SCIENTIFIC REPORTS

OPEN

Multiple Trait Covariance Association Test Identifies Gene Ontology Categories Associated with Chill Coma Recovery Time in *Drosophila melanogaster*

Received: 17 January 2017
Accepted: 10 April 2017
Published online: 25 May 2017

Izel Fourie Sørensen[1], Stefan M. Edwards[1,2], Palle Duun Rohde[1,3,4] & Peter Sørensen[1]

The genomic best linear unbiased prediction (GBLUP) model has proven to be useful for prediction of complex traits as well as estimation of population genetic parameters. Improved inference and prediction accuracy of GBLUP may be achieved by identifying genomic regions enriched for causal genetic variants. We aimed at searching for patterns in GBLUP-derived single-marker statistics, by including them in genetic marker set tests, that could reveal associations between a set of genetic markers (genomic feature) and a complex trait. GBLUP-derived set tests proved to be powerful for detecting genomic features, here defined by gene ontology (GO) terms, enriched for causal variants affecting a quantitative trait in a population with low degree of relatedness. Different set test approaches were compared using simulated data illustrating the impact of trait- and genomic feature-specific factors on detection power. We extended the most powerful single trait set test, covariance association test (CVAT), to a multiple trait setting. The multiple trait CVAT (MT-CVAT) identified functionally relevant GO categories associated with the quantitative trait, chill coma recovery time, in the unrelated, sequenced inbred lines of the *Drosophila melanogaster* Genetic Reference Panel.

The genomic best linear unbiased prediction (GBLUP) model has proven to be useful for estimation of population genetic parameters (e.g. heritability) as well as prediction of complex traits[1,2]. GBLUP is a "black box" modelling approach fitting fixed and random effects simultaneous, utilizing the genetic relationship between individuals based on the correlation structure among genetic markers. Typically, GBLUP ignores prior biological information. Although models ignoring the underlying biology can serve as useful tools for prediction of genetic values or phenotypes, models utilizing known biological mechanisms provide a functional tool for testing our understanding of those mechanisms, and potentially improve inference and prediction accuracy.

It appears that markers associated with trait variation are not uniformly distributed throughout the genome, but enriched in genes that are connected in biological pathways[3–7]. Such knowledge could be utilized to build models that quantify the joint effect of a set of markers located in a genomic feature, i.e. genomic regions defined by e.g. genes, biological pathways, sequence annotation or other external evidence[8–10]. Improved inference and prediction accuray of GBLUP may be achieved by identifying genomic regions enriched for causal genetic variants.

The GBLUP approach can be modified in several ways to utilize genomic features as prior information. One approach is to extend the traditional GBLUP model to include additional genomic effects based on genetic markers located within a genomic feature[11–16]. Applying the genomic feature best linear unbiased prediction (GFBLUP) model to the *Drosophila* Genetic Reference Panel (DGRP)[17,18], we have previously demonstrated, that GFBLUP

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830, Tjele, Denmark. [2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian, Scotland, UK. [3]Centre for Integrative Sequencing, iSEQ, Aarhus University, 8000, Aarhus, Denmark. [4]iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, 8000, Aarhus, Denmark. Correspondence and requests for materials should be addressed to I.F.S. (email: izel.sorensen@gmail.com)

models can increase prediction accuracy for quantitative traits[15]. These results were further supported by simulation studies illustrating the impact of trait- and genomic feature-specific factors on prediction accuracy[15]. The GFBLUP model approach is, however, computationally intensive. An alternative approach is to search for patterns in GBLUP-derived single-marker statistics that can reveal associations between a genomic feature and a complex trait. We have previously evaluated a number of GBLUP-derived set tests on a binary outcome (i.e. disease trait) using high-density single nucleotide polymorphisms (SNPs) from genotyping arrays[19]. These GBLUP-derived set tests proved to be computationally fast and powerful compared to existing set test approaches[19].

Here, we evaluated GBLUP-derived set tests on a quantitative trait as opposed to the binary outcome in the study of Rohde *et al.*[19], and applied it to whole genome sequence data contrary to the genotypes derived from SNP arrays as previously shown[19]. Whole genome sequence data greatly exacerbate the true genomic signal to non-causal marker noise problem and may influence the power of set tests. Extending GBLUP-derived set tests could potentially increase detection power and contribute to a better understanding of complex traits' underlying genetic architecture. First, multiple feature sets can be fitted in the model (e.g. a GFBLUP model), such as grouping markers based on their minor allele frequency[19,20] or prior QTL information[16]. By fitting multiple feature sets, genetic effects are estimated based on a mixture of normal distributions enabling further differential shrinkage of single marker effects across feature sets. Second, a multiple trait GBLUP model[21,22] can be fitted. This can potentially increase the accuracy of the total genomic value[21,22] and thereby the single marker effect, which in turn will lead to more accurate test statistics for genetic marker sets, thereby increasing detection power of the set test.

The aim of the study was to evaluate and compare genetic marker set tests derived from GBLUP on a quantitative trait using whole genome sequence data. Different set tests were evaluated and compared using simulated data generated from DGRP genotypes, focussing on factors specific to genomic features (e.g. the number, location and effect sizes of the true causal variants in the feature) that influence the power of set tests to detect genomic features affecting the trait phenotype. Furthermore, we investigated whether the results obtained using the GBLUP-derived set tests can be used to develop more accurate GFBLUP prediction models. Finally, we derived a multiple trait GBLUP set test (MT-CVAT) and used it to identify genomic features associated with a quantitative trait phenotype, chill coma recovery time (CCRT), in the unrelated, sequenced inbred lines of the DGRP.

Methods

In the following a range of different GBLUP-derived set test approaches will be described in detail. The general procedure is to obtain single marker effects based on a standard GBLUP model, from which it is possible to compute and evaluate a test statistic for a set of genetic markers, measuring the degree of association between the genomic feature and the complex trait phenotype. This includes the statistical model and the underlying assumptions, test statistics for the set of genetic markers, and statistical procedures for assessing the statistical significance of the observed test statistic under a specific null hypothesis.

Set test approach. The GBLUP-derived set test approach is based on two steps: First a standard linear mixed model is fitted, and then a test statistic for the marker set is computed.

Linear mixed model. GBLUP is based on a linear mixed model including only one random genomic effect:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is the vector of phenotypic observations, $\mathbf{X}$ and $\mathbf{Z}$ are design matrices for the fixed and random effects, $\mathbf{b}$ is a vector of fixed effects, $\mathbf{g}$ is the vector of genomic values captured by all genetic markers, and $\mathbf{e}$ is the vector of residuals. 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 $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. Thus, we assume that the observed phenotypes $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{V})$ where $\mathbf{V} = \mathbf{ZGZ}'\sigma_g^2 + \mathbf{I}\sigma_e^2$.

The additive genomic relationship matrix $\mathbf{G}$ is constructed[23] 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: $\mathbf{w_i} = \frac{\mathbf{a_i} - 2p_i}{\sqrt{2p_i(1-p_i)}}$, where p_i is the minor allele frequency of the i^{th} genetic marker and $\mathbf{a_i}$ is the i^{th} column vector of the allele count matrix, $\mathbf{A}$ which contains the genotypes coded as 0, 1 or 2 counting the number of the minor allele.

Single marker statistics. Single marker effects $\hat{\mathbf{s}}$ can be computed from the predicted total genomic value $\hat{\mathbf{g}} = \mathbf{G}\hat{\sigma}_g^2\mathbf{V}^{-1}(\mathbf{y} - \hat{\mathbf{Xb}})$ obtained from the GBLUP model as:

$$\hat{\mathbf{s}} = \mathbf{W}'(\mathbf{WW}')^{-1}\hat{\mathbf{g}}, \tag{2}$$

and the (co)variance of the single marker effects can be computed as:

$$\widehat{\mathrm{Var}}(\hat{\mathbf{s}}) = \mathbf{W}'(\mathbf{WW}')^{-1}\widehat{\mathrm{Var}}(\hat{\mathbf{g}})(\mathbf{WW}')^{-1}\mathbf{W}'. \tag{3}$$

In this expression, the (co)variance of the predicted genomic value $\widehat{\mathrm{Var}}(\hat{\mathbf{g}}) = \mathbf{G}\hat{\sigma}_g^2 + \mathbf{C}^{gg}$ can be derived from the inverse of the coefficient matrix of the mixed model equations[24,25] for GBLUP where $\mathbf{C}^{gg}$ is the part of this equation system that corresponds to the total genomic value.

Assessing association of individual markers is based on a single marker test statistic such as the t-statistic and a threshold for this statistic.

$$t_{\hat{s}_j} = \frac{\hat{s}_j}{\sqrt{\mathrm{Var}(\hat{s}_j)}}, \tag{4}$$

where $\mathrm{Var}(\hat{s}_j)$ is the estimate of variance of the j'th element of $\hat{\mathbf{s}}$ obtained from the j'th element of the diagonal of the (co)variance matrix of the single marker effects. Under the null hypothesis that $\hat{s}_j = 0$, it is assumed that $t_{\hat{s}_j}$ follows a t distribution with df_e residual degrees of freedom. The residual degrees of freedom df_e is computed as $\mathrm{tr}(\mathbf{I} - \mathbf{H})$, which is equivalent to $n - \mathrm{tr}(\mathbf{H})$ where n is the total number of phenotypic observations and $\mathrm{tr}(\mathbf{H})$ represents the degrees of freedom occupied by the penalised fit (e.g. the linear mixed model fit). The hat matrix $\mathbf{H}$ transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$.

Set tests for genomic features. The set test statistics for the feature set can be computed in a number of ways. Below is described four different approaches all derived from the GBLUP model.

The **first set test statistic** is the covariance association test (CVAT)[19], which considers the covariance between the total genomic effect for all markers ($\hat{\mathbf{g}} = \sum_{i=1}^{m} \mathbf{w}_i \hat{s}_i$) and the genomic effect for the feature ($\hat{\mathbf{g}}_f = \sum_{i=1}^{m_f} \mathbf{w}_i \hat{s}_i$):

$$T_{CVAT} = \hat{\mathbf{g}}'\hat{\mathbf{g}}_f = (\hat{\mathbf{g}}_f' + \hat{\mathbf{g}}_r')\hat{\mathbf{g}}_f = \hat{\mathbf{g}}_f'\hat{\mathbf{g}}_f + \hat{\mathbf{g}}_r'\hat{\mathbf{g}}_f. \tag{5}$$

In this expression $\hat{\mathbf{g}}_r = \sum_{i=1}^{m_r} \mathbf{w}_i \hat{s}_i$ is the genomic effect for the remaining set of markers. The number of markers in feature and in the remaining set of markers is given by m_f and m_r respectively.

The distribution of this set test statistic under the competitive null hypothesis (genomic feature comprises randomly sampled markers) is unknown and an empirical distribution is required. An empirical distribution for the competitive null hypothesis can be obtained by sampling m_f columns in $\mathbf{W}$ at random.

The **second set test statistic** considered is a commonly used score based approach. It is derived from the first derivative of the likelihood as is Rao's score test[26]. A key difference compared to Rao's score test is that only the quadratic term in the first derivative form the basis of this test statistic[27–29] from an argument that this is the only part that involves the data[30]. The score based approach used here is thus equivalent to the sequence kernel association test (SKAT)[28]. The score statistic can therefore be written as:

$$T_{Score} = \frac{1}{2}(\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}\mathbf{G}_f\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}), \tag{6}$$

where the fixed effects $\mathbf{b}$ and the phenotypic covariance matrix $\mathbf{V}$ are estimated under a null model. The purpose of the null model is to adjust for environmental non-genetic factors, and for genetic factors not part of the genomic feature, including population structure. Several alternative null models can be used in the score test approach. If the GBLUP model is used as the null model the genomic effects can either be defined as $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$ or alternatively $\mathbf{g} \sim N(0, \mathbf{G}_r\sigma_r^2)$. In the first case the genomic relationship matrix is computed using all genetic markers and therefore the null model needs only to be fitted once. In the latter case, it is computed using only the genetic markers not included in the genomic feature which requires us to fit a different null model for each genomic feature. The set test statistic for the score approach can be re-written as:

$$T_{Score} = \frac{1}{2}\hat{\mathbf{e}}'\mathbf{Z}\mathbf{G}_f\mathbf{Z}'\hat{\mathbf{e}} = \frac{1}{2}\hat{\mathbf{e}}'\mathbf{Z}\frac{\mathbf{W}_f\mathbf{W}_f'}{m_f}\mathbf{Z}'\hat{\mathbf{e}}, \tag{7}$$

where $\hat{\mathbf{e}} = \hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$. The empirical distribution of the score set test statistic under the competitive null hypothesis is obtained by randomly sampling m_f columns in $\mathbf{W}$. It is also possible to derive an approximate distribution using the Satterthwaite's procedure of moment matching to approximate the null distribution of T_{Score} by a Gamma distribution[29]. The two parameters in the approximate distribution are calculated by matching the first and second moments (mean and variance) with those of the score set test statistic.

The **third test statistic** is based on the sum of the test statistic for all genetic markers belonging to the same genomic feature such as:

$$T_{sum} = \sum_{i=1}^{m_f} t_i^2, \tag{8}$$

where t_i represents the i'th single variant test statistic, e.g. marker effects ($\hat{s}$) or t-statistics. The distribution of this test statistic under the null hypothesis (associated markers are picked at random from the total number of tested genetic markers) is unknown and an empirical distribution is required. In this study both $\hat{s}$ and the t-statistic in equation 4 were used to compute T_{sum}.

The **fourth set test statistic** is based on counting the number of genetic markers in the feature that are associated with the trait phenotype and is computed as:

$$T_{count} = \sum_{i=1}^{m_f} I(t_i > t_0), \tag{9}$$

where m_f is the number of markers in the feature, t_i is the i'th single marker test statistic (e.g. t-statistic), t_0 is an arbitrary chosen threshold for the single marker test statistics, and I is an indicator function that takes the value one if the argument $(\mathrm{abs}(t_i) > t_0)$ is satisfied. Under the null hypothesis (i.e. individually associated markers are

distributed randomly, thus, the number of associated markers within a feature is indifferent compared to a random set of markers) it is assumed that $T_{count} \sim Hyper(m, m_a, m_f)$ is a realization from a hypergeometric distribution with parameters m (total number of genetic marker tested), m_a (total number of associated genetic markers amongst all markers) and m_f (total number of genetic markers in the feature). Alternatively, the statistical significance of the T_{count} statistic can be assessed using a χ^2 test for independence[31] or by obtaining an empirical distribution under a specific null hypothesis.

Extensions to GBLUP-derived CVAT. The CVAT is a flexible set test approach which can be extended in a number of ways facilitating further investigation of the underlying genetic architecture of complex traits. E.g. it can be decomposed at different levels of a hierarchy of gene sets, genes and markers; it can be derived from a model with multiple genetic components; or it can be derived from multiple trait models.

First, the CVAT statistic can be decomposed at different levels of a hierarchical genomic feature classification scheme, such as decomposing the covariance between the total genomic value and the genomic value defined by a genomic feature at the pathway level (e.g. group of genes) into the contribution from individual genes ($\hat{\mathbf{g}}_f = \sum_{i=1}^{n_{genes}} \hat{\mathbf{g}}_{f_i}$) to the covariance test statistics and even single markers ($\hat{\mathbf{g}}_{f_i} = \sum_{j=1}^{m_{f_i}} \mathbf{w}_j \hat{s}_j$) within a gene. The number of SNPs m_{f_i} located within genes varies (due to gene size etc.) and therefore partitioned covariance test statistics at the gene level are presented "per SNP".

Second, the CVAT statistic can be derived from a GFBLUP model with multiple genetic components[14–16, 19]. The total genomic values in the GBLUP model are assumed to be drawn from the same distribution $\mathbf{g} \sim N(0, \mathbf{G}\sigma^2_g)$. It is, however, very likely that the genomic values come from a mixture of distributions, e.g. groups of genetic markers having different effects based on their minor allele frequency (MAF)[20] or genetic markers known a priori to have large effects (e.g. discovered in previous GWAS). Such prior information can be used by fitting multiple genetic components in the linear mixed model:

$$\mathbf{y} = \mathbf{Xb} + \sum_{i=1}^{n_f} \mathbf{Z}\mathbf{g}_i + \mathbf{e}. \tag{10}$$

The notation is similar to the GBLUP model presented above except $\mathbf{g}_i$ is the vector of genetic values captured by the i'th genetic marker set. The random genetic effects and residuals were assumed to be independent and distributed as $\mathbf{g}_i \sim N(0, \mathbf{G}_i\sigma^2_{g_i})$, and $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$ where $\mathbf{G}_i = \mathbf{W}_i'\mathbf{W}_i/m_i$ is the additive genomic relationship matrix for the i'th genetic marker set. The single marker effects derived from the GFBLUP model are computed as: $\hat{\mathbf{s}}_i = \mathbf{W}_i'(\mathbf{W}_i\mathbf{W}_i')^{-1}\hat{\mathbf{g}}_i$, thus $\hat{\mathbf{s}} = [\hat{\mathbf{s}}_1' \ldots \hat{\mathbf{s}}_{n_f}']$.

Third, the CVAT statistic can be derived from a multiple trait GBLUP model (or GFBLUP model)[21, 22]. This can be important if we have records on correlated traits, for example a high heritability trait (or a trait with many observations) correlated with a low heritability trait (or a trait with few observations). In such a situation using a multiple trait model is likely to increase the accuray of the predicted total genetic value and single marker effects for the low heritability trait which in turn will increase the power of the set test. This becomes highly relevant for borrowing information across traits or same trait recorded in different breeds or study populations. The linear mixed model for multiple traits (2 traits in this example) can be expressed as:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1\mathbf{b}_1\\ \mathbf{X}_2\mathbf{b}_2\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_1\mathbf{g}_1\\ \mathbf{Z}_2\mathbf{g}_2\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}. \tag{11}$$

The notation is similar to the GBLUP model presented above except that $\mathbf{y}_1$ and $\mathbf{y}_2$ are vectors of phenotypes for trait 1 and 2, respectively. $\mathbf{X}_1$ and $\mathbf{X}_2$ are design matrices for the fixed effects and $\mathbf{b}_1$ and $\mathbf{b}_2$ are the vectors of these fixed effects. $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are design matrices for the random effects, $\mathbf{g}_1$ and $\mathbf{g}_2$ are vectors of total genetic values and $\mathbf{e}_1$ and $\mathbf{e}_2$ are vectors of residuals for trait 1 and 2.

The random genetic effects, $\mathbf{g} = \begin{bmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{bmatrix}$, and residuals, $\mathbf{e} = \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}$, were assumed to be independent and distributed as $\mathbf{g} \sim N\left(0, \mathbf{G} \otimes \begin{bmatrix}\sigma^2_{g_{11}} & \sigma^2_{g_{12}}\\ \sigma^2_{g_{21}} & \sigma^2_{g_{22}}\end{bmatrix}\right)$, and $\mathbf{e} \sim N\left(0, \mathbf{I} \otimes \begin{bmatrix}\sigma^2_{e_{11}} & \sigma^2_{e_{12}}\\ \sigma^2_{e_{21}} & \sigma^2_{e_{22}}\end{bmatrix}\right)$. Furthermore, T_{CVAT} can be used to identify features associated with the covariance between total genetic values in different traits expressed as:

$$T_{CVAT} = \hat{\mathbf{g}}_1'\hat{\mathbf{g}}_{f_2} = \hat{\mathbf{g}}_{f_1}'\hat{\mathbf{g}}_{f_2} + \hat{\mathbf{g}}_{r_1}'\hat{\mathbf{g}}_{f_2}, \tag{12}$$

which consider the covariance between the total genomic effect for all markers $\left(\hat{\mathbf{g}}_1 = \sum_{i=1}^{m} \mathbf{w}_i \hat{s}_{1_i}\right)$ of trait 1 (or trait 2) and the genomic effect for a feature $(\hat{\mathbf{g}}_{f_2} = \sum_{i=1}^{m_f} \mathbf{w}_i \hat{s}_{2_i})$ of trait 2 (or trait 1).

Fitting linear models and estimation of variance components. Estimates of the variance components (i.e. $\hat{\sigma}^2_g$, $\hat{\sigma}^2_{g_1}$, $\hat{\sigma}^2_{g_2}$, $\hat{\sigma}^2_{g_{12}}$ and $\hat{\sigma}^2_e$) defined in the models described above were obtained using an average information restricted maximum likelihood (AI-REML) procedure[32, 33] as implemented in the software DMU. In this procedure, matrices were not full rank due to centering of the observed genotypes, which necessitated a generalized inverse of the genomic relationship matrices.

Testing for association between a genomic feature and a phenotype. The test for association was based on a competitive null hypothesis, i.e. that the degree of association of the feature set was the same as that of a random marker set[27,34].

A null hypothesis is only competitive if the parameters influencing the test statistic are identical to the alternative hypothesis. Thus, there must be an equal number of markers for the random set and the true set, and the correlation structure among markers (due to linkage disequilibrium) should be retained. The empirical distribution of the test statistics was therefore obtained using the circular permutation procedure as described in Cabrera *et al.*[35]. The genome was considered to be circular, ordered from chromosome 2 L to chromosome X and restarting again at chromosome 2 L. Then the complete set of observed test statistics are permuted by rotation with respect to their genomic locations, i.e. a random number between 1 and the total number of SNPs is drawn, and the observed test statistic for the first SNP in the genome rotates to that of the random number-th SNP and all other test statistics rotate to the same degree to the corresponding SNPs. Thus, SNPs retain the same original order but, at each permutation, gain new random test statistics. This uncouples any associations between SNPs and the genomic feature, while retaining similar patterns of the correlation structure among test statistics. A new set test statistic was then computed based on the original position of the genomic features. The permutation was repeated 10,000 times for each set in the feature class, and empirical p-values were obtained through one-tailed tests of the proportion of randomly sampled test statistics larger than that observed.

Implementation. The GBLUP-derived set test approaches described above were 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 methods such as REML, computing the test statistic for the set of genetic markers, and testing the statistical significance of the observed test statistic under a specific null hypothesis. Example scripts and data sets are provided for illustrating the GBLUP model derived set test approaches. For specific experimental design with replicated phenotypes within line such as DGRP it is more efficient to use the AI-REML procedure[32,33] implemented in DMU[32]. The AI-REML function in the qgg package provides an R interface to the DMU which can be downloaded from http://dmu.agrsci.dk/DMU/. The CVAT approach can also be derived from the REML procedures implemented in existing software packages commonly used in genomics such as GCTA[36], LDAK[37], DISSECT[38] and MTG2[39].

Simulation study comparing set test approaches. To compare the different set test approaches described above, and to investigate different factors that might influence the power to detect causal sets of SNPs, a series of phenotypic simulations were established. The factors varied in the simulations should imitate different genetic architectures and included genomic heritability (h^2), proportion of genomic variance explained by causal SNPs in the genomic feature (h_f^2), proportion of non-causal SNPs in the genetic marker set defined by the genomic feature (*dilution*), genome distribution of causal SNPs (*causal model*, i.e. whether the causal SNPs were distributed in the genome randomly or clustered in groups) and the number of phenotypic records for each genotype (N_{rep}). For each data set and replicate we estimated variance components for the GBLUP and GFBLUP models using AI-REML and applied the different set tests (not including the extensions to CVAT).

Simulated data. The simulations were based on the real genotype DGRP data set of 205 lines. Genotypes were originally obtained from whole genome sequences using an integrative genotyping procedure[18]. All simulations were based on segregating biallelic single nucleotide polymorphisms (SNPs) with minor allele frequencies $\geq$0.05 and for which the Phred scaled variant quality was greater than 500 and the genotype call rate was $\geq$0.8, resulting in a total of 1,725,755 SNPs.

Causal sets. In all scenarios, there were 1,000 causal SNPs, which were divided into two subsets. The first subset, C_1, contained 100 SNPs and was used as the causal SNP set in the genomic feature that explained 10%, 20%, 30%, or 50% of the genomic variance. The second subset, C_2, contained 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 non-causal SNPs within the causal sets, we added an increasing number of non-causal SNPs (100, 200, …, 1,900, 2,000), to the causal set C_1, in a process referred to as *dilution*. To determine the false-positive rate, 50 non-causal SNP sets of varying sizes (10 sets each containing 0.1 k, 0.5 k, 1 k, 5 k or 10 k SNPs) were sampled, none of which were contained in the causal sets of SNPs.

Phenotypes. Phenotypes were simulated using the following linear model: $\mathbf{y} = \mathbf{Z}\mathbf{g}_1 + \mathbf{Z}\mathbf{g}_2 + \mathbf{e}$, where $\mathbf{g}_1 \sim N(0, \mathbf{G}_1\sigma^2_{g1})$, $\mathbf{g}_2 \sim N(0, \mathbf{G}_2\sigma^2_{g2})$, and $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$. $\mathbf{G}_1$ and $\mathbf{G}_2$ are the genomic relationship matrices for causal SNPs in C_1 and C_2, respectively. $\mathbf{Z}$ is a design matrix linking DGRP lines to individual phenotypes. The total phenotypic variance $\sigma^2_P = \sigma^2_{g1} + \sigma^2_{g2} + \sigma^2_e$ was 100 in all scenarios. We simulated data with additive genomic heritabilities $\left(h^2 = \frac{\sigma^2_{g1} + \sigma^2_{g2}}{\sigma^2_{g1} + \sigma^2_{g2} + \sigma^2_e}\right)$ of 0.1, 0.3, or 0.5, to analyse scenarios with low to intermediate heritabilities, reflecting those observed in the real data. To analyse scenarios with non-uniform SNP effects, the proportion of

additive genomic variance explained by the causal SNPs in $C_1\left(h_f^2 = \frac{\sigma_{g1}^2}{\sigma_{g1}^2 + \sigma_{g2}^2}\right)$ was varied across scenarios: 0.1, 0.2, 0.3, or 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 1,440 individual simulated data sets [3 (N_{rep}) × 3 (h^2) × 4 (h_f^2) × 2 (causal model) × 20 (dilution)]. For each possible combination of factors 50 independent data sets were obtained.

Assessing the power of set test statistics. To measure the performance of the different test statistics we used the F_1 score:

$$F_1 = 2\frac{p \cdot r}{p + r}, \tag{13}$$

where $p = TP/(TP + FP)$ is the precision and $r = TP/(TP + FN)$ is the recall. The F_1 score is the harmonic mean of precision and recall[40]. The recall *r* is the proportion of true positives (TP) that are correctly identified, i.e. the ratio between the number of identified causal sets and the number of sets that should have been identified, thus, the sum of TP and false negatives (FN). Contrary, the precision p is the proportion of positives that truly are positives, i.e. the proportion of true causal sets of all sets identified, thus, the sum of TP and false positives (FP). The F_1 score can take values between 0 and 1, with maximum performance at the value of 1. The F_1 score was calculated for each test staticstic under each combination of factors, using a p-value cut-off of 0.05 for a positive detection of a genomic feature.

Comparing set test results with the predictive ability of the GFBLUP model. We investigated whether the results obtained using the GBLUP derived set tests can be used to develop more accurate GFBLUP prediction models. The GFBLUP model is an extension of the traditional GBLUP model, where an additional genomic effect (defined by the genomic feature) is included in the linear mixed model[15]. The predictive ability of the GFBLUP model was assessed using a cross validation procedure[15]. In the GFBLUP model the total genomic value is $\hat{\mathbf{g}}_{total} = \hat{\mathbf{f}} + \hat{\mathbf{r}}$, where $\hat{\mathbf{f}}$ is a vector of genomic values captured by genetic markers linked to the genomic feature of interest, $\hat{\mathbf{r}}$ is a vector of genomic values captured by genetic markers outside the genomic feature. In the cross validation procedure, we estimated genomic parameters using the phenotypes from the DGRP lines in the training data (90% of the lines) and predicted the total genomic value of DGRP lines in the validation data (10% of the lines). We then calculated Spearman correlations 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. For comparing the GBLUP-derived set tests with the predictive ability of the GFBLUP model, we calculated the Spearman rank based correlation between the level of significance of the set test statistic and the predictive ability.

CVAT (and its extensions) exemplified on CCRT. We applied the GBLUP-derived CVAT on CCRT measured in the DGRP. The CVAT test statistic was chosen based on its good performance in the simulation studies (see first section of Results). Individual genes and gene ontology (GO) terms defined genetic marker sets (genomic features) for which T_{CVAT} was computed. The relationship between this test statistic and the predictive ability of incorporating these GO terms as features in the GFBLUP model[15] was considered.

DGRP data. *Drosophila lines.* The phenotypic and genotypic data originate from the *Drosophila melanogaster* Genetic Reference Panel (DGRP)[17,18]. All data can be accessed via the website: 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, NC, USA population, and which have full genome sequence data available[17,18]. All flies were reared under standard culture conditions (cornmeal-molasses-agar-medium, 25°C, 60–75% relative humidity, 12-hr light-dark cycle). The DGRP is polymorphic for common inversions and *Wolbachia pipientis* infection status[18]. These factors were included in the models described below as fixed effects.

Quantitative trait phenotype. Chill coma recovery time (CCRT) for 159 DGRP lines was measured by transferring three to seven day old flies without anesthesia to empty vials, and placing them on ice for three hours. Flies were transferred to room temperature, and the time it took for each individual to right itself and stand on its legs was recorded[41]. There were two replicates of ~50 flies/sex/line (total $N = 32{,}231$; female $N = 16{,}170$; male $N = 16{,}061$).

Genotypes. Genotypes were obtained from whole genome sequences using an integrative genotyping procedure[18]. All analyses were based on segregating biallelic single nucleotide polymorphisms (SNPs) with minor allele frequencies ≥0.05 and for which the Phred scaled variant quality was greater than 500 and the genotype call rate was ≥0.8, for a total of 1,725,755 SNPs distributed on six chromosome arms (*2L*, *2R*, *3L*, *3R*, *4* and *X*).

Genomic features. Genomic features were defined at gene-level and GO level. Genes grouped according to a specific GO term were considered a genomic feature. Genes were linked to the 'Biological Processes' (BP), 'Molecular Function' (MF), and 'Cellular Component' (CC) GO terms[42] using the BioConductor package 'org.Dm.eg.db' v. 2.14[43]. 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[17,18,44]. Only the 963,235 SNPs located within genes (i.e. within open reading frames) were used for the genomic feature. In total the markers

were annotated to 10,517 genes and 1,117 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 23–163,938.

Single and multiple trait CVAT. We applied CVAT to the CCRT data, and considered CCRT in males and females as two different, but correlated traits. The multiple trait CVAT analysis was based on phenotypic records of the quantitative trait adjusted for relevant factors using the following multi-trait linear mixed model:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix}=\begin{bmatrix}\mathbf{X}_1\mathbf{b}_1\\ \mathbf{X}_2\mathbf{b}_2\end{bmatrix}+\begin{bmatrix}\mathbf{Z}_1\mathbf{g}_1\\ \mathbf{Z}_2\mathbf{g}_2\end{bmatrix}+\begin{bmatrix}\mathbf{Q}_1\mathbf{l}_1\\ \mathbf{0}\end{bmatrix}+\begin{bmatrix}\mathbf{0}\\ \mathbf{Q}_2\mathbf{l}_2\end{bmatrix}+\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}, \tag{14}$$

$\mathbf{y}_1$ and $\mathbf{y}_2$ are vectors of phenotypes for trait 1 (males) and 2 (females), $\mathbf{X}_1$ and $\mathbf{X}_2$ are design matrices for fixed effects of inversion karyotypes and *Wolbachia* infection status and $\mathbf{b}_1$ and $\mathbf{b}_2$ are the vectors of these fixed effects. $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are design matrices linking observations to genomic values, $\mathbf{g}_1$ and $\mathbf{g}_2$ are vectors of total genetic values. $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are design matrices for replicate within line effects, $\mathbf{l}_1$ and $\mathbf{l}_2$ the vectors of replicate within line effects, and $\mathbf{e}_1$ and $\mathbf{e}_2$ are vectors of residuals for trait 1 and 2.

The random genetic effects, $\mathbf{g}=\begin{bmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{bmatrix}$, line effects, $\mathbf{l}=\begin{bmatrix}\mathbf{l}_1\\ \mathbf{l}_2\end{bmatrix}$ and residuals, $\mathbf{e}=\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}$ were assumed to be independent and distributed as $\mathbf{g}\sim \mathrm{N}\left(0,\ \mathbf{G}\otimes\begin{bmatrix}\sigma^2_{g_{11}} & \sigma^2_{g_{12}}\\ \sigma^2_{g_{21}} & \sigma^2_{g_{22}}\end{bmatrix}\right)$, $\mathbf{l}_1\sim \mathrm{N}\left(0,\ \mathbf{I}_{l_1}\sigma^2_{l_1}\right)$, $\mathbf{l}_2\sim \mathrm{N}\left(0,\ \mathbf{I}_{l_2}\sigma^2_{l_2}\right)$ and $\mathbf{e}\sim \mathrm{N}\left(0,\ \mathbf{I}\otimes\begin{bmatrix}\sigma^2_{e_{11}} & \sigma^2_{e_{12}}\\ \sigma^2_{e_{21}} & \sigma^2_{e_{22}}\end{bmatrix}\right)$. Since the phenotypes for males and females were recorded in different environments, we assume that $\sigma^2_{e_{12}}=\sigma^2_{e_{21}}=0$.

The CVAT test statistic, $\mathrm{T}_{\mathrm{CVAT}}$, was computed using the vectors of total genomic values in males and females, $\mathbf{g}_1$ and $\mathbf{g}_2$ from the multiple trait analyses. The within trait CVAT test statistics were computed as $\mathrm{T}_{\mathrm{CVAT_M}}=\hat{\mathbf{g}}_1'\hat{\mathbf{g}}_{f_1}$ for males (trait 1) and $\mathrm{T}_{\mathrm{CVAT_F}}=\hat{\mathbf{g}}_2'\hat{\mathbf{g}}_{f_2}$ for females (trait 2). The across trait CVAT test statistics were computed as $\mathrm{T}_{\mathrm{CVAT_{MF}}}=\hat{\mathbf{g}}_1'\hat{\mathbf{g}}_{f_2}$ or $\mathrm{T}_{\mathrm{CVAT_{FM}}}=\hat{\mathbf{g}}_2'\hat{\mathbf{g}}_{f_1}$ which consider the covariance between the total genomic effect for all markers ($\hat{\mathbf{g}}_1=\sum_{i=1}^{m}\mathbf{w}_i\hat{s}_{1_i}$) of trait 1 (or trait 2) and the genomic effect for a feature ($\hat{\mathbf{g}}_{f_2}=\sum_{i=1}^{m_f}\mathbf{w}_i\hat{s}_{2_i}$) of trait 2 (or trait 1).

The single trait CVAT was done by analysing phenotypes for males and females separately using the same fixed and random factors as in the multiple trait model presented above.

An empirical distribution of these test statistics was based on a competitive null hypothesis using the permutation procedure described earlier. Two competive null hypotheses were used to test if the observed test statistics of the genomic feature differs from the test statistics obtained by randomly sampling genetic markers from (a) exclusively genic regions or (b) the whole genome (i.e. genic and intergenic regions). Thus empirical distributions were obtained by either sampling genetic markers randomly from gene regions or the whole genome.

Results

Comparison of set test statistics on simulated data. *Comparison of power for set test statistics.* The covariance association test, $\mathrm{T}_{\mathrm{CVAT}}$, was generally more powerful (i.e. highest F_1 score) than other set test statistics, across all scenarios (Fig. 1) under the random model. The figure displays estimated power for different set test statistics across 3 different trait heritabilities (h^2), three number of replicates (N_{rep}), and 4 levels of proportion of additive genetic variance explained by causal SNPs (h^2_f). The F_1 score was calculated for the average of each set test result over a dilution range of 0 to 2,000 non-causal SNPs added to the C_1 causal set. The superior performance of CVAT becomes more pronounced as the genomic heritability increases (left to right column of Fig. 1), genomic variance explained by feature increases (darker colour of points in Fig. 1), and number of replicates increase (top to bottom row of Fig. 1). Slightly less power was observed for the score based test statistic $\mathrm{T}_{\mathrm{Score}}$ followed closely by the set test statistic $\mathrm{T}_{\mathrm{Sum}}$ based on sums of single marker test statistics (marker effects $\hat{s}$ or t-statistics). This trend is observed across all scenarios. All of the aforementioned set test statistics mostly outperform the count based set test statistic $\mathrm{T}_{\mathrm{Count}}$ at p-value cut-offs of 0.05 and 0.01. However, when the feature explains 50% of the genomic variance (i.e. $h^2_f=0.5$) the power of $\mathrm{T}_{\mathrm{Count}}$ (using a stringent single marker p-value cut-off ($p<0.01$)) improves, as heritability and number of replicates increase, such that it's power reaches levels comparable to the score based set test statistic (Fig. 1).

Relationship between set test statistics. The p-values of set test statistics $\mathrm{T}_{\mathrm{CVAT}}$ and $\mathrm{T}_{\mathrm{Score}}$ were highly correlated (0.96) with each other (Fig. 2). The figure shows the relationship between the minus logarithm of the p-values for the observed set test statistics. The results represented is for a genomic heritability of 30% and where the genomic feature explains 30% of the genomic variance (i.e. $h^2=0.3$ and $h^2_f=0.3$). Although less pronounced, $\mathrm{T}_{\mathrm{CVAT}}$ and $\mathrm{T}_{\mathrm{Score}}$ also showed a high correlation with $\mathrm{T}_{\mathrm{Sum}}$ of single marker effects $\hat{s}$ (0.87 and 0.85, respectively) and t-statistics (0.87 and 0.85, respectively). Lower correlations were observed between $\mathrm{T}_{\mathrm{CVAT}}$ and $\mathrm{T}_{\mathrm{Count}}$ at p-value cut-offs of 0.05 (0.76) and 0.01 (0.52). This was also the case for $\mathrm{T}_{\mathrm{Score}}$ and $\mathrm{T}_{\mathrm{Count}}$ showing a correlation of 0.74 at a p-value cut-off of 0.05 and 0.48 at p-values less than 0.01.

Relationship between set test statistics and predictive ability of the GFBLUP model. The three set test statistics ($\mathrm{T}_{\mathrm{CVAT}}$, $\mathrm{T}_{\mathrm{Score}}$, and $\mathrm{T}_{\mathrm{Sum}}$) were all highly correlated to the predictive ability of the GFBLUP model (ranging from 0.59 to 0.62, respectively, Fig. 3). The pair-wise plots presented in Fig. 3 show the relationship between the minus logarithm of the p-value for the observed set test statistic and the predictive ability of the GFBLUP model. The correlation between predictive ability and the count based set test statistics was slightly lower ranging from 0.34 to 0.48.

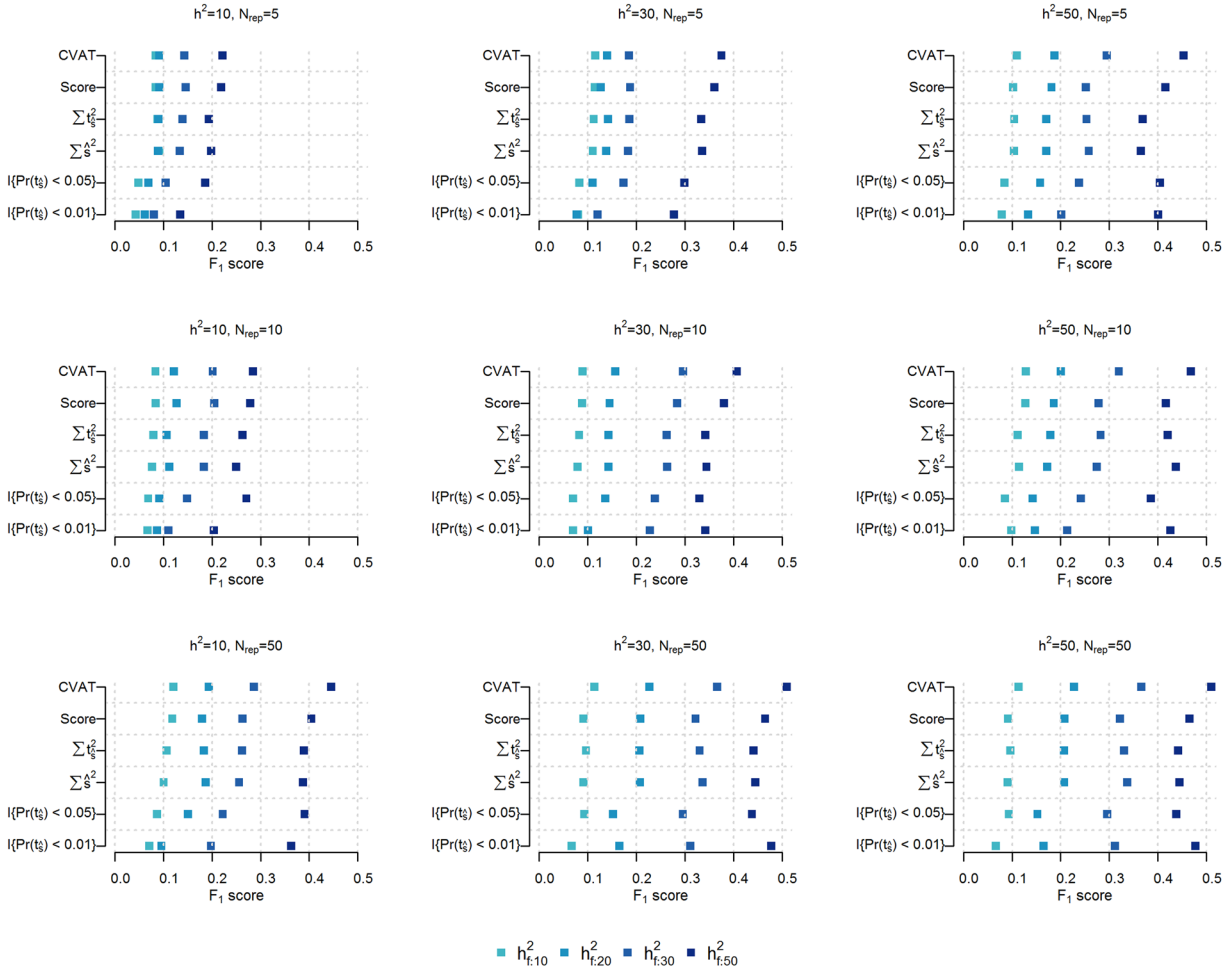

Figure 1. Comparison of detection power between set test statistics. The F_1 score (x-axis) was used to measure the performance of the GBLUP derived set test statistics (y-axis), i.e. T_{CVAT} (CVAT), T_{Score} (Score), T_{Sum} using single marker effects ($\sum\hat{s}^2$) or using single marker t-statistics ($\sum t_{\hat{s}}^2$), and T_{Count} with a threshold of p-value < 0.05 ($I\{Pr(t_{\hat{s}}) < 0.05\}$) and p-value < 0.01 ($I\{Pr(t_{\hat{s}}) < 0.01\}$). The F_1 score was calculated using the average set test statistic results over a dilution range of adding 0 to 2000 non-causal SNPs to the C_1 causal set. P-value cut-off for the set test statistic was 0.05. Each panel represent a different combination of genomic heritability (h^2) and number of replicates within lines (N_{rep}), whereas h_f^2 is visualized by the colour gradient. Results are for the scenarios with three different levels of genomic heritability ($h^2 = 0.1$, 0.3 or 0.5, columns left to right), four different levels of proportion of genomic variance explained by the causal markers in the genomic feature ($h_f^2 = 0.1$, 0.2, 0.3 or 0.5, light to dark colour), and three different levels of number of replicates within lines ($N_{rep} = 5$, 10, or 50, rows top to bottom). Causal sets, including SNPs in feature (C_1) and not in feature (C_2), consisted of SNPs randomly selected from the complete SNP set (random causal model).

Influence of genomic feature and trait specific factors on detection power. Here we present the results for the GBLUP-derived CVAT set test statistic (Fig. 4). We focus on the results of the CVAT test statistic since it had the best performance (i.e. highest F_1 score across all simulation scenarios, Fig. 1). The patterns observed are very similar for the other set test statistics (results not shown). Power to detect genomic features affecting the phenotypes was influenced both by trait and genomic feature specific factors. The proportion of the genomic variance explained by the genomic feature (h_f^2) greatly impacted detection power (higher levels of power from left to right columns of Fig. 4) and robustness towards dilution, i.e. increasing the proportion of non-causal SNPs in the genomic feature (curves as a function of dilution are more steep with decreasing h_f^2 in Fig. 4). Power to detect genomic features was low if both genomic heritability and proportion of genomic variance explained by genomic feature was low ($h_f^2 = 0.1$ and $h^2 = 0.1$), even without dilution. Impact of dilution was less severe when the proportion of genomic variance explained by genomic feature was highest ($h_f^2 = 0.5$). This increased robustness towards dilution resulted in power above 40% 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 (Fig. 4). However, at high h_f^2 and in absence of dilution all genomic features were detected regardless of overall genomic heritability, but with some false positives. Furthermore, if h_f^2 was high, the detection power of CVAT for high heritability traits were less affected by dilution than low heritability traits (steeper slope of upper-right panel,

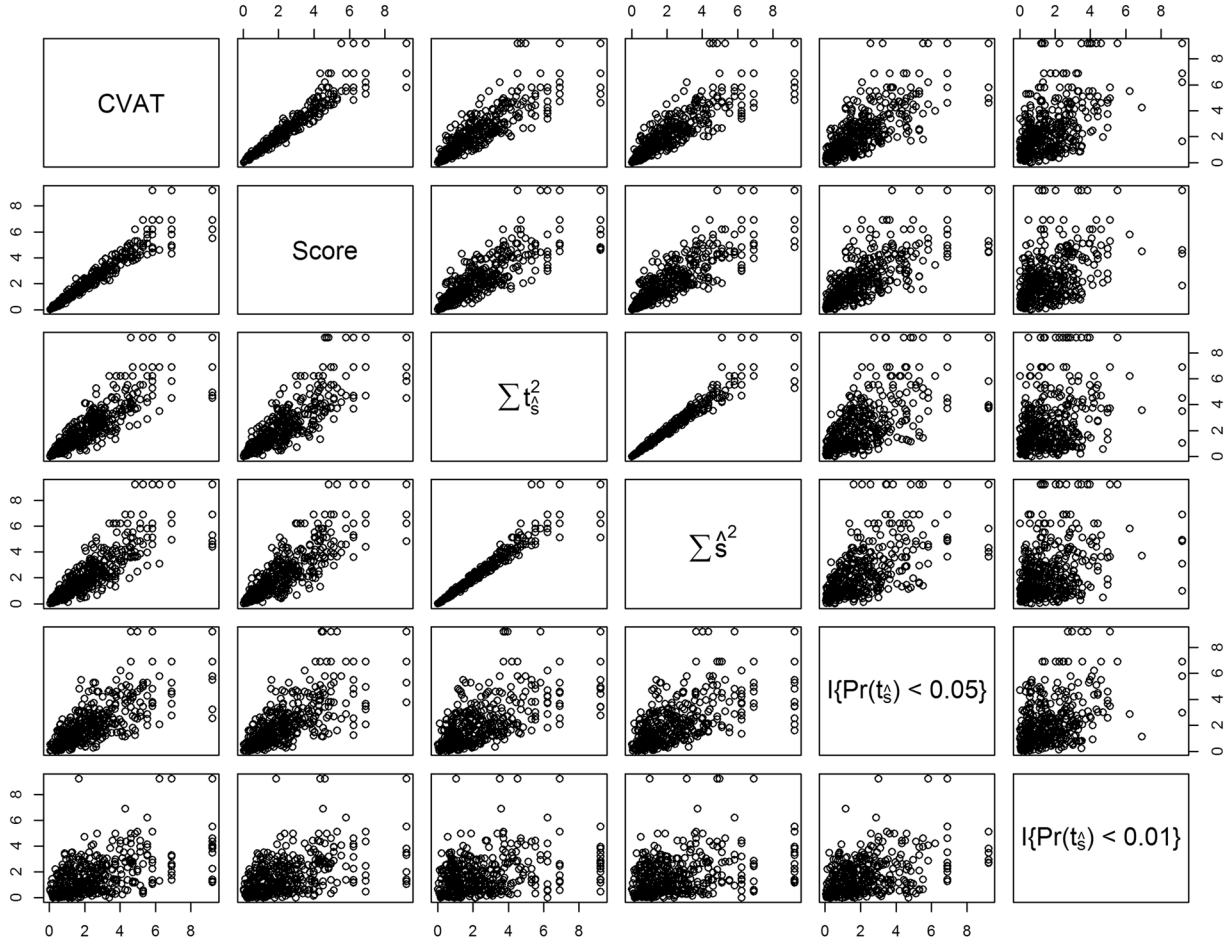

Figure 2. Relationship between the significance levels of different set test statistics. Scatter plots of all pairwise combinations of significance between the set test statistics, i.e. T_{CVAT} (CVAT), T_{Score} (Score), T_{Sum} using single marker effects ($\sum \hat{s}^2$) or using single marker t-statistics ($\sum \hat{t}_s^2$), and T_{Count} with a threshold p-value < 0.05 ($I\{Pr(\hat{t}_s) < 0.05\}$) and p-value < 0.01 ($I\{Pr(\hat{t}_s) < 0.01\}$). Significance, shown as $-\log(p)$, was measured for the association of simulated phenotype with genomic feature over a dilution range of adding 0 to 2000 non-causal SNPs to the C_1 causal set. Plots are arranged such that all plots in a row share a common y-axis, and all plots in a column share a common x-axis. The names of the x- and y-axes are shown in the diagonal boxes. Genomic heritability was set to 30% ($h^2 = 0.3$), and the proportion of genomic variance explained by the feature was 30% ($h_f^2 = 0.3$). The random causal model was used, randomly selecting causal SNPs (C_1 and C_2) from the complete set of SNPs. Five replicates were used within each line ($N_{rep} = 5$).

compared to lower-right panel of Fig. 4). Dilution decreased power in all simulation scenarios (decreasing curves on all panels of Fig. 4). Detection power was slightly higher if causal SNPs in the genomic feature were clustered in smaller regions as compared to distributed randomly on the genome (results not shown). Furthermore, detection power increases with increasing numbers of replicates within line ($N_{rep} = 5$, 10, or 50).

Application of CVAT on CCRT data. Since the simulation study suggested that CVAT was the most powerful set test statistic, we applied CVAT and its extensions to CCRT data.

Determination of linear mixed model to be used for CVAT analysis. Initially we fitted a series of linear mixed models (GBLUP or GFBLUP) to determine the final model to be used in the subsequent CVAT analyses of the CCRT trait in DGRP. Models were fitted for single and multiple traits, including one or two features (in this case genes and inter-genic regions) and considering gene based and genome based null hypotheses. Trait heritability for CCRT estimated using a multiple trait GBLUP model was 0.42 for males and 0.48 for females. The genetic correlation between males and females was 0.97. Partitioning genomic variance into genes and inter-genic regions, using a two-component genomic feature model, did not significantly improve the model fit (likelihood ratio test statistic less than one; p-value > 0.1). The empirical distribution of the CVAT test statistic was determined under two null hypotheses: One involving random sampling of genetic markers from gene regions (gene based), and one based on random sampling of genetic markers from the whole genome (whole genome based). Under the gene based null hypothesis GO terms were slightly more significant compared to the whole genome based null hypothesis both in the case of females and males (Fig. S1 panels (**a**) and (**b**) respectively). Furthermore, the association

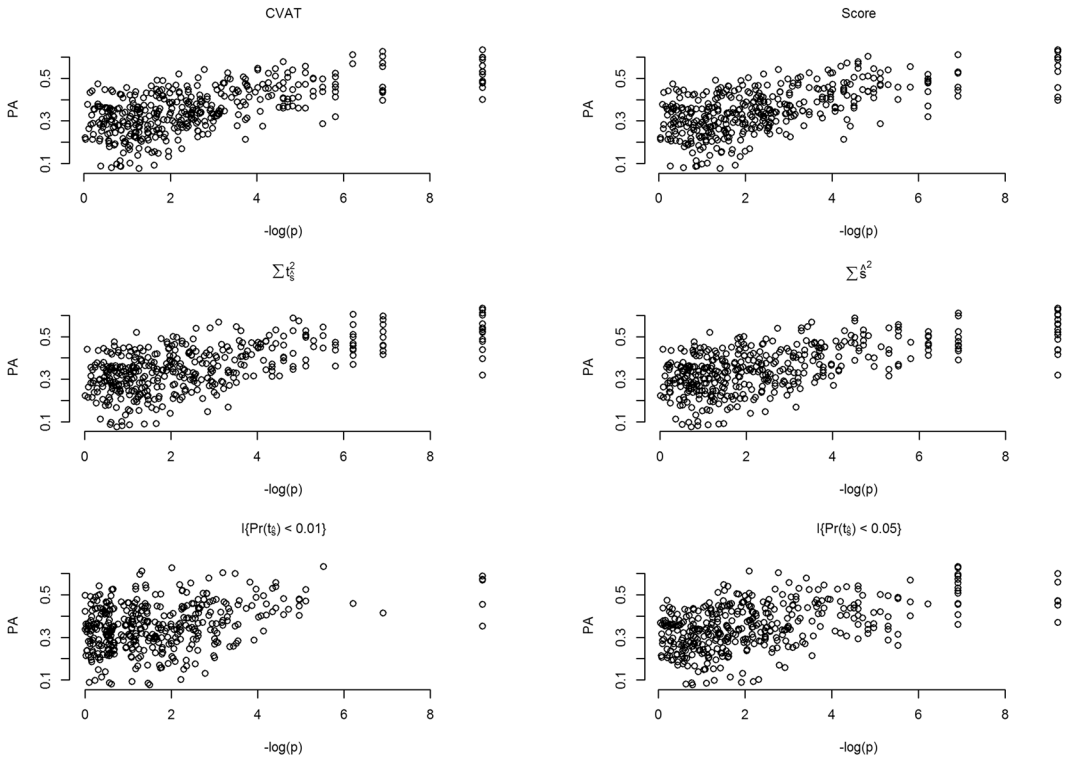

Figure 3. Relationship between the predictive ability of GFBLUP and significance levels of different set test statistics. Scatter plots showing the relationship between significance of set test statistics (x-axis) and predictive ability (PA, y-axis) of GFBLUP. Significance is expressed as $-\log(p)$. The different panels show results for the different set test statistics: T_{CVAT} (CVAT), T_{Score} (Score), T_{Sum} using single marker effects ($\sum \hat{s}^2$) or using single marker t-statistics ($\sum t_{\hat{s}}^2$), and T_{Count} with a threshold p-value < 0.05 ($I\{Pr(t_{\hat{s}}) < 0.05\}$) and p-value < 0.01 ($I\{Pr(t_{\hat{s}}) < 0.01\}$) Genomic heritability was set to 50% ($h^2 = 0.5$), and the proportion of genomic variance explained by the feature was 30% ($h_f^2 = 0.3$). The random causal model was used, randomly selecting causal SNPs (C_1 and C_2) from the complete set of SNPs. Five replicates were used within each line ($N_{rep} = 5$).

of GO terms with CCRT was highly correlated between males and females (Fig. S1 panel (**c**) and (**d**)) under both the gene based or whole genome null hypothesis (correlation = 0.95 and 0.98 respectively).

The significance of T_{CVAT} did not show a considerable difference whether using the two trait or single trait models (the third extension to T_{CVAT}, Fig. S1 panel (**e**) and (**f**)). Therefore, results for the CCRT trait reported here are from a one-component, two trait model using a gene based null hypothesis.

CCRT associated GO terms and genes detected using CVAT. Several GO terms were significantly associated to CCRT in both males and females (p-value adjusted for multiple tests ≤ 0.001, Table 1). Table 1 shows the highest-ranking GO terms for males and females (MT-CVAT within trait), as well as the significance of GO terms when considering the covariance between the total genomic effect for *all markers* in males, and the genomic effect for *markers* in the feature for females and vice versa (MT-CVAT across traits). Eight GO terms for females and nine GO terms for males were significantly associated with CCRT. Males and females shared all but one of the most significant GO terms (p-value ≤ 0.001). The GO term "ATP-dependent DNA helicase activity" (GO:0004003, $p = 0.0015$) for females being only slightly above the 0.001 p-value cut-off. The *across trait* (traits being females and males) MT-CVAT set test results showed similar patterns as the *within trait* CVAT set test.

However, all GO terms reported in Table 1 had unadjusted p-values below 0.01, suggesting that these may be biologically relevant for CCRT in both sexes. In addition, the top-ranking GO terms significantly associated with CCRT were also predictive of the phenotypes as assessed in a cross validation study (Fig. 5).

There was a substantial overlap among the SNPs associated with each of the GO terms (Fig. 6). In particular, "Rho GTPase activator activity" (GO:0005100) and "Rho protein signal transduction" (GO:0007266) shared more than 98% of the SNPs. These two GO terms also shared a substantial number of SNPs (59–67%) with the remaining GO terms except for "ATP-dependent DNA helicase activity" (GO:0004003) which did not share any SNPs with the other GO terms.

Considering the overlap of SNPs between GO terms, further investigations were required to better understand the biological relevance of the CVAT results obtained at the GO term level. Therefore, we applied the CVAT set test at the individual gene level using only the genes that were part of the significantly associated GO terms. This enabled us to identify a number of genes that were associated with CCRT (Fig. 7). In particular, we found that *RhoGAP88C* (FBgn0086901) was significantly associated with CCRT and that this gene is

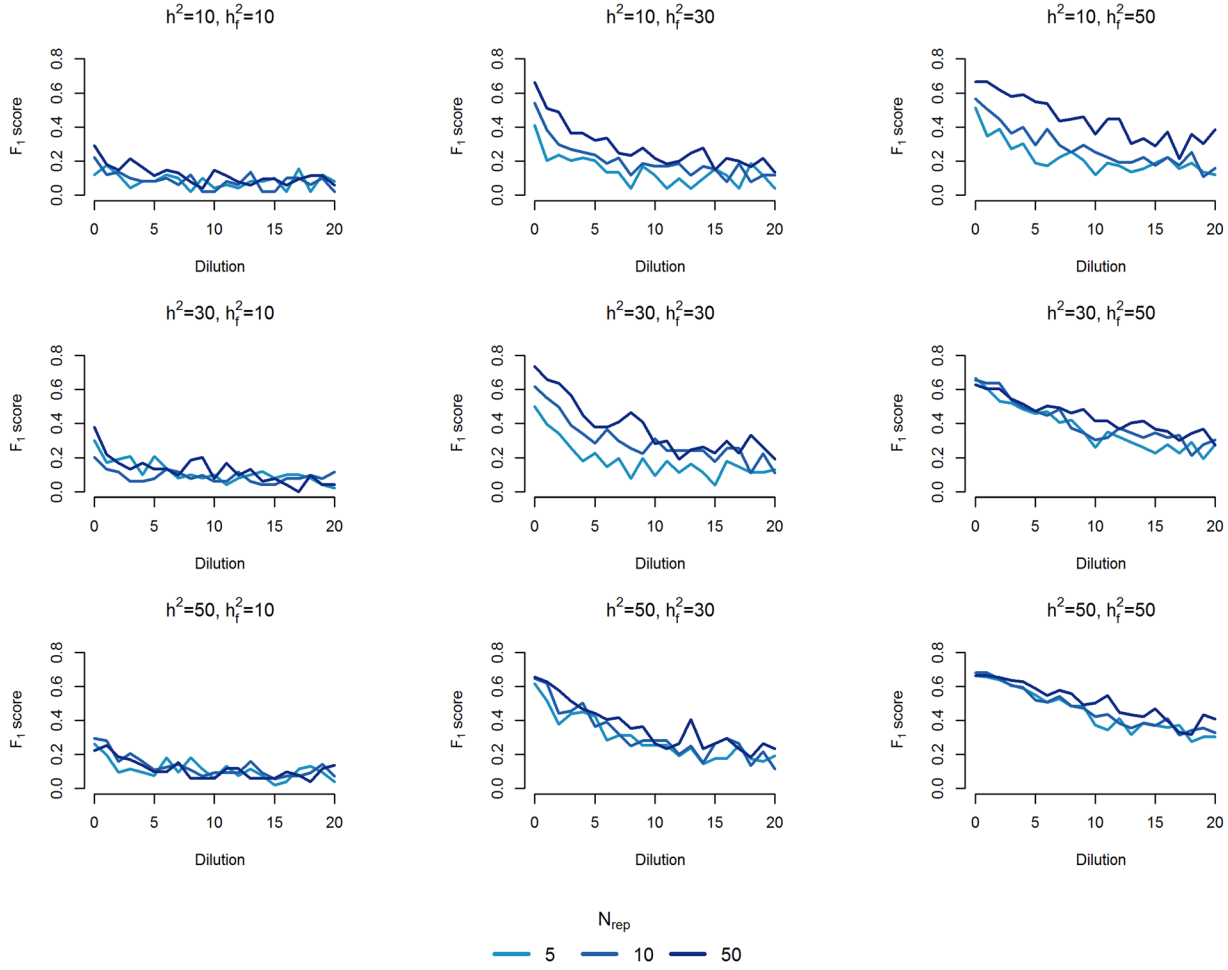

Figure 4. Influence of genomic feature, trait specific factors and dilution on detection power. In each row the heritability (h^2) is kept constant while the proportion of genomic variance explained by the feature increases ($h^2_f = 0.1, 0.2, 0.3, 0.5$). Moving down each column h^2 increases from 0.1 to 0.2 and 0.5 while h^2_f is kept constant. The power to detect features enriched for causal variants was quantified by the F_1 score shown on the y-axis of each panel. P-value cut-off for the set test statistic was 0.05. F_1 score is shown as a function of dilution, i.e. adding up to 2000 non-causal SNPs to the feature, on the x-axis. The number of replicates ($N_{rep} = 5$, 10 and 50) within line is depicted by the colour scale.

part of all but one of the significant GO terms (Fig. 7). In addition, we found evidence that several other genes including *antennapedia* (FBgn0260642), *ultrabithorax* (FBgn0003944), and *extra macrochaetae* (FBgn0000575) contributed to the significance of the GO term "Midgut development" (GO:0007494), and the genes *mago nashi* (FBgn0002736) and *roughoid* (FBgn0003295) contributed to the significance of the GO term "Epidermal growth factor receptor signaling pathway" (GO:0007173). Finally, we found that the genes *Chd3* (FBgn0023395) and *helicase 89B* (FBgn0022787) contributed to the significance of the GO term "ATP-dependent DNA helicase activity" (GO:0004003).

Discussion

We demonstrated that GBLUP-derived set tests are powerful for detecting genomic features enriched for causal variants affecting a quantitative trait in populations with a low degree of linkage disequilibrium. The different set tests were compared using simulated data generated from DGRP genotypes further illustrating the impact of trait- and genomic feature-specific factors on detection power. These set tests 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. The methods are computationally fast allowing us to rapidly analyze many different classes of genomic features. This will help to discover genomic features enriched for causal variants that can be used to develop more accurate predictions using GFBLUP models. GBLUP-derived set tests are based on a flexible linear mixed modelling framework that allows us to adjust for other known genetic and non-genetic factors, while using existing standard software. Importantly, the GBLUP models can be extended in several ways that potentially can increase detection power.

GO id[a]	Empirical p-values[b,c,d]				Ontology[e]	Gene Ontology term
	Female	Male	Male/Female covariance	Female/Male covariance		
GO:0007266	$<1\times10^{-4}$	$<1\times10^{-4}$	$<1\times10^{-4}$	$<1\times10^{-4}$	BP	Rho protein signal transduction
GO:0035160	$<1\times10^{-4}$	$<1\times10^{-4}$	$<1\times10^{-4}$	$<1\times10^{-4}$	BP	Maintenance of epithelial integrity, open tracheal system
GO:0005100	$<1\times10^{-4}$	$<1\times10^{-4}$	$<1\times10^{-4}$	$<1\times10^{-4}$	MF	Rho GTPase activator activity
GO:0016323	2×10^{-4}	3×10^{-4}	1×10^{-4}	4×10^{-4}	CC	Basolateral plasma membrane
GO:0007173	3×10^{-4}	7×10^{-4}	7×10^{-4}	6×10^{-4}	BP	Epidermal growth factor receptor signaling pathway
GO:0035277	3×10^{-4}	$<1\times10^{-4}$	$<1\times10^{-4}$	3×10^{-4}	BP	Spiracle morphogenesis, open tracheal system
GO:0008289	5×10^{-4}	4×10^{-4}	3×10^{-4}	4×10^{-4}	MF	Lipid binding
GO:0007494	1×10^{-3}	6×10^{-4}	4×10^{-4}	2×10^{-4}	BP	Midgut development
GO:0004003	1.5×10^{-3}	2×10^{-4}	5×10^{-4}	8×10^{-4}	MF	ATP-dependent DNA helicase activity

Table 1. Gene ontology terms significantly associated with CCRT for males and females. [a]GO id = gene ontology id. [b]p-values smaller than or equal to 0.001 were included for each sex. [c]The empirical distributions were obtained by randomly sampling from gene regions on the genome. [d]10,000 permutations were performed for each GO term. For p-values less than 1×10^{-4}, more permutations would yield more precise p-values. [e]BP = Biological Process, MF = Molecular Function and CC = Cellular Component.

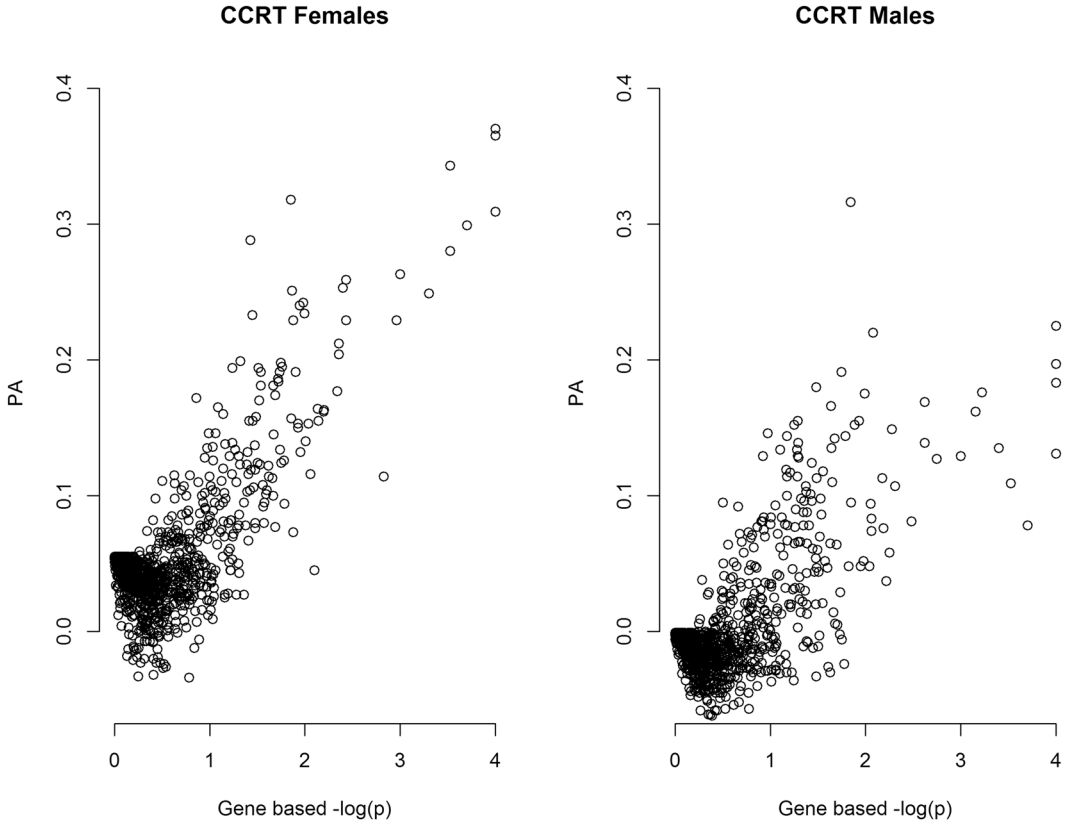

Figure 5. Relationship between gene ontology (GO) term CVAT test statistic and predictive ability of the GFBLUP model for chill coma recovery time (CCRT). The significance of GO terms related to CCRT in *Drosophila melanogaster* in females and males as determined by the CVAT test statistic (expressed as gene based −log(p), x-axis) from single trait analyses, plotted against the predictive ability (PA) of the single trait GFBLUP model (y-axis).

Comparison of set tests. Several GBLUP-derived set tests were compared in terms of statistical power to detect genomic features enriched for causal variants. Despite GBLUP being considered a "black box" modeling approach we showed that it is possible to derive powerful set tests from it. In particular, in all scenarios evaluated we showed that the covariance association test (CVAT) had similar power to a commonly used score based approach[28] (also known as the sequence kernel association test, SKAT), and that both CVAT and SKAT

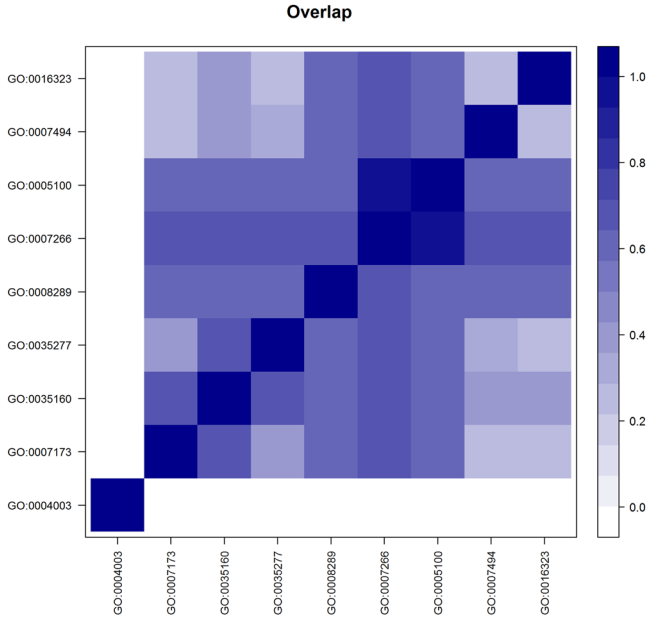

Figure 6. Heatmap showing the overlap between SNPs of significant GO terms. Each square [*i*, *j*] shows the proportion of SNPs associated with GO term *i*, as well as GO term *j*. Where *i* indexes rows and *j* indexes columns. Darker colours represent larger proportions of SNPs that overlap between GO terms. Only the most significant GO terms, presented in Table 1, are included.

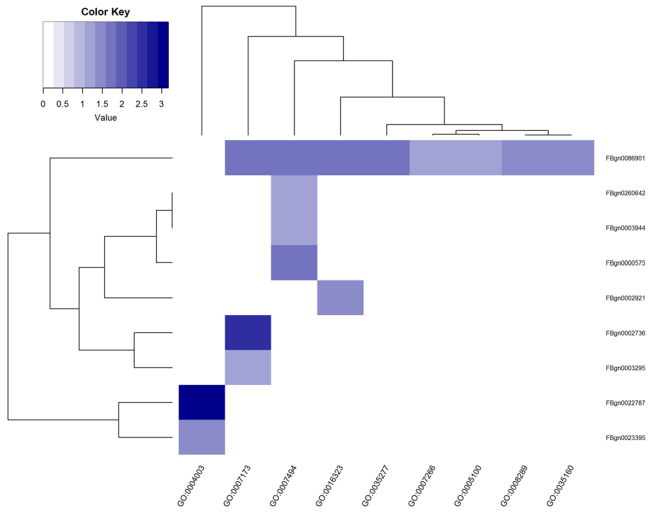

Figure 7. Heatmap showing the individual genes associated to CCRT for each of the top GO terms. The colour scale indicates the degree of association (expressed as $-\log(p)$). The darkest blue colour indicates $-\log(p) = 4$ and as the colour fades p-values increase.

outperformed the methods based on summing the number of single marker statistics in the feature. Both CVAT and SKAT are fast and powerful methods to identify genomic features enriched for causal variants and thereby contribute to develop more accurate prediction models. One advantage of the CVAT approach is that it builds on a flexible linear mixed modelling framework that can be extended in several ways that potentially can increase detection power. Extensions include the consideration of different levels of a hierarchical feature, multiple genetic components having different genomic value distributions and a multiple trait GBLUP.

Set tests based on counting test statistics (T_{Count}) appear to have lower power compared to test statistics based on summing the squared single marker statistics (T_{Sum}). This may in part be explained by the simulated genetic architecture which were enriched for causal variants with small to moderate effects. In general, methods based on a count test statistic are likely to have high power to detect association if the genomic feature harbours genetic markers with large effects, but it will not detect a genomic feature with many genetic markers having small to moderate effects[45]. Our results show that in such cases, it is more powerful to use a test statistic, such as the mean or sum of the single marker statistics for the genomic feature.

Finally, we have shown a clear link between the significance levels of the set test statistics and the level of predictive ability using these sets as features in the GFBLUP model. This link could be exploited to build more accurate GFBLUP models in a computationally efficient way. That is, using the GBLUP model to identify genomic features enriched for associated variants and subsequently apply the identified sets as features in the GFBLUP model.

Factors influencing detection power. Several trait and feature specific factors can influence the power to determine whether a genomic feature is enriched for causal variants. Power is positively correlated with the proportion of genomic variance explained by the genomic feature, and power decreases with the addition of non-causal SNPs in the feature (dilution). Furthermore, the genetic architecture of the causal variants (distributed randomly or clustered along the genome) also influenced power. The increased detection power and resistance towards dilution in the case where the true causal SNPs are clustered in smaller genomic regions is likely due to larger effect size of individual markers in these regions. Not surprisingly, power is increased if the trait is highly heritable and the number of phenotypic records available is high. These patterns were consistent across the different set tests and are factors that need to be considered in the analyses of real data.

Influence of linkage disequilibrium on detection power. We compared the GBLUP derived set tests based on genotypes obtained from the sequenced inbred lines of the *Drosophila melanogaster* Genetic Reference Panel. The population consist of 205 largely unrelated lines with a low degree of linkage disequlibrium across their genomes. Thus, our results suggest that GBLUP-derived set tests may have high power in situations where individuals are largely unrelated such as human study populations. In a population of highly related individuals the general genomic relationship will be a good approximation of the genomic relationship at the true causal variants[2]. This will lead to more accurate estimates of overall genomic value. On the other hand due to extensive linkage disequilibrium it may be difficult to accurately estimate single marker effects and this will in turn influence the feature set test statistic. Therefore more research is required to understand the influence of genetic relatedness and degree of linkage disequilibrium on detection power of the GBLUP derived set tests.

Influence of null hypothesis on detection power. In this study we compared the set tests using a competitive null hypothesis. The competitive null hypothesis states that the degree of association within a genomic feature is equal to that of a random set of genetic markers. An alternative is the *self-contained* null hypothesis[46–48]. The self-contained null hypothesis states that the genomic feature, by it self, does not display any association to the phenotypic trait. This is usually done by testing whether the variance component or the test statistics for the genomic feature are zero. The self-contained may be preferable over a competitive, as it has more power in general[46], and the interpretation is simpler, as it determines whether there is association or not. On the other hand the competitive null hypothesis is perhaps more biologically relevant as it is in agreement with the infinitesimal model[49,50] which is a commonly used genetic model assuming that there are many causal variants each with small to moderate effects underlying the complex trait.

Further extensions of GBLUP-derived set tests and alternative methods. The GBLUP-derived set test modeling framework can be extended in several ways that potentially can increase detection power. First, multiple feature sets can be fitted in the model (e.g. a GFBLUP model), such as grouping markers based on their minor allele frequency[19,20] or prior QTL information[16]. By fitting multiple feature sets genetic effects are estimated based on a mixture of normal distributions enabling further differential shrinkage of single marker effects across feature sets. Second, further shrinkage of single marker effects within features may be achieved by using a weighted genomic relationship matrix[11,51] for each feature set. Third, a multiple trait GBLUP model[21,22] can be fitted. This can increase the accuracy of the overall genomic effect[21,22] and thereby the single marker effect which in turn will lead to a more accurate test statistic for the genetic marker set. Fourth, in animal and plant populations with extended pedigrees we might use information on inviduals without genotype information[51] to increase accuracy of the overall genomic value. We are currently investigating these extensions hypothesizing that they, in some situations, may lead to increased power of the GBLUP-derived set test.

Comparison to Bayesian methods. The GBLUP and GFBLUP models used in this study can also be implemented using Bayesian methods[52–55]. In particular Bayesian mixture models such as BayesB[56], BayesR[56,57] or Bayesian Lasso models[58] are relevant alternative methods. For these methods it is also possible to derive test statistics that quantify the joint effect of the markers in the feature set. Furthermore, they also allow for differential shrinkage of marker effects within feature sets and can be used to fit multiple feature sets. More investigations are required to compare these methods to the GBLUP-derived set test and investigate to what extend these methods will increase detection power.

Application of CVAT on CCRT. Although the main objective of this paper was to compare different GBLUP derived set tests, we would like to discuss, albeit very condensed, the plausible biological relevance of our results.

CCRT was strongly associated with the GO terms 'Rho protein signal transduction' (GO:0007266) and 'Rho GTPase activator activity' (GO:0005100) in both males and females. Rho genes' functional relevance, with regards to CCRT, has been implied by their involvement in (a) intracellular signal transduction pathways[59], (b) indirectly mediating circadian rhythm, through actin regulation[60] as well as (c) contributing to ion homeostasis by regulating K^+ channel cell surface expression[61].

"Midgut development" (GO:0007494) was also among the high-ranking GO terms for association with CCRT. It is well established that the midgut of insects is an important site for the exchange of ions with the hemolymph[62,63]. Insect cold resistance is directly related to maintenance of water and ion homeostasis[64–66]. In the fall field cricket,

Gryllus pennsylvanicus, the midgut has been shown to be the most sensitive site for the exchange of ions and water during cold exposure[67]. In the midgut cold exposure caused rising Na^+ levels causing a disruption in water homeostasis ultimately leading to an increased K^+concentration in the hemolymph[65]. It is ultimately this increased K^+ concentration that causes an electrophysiological failure of the neuromuscular system and subsequent chill-coma[68–70].

There was a substantial overlap among the SNPs associated with each of the top-ranking GO terms. In order to zoom in on relevant genes underlying these GO terms, the CVAT set test was also applied at the individual gene level of the significant GO terms. This enabled us to identify a number of the genes that was associated to CCRT. In particular, we found that the *crossveinless-c* gene (FBgn0086901), was highly significant and is a part of both Rho protein signal transduction and Rho GTPase activator activity GO terms[42]. *Crossveinless-c* is an important regulator of Rho GTPase activity[71]. The Rho-family of GTPases are in turn associated with the direct regulation of the actin cytoskeleton[72]. Chilling has been shown to disrupt cytoskeletal organization in primary embryonic cultures of *Drosophila* cells[73]. Interestingly, diapausing mosquitos (*Culex pipens*) have greater abundance of polymerized actin at muscle fiber intersections in the midgut[74]. Thus, regulation of cytoskeletal function may be implicated as an important component of cold acclimation.

In general, 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[75].

Conclusion

GBLUP-derived set tests are powerful compared to existing methods for detecting genomic features enriched for causal variants in populations with a low degree of linkage disequilibrium. The tests can be implemented using standard BLUP models, and can be extended in several ways that potentially can increase detection power. The methods are computationally fast allowing us to rapidly analyze many different classes of genomic features. This will help to discover genomic features enriched for causal variants that can be used to develop more accurate predictions using GFBLUP models.

References

1. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
2. de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C. & Sorensen, D. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLOS Genetics* **9**, e1003608, doi:10.1371/journal.pgen.1003608 (2013).
3. Lango Allen, H. *et al*. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838, doi:10.1038/nature09410 (2010).
4. O'Roak, B. J. *et al*. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250, doi:10.1038/nature10989 (2012).
5. Lage, K. *et al*. 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, doi:10.1073/pnas.1210730109 (2012).
6. Maurano, M. T. *et al*. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195, doi:10.1126/science.1222794 (2012).
7. Peñagaricano, F., Weigel, K. A., Rosa, G. J. M. & Khatib, H. Inferring quantitative trait pathways associated with bull fertility from a genome-wide association study. *Front. Genet.* **3**, doi:10.3389/fgene.2012.00307 (2013).
8. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat. Rev.* **11**, 843–854, doi:10.1038/nrg2884 (2010).
9. Listgarten, J. *et al*. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* **29**, 1526–1533, doi:10.1093/bioinformatics/btt177 (2013).
10. Sørensen, I. F. *et al*. Pharmacogenetic effects of "candidate gene complexes" on stroke in the GenHAT study. *Pharmacogenet Genomics* **24**, 556–563, doi:10.1097/FPC.0000000000000088 (2014).
11. Sørensen, P., Edwards, S. M. & Jensen, P. Genomic feature models. In *10th World Congress of Genetics Applied to Livestock Production* (2014).
12. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557, doi:10.1101/gr.169375.113 (2014).
13. Gusev, A. *et al*. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552, doi:10.1016/j.ajhg.2014.10.004 (2014).
14. Edwards, S. M., Thomsen, B., Madsen, P. & Sørensen, P. Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. *Genet. Sel. Evol.* **47**, 60, doi:10.1186/s12711-015-0132-6 (2015).
15. Edwards, S. M., Sørensen, I. F., Sarup, P., Mackay, T. F. C. & Sørensen, P. Genomic prediction for quantitative traits is improved by mapping variants to Gene Ontology categories in *Drosophila melanogaster*. *Genetics* **203**, 1871–1883, doi:10.1534/genetics.116.187161 (2016).
16. Sarup, P., Jensen, J., Ostersen, T., Henryon, M. & Sørensen, P. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet.* **17**, 11, doi:10.1186/s12863-015-0322-9 (2016).
17. Mackay, T. F. C. *et al*. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**, 173–178, doi:10.1038/nature10811 (2012).
18. Huang, W. *et al*. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* **24**, 1193–1208, doi:10.1101/gr.171546.113 (2014).
19. Rohde, P. D. *et al*. Covariance association test (CVAT) identify genetic markers associated with schizophrenia in functionally associated biological processes. *Genetics* **203**, 1901–1913, doi:10.1534/genetics.116.189498 (2016).
20. Loh, P.-R. *et al*. Contrasting regional architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
21. Calus, M. P. & Veerkamp, R. F. Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* **43**, 26, doi:10.1186/1297-9686-43-26 (2011).
22. Maier, R. *et al*. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.* **96**, 283–294, doi:10.1016/j.ajhg.2014.12.006 (2015).
23. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423, doi:10.3168/jds.2007-0980 (2008).
24. Henderson, C. R. Sire evaluation and genetic trends. *J. Anim. Sci.* 10–14 (1973).

25. Robinson, G. K. That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* **6**, 15–32, doi:10.1214/ss/1177011926 (1991).
26. Rao, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cambridge Philos. Soc.* **44**, 50–57, doi:10.1017/S0305004100023987 (1948).
27. Goeman, J. J., van de Geer, S. A., de Kort, F. & van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93–99, doi:10.1093/bioinformatics/btg382 (2004).
28. Wu, M. C. *et al*. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93, doi:10.1016/j.ajhg.2011.05.029 (2011).
29. Wang, X., Morris, N. J., Zhu, X. & Elston, R. C. A variance component based multi-marker association test using family and unrelated data. *BMC Genet.* **14**, 17, doi:10.1186/1471-2156-14-17 (2013).
30. Huang, Y.-T. & Lin, X. Gene set analysis using variance component tests. *BMC Bioinformatics* **14**, 210, doi:10.1186/1471-2105-14-210 (2013).
31. Rivals, I., Personnaz, L., Taing, L. & Potier, M.-C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**, 401–407, doi:10.1093/bioinformatics/btl633 (2007).
32. Madsen, P., Jensen, J. & Thompson, R. Estimation of (co)variance components by REML in multivariate mixed linear models using average of observed and expected information. In *5th WCGALP* 455–462 (1994).
33. Johnson, D. L. & Thompson, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* **78**, 449–456, doi:10.3168/jds.S0022-0302(95)76654-1 (1995).
34. Goeman, J. J., Van De Geer, S. A. & Van Houwelingen, H. C. Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 477–493, doi:10.1111/rssb.2006.68.issue-3 (2006).
35. Cabrera, C. P. *et al*. Uncovering networks from genome-wide association studies via circular genomic permutation. *G3* **2**, 1067–1075, doi:10.1534/g3.112.002618 (2012).
36. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82, doi:10.1016/j.ajhg.2010.11.011 (2011).
37. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research* **24**(9), 1550–1557, doi:10.1101/gr.169375.113 (2014).
38. Canela-Xandri, O., Law, A., Gray, A., Woolliams, J. A. & Tenesa, A. A new tool called DISSECT for analysing large genomic data sets using a big data approach. *Nat. Commun.* **6**, 10162, doi:10.1038/ncomms10162 (2015).
39. Lee, S. H. & van der Werf, J. H. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* **32**, 1420–1422, doi:10.1093/bioinformatics/btw012 (2016).
40. Powers, D. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2**, 37–63 (2011).
41. Morgan, T. J. & Mackay, T. F. C. Quantitative trait loci for thermotolerance phenotypes in *Drosophila melanogaster*. *Heredity* **96**, 232–242, doi:10.1038/sj.hdy.6800786 (2006).
42. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29, doi:10.1038/75556 (2000).
43. Carlson, M. *org.DM.eg.db*: *Genome wide annotation for Fly*. (2013).
44. Tweedie, S. *et al*. FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.* **37**, D555–559, doi:10.1093/nar/gkn788 (2009).
45. Newton, M. A., Quintana, F. A., den Boon, J. A., Sengupta, S. & Ahlquist, P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.* **1**, 85–106, doi:10.1214/07-AOAS104 (2007).
46. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987, doi:10.1093/bioinformatics/btm051 (2007).
47. Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.* **15**, 504–518, doi:10.1093/bib/bbt002 (2014).
48. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364, doi:10.1038/nrg.2016.29 (2016).
49. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Philos. Trans. R. Soc. Edinb.* **52**, 399–433, doi:10.1017/S0080456800012163 (1918).
50. Bulmer, M. G. The effect of selection on genetic variability. *Am. Nat.* **105**, 201–211, doi:10.1086/282718 (1971).
51. Wang, H., Misztal, I., Aguilar, I., Legarra, A. & Muir, W. M. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res.* **94**, 73–83, doi:10.1017/S0016672312000274 (2012).
52. Ehsani, A., Sørensen, P., Pomp, D., Allan, M. & Janss, L. Inferring genetic architecture of complex traits using Bayesian integrative analysis of genome and transcriptome data. *BMC Genomics* **13**, 456, doi:10.1186/1471-2164-13-456 (2012).
53. Sørensen, P., de Los Campos, G., Morgante, F., Mackay, T. F. C. & Sorensen, D. Genetic control of environmental variation of two quantitative traits of *Drosophila melanogaster* revealed by whole-genome sequencing. *Genetics* **201**, 487–497, doi:10.1534/genetics.115.180273 (2015).
54. Loh, P.-R. *et al*. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290, doi:10.1038/ng.3190 (2015).
55. Ehsani, A., Janss, L., Pomp, D. & Sørensen, P. Decomposing genomic variance using information from GWA, GWE and eQTL analysis. *Anim. Genet.* **47**, 165–173, doi:10.1111/age.2016.47.issue-2 (2016).
56. Meuwissen, T. H., Solberg, T. R., Shepherd, R. & Woolliams, J. A. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol. GSE* **41**, 2, doi:10.1186/1297-9686-41-2 (2009).
57. Erbe, M. *et al*. 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, doi:10.3168/jds.2011-5019 (2012).
58. Park, T. & Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **103**, 681–686, doi:10.1198/016214508000000337 (2008).
59. Tcherkezian, J. & Lamarche-Vane, N. Current knowledge of the large RhoGAP family of proteins. *Biol. Cell* **99**, 67–86, doi:10.1042/BC20060086 (2007).
60. Rao, N. V. Role of the RHO1 GTPase signaling pathway in regulating the circadian clock. In *Drosophila melanogaster*. Doctoral Dissertation. (University of Virginia, 2013).
61. Stirling, L., Williams, M. R. & Morielli, A. D. Dual roles for RHOA/RHO-kinase in the regulated trafficking of a voltage-sensitive potassium channel. *Mol. Biol. Cell* **20**, 2991–3002, doi:10.1091/mbc.E08-10-1074 (2009).
62. Dow, J. A. T. Insect midgut function. In *Advances in Insect Physiology* (ed. Evans, P. D. and Wigglesworth, V. B.) 187–328 (Academic Press, 1987).
63. Zeiske, W. Insect ion homeostasis. *J. Exp. Biol.* **172**, 323–334 (1992).
64. Hochachka, P. W. Defense strategies against hypoxia and hypothermia. *Science* **231**, 234–241, doi:10.1126/science.2417316 (1986).
65. Koštál, V., Vambera, J. & Bastl, J. On the nature of pre-freeze mortality in insects: water balance, ion homeostasis and energy charge in the adults of *Pyrrhocoris apterus*. *J. Exp. Biol.* **207**, 1509–1521, doi:10.1242/jeb.00923 (2004).
66. Koštál, V., Yanagimoto, M. & Bastl, J. Chilling-injury and disturbance of ion homeostasis in the coxal muscle of the tropical cockroach (*Nauphoeta cinerea*). *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **143**, 171–179, doi:10.1016/j.cbpb.2005.11.005 (2006).
67. MacMillan, H. A. & Sinclair, B. J. The role of the gut in insect chilling injury: cold-induced disruption of osmoregulation in the fall field cricket, *Gryllus pennsylvanicus*. *J. Exp. Biol.* **214**, 726–734, doi:10.1242/jeb.051540 (2011).

68. Wareham, A. C., Duncan, C. J. & Bowler, K. The resting potential of cockroach muscle membrane. *Comp. Biochem. Physiol. A Physiol.* **48**, 765–797, doi:10.1016/0300-9629(74)90619-7 (1974).
69. Hosler, J. S., Burns, J. E. & Esch, H. E. Flight muscle resting potential and species-specific differences in chill-coma. *J. Insect Physiol.* **46**, 621–627, doi:10.1016/S0022-1910(99)00148-1 (2000).
70. Findsen, A., Andersen, J. L., Calderon, S. & Overgaard, J. Rapid cold hardening improves recovery of ion homeostasis and chill coma recovery time in the migratory locust, *Locusta migratoria*. *J. Exp. Biol.* **216**, 1630–1637, doi:10.1242/jeb.081141 (2013).
71. Denholm, B. *et al*. *crossveinless-c* is a RhoGAP required for actin reorganisation during morphogenesis. *Development* **132**, 2389–2400, doi:10.1242/dev.01829 (2005).
72. Hall, A. Rho GTPases and the actin cytoskeleton. *Science* **279**, 509–514, doi:10.1126/science.279.5350.509 (1998).
73. Cottam, D. M. *et al*. Non-centrosomal microtubule-organising centres in cold-treated cultured *Drosophila* cells. *Cell Motil. Cytoskeleton* **63**, 88–100, doi:10.1002/cm.20103 (2006).
74. Kim, M., Robich, R. M., Rinehart, J. P. & Denlinger, D. L. Upregulation of two actin genes and redistribution of actin during diapause and cold stress in the northern house mosquito, *Culex pipiens*. *J. Insect Physiol.* **52**, 1226–1233, doi:10.1016/j.jinsphys.2006.09.007 (2006).
75. Skarman, A., Shariati, M., Jans, L., Jiang, L. & Sørensen, P. A Bayesian variable selection procedure to rank overlapping gene sets. *BMC Bioinformatics* **13**, 73, doi:10.1186/1471-2105-13-73 (2012).

Acknowledgements

This study was in part funded by the Danish Strategic Research Council (GenSAP: Centre for Genomic Selection in Animals and Plants, contract no. 12-132452) to P.S. Furthermore funded by the MRC to S.M.E. (FARSPhase: a Flexible, widely Applicable, Robust, and Scalable phasing algorithm for human genetics, grant no. MR/M000370/1), and to P.D.R. by the Lundbeck Foundation (grant number R155-2014-1724).

Author Contributions

I.F.S. analysed the data, evaluated the results, and drafted the manuscript. S.M.E. and P.D.R. evaluated the results and contributed to the manuscript. P.S. conceived the study, designed the experiments, analysed the data, evaluated the results, and drafted the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-02281-3

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.