Genetics and population analysis

pgainsim: an R-package to assess the mode of inheritance for quantitative trait loci in GWAS

Nora Scherer, Peggy Sekula, Peter Pfaffelhuber and Pascal Schlosser

Abstract

Motivation: When performing genome-wide association studies conventionally the additive genetic model is used to explore whether a single nucleotide polymorphism (SNP) is associated with a quantitative trait. But for variants, which do not follow an intermediate mode of inheritance (MOI), the recessive or the dominant genetic model can have more power to detect associations and furthermore the MOI is important for downstream analyses and clinical interpretation. When multiple MOIs are modelled the question arises, which describes the true underlying MOI best.

Results: We developed an R-package allowing for the first time to determine study specific critical values when one of the three models is more informative than the other ones for a quantitative trait locus. The package allows for user-friendly simulations to determine these critical values with predefined minor allele frequencies and study sizes. For application scenarios with extensive multiple testing we integrated an interpolation functionality to determine critical values already based on a moderate number of random draws.

Availability and implementation: The R-package pgainsim is freely available for download on Github at https://github.com/genepi-freiburg/pgainsim.

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

When performing genome-wide association studies (GWAS) additive models are state-of-the-art to explore associations of SNPs and quantitative traits regardless of the actual mode of inheritance (MOI). Recessive and dominant models are able to improve statistical power to identify non-additive variants and for example a non-additive quantitative trait locus (QTL) has been observed in FTO (Wood et al., 2016). The knowledge of the MOI of a QTL is important for clinical interpretation and subsequent analyses.

After computing additive, recessive and dominant models in a GWAS and the rejection of multiple null hypotheses for a QTL the question arises which assumed MOI describes the data best. In studies that do consider multiple MOIs, this question is rarely examined in detail (Kraus et al., 2015). To answer this we adapted the p-gain concept introduced by Petersen et al. (2012) in the context of metabolome-wide GWAS (mGWAS). We transferred this to the difference between genetic models in a similar genome-wide manner.

2 Approach

We define the p-gains for a locus as follows:

\[
p_{\text{gain recessive}}(y) := \min(p_{\text{value recessive}}, p_{\text{value dominant}}) / p_{\text{value recessive}}(y) \tag{1}
\]

\[
p_{\text{gain additive}}(y) := \min(p_{\text{value recessive}}, p_{\text{value dominant}}) / p_{\text{value additive}}(y) \tag{2}
\]

\[
p_{\text{gain dominant}}(y) := \min(p_{\text{value recessive}}, p_{\text{value dominant}}) / p_{\text{value dominant}}(y) \tag{3}
\]

Here, we use \( p_{\text{value MOI}}(y) \) to reference the P-value corresponding to a t-test for association between a locus and a trait with observed values \( y \) based on a linear regression modelled with the specific MOI.

To assess for which critical value of the p-gain the model can be viewed as more informative than the others we derive critical values corresponding to a t-test for association between a locus and a trait with observed values \( y \) based on a linear regression modelled with the specific MOI.
through a simulation. By default the simulation is performed under the assumption of no genetic association (Supplementary Material).

3 Materials and methods

3.1 Properties of the p-gain

By definition (1)-(3) only the p-gain of the model with the lowest $P$-value can be greater than 1. The definition of the three genetic models for a biallelic SNP and a quantitative trait is given in the Supplementary Material. Because of the relation between these models the recessive and dominant p-gains are the same when the coded allele is flipped and the additive p-gain remains unaltered for flipped alleles [Supplementary Material (4, 5)].

Considering the p-gain as a random variable depending on a trait $Y$ and a SNP in Hardy-Weinberg equilibrium under the null hypothesis of no genetic association the p-gain depends only on $Y$ and the allele frequency AF at the SNP [Supplementary Material (6)]. Thus, for the cumulative distribution function (CDF) of the p-gain under the null hypothesis we have

$$\text{CDF}(p \text{-gain}_{\text{dominant}}(Y, \text{AF})) = \text{CDF}(p \text{-gain}_{\text{recessive}}(Y, 1 - \text{AF})), \quad \text{(4)}$$

$$\text{CDF}(p \text{-gain}_{\text{additive}}(Y, \text{AF})) = \text{CDF}(p \text{-gain}_{\text{additive}}(Y, 1 - \text{AF})). \quad \text{(5)}$$

3.2 R-package pgainsim

In the R-package pgainsim the traits are normally distributed with a user-specified study size. By default the genotypes of the SNPs are simulated independently from the trait according to the Hardy-Weinberg assumption of no genetic association (Supplementary Material). With the function $p\text{-gainsim}$ simulated independently from the trait according to the Hardy-Weinberg assumption of no genetic association (Supplementary Material). By default the genotypes of the SNPs are simulated independently from the trait according to the Hardy-Weinberg assumption of no genetic association (Supplementary Material). By default the genotypes of the SNPs are simulated independently from the trait according to the Hardy-Weinberg assumption of no genetic association (Supplementary Material). By default the genotypes of the SNPs are simulated independently from the trait according to the Hardy-Weinberg assumption of no genetic association (Supplementary Material). By default the genotypes of the SNPs are simulated independently from the trait according to the Hardy-Weinberg assumption of no genetic association (Supplementary Material).

Additionally, p-gains can be simulated with an assumed true effect of random draws simulated under no genetic association, MAF $= 33\%$ and a study size of 1627 and hence were able to reject the additive and dominant MOI (Supplementary Material).

4 Discussion

State-of-the-art GWAS model SNPs in an additive fashion. If considered at all the MOI is determined by the lowest $P$-value, which corresponds to a p-gain statistic greater one, or by the graphical representation of the gene dosages [Kraus et al., 2015; Schlosser et al., 2020]. This not only leads to a loss in power for associations with a true recessive or dominant MOI but also to false positive detections of non-additive MOIs.

An application of particular importance is mGWAS. Here we illustrate the application of the p-gain by a GWAS of glutamate concentrations and identified the recessive association with a variant in DGLUCY, which encodes D-glutamate cyclase that converts D-glutamate to 5-oxo-D-proline. When such an application of the critical values of the p-gain is extended to the metabolome-wide fashion of an mGWAS we suggest the binning of AFs in 5% intervals, determination of critical values for all interval limits and application of the more stringent of the two critical values for AFs within the interval.

We published an R-package rather than a reference set of critical values based on a large simulation as study size will influence the correlation between t-statistics of the three models and the appropriate study size should be used in the simulations. The R-package is based on the assumption of a linear regression model with a biallelic SNP and a continuously distributed outcome variable, as usually done in GWAS. Should one be interested in non-normal traits, such as binary traits, an adapted simulation design would be needed.

5 Conclusion

The R-package pgainsim allows for the first time the differentiation of MOIs in a study and allele frequency specific manner. This will lead to increased power for detection of non-additive genetic associations and inform downstream analyses and clinical interpretation.
Funding

Pascal Schlosser was supported by EQUIP—Funding for Medical Scientists, Faculty of Medicine, University of Freiburg.

Conflict of Interest: none declared.

References

Krauss, W.E. et al. (2015) Metabolomic quantitative trait loci (mQTL) mapping implicates the ubiquitin proteasome system in cardiovascular disease pathogenesis. *PLoS Genet.*, 11, e1005553.

Petersen, A. et al. (2012) On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC Bioinformatics*, 13, 120.

Schlosser, P. et al. (2020) Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nature Genetics*, 52, 167–176. 10.1038/s41588-019-0567-8 31959995

Tam, V. et al. (2019) Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.*, 20, 467–484.

Wood, A. R. et al. (2016) Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively. *Diabetologia*, 59, 1214–1221. 10.1007/s00125-016-3908-5