# Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in *Drosophila melanogaster*

**Stefan M. Edwards,\*,† Izel F. Sørensen,\* Pernille Sarup,\* Trudy F. C. Mackay,‡ and Peter Sørensen\*,1**

\*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark, †The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG Scotland, United Kingdom, and ‡Department of Biological Sciences and Program in Genetics, North Carolina State University, Raleigh, North Carolina 27695

**ABSTRACT** Predicting individual quantitative trait phenotypes from high-resolution genomic polymorphism data is important for personalized medicine in humans, plant and animal breeding, and adaptive evolution. However, this is difficult for populations of unrelated individuals when the number of causal variants is low relative to the total number of polymorphisms and causal variants individually have small effects on the traits. We hypothesized that mapping molecular polymorphisms to genomic features such as genes and their gene ontology categories could increase the accuracy of genomic prediction models. We developed a genomic feature best linear unbiased prediction (GFBLUP) model that implements this strategy and applied it to three quantitative traits (startle response, starvation resistance, and chill coma recovery) in the unrelated, sequenced inbred lines of the *Drosophila melanogaster* Genetic Reference Panel. Our results indicate that subsetting markers based on genomic features increases the predictive ability relative to the standard genomic best linear unbiased prediction (GBLUP) model. Both models use all markers, but GFBLUP allows differential weighting of the individual genetic marker relationships, whereas GBLUP weighs the genetic marker relationships equally. Simulation studies show that it is possible to further increase the accuracy of genomic prediction for complex traits using this model, provided the genomic features are enriched for causal variants. Our GFBLUP model using prior information on genomic features enriched for causal variants can increase the accuracy of genomic predictions in populations of unrelated individuals and provides a formal statistical framework for leveraging and evaluating information across multiple experimental studies to provide novel insights into the genetic architecture of complex traits.

**KEYWORDS** genomic feature models; best linear unbiased prediction; *Drosophila* Genetic Reference Population; startle response; starvation resistance; chill coma recovery time; genomic selection; GenPred; shared data resource

**V**ARIATION for complex traits is due to many interacting loci with small individual effects on each trait as well as environmental influences (Falconer and Mackay 1996). Knowledge of the underlying causal polymorphisms and their effects is thus critical for predicting disease susceptibility in humans, improving production traits in plants and animals, and predicting adaptive evolution. Genetic mapping approaches to dissect the genotype–phenotype map

one locus at a time in outbred populations are successful in identifying quantitative trait loci (QTL) with the largest effects, but together these loci typically account for only a small fraction of the total genetic variance (Visscher 2008, 2012; Manolio *et al.* 2009; Vinkhuyzen *et al.* 2013; Caballero *et al.* 2015).

The realization that the effects of the majority of loci affecting complex traits are too small to be individually detected unless sample sizes are huge motivated the development of statistical methods to predict complex trait phenotypes, using all molecular markers simultaneously (Meuwissen *et al.* 2001). These genomic prediction methods are very successful in predicting phenotypes from marker genotypes in populations with a large amount of linkage disequilibrium (LD), such as selectively bred animals and plants (de Roos *et al.* 2009;

Hayes *et al.* 2009, 2010; Crossa *et al.* 2010; Daetwyler *et al.* 2013). However, genomic prediction does not work well when LD is low, such as across breeds or strains or in outbred populations of largely unrelated individuals (Habier *et al.* 2007; Makowsky *et al.* 2011; de los Campos *et al.* 2013), because many causal polymorphisms will not be in LD with the genotyped markers.

Genomic predictions currently utilize high-density single-nucleotide polymorphism (SNP) genotyping arrays. Given the rapid advances in sequencing technologies, genomic prediction soon will be based on whole-genome sequence data. This will greatly exacerbate the true genomic signal to noncausal marker noise problem. Therefore, the key to better prediction models may be to "guess" which markers could be causal by utilizing prior biological findings, as it appears that the markers associated with trait variation are not uniformly distributed throughout the genome, but enriched in genes that are connected in biological pathways (Lango Allen *et al.* 2010; O'Roak *et al.* 2012; Lage *et al.* 2012; Maurano *et al.* 2012; Peñagaricano *et al.* 2013). Here, we extend the commonly used genomic best linear unbiased prediction (GBLUP) model (Meuwissen *et al.* 2001) by incorporating prior information on gene ontologies (Gene Ontology Consortium *et al.* 2000). Our model, which we call genomic feature BLUP (GFBLUP), includes an additional genomic effect that quantifies the collective action of a set of markers located in a genomic feature defined by genes, biological pathways, sequence annotation, or other external evidence. We previously used this approach to partition the genomic variance of pathways in health and milk production traits in Danish Holstein dairy cattle (Edwards *et al.* 2015). Here, we apply it to three quantitative traits (starvation resistance, startle response, and chill coma recovery time) in the largely unrelated inbred, sequenced lines of the *Drosophila melanogaster* Genetic Reference Panel (DGRP) (Mackay *et al.* 2012; Huang *et al.* 2014). Prediction accuracies from previous GBLUP analyses of these traits in the DGRP ranged from zero to very low (Ober *et al.* 2012, 2015).

The premise of the GFBLUP model is that genomic features are enriched for causal variants affecting the traits. However, in reality, the number, location, and effect sizes of the true causal variants in the genomic feature are unknown. Therefore, we also used simulation to investigate feature- and trait-specific factors that may influence predictive ability, using the GFBLUP model. Genomic feature factors include the proportion of the total genomic variance that can be explained by the genomic feature, the number of noncausal variants that is included, and the distribution of the causal variants in the genome (either distributed randomly or clustered in smaller genome regions). The trait-specific factors include the total genomic heritability of the trait and the number of phenotypic records available for analysis.

Our GFBLUP models applied to the DGRP provide better model fits and have a higher predictive ability than the standard GBLUP model. The GFBLUP models provide novel insight into the genetic architecture of starvation resistance, startle response, and chill coma recovery by identifying genomic features that explain large proportions of genomic variance. Finally, our simulated data generated from DGRP genotypes illustrate factors affecting estimation of genomic parameters, model fit, and predictive ability.

## Materials and Methods

### DGRP data

***Drosophila lines:*** The phenotypic and genotypic data originate from the DGRP (Mackay *et al.* 2012; Huang *et al.* 2014). All data can be accessed via the Web site http://dgrp2.gnets.ncsu.edu. The DGRP consists of 205 inbred lines obtained by 20 generations of full-sib mating from the offspring of single wild-caught females collected from the Raleigh, North Carolina population, which have full genome sequences (Mackay *et al.* 2012; Huang *et al.* 2014). All flies were reared under standard culture conditions (cornmeal–molasses–agar medium, 25°, 60–75% relative humidity, 12-hr light–dark cycle). The DGRP is polymorphic for common inversions and *Wolbachia pipientis* infection status (Huang *et al.* 2014). These factors were included in the models described below as fixed effects.

***Quantitative trait phenotypes:*** Starvation resistance for 197 DGRP lines was assessed by placing 10 same-sex, 2-day-old flies in culture vials containing nonnutritive medium (1.5% agar and 5 ml water) and scoring survival every 8 hr until all flies were dead (Harbison *et al.* 2004). There were five replicate vials per sex per line (total $N = 19,361$; female $N = 9672$; male $N = 9689$). Chill coma recovery for 159 DGRP lines was measured by transferring 3- to 7-day-old flies without anesthesia to empty vials and placing them on ice for 3 hr. Flies were transferred to room temperature, and the time it took for each individual to right itself and stand on its legs was recorded (Morgan and Mackay 2006). There were two replicates of 50 flies per sex per line (total $N = 32,231$; female $N = 16,170$; male $N = 16,061$). Startle response for 166 DGRP lines was measured by placing single 3- to 7-day-old adult flies, collected under $CO_2$ exposure, into vials containing 5 ml culture medium and leaving them overnight to acclimate to their new environment. On the next day, between 8 AM and 12 PM (2–6 hr after lights on), each fly was subjected to a mechanical disturbance with a gentle tap, and the total amount of time the fly was active in the 45 sec immediately following the disturbance was recorded (Mackay 2001). There were two replicates of 20 flies per sex per line (total $N = 13,276$; female $N = 6674$; male $N = 6602$).

***Genotypes:*** Genotypes were obtained from whole-genome sequences, using an integrative genotyping procedure (Huang *et al.* 2014). All analyses were based on segregating biallelic SNPs with minor allele frequencies ≥0.05 and for which the Phred scaled variant quality was >500 and the genotype call rate was ≥0.8, for a total of 1,725,755 SNPs distributed on five chromosome arms (*2*L, *2*R, *3*L, *3*R, and *X*).

***Genomic features:*** Genes grouped according to a specific Gene Ontology (GO) term were considered a genomic

feature. Genes were linked to the "biological processes" (BP), "molecular function" (MF), and "cellular component" (CC) GO terms (Gene Ontology Consortium *et al.* 2000), using the BioConductor package "org.Dm.eg.db" v. 2.14 (Carlson 2013). Only GO terms with at least 10 directly evidenced genes were used in the analyses. SNPs were mapped to FlyBase genes using the v5.49 annotations of the *D. melanogaster* reference genome (Tweedie *et al.* 2009; Mackay *et al.* 2012; Huang *et al.* 2014). Only the 963,235 SNPs located within genes (*i.e.*, within open reading frames) were used for the genomic feature. In total markers were associated with 10,517 genes and 1145 GO terms. A total of 1,725,755 markers were used in all analyses, and the number of markers linked to a single GO term ranged from 23 to 163,938.

### Statistical analyses using linear mixed models

Analyses were performed using two different linear mixed models: a standard GBLUP model and a GFBLUP model using prior information on genomic features. In the following we present details of the models and statistical procedures used to compare the models.

***Genomic models:*** For each genomic feature (*i.e.*, GO term) a separate analysis was conducted. In each of the GFBLUP analyses we evaluated a single genomic feature based on a linear mixed model including two random genomic effects,

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{Zf} + \mathbf{Zr} + \mathbf{Z}_l\mathbf{l} + \mathbf{e} \ (M_{\mathrm{GFBLUP}}),$$

where $\tilde{\mathbf{y}}$ is the vector of adjusted phenotypic observations (or simulated phenotypes), $\boldsymbol{\mu}$ is the vector of an overall mean, $\mathbf{Z}$ is the design matrix linking observations to genomic values ($\mathbf{f}$ and $\mathbf{r}$), $\mathbf{Z}_l$ is a design matrix for the replicate within-line effects ($\mathbf{l}$), $\mathbf{f}$ is the vector of line-specific genomic values captured by genetic markers linked to the genomic feature of interest, $\mathbf{r}$ is the vector of line-specific genomic values *not* captured by genetic markers linked to the genomic feature, and $\mathbf{e}$ is the vector of residuals. The random genomic effects and the residuals were assumed to be independent normally distributed values described as follows: $\mathbf{f} \sim N(0, \mathbf{G_f}\sigma_f^2)$, $\mathbf{r} \sim N(0, \mathbf{G}\sigma_r^2)$, $\mathbf{l} \sim N(0, \mathbf{I_l}\sigma_l^2)$, and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$.

GBLUP was based on a linear mixed model including only one random genomic effect,

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{Zg} + \mathbf{Z}_l\mathbf{l} + \mathbf{e} \ (M_{\mathrm{GBLUP}})$$

with the same notation as above except that $\mathbf{g}$ is the vector of genomic values captured by all genetic markers. The random genomic values and the residuals were assumed to be independent normally distributed values described as follows: $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$.

In the analyses of the simulated data the term for the replicate within-line effects was excluded in both the genomic model analyses.

The additive genomic relationship matrix $\mathbf{G}$ (VanRaden 2008) was constructed using all genetic markers as follows: $\mathbf{G} = \mathbf{WW}'/m$, where $\mathbf{W}$ is the centered and scaled genotype matrix, and $m$ is the total number of markers. Each column vector of $\mathbf{W}$ was calculated as follows: $\mathbf{w_i} = (\mathbf{m_i} - 2p_i)/\sqrt{2p_i(1-p_i)}$, where $p_i$ is the minor allele frequency of the $i$th genetic marker and $\mathbf{m_i}$ is the $i$th column vector of the allele count matrix, $\mathbf{M}$, which contains the genotypes coded as 0 or 2 depending on the number of minor alleles. The additive genomic relationship matrix for the genomic feature $\mathbf{G_f}$ was constructed in a similar way, using only the genetic marker set defined by the genomic feature.

***Adjusted phenotypes used in genomic model analyses:*** The phenotypes used in the GBLUP and GFBLUP model analyses were derived from phenotypic records of the quantitative traits adjusted for relevant factors, using the linear mixed model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{Z}_l\mathbf{l} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypic observations, $\mathbf{X}$ is the design matrix, $\mathbf{b}$ is the vector of fixed effects of inversion karyotypes and *Wolbachia* infection status, $\mathbf{Z}$ is the design matrix linking observations to genomic values, $\mathbf{Z}_l$ is a design matrix for the replicate within-line effects, $\mathbf{g}$ is the vector of genomic values captured by all genetic markers, $\mathbf{l}$ is the vector of replicate within-line effects, and $\mathbf{e}$ is the vector of residuals. The random effects ($\mathbf{g}$ and $\mathbf{l}$) and the residuals were assumed to be independent normally distributed values described as follows: $\mathbf{l} \sim N(0, \mathbf{I_l}\sigma_l^2)$, $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$, and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. The adjusted phenotypes used as response variables for genomic model analysis described below were calculated as $\tilde{\mathbf{y}} = \mathbf{Z}\hat{\mathbf{g}} + \mathbf{Z}_l\hat{\mathbf{l}} + \hat{\mathbf{e}}$.

***Estimation of variance components:*** Estimates of the variance components ($\hat{\sigma}_f^2, \hat{\sigma}_r^2, \hat{\sigma}_g^2, \hat{\sigma}_l^2$, and $\hat{\sigma}_e^2$) defined in the models described above were obtained using an average information restricted maximum-likelihood (REML) procedure (Madsen *et al.* 1994; Johnson and Thompson 1995) as implemented in the software DMU (Madsen *et al.* 1994). In this procedure, we used a generalized inverse of the genomic relationship matrices. This was necessary because these matrices were not full rank due to centering, as well as in cases where the number of genetic markers was smaller than the number of lines.

***Model comparisons:*** The models were evaluated and compared based on model fit, model predictive ability, and precision of estimated genomic parameters.

*Genomic parameters:* Genomic parameters were derived from the estimates of the variance components. Inferences on genomic heritability were based on the following ratios: $\hat{h}_{\mathrm{GBLUP}}^2 = \hat{\sigma}_g^2/(\hat{\sigma}_g^2 + \hat{\sigma}_e^2)$ for GBLUP and $\hat{h}_{\mathrm{GFBLUP}}^2 = (\hat{\sigma}_f^2 + \hat{\sigma}_r^2)/(\hat{\sigma}_f^2 + \hat{\sigma}_r^2 + \hat{\sigma}_e^2)$ for GFBLUP. Inference on partitioning of genomic variance in GFBLUP was

based on the following ratios: $\hat{h}_f^2 = \hat{\sigma}_f^2/(\hat{\sigma}_f^2 + \hat{\sigma}_r^2)$ and $\hat{h}_r^2 = \hat{\sigma}_r^2/(\hat{\sigma}_f^2 + \hat{\sigma}_r^2)$. These ratios quantify the proportion of total genomic variance captured $(\hat{h}_f^2)$ and not captured $(\hat{h}_r^2)$ by the genetic markers in the genomic feature.

*Model fit:* The model fit was assessed using a likelihood ratio (LR) here defined as $2\log lh_{M_G} - 2\log lh_{M_{GF}}$, where $lh$ is the likelihood of the fitted model. Standard theory shows that the LR test statistic is asymptotically distributed as $\chi_\kappa^2$, where $\kappa$, the degrees of freedom, is the difference in the number of parameters between the two models. The P-values calculated for assessing the significance of the likelihood-ratio test were based on a $\chi^2$-distribution with 1 d.f. However, it has previously been shown that if the null hypothesis value is on the boundary of the parameter space (*e.g.*, $\sigma_f^2 = 0$), the asymptotic distribution of LR test statistics may be approximated by a 50:50 mixture of $\chi^2$-distributions with 0 and 1 d.f. (Self and Liang 1987). Alternatively it is possible to derive an empirical distribution of the LR test statistic as we have shown previously (Edwards *et al.* 2015).

*Model predictive ability:* The ability of the models to predict total genomic value was assessed using a cross-validation procedure. In the GFBLUP model the total genomic value is $\hat{g}_{total} = \hat{f} + \hat{r}$ and in GBLUP it is $\hat{g}_{total} = \hat{g}$. In the cross-validation procedure, we estimated genomic parameters using the phenotypes from the lines in the training data (90% of the lines) and predicted the total genomic value of lines in the validation data (10% of the lines). BLUP is used to predict total genomic values in both the GFBLUP and GBLUP models as described below. We then calculated a correlation between the total genomic values predicted with or without the observed phenotypes set to missing. For the simulated data and for the observed DGRP data we defined 50 cross training (validation) data subsets and applied these to each genomic feature. For each genomic feature, the predictive ability was defined as the average correlation of the 50 cross validations. A similar procedure was applied to the GBLUP model serving as a reference. For each genomic feature Welch's t-test (*i.e.*, unequal variance t-test) was used to test the difference in mean predictive ability of the two models (Welch 1947). For each trait 1145 GO terms were tested. Therefore P-values from Welch's t-test were adjusted for multiple testing by controlling the false discovery rate as implemented in R (Benjamini and Hochberg 1995; R Core Team 2015).

*Prediction of total genomic value using BLUP:* The total genomic value of lines in the validation data was predicted conditional on the observed phenotypes for the lines in the training data. In the GFBLUP model the conditional expectation of the total genomic values for the lines in the validation data ($\hat{g}_1 = \hat{f}_1 + \hat{r}_1$) given the observed phenotypes for the lines in the training data ($y_2$) can be written as

$$\hat{g}_1 = E(g_1|y_2) = (G_{f_{12}}\hat{\sigma}_f^2 + G_{r_{12}}\hat{\sigma}_r^2)[G_{f_{22}}\hat{\sigma}_f^2 + G_{r_{22}}\hat{\sigma}_r^2 + I_{22}\hat{\sigma}_e^2]^{-1}(y_2 - X_2\hat{b}_2),$$

where the genomic relationship matrix for the genomic feature $G_f = \begin{pmatrix} G_{f_{11}} & G_{f_{12}} \\ G_{f_{21}} & G_{f_{22}} \end{pmatrix}$ is partitioned according to relationships between the lines in the training data ($G_{f_{11}}$), between the lines in the validation data ($G_{f_{22}}$), and between the lines in the training and validation data ($G_{f_{12}}$). A similar partitioning is applied to $G_r$ and $I$ in the GFBLUP model. For the sake of simplicity we have ignored the design matrix $Z$ and the replicate within-line effects in the expression for the conditional expectation. Thus the total genomic value is predicted using the estimated variance components ($\hat{\sigma}_f^2$, $\hat{\sigma}_r^2$, and $\hat{\sigma}_e^2$) in the training data. The rightmost term, ($y_2 - X_2\hat{b}_2$), constitutes the phenotypes corrected for fixed effects for the lines in the training data. The inverse term $[G_{f_{22}}\hat{\sigma}_f^2 + G_{r_{22}}\hat{\sigma}_r^2 + I_{22}\hat{\sigma}_e^2]$ is essentially the variance–covariance structure for the corrected phenotypes. These two terms together are the standardized and corrected phenotypes for the individuals in the training data, which are projected onto the total genetic covariance structure between the training and the validation data, ($G_{f_{12}}\hat{\sigma}_f^2 + G_{r_{12}}\hat{\sigma}_r^2$).

In the GBLUP model a similar expression for the conditional expectation of the total genomic value for the lines in the validation data given the observed phenotypes for the lines in the training data can be written as

$$\hat{g}_1 = E(g_1|y_2) = (G_{12}\hat{\sigma}_g^2)[G_{22}\hat{\sigma}_g^2 + I_{22}\hat{\sigma}_e^2]^{-1}(y_2 - X_2\hat{b}_2).$$

***Implementation:*** The GFBLUP and GBLUP modeling approaches are implemented in the R package qgg, which is available at http://psoerensen.github.io/qgg/. This includes fitting a series of linear mixed models, estimating variance components using REML, prediction using BLUP, and cross-validation procedures. Example scripts and data sets are provided for illustrating the GFBLUP and GBLUP modeling approaches. For a specific experimental design with replicated phenotypes within line such as DGRP it is more efficient to use the average information REML procedure (Madsen *et al.* 1994; Johnson and Thompson 1995) implemented in DMU (Madsen *et al.* 1994). The aireml function in the qgg package provides an R interface to the DMU that can be downloaded from http://dmu.agrsci.dk/DMU/.

### Simulated data

We established a series of simulation studies to investigate factors influencing the power to detect genomic features affecting the trait phenotype, estimation of genomic parameters, and prediction ability of the two tested linear mixed models. The factors varied in the simulations included genomic heritability ($h^2$), proportion of genomic variance explained by causal SNPs in the genomic feature ($h_f^2$),

proportion of noncausal SNPs in the genetic marker set defined by the genomic feature (*dilution*), genome distribution of causal SNPs (*causal model*) (*i.e.*, how the causal SNPs were physically distributed on the genome: random or clustered), and the number of replicates (*i.e.*, the number of phenotypic records for each line) within lines ($N_{rep}$).

***Genotypes:*** The simulations were based on the real genotype DGRP data set of 205 lines and 1,725,755 SNPs. In all scenarios, there were 1000 causal SNPs. Causal sets were divided into two subsets. The first subset $C_1$ included 100 SNPs and was used as the causal SNP set in the genomic feature that explains 10%, 20%, 30%, or 50% of the total genomic variance. The second subset $C_2$ included 900 SNPs and explained the remaining genomic variance. To mimic relevant genetic scenarios, the genome distribution of the causal SNPs in the genomic feature was simulated using two different causal models: a *random* and a *cluster* model. The cluster model simulates the situation in which multiple causal SNPs occur in a limited number of genes, whereas in the random model single causal SNPs occur in a larger number of genes. The main difference is that the genomic variance is associated with a smaller genome region in the cluster model compared to the random model. For the clustered causal model, the 100 causal SNPs in $C_1$ were chosen from 20 randomly selected genome regions spanning 50 SNPs each, and the remaining 900 SNPs in $C_2$ were randomly selected from the complete SNP set (excluding the SNPs in $C_1$). For the random causal model, the SNPs in $C_1$ and $C_2$ were randomly selected from the complete SNP set. To investigate the effects of noncausal SNPs within the causal sets, we added an increasing number of noncausal SNPs (200, 400, 800, . . . , 2000) to the causal sets, in a process referred to as dilution. The noncausal SNPs were picked either at random in the genome or by sampling SNPs located directly up- and downstream of the causal SNP (referred to as local SNPs).

***Phenotypes:*** Phenotypes were simulated using the following linear model: $\mathbf{y} = \mathbf{g}_1 + \mathbf{g}_2 + \mathbf{e}$, where $\mathbf{g}_1 \sim N(0,\mathbf{G}_1{*}\sigma_{g1}^2)$, $\mathbf{g}_2 \sim N(0,\mathbf{G}_2{*}\sigma_{g2}^2)$, and $\mathbf{e} \sim N(0,\mathbf{I}{*}\sigma_e^2)$. $\mathbf{G}_1$ and $\mathbf{G}_2$ are the genomic relationship matrices for causal SNPs in $C_1$ and $C_2$, respectively. The total phenotypic variance $\sigma_p^2 = \sigma_{g1}^2 + \sigma_{g2}^2 + \sigma_e^2$ was 100 in all scenarios. We simulated data under additive genomic heritabilities $\left(h^2 = (\sigma_{g1}^2 + \sigma_{g2}^2)/(\sigma_{g1}^2 + \sigma_{g2}^2 + \sigma_e^2)\right)$ of 0.1, 0.3, and 0.5, to analyze scenarios with low to intermediate heritabilities, reflecting those observed in the real data. To analyze scenarios with nonuniform SNP effects, the proportion of additive genomic variance explained by the causal SNPs in $C_1$ $\left(h_f^2 = \sigma_{g1}^2/(\sigma_{g1}^2 + \sigma_{g2}^2)\right)$ was varied across scenarios: 0.1, 0.2, 0.3, and 0.5. These parameters were investigated for $N_{rep}$ of 5, 10, and 50. Increasing the number of replicates per line decreases the variance of the phenotypic value for each line.

Combining these factors yielded a total of 1440 individual simulated data sets [3 ($N_{rep}$) × 3 ($h^2$) × 4 ($h_f^2$) × 2 (causal model) × 20 (dilution)], which were each replicated 50 times. For each data set and replicate we estimated variance components for the GBLUP and GFBLUP models, using REML. Model fit was assessed using the likelihood-ratio test and model predictive ability using the cross-validation procedure described above. The statistics $\hat{h}_f^2$, $\hat{h}_r^2$, $\hat{h}_{GFBLUP}^2$, $\hat{h}_{GBLUP}^2$, LR, power, and predictive ability were calculated and are summarized in *Results*.

***Detection power:*** Power was calculated for each of the simulated scenarios defined in detail above. The *P*-values used for determining the power of the likelihood-ratio test were calculated based on the theoretical $\chi^2$-distribution with 1 d.f. For each of the simulated scenarios, power was calculated as the fraction (of 50 replicates) of the analyses that led to a significantly better model fit using the GFBLUP model compared to using the GBLUP model (*i.e.*, the observed *P*-value was <0.05).

### Data availability

The DGRP data can be accessed via the website at http://dgrp2.gnets.ncsu.edu.

## Results

The DGRP genotypes include ~1.7 million common (minor allele frequencies ≥0.05) SNPs derived from genomic sequences of 205 largely unrelated inbred lines (Mackay *et al.* 2012; Huang *et al.* 2014). We evaluated and compared GFBLUP and GBLUP models based on model fit, model predictive ability, and precision of estimated genomic parameters, using both observed genotypes and phenotypes for three quantitative traits in the DGRP (chill coma recovery time, starvation resistance, and startle response) (Mackay *et al.* 2012; Ober *et al.* 2012, 2015; Huang *et al.* 2014) and simulated genotypic and phenotypic data.

### GBLUP and GFBLUP analyses in the DGRP

We performed GBLUP and GFBLUP prediction analyses for each of the three traits, using 10-fold cross-validation; *i.e.*, the training data consisted of 90% of the lines and the total genomic value was validated in 10% of the lines. Males and females were analyzed separately.

***Predictive ability:*** The predictive ability of the GBLUP models was low. The predictive ability of GBLUP for females (males) was 0.055±0.029 (0.00±0.032) for chill coma recovery, 0.25±0.029 (0.27±0.027) for starvation resistance, and 0.25±0.033 (0.28±0.029) for startle response. The low values are not statistically different from those previously reported using fivefold cross-validation GBLUP models, which were 0.24 for starvation resistance and 0.23 for startle response in both sexes (Ober *et al.* 2012) and 0.1 for female and zero for male chill coma recovery time (Ober *et al.* 2015). Furthermore, predictive ability for starvation resistance and startle response was not improved using either a Bayesian mixture model that

**Table 1 Ten most significant predictions for chill coma recovery with the GFBLUP model**

| Sex | GO ID[a] | PA[b] ± SEM[c] | Padj[d] | LR[e] | Hf[f] | Nsets[g] | Gene Ontology term |
|---|---|---|---|---|---|---|---|
| Female | GO:0007266 | 0.370 ± 0.022 | $1.8 \times 10^{-10}$ | 11.39 | 0.31 | 3,139 | Rho protein signal transduction |
| | GO:0005100 | 0.365 ± 0.023 | $4.0 \times 10^{-10}$ | 12.67 | 0.37 | 3,886 | Rho GTPase activator activity |
| | GO:0007173 | 0.343 ± 0.026 | $1.9 \times 10^{-8}$ | 15.96 | 0.92 | 9,674 | Epidermal growth factor receptor signaling pathway |
| | GO:0030031 | 0.318 ± 0.027 | $6.7 \times 10^{-7}$ | 11.49 | 0.74 | 2,700 | Cell projection assembly |
| | GO:0035160 | 0.309 ± 0.027 | $1.5 \times 10^{-6}$ | 9.65 | 0.39 | 5,011 | Maintenance of epithelial integrity; open tracheal system |
| | GO:0016323 | 0.299 ± 0.026 | $2.7 \times 10^{-6}$ | 9.13 | 0.47 | 7,761 | Basolateral plasma membrane |
| | GO:0035277 | 0.280 ± 0.027 | $2.3 \times 10^{-5}$ | 11.12 | 0.56 | 7,582 | Spiracle morphogenesis; open tracheal system |
| | GO:0007494 | 0.263 ± 0.025 | $5.8 \times 10^{-5}$ | 8.86 | 0.58 | 9,614 | Midgut development |
| | GO:0006406 | 0.288 ± 0.033 | $8.3 \times 10^{-5}$ | 12.53 | 0.52 | 1,530 | mRNA export from nucleus |
| | GO:0005089 | 0.253 ± 0.026 | $2.2 \times 10^{-4}$ | 9.54 | 0.67 | 11,922 | Rho guanyl-nucleotide exchange factor activity |
| Male | GO:0030031 | 0.316 ± 0.023 | $6.5 \times 10^{-9}$ | 9.26 | 0.62 | 2,700 | Cell projection assembly |
| | GO:0035160 | 0.225 ± 0.030 | $7.0 \times 10^{-4}$ | 6.55 | 0.29 | 5,011 | Maintenance of epithelial integrity; open tracheal system |
| | GO:0009612 | 0.220 ± 0.028 | $7.0 \times 10^{-4}$ | 8.50 | 0.49 | 3,140 | Response to mechanical stimulus |
| | GO:0032039 | 0.191 ± 0.022 | $1.5 \times 10^{-3}$ | 5.61 | 0.25 | 561 | Integrator complex |
| | GO:0005100 | 0.197 ± 0.030 | $4.6 \times 10^{-3}$ | 6.24 | 0.22 | 3,886 | Rho GTPase activator activity |
| | GO:0007494 | 0.176 ± 0.026 | $1.1 \times 10^{-2}$ | 7.09 | 0.46 | 9,614 | Midgut development |
| | GO:0007266 | 0.183 ± 0.031 | $1.2 \times 10^{-2}$ | 5.46 | 0.18 | 3,139 | Rho protein signal transduction |
| | GO:0016887 | 0.175 ± 0.027 | $1.2 \times 10^{-2}$ | 3.73 | 0.33 | 3,173 | ATPase activity |
| | GO:0001673 | 0.181 ± 0.032 | $1.4 \times 10^{-2}$ | 6.22 | 0.25 | 334 | Male germ cell nucleus |
| | GO:0003015 | 0.180 ± 0.031 | $1.4 \times 10^{-2}$ | 4.83 | 0.43 | 3,394 | Heart process |

[a] Gene Ontology ID.
[b] Predictive ability.
[c] Standard error of the mean.
[d] False discovery rate adjusted *P*-values.
[e] Likelihood-ratio statistics.
[f] Proportion of genomic variance explained by feature.
[g] Number of SNPs within feature.

allowed for differential shrinkage estimation of SNP effects or a preselection of markers with the highest absolute additive genetic effect or genetic variance explained (Ober *et al.* 2012).

Compared to the GBLUP model, several GO terms led to significantly greater predictive abilities of the GFBLUP model (*P*-value adjusted for multiple tests <0.05), and these terms give novel insights regarding the biology of the traits (Table 1, Table 2, and Supplemental Material, Table S1). Further, the GO term with the highest predictive ability within each trait and sex was significantly higher than the predictive ability obtained from the GBLUP model. For chill coma recovery time, 32 GO terms in females and 16 in males had predictive values that were significantly higher; 7 GO terms were the same for males and females. The top GO term in females was "Rho protein signal transduction," with a predictive ability of 0.37±0.022; the top GO term in males was "cell projection assembly," with a predictive ability of 0.32±0.023 (Table 1, Table 2, and Table S1). For startle response, 11 GO terms in females and 4 in males had predictive values that were significantly higher; 2 GO terms were the same for males and females. The top GO term in females was "retrograde vesicle-mediated transport, Golgi to ER," with a predictive ability of 0.52±0.026; the top GO term in males was "spermatogenesis," with a predictive ability of 0.47±0.025 (Table 1, Table 2, and Table S1). For starvation resistance, the top GO term in females was "alpha-glucosidase activity," with a predictive ability of 0.37±0.027, but this was not significant after adjusting for multiple tests (Table S1). The top GO term for male starvation resistance, "foregut morphogenesis," was the only sig-

nificant GO term for this trait following the multiple-testing correction and had a predictive ability of 0.43±0.022 (Table S1).

***Genomic parameters:*** The range of genomic variance explained by significant feature ($\hat{h}_f^2$) was between 31% and 92% for chill coma recovery and between 23% and 70% for startle response in females. Males showed a similar range of $\hat{h}_f^2$ for the latter traits and 45% for starvation resistance (Table 1, Table 2, and Table S1). Notably this range of total genomic variance in females and males was explained by only 0.09–0.7% of the total SNPs for chill coma recovery, 0.01–0.6% for startle response, and 0.2% for starvation resistance. These results suggest that the genomic variance is not evenly distributed throughout the genome (as would be the case if the genetic architecture of the three traits approximated an infinitesimal model), but instead appears to be associated with a subset of the genome annotated with different GO terms for each trait. The genomic parameters estimated using the GFBLUP model allow us to put different weights on the individual genetic marker relationships used in the prediction equations, in contrast to the GBLUP model, which weights each element of the genetic marker relationship equally. However, the genomic heritabilities estimated using GBLUP or GFBLUP are very similar. The estimated genomic heritabilities using GBLUP were moderate and similar for males and females for each trait: 0.41 and 0.45 for chill coma recovery, 0.49 and 0.47 for startle response, and 0.55 and 0.57 for starvation resistance for females and males, respectively. In the GFBLUP model analyses, the overall genomic heritability

**Table 2 Ten most significant predictions for startle response with the GFBLUP model**

| Sex | GO ID[a] | PA[b] ± SEM[c] | Padj[d] | LR[e] | Hf[f] | Nsets[g] | Gene Ontology term |
|---|---|---|---|---|---|---|---|
| Female | GO:0006890 | 0.520 ± 0.026 | $9.1 \times 10^{-6}$ | 7.42 | 0.39 | 232 | Retrograde vesicle-mediated transport; Golgi to ER |
| | GO:0007436 | 0.509 ± 0.026 | $1.9 \times 10^{-5}$ | 8.81 | 0.70 | 2,740 | Larval salivary gland morphogenesis |
| | GO:0042826 | 0.431 ± 0.023 | $1.2 \times 10^{-2}$ | 5.47 | 0.23 | 404 | Histone deacetylase binding |
| | GO:0051537 | 0.433 ± 0.026 | $1.5 \times 10^{-2}$ | 5.56 | 0.28 | 683 | 2 iron; 2 sulfur cluster binding |
| | GO:0072499 | 0.432 ± 0.027 | $1.5 \times 10^{-2}$ | 6.44 | 0.48 | 4,329 | Photoreceptor cell axon guidance |
| | GO:0008237 | 0.418 ± 0.022 | $1.5 \times 10^{-2}$ | 4.79 | 0.48 | 3,873 | Metallopeptidase activity |
| | GO:0019898 | 0.426 ± 0.027 | $1.8 \times 10^{-2}$ | 4.70 | 0.34 | 1,635 | Extrinsic component of membrane |
| | GO:0051015 | 0.417 ± 0.025 | $2.5 \times 10^{-2}$ | 4.43 | 0.49 | 1,682 | Actin filament binding |
| | GO:0043195 | 0.412 ± 0.024 | $2.5 \times 10^{-2}$ | 4.08 | 0.45 | 5,476 | Terminal bouton |
| Male | GO:0007283 | 0.473 ± 0.025 | $7.4 \times 10^{-4}$ | 7.21 | 0.73 | 6,514 | Spermatogenesis |
| | GO:0007436 | 0.473 ± 0.025 | $7.4 \times 10^{-4}$ | 6.73 | 0.50 | 2,740 | Larval salivary gland morphogenesis |
| | GO:0042331 | 0.458 ± 0.029 | $7.2 \times 10^{-3}$ | 6.23 | 0.42 | 1,574 | Phototaxis |
| | GO:0051537 | 0.432 ± 0.023 | $1.3 \times 10^{-2}$ | 5.41 | 0.30 | 683 | 2 iron; 2 sulfur cluster binding |
| | GO:0072499 | 0.411 ± 0.029 | $2.3 \times 10^{-2}$ | 5.35 | 0.38 | 4,329 | Photoreceptor cell axon guidance |
| | GO:0046854 | 0.401 ± 0.028 | $3.7 \times 10^{-2}$ | 4.92 | 0.39 | 842 | Phosphatidylinositol phosphorylation |
| | GO:0035075 | 0.392 ± 0.025 | $3.7 \times 10^{-2}$ | 3.52 | 0.26 | 3,637 | Response to ecdysone |
| | GO:0008152 | 0.387 ± 0.025 | $4.2 \times 10^{-2}$ | 2.86 | 0.61 | 10,912 | Metabolic process |
| | GO:0051015 | 0.383 ± 0.022 | $4.2 \times 10^{-2}$ | 3.19 | 0.51 | 1,682 | Actin filament binding |
| | GO:0045167 | 0.398 ± 0.031 | $4.6 \times 10^{-2}$ | 4.45 | 0.35 | 2,409 | Asymmetric protein localization involved in cell fate determination |

[a] Gene Ontology ID.
[b] Predictive ability.
[c] Standard error of the mean.
[d] False discovery rate adjusted *P*-values.
[e] Likelihood-ratio statistics.
[f] Proportion of genomic variance explained by feature.
[g] Number of SNPs within feature.

$(\hat{h}^2_{\text{GFBLUP}})$, within trait and sex, was similar across all the GO terms used for partitioning the genomic variance.

### GBLUP and GFBLUP analyses of simulated data based on DGRP genotypes

The GBLUP and GFBLUP models were compared using simulated data in terms of model fit, model predictive ability, and accuracy of the estimated genomic parameters. The power to detect the genomic feature affecting trait phenotypes was determined based on a likelihood-ratio test (*i.e.*, testing whether the GFBLUP model provides a better fit than the GBLUP model). The simulated data sets were based on the observed DGRP genotypes. We varied several feature- and trait-specific factors that are likely to influence the accuracy of parameter estimates and predictive ability.

*Estimation of genomic parameters:* Estimates of genomic heritability ($\hat{h}^2_{\text{G}}$ and $\hat{h}^2_{\text{GF}}$) were unbiased in all simulation scenarios for both the GBLUP and GFBLUP models (Figure 1). Estimates for GBLUP were centered on the true values of genomic heritability with a similar level of accuracy to that of GFBLUP (results not shown). Increasing the proportion of noncausal SNPs in the genetic marker set defined by the genomic feature (dilution) led to decreased accuracy (larger variance) and bias of the estimated genomic parameters $\hat{h}^2_f$ and $\hat{h}^2_r$ (Figure 1). This pattern was observed in all simulation scenarios. To illustrate, for the case where the genomic heritability was 50% and the genomic feature explained 30% of the genomic variance (*i.e.*, $h^2 = 0.5$ and $h^2_f = 0.3$) the estimated value $\hat{h}^2_f$ was centered on the true value at a lower level of dilution. However, the standard deviation increased

from 0.092 to 0.15 and we observed a downward bias following dilution (with 2000 noncausal SNPs). A similar pattern was observed for $\hat{h}^2_r$ (*i.e.*, the proportion of genomic variance captured by genetic markers not included in the genomic feature) except that the estimates tended to be biased upward.

*Predictive ability:* GFBLUP had higher predictive ability (up to 0.62) than GBLUP (0.32 in all scenarios), provided the proportion of the genomic variance explained by the genomic feature was high, with few noncausal SNPs included (Figure 2). The predictive ability of the GFBLUP model is positively correlated with the proportion of genomic variance explained by the genomic feature and negatively correlated with increased dilution (Figure 2). Our results for a genomic heritability of 50% illustrate the general patterns observed across the different simulation scenarios. When the feature explains 10% of the genomic variance, the predictive ability is 0.34 if there is no dilution. Increasing $h^2_f$ from 0.2 to 0.3 to 0.5 increases the predictive ability from 0.42 to 0.46 until a maximum of 0.62 (twice the value obtained using GBLUP). The first dilution, adding 100 noncausal SNPs, has the most prominent effect on predictive ability for all the tested levels of $h^2_f$. Thereafter predictive ability of the GFBLUP model rapidly declines toward the predictive ability obtained using the GBLUP model. The predictive ability was slightly higher if causal SNPs in the genomic feature were clustered in smaller genome regions rather than distributed randomly across the genome. For both models, predictive abilities were generally independent of the level of genomic heritability ($h^2 = 0.1$, 0.3, or 0.5) and the number of replicates within line ($N_{\text{rep}} = 5$, 10, or 50).
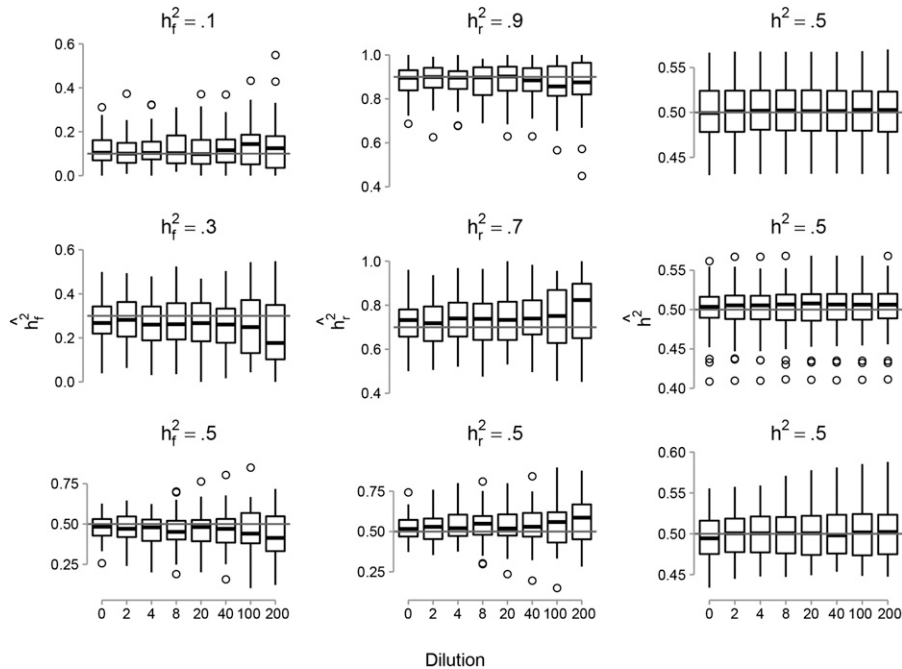
**Figure 1** Boxplots showing estimates from the GFBLUP model analyses of genomic parameters as a function of the proportion of noncausal SNPs in the genomic feature (dilution). Estimates are proportion of genomic variance captured by the genetic markers in the genomic feature ($h_f^2 = 0.1, 0.3$, or $0.5$), proportion of genomic variance captured by genetic markers not included in the genomic feature ($h_r^2 = 0.9$, $0.7$, or $0.5$), and genomic heritability ($h^2 = 0.5$). Results are for the scenarios where the causal SNPs in the genomic feature are clustered in certain genome regions, adding noncausal SNPs located directly up- and downstream of the causal SNPs, and the number of replicates within lines = 50. The light line with light shading corresponds to the true genomic parameter.

**Detection power:** The power to detect genomic features affecting the trait phenotypes was influenced both by trait-specific and by genomic feature-specific factors. The proportion of the genomic variance explained by the genomic feature ($h_f^2$) greatly affected detection power (Figure 3) and robustness to dilution. At low values ($h_f^2 = 0.1$), the power to detect the genomic features was low if the overall trait heritability was low ($h^2 = 0.1$), even without dilution. At the highest values ($h_f^2 = 0.5$) the impact of dilution was less severe. This increased robustness to dilution resulted in power >72% in all cluster model scenarios with $N_{rep} = 50$ replicates within line and a genomic heritability of 50%. The level of genomic heritability ($h^2$) was positively correlated with power (Figure 3). However, at high $h_f^2$ and in the absence of dilution, all genomic features were detected regardless of overall genomic heritability. Furthermore, if $h_f^2$ was high, high heritability traits were less affected by dilution than low heritability traits (Figure 3). Dilution decreased power in all simulation scenarios (Figure 3). Furthermore, detection power increases with increasing numbers of replicates within line ($N_{rep} = 5, 10$, or $50$).

**Genomic relationships:** Estimation of genomic parameters, detection power, and predictive ability using the GFBLUP model analyses were based on two genomic relationship matrices: $\mathbf{G}_f$ for the genetic marker set defined by the genomic feature and $\mathbf{G}_r$ for the remaining set of markers. These "fitted" genomic relationships differ from the true causal relationships that in practice are unknown. Dilution of the true causal relationship by increasing the proportion of noncausal SNPs in the genomic feature decreases the correlation between $\mathbf{G}_f$ (the genomic relationships calculated using all genetic markers defined by the genomic feature including causal and noncausal SNPs) and $\mathbf{G}_1$ (the true causal genomic relationships calculated using only the true causal SNPs in

the genomic feature). If the true causal SNPs are clustered in a smaller number of genome regions, then the effect of dilution is more "extreme" (Figure 4). The correlation between $\mathbf{G}_f$ and $\mathbf{G}_1$ for the clustered causal model quickly decreases as dilution with local (or random) noncausal SNPs increases, leveling off at a value of 0.71 (0.56). In contrast, dilution leads to an increasing correlation between $\mathbf{G}_f$ and $\mathbf{G}_r$ (the genomic relationships for the set of markers not included in the feature) (Figure 4). For $\mathbf{G}_f$ comprising only true causal SNPs, the correlation between $\mathbf{G}_f$ and $\mathbf{G}_r$ is 0.53 for the cluster causal model. Following dilution of $\mathbf{G}_f$ with local (or random) noncausal SNPs, the correlation rapidly increases toward a value of 0.75 (0.99).

## Discussion

We applied and evaluated a GFBLUP model, using prior information on genomic features. Genomic features are regions of the genome that are linked to external information. This modeling approach is predicated on the assumption that these regions are enriched for causal variants affecting the trait. Several genomic feature classes can be formed based on different sources of prior information, for example, genes, chromosomes, biological pathways, gene ontologies, sequence annotation, prior QTL regions, or other types of external evidence. We demonstrated that the GFBLUP model using prior information on Gene Ontology categories can increase the predictive ability of the genomic value for three quantitative traits (starvation resistance, startle response, and chill coma recovery) in *D. melanogaster*. These results were supported by using simulated data generated from DGRP genotypes, further illustrating the impact of trait-specific and genomic feature-specific factors on predictive ability.
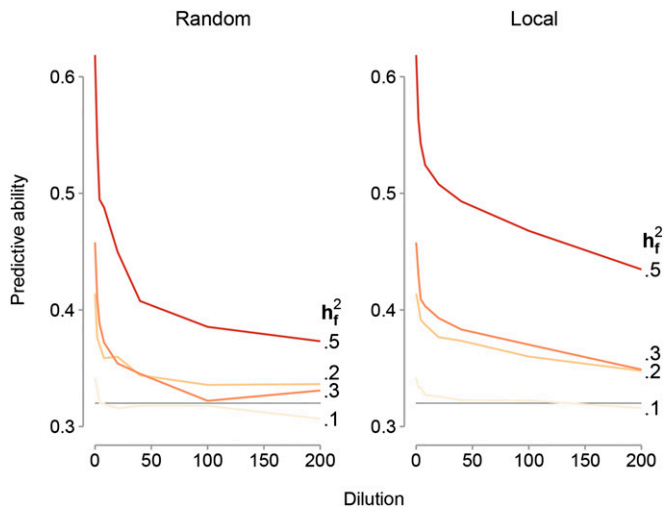
**Figure 2** Plots showing predictive ability of the GFBLUP model as a function of the proportion of noncausal SNPs in the genomic feature (dilution). Results are for four different levels of the proportion of genomic variance captured by the genetic markers in the genomic feature ($h_f^2$ = 0.1, 0.2, 0.3, or 0.5) for the scenarios where the causal SNPs in the genomic feature are clustered in smaller genome regions, adding noncausal SNPs picked at random in the genome (left) or located directly up- and downstream of the causal SNPs (right), and the number of replicates within lines = 50. The light gray line corresponds to the predictive ability of the GBLUP model.

### The GFBLUP model improves predictive ability in the DGRP

The increase in predictive ability using the GFBLUP model compared to the commonly used GBLUP model was substantial for all traits and both sexes. For females (males) the increase was 0.12 (0.16) for starvation resistance, 0.27 (0.19) for startle response, and 0.33 (0.32) for chill coma recovery time. These differences between the two models correspond to a 48–89% relative increase in predictive ability of genomic values for startle response and starvation resistance and even higher for chill coma recovery time. Our predictive ability using GBLUP was similar to estimates from previous studies (Ober *et al.* 2012, 2015). Predictive ability using GBLUP decreased when smaller numbers of markers were used; for example, the predictive ability for starvation resistance and startle response dropped to 0.1 when 4863 randomly selected markers were used (Ober *et al.* 2012). In contrast, the highest-ranked GO term for chill coma recovery in females was associated with a predictive ability of 0.37, using only 3129 markers. We hypothesize that the difference in predictive ability between the two models is that the assumption of the GBLUP model, that the genomic variance is evenly distributed throughout the genome (*i.e.*, the underlying genetic architecture of the trait approaches an infinitesimal model), is not met. Rather, the genomic variance for the three traits assessed seems to be associated with subsets of the genome annotated with specific biological processes that differ among the traits. The markers located in these genome regions have greater weights than the

remaining markers in the GFBLUP model analyses, leading to increased predictive ability. Note that the genetic marker relationship matrix used for the GBLUP model is the same for all traits, because of the underlying infinitesimal model assumption of genetic architecture. However, the GFBLUP model permits a different genetic architecture for each of these (genetically uncorrelated) traits, which is more biologically plausible.

Our empirical results were further supported by simulation studies investigating the influence of genomic feature- and trait-specific factors on the predictive ability of GFBLUP. The simulations revealed that it is possible, even under an additive genetic model, to further increase the predictive ability of genomic value for quantitative traits in the DGRP, using the GFBLUP model. This requires that prior information on genomic features highly enriched for causal variants is used. Such information is rapidly becoming available and being refined, given advances in functional and genetic studies of complex traits that continue to increase our understanding of how the putative causal variants are distributed over the genome. Furthermore, improvement in predictive ability for genomic value of complex trait phenotypes may be achieved by accounting for other types of genetic variation due to different types of variants (rare and structural) and nonadditive gene action (dominance and epistasis).

***Genomic features predictive of organismal quantitative trait phenotypes:*** Several of the high-ranking GO terms in our study have previously been associated with correlated transcriptional modules associated with chill coma recovery time and starvation resistance (Ayroles *et al.* 2009). These modules are plausibly enriched for causal variants affecting the phenotype (Cookson *et al.* 2009) that could affect expression of the genes in the module, such as mutations in promoter motifs, transcription enhancers, or silencers in introns or regulatory microRNAs. In addition, if the gene products of the differentially expressed genes are associated with the phenotype, variants that change the structure of the expressed RNA and the transcribed protein could also affect the phenotype.

The GO terms Rho protein signal transduction (GO:0007266) and Rho GTPase activator activity (GO:0005100) had the highest prediction accuracies for male and female chill coma recovery time. There are several ways in which Rho genes may functionally affect the time to recover from a chill-induced coma. Rho proteins function as molecular switches, conducting cues from the external environment to intracellular signal transduction pathways (Tcherkezian and Lamarche-Vane 2007). In addition, members of the Rho family of GTPases are among the important modulators of actin dynamics and neuronal as well as behavioral plasticity. By playing a role in the regulation of actin, these proteins are important in mediating the circadian rhythm and other behaviors in *Drosophila* (Rao 2013); and circadian rhythm in turn has been associated with chill coma recovery time (Pegoraro *et al.* 2014). Rho activity also plays a role in the maintenance of ion homeostasis. Chill coma is the result of an inability to maintain ion homeostasis (MacMillan and Sinclair 2011), particularly extracellular [K⁺], and an additional effect of low temperature
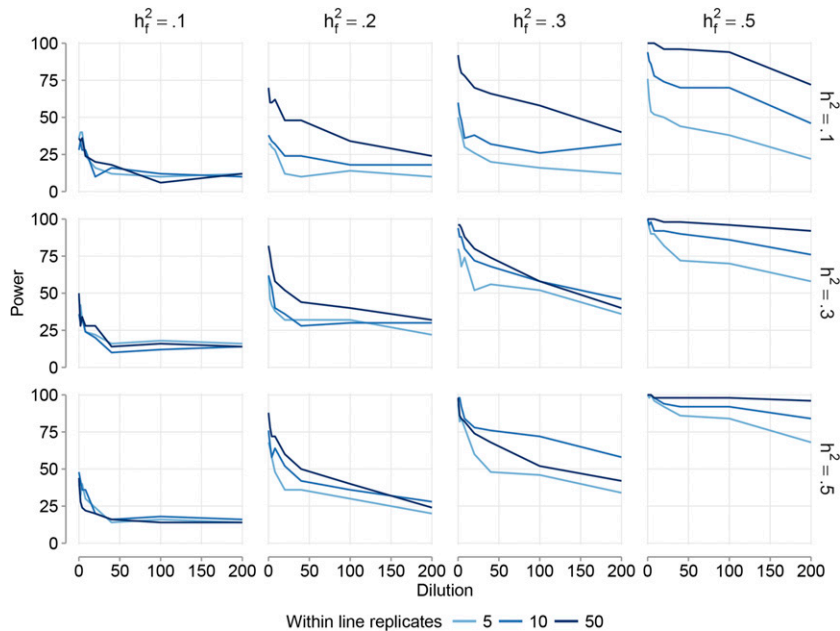
**Figure 3** Plots showing detection power of the GFBLUP model as a function of the proportion of non-causal SNPs in the genomic feature (*e.g.*, dilution by adding noncausal SNPs located directly up- and downstream of the causal SNPs). Results are for the scenarios where the causal SNPs in the genomic feature are clustered in certain genome regions for three different levels of genomic heritability ($h^2$ = 0.1, 0.3, or 0.5), four different levels of the proportion of genomic variance captured by the genetic markers in the genomic feature ($h^2_f$ = 0.1, 0.2, 0.3, or 0.5), and three different levels of the number of replicates within lines ($N_{rep}$ = 5, 10, or 50).

(Findsen *et al.* 2014). A RHO activator has been linked to the regulation of [K$^+$] channel cell surface expression and thus activity in human cell cultures (Stirling *et al.* 2009). Finally, analyses of whole-genome sequences of DNA pools from *Drosophila* populations collected along the North American East Coast reveal patterns of selection in genes involved in major functional pathways such as circadian rhythm and the epidermal growth factor pathway (Fabian *et al.* 2012), genes in both of which harbor variants that are predictive of chill coma recovery in our study. These examples highlight a possible direct functional link between Rho protein activity and chill coma recovery that can be tested in the future. Similar hypotheses can be developed for the other GO terms that are predictive of the traits investigated.

*Genomic feature classes helping biological interpretation:* Applying the GFBLUP model using prior information on genomic features may help open the black box that is the genetic architecture of complex traits. This approach provides novel insight into the biological mechanisms causing trait variation and simultaneously improves predictive ability relative to a commonly used prediction model. Several genomic feature classes can be formed based on different sources of prior information (*e.g.*, genes, chromosomes, biological pathways, gene ontologies, sequence annotation, prior QTL regions, or other types of external evidence). The gain in knowledge depends highly on the quality and complexity of the genomic feature classification scheme upon which the genetic marker sets are based. Genomic features based on physical genome regions, such as chromosomes or single genes, might not increase the information level; however, as additional layers of complexity such as pathways are added, the biological interpretation might become more informative. On the other hand, biological interpretation might be hampered by the definition (or misspecification) of the genomic feature and a potential large overlap in the genetic marker sets between
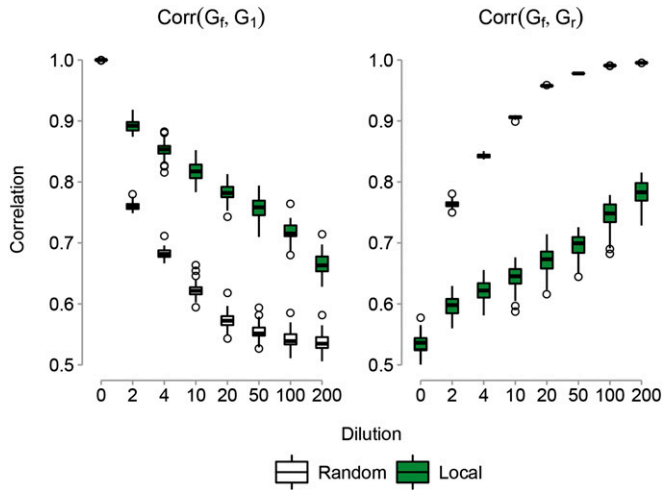
the different genomic feature classes. In the latter case, biological interpretation may be improved by using methods that take the overlap into account (Skarman *et al.* 2012).

### Factors influencing GFBLUP model performance

The simulation study clarified the conditions needed to make genomic partitioning "work" (*i.e.*, harvest the benefits from prior information in terms of model fit and predictive ability). The simulations showed that a GFBLUP model can increase predictive ability compared to a standard GBLUP model, highlighted the importance of maximizing the proportion of causal variants in $\mathbf{G}_f$ (and $\mathbf{G}_r$), and indicated some limitations of the GFBLUP modeling approach.

Predictive ability of the GFBLUP model is influenced both by the proportion of genomic variance explained by the genomic feature and by the addition of noncausal SNPs in the feature (dilution). Predictive ability (and detection power) was positively correlated with the proportion of genomic variance explained by the genomic feature and negatively correlated with dilution. Estimates of the proportion of genomic variance explained by the genomic feature ($\hat{h}^2_f$) were generally unbiased. However, increased dilution led to decreased accuracy (a larger variance) of the estimated genomic parameters $\hat{h}^2_f$ and $\hat{h}^2_r$. It is important to note that models were compared based on their ability to predict the genomic values (and not phenotypes). Therefore, the predictive abilities reported in this study were generally independent of the level of genomic heritability and the number of replicates within lines. If we were to predict the trait phenotypes, we would expect an influence of these trait-specific factors on predictive ability, as was the case for detection power.

The GFBLUP model is mos beneficial when the genomic feature is highly enriched for true causal variants. To better understand this phenomenon, it is useful to examine the details of the GFBLUP model. The estimation of genomic

**Figure 4** Boxplots showing correlations between different genomic relationship matrices used in the GFBLUP model analyses as a function of the proportion of noncausal SNPs in the genomic feature (dilution). Correlations are between $\mathbf{G}_f$ (the genomic relationships calculated using all genetic markers defined by the genomic feature including causal and noncausal SNPs) and $\mathbf{G}_1$ (the true causal genomic relationships calculated using only the true causal SNPs in the genomic feature) or between $\mathbf{G}_f$ and $\mathbf{G}_r$ (the genomic relationships for the set of markers not included in the feature). Correlations are for the scenarios where the causal SNPs in the genomic feature are clustered in smaller genome regions and adding noncausal SNPs picked at random in the genome (Random) or located directly up- and downstream of the causal SNPs (Local).

parameters in the GFBLUP model was based on two genomic relationship matrices, $\mathbf{G}_f$ for the genetic marker set defined by the genomic feature and $\mathbf{G}_r$ for the remaining set of markers. The decrease in accuracy of genomic parameter estimates following dilution is caused by the increased correlation between these two genomic relationship matrices. The high correlation between the elements in these matrices makes it difficult for the REML method to estimate and thereby reliably partition the corresponding genomic variances ($\hat{\sigma}_f^2$ and $\hat{\sigma}_r^2$). It is also clear from the BLUP equations used in the GFBLUP model (described in *Materials and Methods*) that inaccurate estimates of genomic variances would affect predictive ability. The estimates of $\hat{\sigma}_f^2$ and $\hat{\sigma}_r^2$ determine the relative contribution of the two genomic relationship matrices in the prediction of the total genomic value. If these estimates deviate from the true value of the parameters, it will lead to less accurate predictions, because there is too much weight on the "wrong" genomic relationships in the prediction equations. Obviously this will also occur if the two genomic relationships differ from the true causal relationships.

In this study we used a GFBLUP model with two genomic effects ($\mathbf{f}$ and $\mathbf{r}$ in model $M_{\text{GFBLUP}}$), but in principle it is possible to include multiple genomic feature effects (Speed and Balding 2014; Sørensen *et al.* 2015). However, as our simulations with only two genomic effects indicate, the complex correlation patterns that are likely to exist between different parts of the genome may lead to inaccurate estimates of the genomic variances and consequently decreased predictive

ability. It would also explain why GFBLUP sometimes led to a decrease in predictive ability. GBLUP depends only on the total genomic variance parameter that is reliably estimated and therefore more robust. The linear mixed model used by Gusev *et al.* (2014) is very similar to our approach. In their analyses they fitted multiple random genomic effects defined by sequence ontologies. They found little improvement in polygenic risk prediction using their model and argue this is because of pervasive LD across categories (*e.g.*, feature classes). This is exactly what we observe in our simulations—when the correlation between the two (or more) genomic relationship matrices is high, we cannot reliably estimate the variance components and therefore there is no improvement in predictive ability.

***Predictive ability and genetic relatedness in the study population:*** The extent to which utilizing prior biological information will increase the predictive ability of the statistical model depends on the degree of genetic relatedness among the individuals in the mapping population (de los Campos *et al.* 2013). In a population of highly related individuals the general genomic relationship will be a good approximation for the genomic relationship at the true causal variants. Our results indicate that using an informed choice of subsets of markers with the GBLUP model increases the predictive ability in a population of largely unrelated individuals. In such situations it is important to model the genomic relationship due to the causal variants differently from the overall genomic relationship. Thus the GFBLUP model may also be useful in improving predictive ability in other situations where individuals are largely unrelated, such as human populations and across animal breeds and plant strains. Further work is required to better understand the impact of varying degrees of genetic relatedness on the performance of the GFBLUP model.

### GFBLUP model and alternatives

GFBLUP is based on a linear mixed-modeling framework that allows us to adjust for other known genetic and nongenetic factors. In this study we implemented the GFBLUP model, using a REML method (Gusev *et al.* 2014; Sørensen *et al.* 2014; Speed and Balding 2014). Bayesian mixture models ignoring prior genomic feature information such as BayesB or BayesR (Meuwissen *et al.* 2009; Erbe *et al.* 2012) are relevant alternative methods. Both of these methods allow the contribution from each marker to come from different distributions. However, these models are not necessarily computationally less demanding; they require the same considerations with regard to formulating the models and do not necessarily perform better (Ober *et al.* 2012). This study adds evidence that an *externally informed* subset of markers is necessary for a successful partitioning of the genomic variance, as the data themselves may not necessarily indicate which markers should have greater weights. Whereas Bayesian mixture models attempt to assign and estimate marker effects from different distributions, we use prior knowledge to assign a marker set (defined by the genomic feature) to a distribution [*i.e.*, $\mathbf{f} \sim N(0, \mathbf{G}_f \sigma_f^2)$ and

$\mathbf{r} \sim N(0, \mathbf{G}\sigma_r^2)$]. We then estimate the parameters for these distributions ($\sigma_f^2$ and $\sigma_r^2$) conditional on the observed data and evaluate whether this is a sensible assignment, using standard model comparison techniques such as cross-validations or likelihood-ratio tests. We therefore conclude that the GFBLUP approach described here provides a general framework for estimating and evaluating the association of genomic features with complex trait phenotypes.

The GFBLUP model can be implemented using a Bayesian mixed model (*e.g.*, Sørensen *et al.* 2015) or Bayesian mixture models such as BayesRC (MacLeod *et al.* 2014). BayesRC is the same as BayesR except that, *a priori*, each SNP is identified as belonging to a specific genomic feature and it allows for differential shrinkage within each genomic feature. The advantage of the BayesRC approach is in cases where enough information is available in the data to reliably allocate the SNPs in the different variance classes defined in the Bayesian mixture model. If this is not the case, we do not expect a major difference in performance in terms of predictive ability.

A key element of the GFBLUP and Bayesian GF mixture models is the use of prior information to reliably partition the genomic variance. These approaches are computationally intensive. However, there are computationally simple approaches to obtain approximate measures of genomic variance (heritability), using only genome-wide association (GWA) summary statistics (Finucane *et al.* 2015). In this approach, genomic variance is partitioned using single-marker effects obtained from GWA and LD information from a population similar to the one used for obtaining the single-marker effects. In contrast, partitioning of genomic variance using the GFBLUP model requires both the phenotypes and genotypes of the study population. The approximate measures of genomic variance obtained for each genomic feature in the Finucane *et al.* (2015) study can in principle also be used in the GFBLUP prediction equations or as prior scale parameters in Bayesian models, but was not pursued in their study.

### Conclusion

Our GFBLUP modeling approach using prior information on genomic features enriched for causal variants can increase the accuracy of genomic predictions for complex traits in a population of largely unrelated individuals. The simulations revealed that it is possible to further increase the accuracy of genomic prediction for complex traits with a quasi-infinitesimal genetic architecture with many causal polymorphisms each with a small effect, using the GFBLUP model, provided that prior information on genomic features is highly enriched for causal variants. These models provide a formal statistical modeling framework for borrowing and evaluating information across a wide range of experimental studies that may help provide novel insights into genetic and biological mechanisms underlying complex traits.

### Acknowledgments

S.M.E. conceived the study; designed, performed, and evaluated the experiments; analyzed the data; and drafted the manuscript. I.F.S. evaluated the experiments and drafted the manuscript. P. Sarup evaluated the experiments and drafted the manuscript. T.F.C.M. conceived the study, evaluated the experiments, and drafted the manuscript. P. Sørensen conceived the study; designed, performed, and evaluated the experiments; analyzed the data; and drafted the manuscript. All authors read and approved the final manuscript.

### Literature Cited

Ayroles, J. F., M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman et al., 2009 Systems genetics of complex traits in *Drosophila melanogaster*. Nat. Genet. 41: 299–307.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57: 289–300.

Caballero, A., A. Tenesa, and P. D. Keightley, 2015 The nature of genetic variation for complex traits revealed by GWAS and regional heritability mapping analyses. Genetics 201: 1601–1613.

Carlson M., 2013 *org.DM.eg.db: Genome Wide Annotation for Fly*. R package version 3.2.3. Available at: http://bioconductor.org/packages/org.Dm.eg.db/.

Cookson, W., L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, 2009 Mapping complex disease traits with global gene expression. Nat. Rev. Genet. 10: 184–194.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño et al., 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193: 347–365.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen, 2013 Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9: e1003608.

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard, 2009 Reliability of genomic predictions across multiple populations. Genetics 183: 1545–1553.

Edwards, S. M., B. Thomsen, P. Madsen, and P. Sørensen, 2015 Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. Genet. Sel. Evol. 47: 60.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman et al., 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95: 4114–4129.

Fabian, D. K., M. Kapun, V. Nolte, R. Kofler, P. S. Schmidt et al., 2012 Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Mol. Ecol. 21: 4748–4769.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Benjamin-Cummings, Menlo Park, CA.

Findsen, A., T. H. Pedersen, A. G. Petersen, O. B. Nielsen, and J. Overgaard, 2014 Why do insects enter and recover from chill coma? Low temperature and high extracellular potassium compromise muscle function in *Locusta migratoria*. J. Exp. Biol. 217: 1297–1306.

Finucane, H. K., B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef et al., 2015 Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47: 1228–1235.

Gene Ontology ConsortiumAshburner, M., C. A. Ball, J. A. Blake, D. Botstein et al., 2000 Gene Ontology: tool for the unification of biology. Nat. Genet. 25: 25–29.

Gusev, A., S. H. Lee, G. Trynka, H. Finucane, B. J. Vilhjálmsson et al., 2014 Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. 95: 535–552.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397.

Harbison, S. T., A. H. Yamamoto, J. J. Fanara, K. K. Norga, and T. F. C. Mackay, 2004 Quantitative trait loci affecting starvation resistance in Drosophila melanogaster. Genetics 166: 1807–1823.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009 Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard, 2010 Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6: e1001139.

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ràmia et al., 2014 Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. 24: 1193–1208.

Johnson, D. L., and R. Thompson, 1995 Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. J. Dairy Sci. 78: 449–456.

Lage, K., S. C. Greenway, J. A. Rosenfeld, H. Wakimoto, J. M. Gorham et al., 2012 Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development. Proc. Natl. Acad. Sci. USA 109: 14035–14040.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon et al., 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.

Mackay, T. F. C., 2001 The genetic architecture of quantitative traits. Annu. Rev. Genet. 35: 303–339.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles et al., 2012 The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178.

MacLeod, I. M., B. J. Hayes, C. J. Vander Jagt, K. E. Kemper, M. Haile-Mariam et al., 2014 A Bayesian analysis to exploit imputed sequence variants for QTL discovery. 10th World Congress of Genetics Applied to Livestock Production, Vancouver, British Columbia, Canada.

MacMillan, H. A., and B. J. Sinclair, 2011 Mechanisms underlying insect chill-coma. J. Insect Physiol. 57: 12–20.

Madsen, P., J. Jensen, and R. Thompson, 1994 Estimation of (co)-variance components by REML in multivariate mixed linear models using average of observed and expected information. Fifth World Congress of Genetics Applied to Livestock Production, Guelph, Ontario, Canada, pp. 455–462.

Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte et al., 2011 Beyond missing heritability: prediction of complex traits. PLoS Genet. 7: e1002051.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff et al., 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen et al., 2012 Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190–1195.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Meuwissen, T. H., T. R. Solberg, R. Shepherd, and J. A. Woolliams, 2009 A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet. Sel. Evol. 41: 2.

Morgan, T. J., and T. F. C. Mackay, 2006 Quantitative trait loci for thermotolerance phenotypes in Drosophila melanogaster. Heredity 96: 232–242.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu et al., 2012 Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. 8: e1002685.

Ober, U., W. Huang, M. Magwire, M. Schlather, H. Simianer et al., 2015 Accounting for genetic architecture improves sequence based genomic prediction for a Drosophila fitness trait. PLoS One 10: e0126880.

O'Roak, B. J., L. Vives, S. Girirajan, E. Karakoc, N. Krumm et al., 2012 Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485: 246–250.

Pegoraro, M., J. S. Gesto, C. P. Kyriacou, and E. Tauber, 2014 Role for circadian clock genes in seasonal timing: testing the Bünning hypothesis. PLoS Genet. 10: e1004603.

Peñagaricano, F., K. A. Weigel, G. J. M. Rosa, and H. Khatib, 2013 Inferring quantitative trait pathways associated with bull fertility from a genome-wide association study. Front. Genet. 3: 307.

R Core Team, 2015 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Rao, N. V., 2013 Role of the RHO1 GTPase signaling pathway in regulating the circadian clock in Drosophila melanogaster. Ph.D. Thesis, University of Virginia.

Self, S. G., and K.-Y. Liang, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 82: 605–610.

Skarman, A., M. Shariati, L. Jans, L. Jiang, and P. Sørensen, 2012 A Bayesian variable selection procedure to rank overlapping gene sets. BMC Bioinformatics 13: 73.

Sørensen, P., S. M. Edwards, and P. Jensen, 2014 Genomic feature models. 10th World Congress of Genetics Applied to Livestock Production, Vancouver, British Columbia, Canada.

Sørensen, P., G. de los Campos, F. Morgante, T. F. C. Mackay, and D. Sorensen, 2015 Genetic control of environmental variation of two quantitative traits of Drosophila melanogaster revealed by whole-genome sequencing. Genetics 201: 487–497.

Speed, D., and D. J. Balding, 2014 MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 24: 1550–1557.

Stirling, L., M. R. Williams, and A. D. Morielli, 2009 Dual roles for RHOA/RHO-kinase in the regulated trafficking of a voltage-sensitive potassium channel. Mol. Biol. Cell 20: 2991–3002.

Tcherkezian, J., and N. Lamarche-Vane, 2007 Current knowledge of the large RhoGAP family of proteins. Biol. Cell 99: 67–86.

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton et al., 2009 FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 37: D555–D559.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Vinkhuyzen, A. A., N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher, 2013 Estimation and partitioning of heritability in human populations using whole genome analysis methods. Annu. Rev. Genet. 47: 75–95.

Visscher, P. M., 2008 Sizing up human height variation. Nat. Genet. 40: 489–490.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang, 2012 Five years of GWAS discovery. Am. J. Hum. Genet. 90: 7–24.

Welch, B. L., 1947 The generalization of "Student's" problem when several different population variances are involved. Biometrika 34: 28–35.

*Communicating editor: S. F. Chenoweth*

# GENETICS

# Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in *Drosophila melanogaster*

Stefan M. Edwards, Izel F. Sørensen, Pernille Sarup, Trudy F. C. Mackay, and Peter Sørensen

Table S1: Predictive GO categories. (.xlsx, 33 KB)

Available for download as a .xlsx file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187161/-/DC1/TableS1.xlsx