qgg: an R package for large-scale quantitative genetic analyses

Palle Duun Rohde*, **Izel Fourie Sørensen*** & **Peter Sørensen***, [1]

*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark.
[1]pso@mbg.au.dk

Abstract

Summary: Studies of complex traits and diseases are strongly dependent on the availability of user-friendly software designed to handle large-scale genetic and phenotypic data. Here, we present the R package **qgg**, which provides an environment for large-scale genetic analyses of quantitative traits and disease phenotypes. The qgg package provides an infrastructure for efficient processing of large-scale genetic data and functions for estimating genetic parameters, performing single and multiple marker association analyses, and genomic-based predictions of phenotypes. In particular, we have developed novel predictive models that use information on functional features of the genome that enables more accurate predictions of complex trait phenotypes. We illustrates core facilities of the qgg package by analysing human standing height from the UK Biobank.

Availability and implementation: The R package qgg is freely available. For latest updates, user guides and example scripts, consult the main page http://psoerensen.github.io/qgg/.

Introduction

Collection of large-scale genotype and phenotype data is fundamental for investigating the genetic basis underlying complex traits and diseases in evolutionary biology, animal and plant breeding, and human health. Furthermore, functional genomics is rapidly accumulating data about DNA function at gene, RNA transcript and protein product levels. We take advantage of the different layers of biological data to improve current prediction models of complex trait phenotypes.

Here, we present the **qgg** package, which provide a range of statistical models that incorporates prior information on genomic features. Genomic features, consisting of a set of genetic markers, are regions on the genome that links to different types of functional genomic information (e.g. genes, biological pathways, gene ontologies, sequence annotation, genome-wide expression or methylation patterns). Our main hypothesis is that these genome regions are enriched for causal variants affecting a specific trait. If this hypothesis is valid, then identifying the genomic features enriched for causal variants will aid in identifying the biological processes underlying trait variation (e.g. Rohde *et al.* (2016b); Sørensen *et al.* (2017)) and increase prediction accuracy of trait phenotypes (e.g. Edwards *et al.* (2016); Sarup *et al.* (2016)). We have previously demonstrated that genomic feature models can provide novel biological knowledge about the genetic basis of complex trait phenotypes in different species, including fruit fly (Edwards *et al.*, 2016; Rohde *et al.*, 2016a, 2017,

2018; Sørensen *et al.*, 2017; Ørseted *et al.*, 2017, 2018), mice (Ehsani *et al.*, 2015), dairy cattle (Edwards *et al.*, 2015; Fang *et al.*, 2017a,b, 2018), pigs (Sarup *et al.*, 2016) and humans (Rohde *et al.*, 2016b).

The qgg package provides a range of genomic feature prediction modelling approaches. Multiple features and multiple traits can be included, different genetic models (e.g. additive, dominance, gene by gene and gene by environment interactions) can be fitted, and a number of genetic marker set tests can be performed. Marker set tests allow for rapid analysis of different layers of genomic feature classes to discover genomic features potentially enriched for causal variants, thereby facilitating more accurate prediction models.

Implementation and main functions

The qgg package is implemented in the R statistical programming language (R Core Team, 2018). This allows users to utilise existing statistical computing and graphic facilities, and to develop efficient workflows that takes advantage of the genomic annotation resources (a key element in genomic feature models) available in e.g. Bioconductor (Huber *et al.*, 2015).

The qgg package provides core functions for performing quantitative genetic analyses including: 1) fitting linear mixed models, 2) constructing marker-based genomic relationship matrices, 3) estimating genetic parameters (e.g. heritability and correlation), 4) prediction of genetic predisposition and phenotypes, 5) single marker association analysis, and 6) gene set enrichment analysis.

Multi-core processing with openMP, multithreaded matrix operations implemented in BLAS libraries (e.g. OpenBLAS, ATLAS or MKL) and fast memory-efficient batch processing of genotype data stored in binary files (e.g. PLINK bedfiles) provide an efficient computational infrastructure for analysing large-scale data. The package compiles under all major platforms (Linux, MS Windows and OS X). Detailed documentation and tutorials are available online (http://psoerensen.github.io/qgg/).

Analysing human height

Core facilities of qgg are illustrated on standing height of the White British cohort in the UK Biobank (Bycroft *et al.*, 2017). Analyses were restricted to unrelated individuals with $<$ 5,000 missing SNP-genotypes and those without chromosomal aneuploidy (n=335,744 individuals). SNPs with minor allele frequency $<$0.01, loci with $>$5,000 missing genotypes, and SNPs associated with the major histocompatibility complex were removed (m=599,297 SNPs). Analyses accounted for age, gender and the first 10 genetic principal components. The scripts are available as supplementary material.

We estimated the heritability of human height (h^2, Fig. 1A) and partitioned the total genetic variance by autosomal chromosomes (h^2_f, Fig. 1B); a simple example of a genomic feature class. Partitioning SNPs in the training population by degree of association, improved the accuracy of genomic prediction (R^2) in the validation population when increasing the number of included top SNPs and the size of the training population (Fig. 1C). The top 200 associated SNPs from the GIANT height study (Lango Allen *et al.*, 2010), was highly enriched in the UK Biobank data (Fig. 1D).

Conclusion

The qgg package provides an efficient computational infrastructure for analysing large-scale genotype-phenotype data, and contains a range of quantitative genetic modelling approaches for investigating the genetic basis of complex traits and diseases.

Funding

The Danish Strategic Research Council supported this work (GenSAP: Centre for Genomic Selection in Animals and Plants, contract no. 12-132452), and we obtained the example data from the UK Biobank Resource (project ID 31269).

Conflict of Interest: Peter Sørensen is an employee of Aarhus University and Genomics plc. The research described in this paper was conducted solely at Aarhus University, with no involvement of Genomics plc.

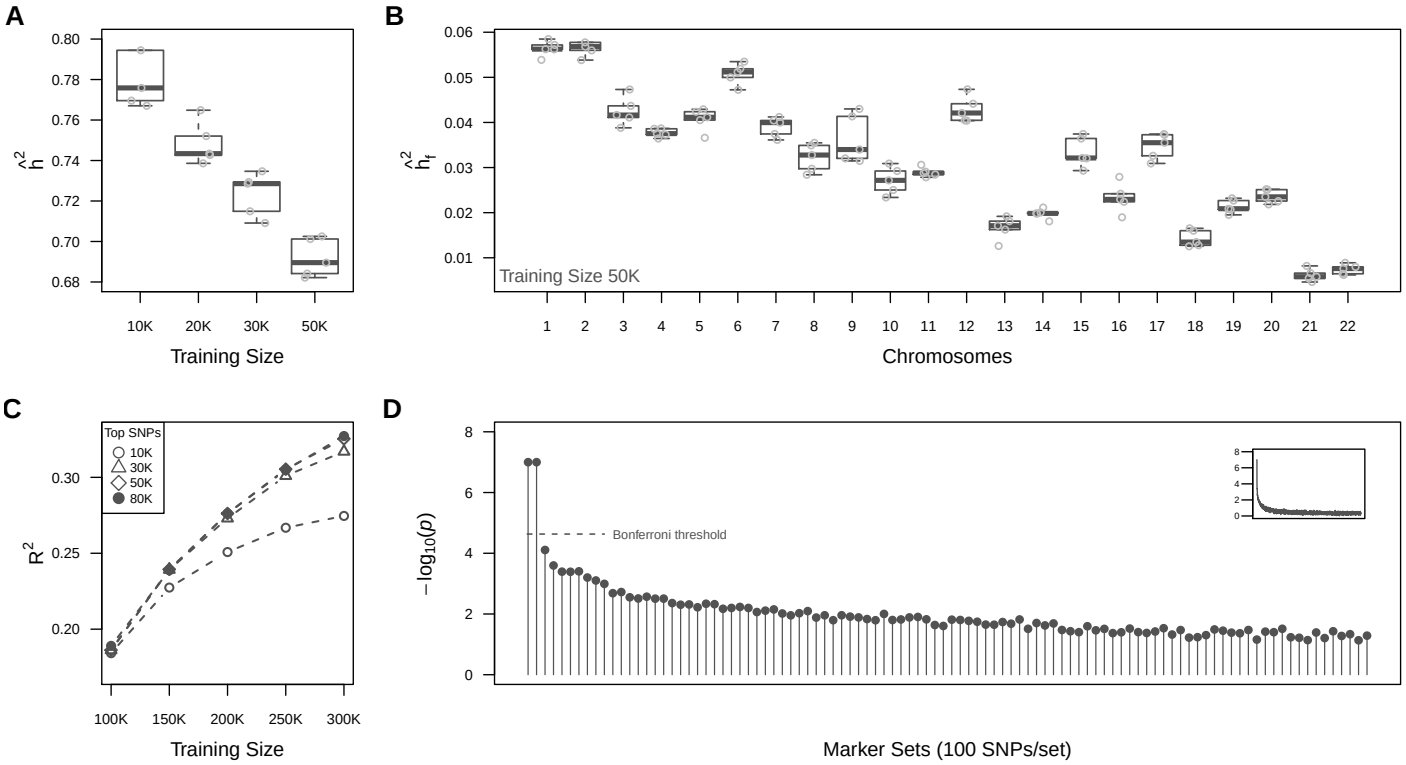

Figure 1. Summary of results from human standing height using qgg functions. (**A**) SNP-based heritability estimates (h^2) as a function of the size of the training population. For each training population the analysis was performed on five different training sets (points). (**B**) Partitioning of genetic variance (h^2_f) using autosomal chromosomes as genomic features. (**C**) Prediction accuracy (R^2) of a genomic feature best linear unbiased prediction model as function of training population size and number of top ranking SNPs used as genomic features. (**D**) Gene set enrichment analysis using marker sets defined by p-value ranked SNP-markers from the GIANT consortium with 100 SNPs within each feature set. Main figure shows the enrichment of the first 100 sets, whereas the small inserted figure contains all 2,145 sets.

References

Bycroft, C., *et al.* (2017). Genome-wide genetic data on 500,000 UK Biobank participants. bioRxiv. eprint. doi: https://doi.org/10.1101/166298.

Edwards, S. M., *et al.* (2015). Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. *Genet. Select. Evol.*, **47**:60.

Edwards, S. M., *et al.* (2016). Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. *Genetics*, **203**:1871-1883.

Ehsani, A., *et al.* (2015). Decomposing genomic variance using information from GWA, GWE and eQTL analysis. *Animal Genetics*, **47**:165-173.

Fang, L., *et al.* (2017a). Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet. Select. Evol.*, **49**:44.

Fang, L., *et al.* (2017b). Integrating sequence-based GWAS and RNA-seq provides novel insights into the genetic basis of mastitis and milk production in dairy cattle. *Scientific Reports*, **7**:45560.

Fang, L., *et al.* (2018). MicroRNA-guided prioritization of genome-wide association signals reveals the importance of microRNA-target gene networks for complex traits in cattle. *Scientific Reports*, **8**:9345.

Huber, W., *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**:115-121.

Lango Allen, H., *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**:832-838.

Ørsted, M., *et al.* (2017). Environmental variation partitioned into separate heritable components. *Evolution*, **72**:136-152.

Ørsted, M., *et al.* (2018). Strong impact of thermal environment on the quantitative genetic basis of a key stress tolerance trait. *Heredity*, **2018**:1-11.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rohde, P. D., *et al.* (2016a). A quantitative genomic approach for analysis of fitness and stress related traits in a *Drosophila melanogaster* model population. *Int. J. Genomics*, **2016**:1-11.

Rohde, P. D., *et al.* (2016b). Covariance Association Test (CVAT) identify genetic markers associated with schizophrenia in functionally associated biological processes. *Genetics*, **203**:1901-1913.

Rohde, P. D., *et al.* (2017). Genomic analysis of genotype-by-social environment interaction for *Drosophila melanogaster*. *Genetics*, **206**:1969-1984.

Rohde, P. D., *et al.* (2018). Functional validation of candidate genes detected by genomic feature models. *G3*, **8**:1659-1668.

Sarup, P., *et al.* (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genetics*, **17**:11.

Sørensen, I. F., *et al.* (2017). Multiple trait covariance association test identifies gene ontology categories associated with chill coma recovery time in *Drosophila melanogaster*. *Scientific Reports*, **7**:2413.