spindle organization), *nAChRα4* (ion transport) and *a10* (response to chemical stimuli) and several genes with unknown functions. For startle response we identified *Rpk* (sensory perception to touch) and numerous genes with unknown functions. Interestingly, 57 of the 58 genes in common for startle response and chill coma recovery are connected in one big network (figure 2, created using GeneMania, an online tool, that uses data from different public databases, and a weighting algorithm so the query genes connect as much as possible but also extend the network with predicted similar genes, see Warde-Farley et al. (2010)). Some of these genes have many connections in the network (*Grip84*, *Rpk*, *Pink1* and *Ifc*). In particular, these genes have known functions in response to perception of touch, oxidative stress and oxidation-reduction processes.

We obtained intermediate broad sense heritabilities for the three traits (table 2). The heritabilities based on genomic information is overall higher than those based on the simple variance structure of independent relationships among lines. However, the differences are minor and the heritabilities computed on genetic and genomic data are interpreted as being equal.

Pairwise correlations between the genetic and genomic values from the BLUP were computed, showing no correlations between traits (table 2). Despite lack of significant correlations, the multi-variant statistical approach showed a significant overlap between enriched genes for startle response and chill coma recovery (table 1). Our results indicate that although a pair of complex traits appears to be genetically uncorrelated they may be influenced by genetic variants in common genes.

Table 2 Genetic- and genomic correlations, below and above diagonal elements respectively.

	Starvation	Startle	Chill coma
Starvation	**0.38 : 0.46**	0.02 (0.77)	-0.01 (0.92)
Startle	0.11 (0.17)	**0.44 : 0.50**	0.00 (0.97)
Chill coma	0.03 (0.67)	0.01 (0.89)	**0.37 : 0.42**

P-values of no correlation are in parenthesis. Diagonal elements are estimated broad sense heritability for each trait; (genetic heritability : genomic heritability).

The strength of the method used in this study are *i*) it may increase the power to detect genetic variants with small effects using prior biological knowledge, *ii*) facilitation of the biological interpretation of significant results when used in collaboration with clustering tools and *iii*) the model can easily be extended to other important biological groupings, such as biological pathways, gene expression data or linkage patterns or other genomic features.

CONCLUSION

We identified a number of genes associated to stress traits in flies. The top ranking genes appear to be part of network of functionally related genes. There is a significant overlap among the topranking genes across traits although these traits are genetically uncorrelated.

ACKNOWLEDGMENTS

This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research.

LITERATURE CITED

Allen, H. L., Estrada, K., Lettre, G., et al. (2010). *Nature*, 467:832-838.

Bates, D., Maechler, M., Bolker, B., et al. (2013). Retrieved from http://cran.r-project.org/package=lme4.

Browning, B. L. and Browning, S. R. (2009). *Am. J. Hum. Genet.*, 84(2):210–223.

Carlson, M. (2013). org.Dm.eg.db: Genome wide annotation for Fly. R package version 2.9.0.

Jiang, Z. and Gentleman, R. (2007). *Bioinformatics*, 23(3):306-313.

Mackay, T. F. C., Richards, S., Stone, E. A., et al., (2012). *Nature*, *482*(7384):173–178.

Madsen, P. and Jensen, J. A. (2012). DMU. Version 6, release 5.2.

Newton, M., Quintana, F. A., Boon, J. A., et al (2007). *Ann. Appl. Stat.*, 1(1):85-106.

R Core Team. (2013). Version 3.0.2.

VanRaden, P. M. (2008). *J. Dairy Sci.*, 91(11):4414-4423.

Warde-Farley, D., Donaldson, S. L., Comes, O. et al., (2010). *Nucleic Acids Res.*, 38:W214-220.

Witte, J. S. (2010). *Annu. Rev. Publ. Health.*, 31:9-20.