Bioinformatics, 35(6), 2019, 1061–1063
doi: 10.1093/bioinformatics/bty728
Advance Access Publication Date: 28 August 2018
Applications Note

OXFORD

Genetics and population analysis

PolyQTL: Bayesian multiple eQTL detection with control for population structure and sample relatedness

Biao Zeng and Greg Gibson *

School of Biological Sciences and Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on January 25, 2018; revised on June 4, 2018; editorial decision on August 19, 2018; accepted on August 23, 2018

Abstract

Motivation: Expression quantitative loci (eQTL) are being used widely to annotate and interpret GWAS hits. Recent studies have demonstrated that individual gene expression is often regulated by multiple independent cis-acting eQTL. Diverse methods, frequentist and Bayesian, have already been developed to simultaneously detect and fine-map such multiple eQTL, but most of these ignore sample relatedness and potential population structure. This can result in false positives and disrupt the accuracy of fine-mapping. Here we introduce PolyQTL software for identifying and estimating eQTL effects. The package incorporates a genetic relatedness matrix to remove the influence of population structure and sample relatedness, while utilizing a Bayesian multiple eQTL detection pipeline to identify the most plausible candidate causal variants at one or more independent loci influencing abundance of a transcript.
Results: Simulations demonstrate that our approach improves the rate of discovery of causal variants relative to methods that do not account for relatedness.
Availability and implementation: The software is written in C++, and freely available for download at https://github.com/jxzb1988/PolyQTL.
Contact: greg.gibson@biology.gatech.edu

1 Introduction

Genome-wide association analysis of gene expression leads to detection of eQTL (Albert and Kruglyak, 2015; Cheung *et al.*, 2005; GTEx Consortium, 2017). More than 100 human studies have been performed, most assuming parsimoniously that each eQTL region contains a single eQTL. However, genes tend to be regulated by numerous regulatory elements often located several hundred kilo-bases from the transcription start site, and consequently multiple regulatory polymorphisms are now thought to contribute to the variance in expression of most genes.

A common approach to causal variant discovery is to first calculate marginal association statistics for each variant, and then perform stepwise conditional analysis including lead associations as covariates in each successive model (Yang *et al.*, 2012). Investigators can then focus on the on the top ranked independent variants for follow-up studies. Bayesian methods have also been shown to be powerful for performing association analysis and fine-mapping, and multiple packages are available, including CAVIAR (Hormozdiari *et al.*, 2014), CAVIARBF (Chen *et al.*, 2015), FINEMAP (Benner *et al.*, 2016), FMQTL (Wen *et al.*, 2015), and DAP (Wen *et al.*, 2016). However, a major caveat that precludes their use on many datasets is that they use only summary statistics, ignoring any population structure and relatedness, or require external ancestry information to control for population structure in meta-analysis. Here we present PolyQTL, a software package for association analysis and fine mapping that addresses the limitations of these existing methods.

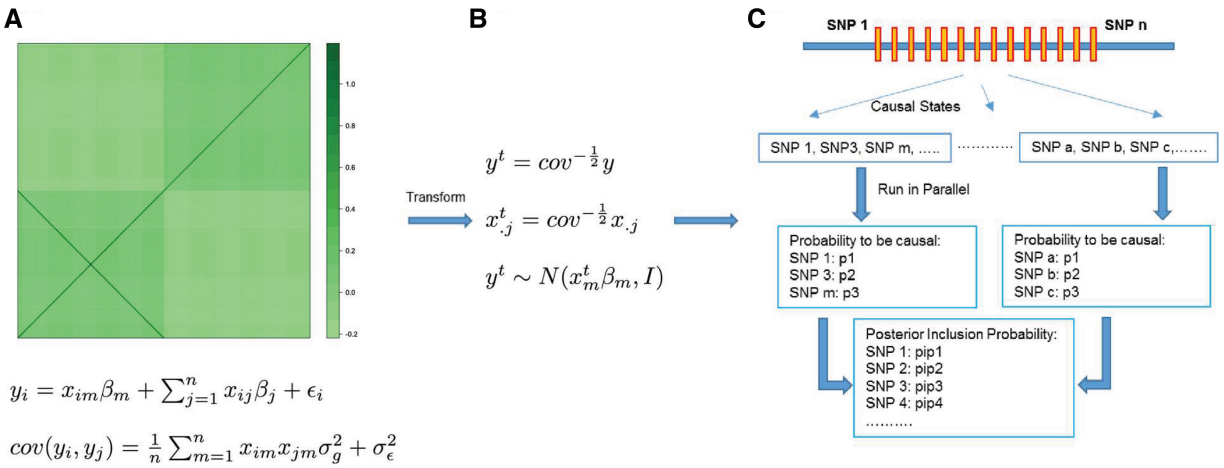

Fig. 1. Pipeline of PolyQTL. (**A**) Illustrates the complexity of simulation of related individuals and population structure. (**B**) Shows transformation of the phenotype and genotype with the square root of the covariance, and (**C**) a Bayesian method is used to compute a posterior inclusion probability (PIP) for each variant, ranking candidate variants

2 Materials and methods

Estimation of genetic contributions to a trait is biased in the presence of genetic covariance, due both to population structure (dark shading in quadrants of Figure 1A imply individuals in lower left and upper right are more genetically similar) and familial relatedness (diagonal lines represent identical twins as an extreme case). To remove the influence of relatedness, we transform the phenotype and genotype with the square root of the covariance computed from a genetic relatedness matrix (Abney *et al.*, 2002), resulting in uncorrelated phenotype residuals (Fig. 1B, Supplementary Methods). Subsequently, a Bayesian algorithm extending previous methods is used to compute a posterior inclusion probability (PIP) for each variant, allowing ranking of candidate causal variants in a QTL interval (Fig. 1C).

Since cis-regions encompassing 1 Mb around the transcription start site can contain thousands of variants, we also provide an option implementing GEMMA's mixed linear regression method (Zhou and Stephens, 2012), firstly performing sequential stepwise regression to isolate independent cis-eQTL at a locus, allowing parallel evaluation of the importance of all variants in each LD block surrounding peak variants. PolyQTL applies a C++ OpenMP library to compute the causal state posterior probabilities, running the computations in parallel when multiple CPUs are available.

We conducted simulations to demonstrate the advantages of our method. Two subpopulations were simulated, with ancestral allele frequencies × for causal variants uniformly distributed on [0.1, 0.9], with subpopulation allele frequencies sampled from a beta distribution with parameters x(1 – Fst)/Fst and (1 – *x*)(1 – Fst)/Fst, where Fst is the population differentiation index (Yang *et* al., 2014) over the range of 0.2 to 0.01. To mimic LD structure in real data, 100 variants in the cis-regulatory regions of randomly chosen genes were sampled from the 1843 non-African individuals in the 1000 Genomes Project (Auton *et al.*, 2015), from which two eQTL each explaining 4%–8% of the variance of a standard normal trait were simulated.

3 Results

We compared the performance of PolyQTL with the established Bayesian method, DAP, exploring the influence of three factors: heritability of gene expression, population structure and relatedness (0% or 20% of samples related as identical twins), resulting in a grid of eight different cases, each tested by 600 simulations. Statistical power was evaluated as the PIP score for the modeled causal variant at three different PIP cutoffs (0.1, 0.3 and 0.5).

Supplementary Figure S1 illustrates the general improvement in performance with PolyQTL relative to DAP for simulations in the presence of considerable population structure (Fst = 0.2) and two cis-eQTL per transcript. The cumulative distribution curves for PolyQTL are consistently below that of DAP, implying that more causal variants are discovered at lower PIP cutoffs. For example, with low background genetic contribution (heritability = 0.3), 73.5% of the causal variants have PIP greater than 0.1 with PolyQTL compared with only 66.8% in DAP (Supplementary Fig. S1A). Slight improvements are seen as heritability increases to 0.6 (Supplementary Fig. S1C) and if relatedness is introduced (Supplementary Fig. S1B), in each case improving the gains relative to DAP, with up to ∼10% more variants included at all PIP cutoffs (Supplementary Fig. S1D). The enhancement due to PolyQTL was more limited in the presence of a single causal variant per trait, or in the presence of more subtle population structure (Supplementary Figs S2 and S3). Simulating lesser relatedness (siblings rather than identical twins) did not markedly affect the results Supplementary Fig. S4). In addition, we find that PolyQTL also provides superior control of Type 1 error relative to DAP.

In summary, our simulation results demonstrate that PolyQTL controls population structure and relatedness, improving statistical power to include true causal variants in the list of high probability eQTL SNPs.

Funding

This work was supported by National Institutes of Health (1R01-HG008146), and China Scholarship Council (CSC) Scholarship (to B.Z).

Conflict of Interest: none declared.

References

Abney,M. *et al.* (2002) Quantitative-trait homozygosity and association mapping and empirical genome-wide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am. J. Hum. Genet.*, 70, 920–934.

Albert,F.W. and Kruglyak,L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Benner,C. *et al.* (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, **32**, 1493–1501.

Chen,W. *et al.* (2015) Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics*, **200**, 719–736.

Cheung,V.G. *et al.* (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365–1369.

GTEx Consortium. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

Hormozdiari,F. *et al.* (2014) Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**, 497–508.

Wen,X. *et al.* (2016) Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.*, **98**, 1114–1129.

Wen,X. *et al.* (2015) Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet*, **11**, e1005176.

Yang,J. *et al.* (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, **44**, 369–375.

Yang,J. *et al.* (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.

Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.