Implementation of a weighted burden test using logistic regression to allow integrated analysis of rare sequence variants, copy number variants and polygenic risk score

David Curtis

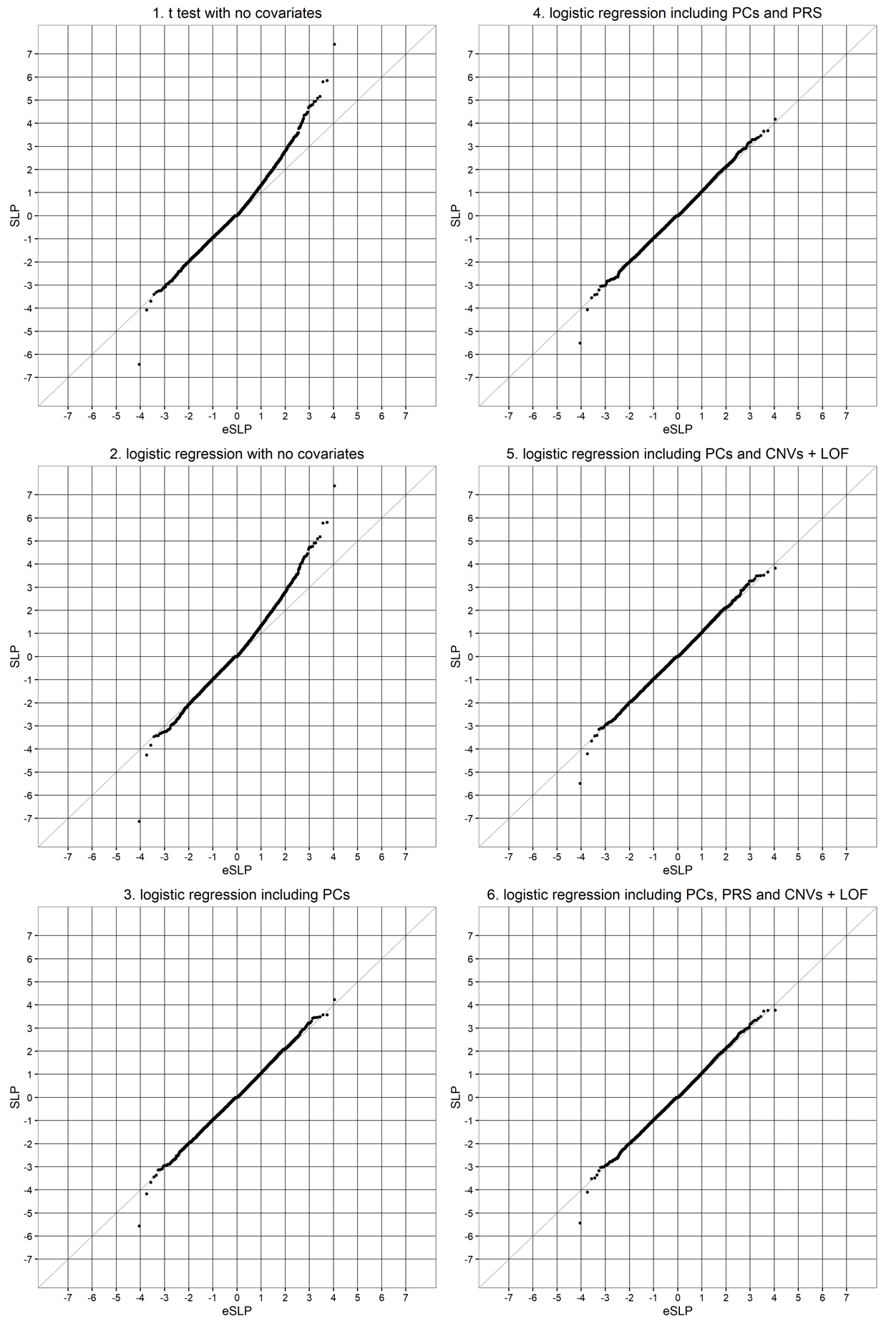
UCL Genetics Institute

Background

Many illnesses have different kinds of genetic variation which contribute to risk. In the case of schizophrenia, common variants, including genome-wide significant SNPs, have individually small effects but can be combined into a polygenic risk score (PRS) with substantial effect size. Some very rare copy number variants have very substantial effects. A small number of loss of function (LOF) variants of individual genes have been implicated and there is statistical evidence that very rare sequence variants in particular sets of genes influence risk. A weighted burden test which uses a t test to compare average gene-wise risk scores obtained from multiple sequence variants within a gene, as previously implemented in the SCOREASSOC program, is not able to incorporate other contributors to risk.

Method

SCOREASSOC was modified to carry out logistic regression with a ridge penalty function to test for a contribution from sequence variants while incorporating as covariates the PRS, the presence of pathogenic CNVs and population principal components (PCs). The main variable of interest is the gene-wise risk score consisting of the weighted sum of contributions from sequence variants within a gene, each variant being weighted by its rarity and by its predicted effect on function. The method was applied to an ethnically heterogeneous exome-sequenced Swedish sample of 6000 controls and 5000 schizophrenia cases. PCs were also included, as was the PRS and an indicator of whether or not each subject carried a pathogenic CNV or a LOF variant in SETD1A. The results are expressed as a signed log10 p value (SLP), positive if the gene carries an excess of rare functional variants in cases rather than controls.



Results

The likelihood ratio test comparing models with or without the first 20 population principal components (PCs) showed that ancestry was strongly associated with caseness (X^2=374, 20 df, MLP>35). Using these principal components as covariates, both the PRS (X^2=156, 1 df, MLP=35) and the presence of a pathogenic CNV or sequence variant (X^2=39.6, 8df, MLP=5.4), were also associated with caseness.

The QQ plots show that the t test and logistic regression without PCs are anticonservative but that including PCs as covariates produces tests which comply well with the null distribution.

No individual genes reached genome wide statistical significance. In the analysis with all covariates included, the set of genes which are FMRP targets has a significant excess of rare, functional variants among cases (p=0.0005).

Conclusion

Using a logistic regression framework means that a gene-wise risk score can be used in a weighted burden test while including other potentially important variables in the analysis. Incorporating population principal components guards against false positive results due to population stratification. Including known genetic and non-genetic risk factors may enhance the ability to detect novel associations. This approach could also be used for developing individualised risk estimates.

Software availability

The code and documentation for SCOREASSOC and GENEVARASSOC is available from *https://github.com/davenomiddlenamecurtis/scoreassoc* and *https://github.com/davenomiddlenamecurtis/geneVarAssoc*.

The author declares no conflict of interest.