Covariance Association Test (CVAT) Identify Genetic Markers Associated with Schizophrenia in Functionally Associated Biological Processes

Palle Duun Rohde [1,2,3], Ditte Demontis [2,3,4], Beatriz Castro Dias Cuyabano [1], The GEMS Group ^, Anders D. Børglum [2,3,4] and Peter Sørensen [1]

[1] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark; [2] Centre for Integrative Sequencing, iSEQ, Aarhus University, Denmark; [3] The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark; [4] Department of Biomedicine, Aarhus University, Denmark; ^ The GEMS (Genomic Medicine for Schizophrenia) Group: Ole Mors, David M. Hougaard, and Preben Bo Mortensen

palle.d.rohde@mbg.au.dk

PROBLEM: IDENTIFYING THE GENETIC CAUSALITY OF COMPLEX DISEASES

WHY Schizophrenia is a disorder with large personal and social costs. Knowledge of the genetic etiology is essential for improving diagnosis and treatment.

HOW Individual genetic markers are tested for association with the disease status (Box 1).

CHALLENGE Single marker association have limited power to detect variants with small effect (Figure 1).

SOLUTION Aggregating genetic markers based on biological information might increase the power to identify causal sets of markers (Box 1).

AIM We propose a new set test, the Covariance Association Test (CVAT), and show how it can be extended to utilize a mixture distribution of marker effects (Figure 2).

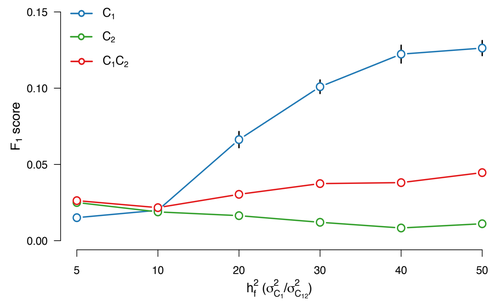


Figure 1 Performance of single marker association (F_1score, see box 2) on simulated data (see Box 2). For each scenario we computed the F_1 score and show here the average F_1 score (with standard errors) within scenario. The F_1 score was computed for the main genetic variants (C_1, blue), the genetic variants with small effects (C_2, green), and for all 1,000 causal variants (C_1C_2, red).

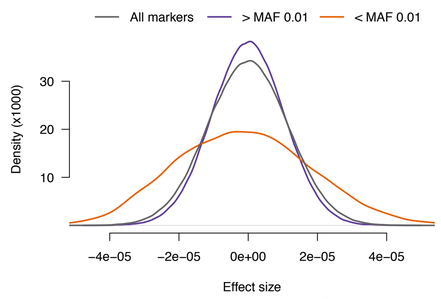


Figure 2 Distribution of marker effects, $\hat{s}$ (observed data, Box 2). The grey line is the distribution of $\hat{s}$ when the minor allele frequencies (MAFs) not were accounted for. Purple and orange lines are the distributions of $\hat{s}$ when they are estimated based on differences in MAFs.

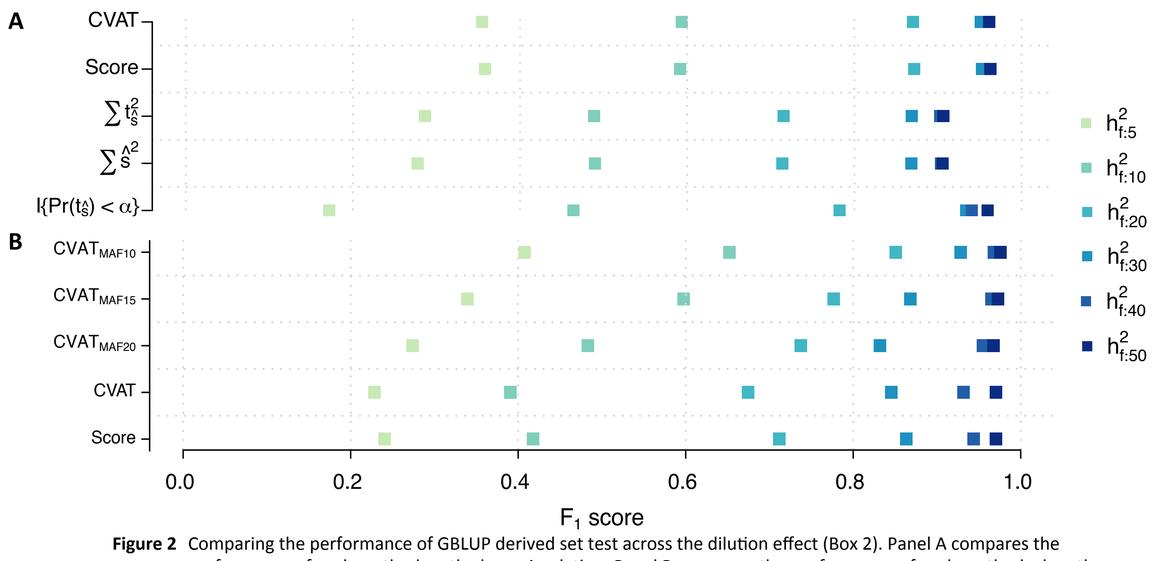


Figure 2 Comparing the performance of GBLUP derived set test across the dilution effect (Box 2). Panel A compares the performance of each method on the base simulation. Panel B compares the performance of each method when the data were enriched for rare, causal variants

COVARIANCE ASSOCIATION TEST (CVAT)

CVAT are derived from Genomic Best Linear Unbiased Prediction, GBLUP: $y = Xb + Zg + e$
The genetic effects, $\hat{g}$, are obtained as: $\hat{g} = G\sigma_g^2\hat{V}^{-1}(y - X\hat{b})$

CVAT captures the covariance between the total genetic effect from all markers ($\hat{g}$), and the genetic effect from the markers within the group ($\hat{g}_f$), and the summary statistic is computed as:

$$CVAT = \hat{g}'\hat{g}_f = (\hat{g}_r' + \hat{g}_f')\hat{g}_f = \hat{g}_f'\hat{g}_f + \hat{g}_r'\hat{g}_f; \quad \hat{g} = \sum_{i=1}^{m} w_i\hat{s}_i, \quad \hat{g}_f = \sum_{i=1}^{m_f} w_i\hat{s}_i;$$

where w_i is the *i*-th marker in the centered and scaled genotype matrix, and $\hat{s}_i$ is the marker effect of the *i*-th marker computed as: $\hat{s} = W'(WW')^{-1}\hat{g}$

EXTENDING CVAT

The predicted genetic effects are assumed to be drawn from the same distribution; $g \sim N(0, G\sigma_g^2)$
However, it is likely that the genetic effects comes from a mixture of distributions, e.g., caused by differences in minor allele frequencies. This can be modeled as:

$$\hat{s} = \begin{bmatrix} \hat{s}_1 \\ \hat{s}_2 \end{bmatrix} = \begin{bmatrix} w_1'(w_1w_1')^{-1}\hat{g}_1 \\ w_2'(w_2w_2')^{-1}\hat{g}_2 \end{bmatrix}; \; y = Xb + Zg_1 + Zg_2 + e; \quad g_1 \sim N(0, G_1\sigma_{g_1}^2); \quad g_2 \sim N(0, G_2\sigma_{g_2}^2).$$

COMPARABLE SET TESTS

Many types of set tests exists and different ways to compute the summary statistic exists. Here, we compare the performance of CVAT to other set tests derived from GBLUP.

$$T_{count} = \sum_{i=1}^{m_f} \mathrm{I}(t_i < t_0) \qquad T_{sum} = \sum_{i=1}^{m_f} t_i^2 \qquad T_{score} = 0.5\frac{1}{m_f}\tilde{y}W_fW_f'\tilde{y}$$

TESTING THE NULL HYPOTHESIS

We applied a permutation approach to test the null hypothesis that the degree of association between a certain set of genetic markers and a phenotype is equal to the degree of association to a random complement set of markers. To test this, the observed summary statistic was compared to an empirical distribution of summary statistics.

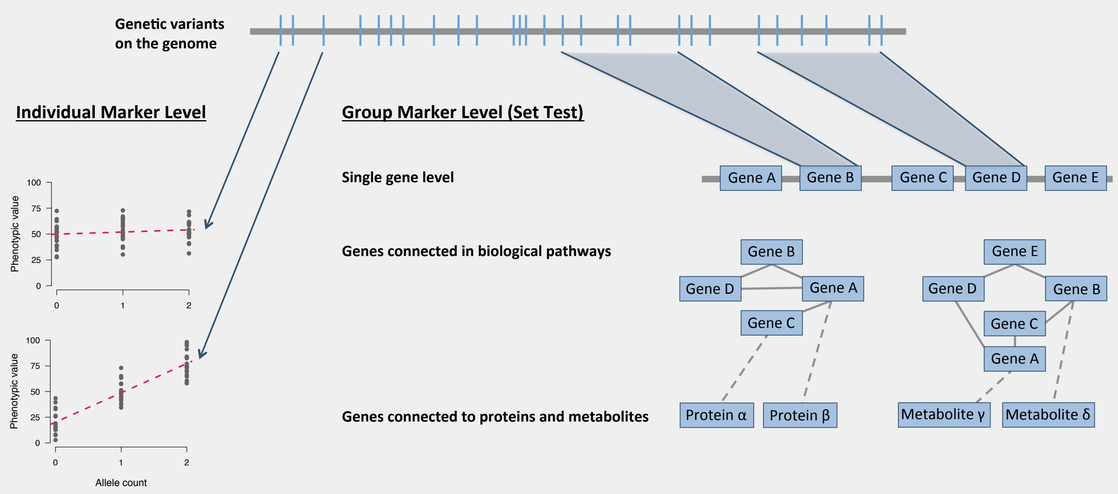
CONCLUSION

CVAT performs as equal as comparable methods. Extending CVAT to utilize a mixture distribution of marker effects, the power of CVAT increased significantly compared to the other methods.

Applying the methods to the observed schizophrenia data we identified several pathways, such as vitamin A metabolism and an immunological signature, which previously have been implicated with schizophrenia based on experimental and observational studies.

BOX 1: ASSOCIATING GENETIC MARKERS WITH COMPLEX DISEASES

Associating genetic polymorphic markers with phenotypic variation can be accomplished on individual genetic marker level (left part below), or grouping the genetic markers accordingly to biological information, such as genes, pathways or expression patterns, namely, set tests (right part below).



BOX 2: OBSERVED AND SIMULATED GENO- AND PHENOTYPE DATA

Observed Data Schizophrenia case-control data (*Børglum et al., 2014, Mol. Psychiatry*) with 882 controls and 888 cases (all from DK), genotyped for 520,897 genetic polymorphic markers distributed on 22 autosomes. We estimated the proportion of phenotypic variation captured by the common genetic markers (on the liability scale) to $h_l^2 = 0.16$.

Simulated Data We simulated 50 independent data sets consisting of 2000 individuals (case-control ratio=0.5). Cases were sampled from a population with a disease prevalence corresponding to that of schizophrenia (K=0.01). 80,000 genotypes on 22 autosomes (with same correlation structure as the human population) were simulated. Phenotypes were simulated based on 1,000 causal markers; 100 main effects (C_1) sampled from 'gene regions', and 900 minor effects (C_2) randomly sampled. The genetic variance explained by C_1 ranged from 0.05 to 0.5: $h_f^2 = \sigma_{C_1}^2 / \left(\sigma_{C_1}^2 + \sigma_{C_2}^2\right)$.The total phenotypic variation explained by genetic variation was set to that of schizophrenia. To investigate the effect of 'diluting' the signal of causal markers by non-causal markers we created sets containing the causal markers and an increasingly number of non-causal SNPs (0.1k-5k). In addition, we created sets containing no causal genetic variants. The power to detect causal sets was described using the F_1 classification score:

$$F_1 = 2\frac{p \cdot r}{p + r}; \quad p = TP/(TP + FP); \quad r = TP/(TP + FN).$$

AARHUS UNIVERSITY · iSEQ · iPSYCH · GenSAP · The Danish Council for Strategic Research