# Expanded utility of the R package, qgg, with applications within genomic medicine

Rohde, Palle Duun; Fourie Sørensen, Izel; Sørensen, Peter

*Published in:*
Bioinformatics

*DOI (link to publication from Publisher):*
10.1093/bioinformatics/btad656

*Publication date:*
2023

*Document Version*
Publisher's PDF, also known as Version of record

*Bioinformatics*, 2023, **39(11)**, btad656
https://doi.org/10.1093/bioinformatics/btad656
Advance Access Publication Date: 26 October 2023
**Applications Note**

OXFORD

Genetics and population analysis

# Expanded utility of the R package, qgg, with applications within genomic medicine

**Palle Duun Rohde [1,*], Izel Fourie Sørensen [2], Peter Sørensen [2,*]**

[1]Genomic Medicine, Department of Health Science and Technology, Aalborg University, 9260 Gistrup, Denmark
[2]Center for Quantitative Genetics and Genomics, Aarhus University, 8000 Aarhus, Denmark

*Corresponding authors. Genomic Medicine, Department of Health Science and Technology, Aalborg University, Selma Langerløfs Vej 249, 9260 Gistrup, Denmark. E-mail: palledr@hst.aau.dk (P.D.R.); Center for Quantitative Genetics and Genomics, Aarhus University, C. F. Møllers Allé 3, 8000 Aarhus, Denmark. E-mail: pso@qgg.au.dk (P.S.)

Associate Editor: Christina Kendziorski

## Abstract

**Summary:** Here, we present an expanded utility of the R package qgg for genetic analyses of complex traits and diseases. One of the major updates of the package is, that it now includes Bayesian linear regression modeling procedures, which provide a unified framework for mapping of genetic variants, estimation of heritability and genomic prediction from either individual level data or from genome-wide association study summary data. With this release, the qgg package now provides a wealth of the commonly used methods in analysis of complex traits and diseases, without the need to switch between software and data formats.

**Availability and implementation:** The methodologies are implemented in the publicly available R software package, qgg, using fast and memory efficient algorithms in C++ and is available on CRAN or as a developer version at our GitHub page (https://github.com/psoerensen/qgg). Notes on the implemented statistical genetic models, tutorials and example scripts are available at our GitHub page https://psoerensen.github.io/qgg/.

## 1 Introduction

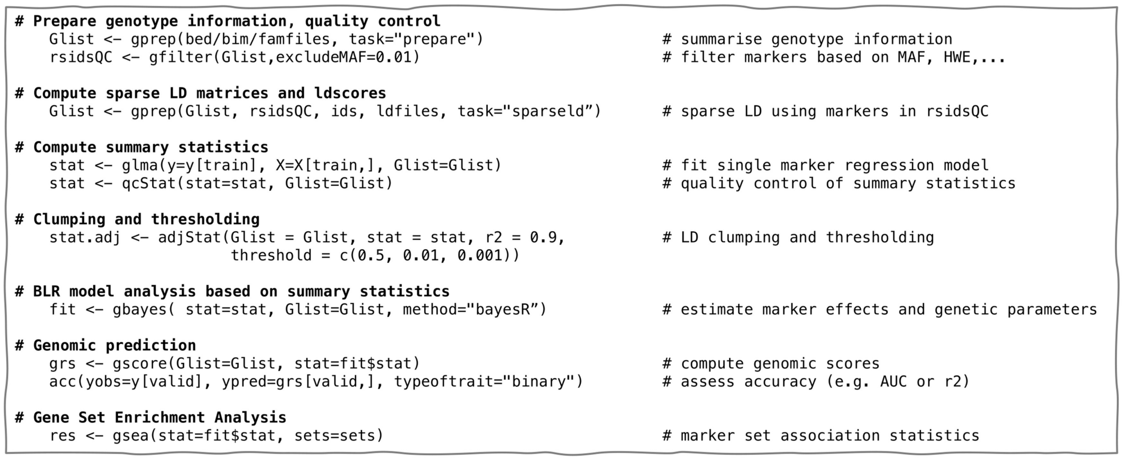
With the increasing availability of large genetic biobanks and a simultaneous increase in number of collected samples by biobanks, there is a demand for software tools that allow for robust and efficient genomic analyses. Tools such as PLINK (Chang *et al.* 2015), GCTA (Yang *et al.* 2011), LDpred (Vilhjálmsson *et al.* 2015, Privé *et al.* 2021), LDAK (Speed and Balding 2014, Speed *et al.* 2017, Zhang *et al.*, 2021), and PRSice (Euesden *et al.* 2015) have changed the way researchers around the globe conduct genetic analyses of human complex traits and diseases. Here we present a major update of the R package qgg (Rohde *et al.* 2020), which has now been expanded to include a range of commonly used methods such as LD Score Regression (LDSC), adjustment of marker effect using correlated trait information (Rohde *et al.* 2022), a range of different Bayesian shrinkage models for gene mapping (Shrestha *et al.* 2023), and construction of single- and multiple trait polygenic scores (PGS). With a user-friendly interface, qgg offers a unified tool for quantitative genetic analysis of complex traits and diseases. Supplementary Tables S1–S4 contains an overview of the main functionalities that are implemented in the qgg package.

Human complex traits and diseases vary greatly in how their genetic architecture is shaped, e.g. in terms of effect size distribution, number of causal variants and non-linear genetic effects (Timpson *et al.* 2018). It is therefore important that the statistical model used for gene mapping or genomic risk prediction account for different types of genetic architectures. Bayesian linear regression (BLR) models have been proposed as a unified framework for gene mapping, estimation of genetic parameters and genomic prediction (Ehsani *et al.* 2012, 2016, Moser *et al.* 2015, Sørensen *et al.* 2015, Vilhjálmsson *et al.* 2015, Lloyd-Jones *et al.* 2019). BLR models account for the underlying genetic architecture by allowing the true genetic signal to be heterogeneously distributed over the genome (i.e. some regions have stronger genetic signal than others). Since BLR models fit all genetic markers simultaneously and account for linkage disequilibrium (LD) between markers, they often have greater power to detect causal associations and find less false negatives (Lloyd-Jones *et al.* 2019). The gain in statistical power to detect causal genetic variants subsequently increases the accuracy of genomic prediction.

## 2 Implementation

The software package qgg is available as an R package with main functions written in C++ taking advantage of fast and memory efficient algorithms. qgg handles large-scale data using efficient algorithms and by taking advantage of multi-core processing using openMP, multithreaded matrix operations implemented in BLAS libraries (e.g. OpenBLAS, ATLAS or MKL), and direct fast and memory-efficient processing of genetic data without the need to re-format the binary PLINK files (Chang *et al.* 2015), as is needed in other R packages for handling large-scale genetic data.

**Received:** 24 March 2023; **Revised:** 17 September 2023; **Editorial Decision:** 20 October 2023; **Accepted:** 24 October 2023



```
# Prepare genotype information, quality control
   Glist <- gprep(bed/bim/famfiles, task="prepare")              # summarise genotype information
   rsidsQC <- gfilter(Glist,excludeMAF=0.01)                     # filter markers based on MAF, HWE,...

# Compute sparse LD matrices and ldscores
   Glist <- gprep(Glist, rsidsQC, ids, ldfiles, task="sparseld")     # sparse LD using markers in rsidsQC

# Compute summary statistics
   stat <- glma(y=y[train], X=X[train,], Glist=Glist)            # fit single marker regression model
   stat <- qcStat(stat=stat, Glist=Glist)                        # quality control of summary statistics

# Clumping and thresholding
   stat.adj <- adjStat(Glist = Glist, stat = stat, r2 = 0.9,     # LD clumping and thresholding
                       threshold = c(0.5, 0.01, 0.001))

# BLR model analysis based on summary statistics
   fit <- gbayes( stat=stat, Glist=Glist, method="bayesR")       # estimate marker effects and genetic parameters

# Genomic prediction
   grs <- gscore(Glist=Glist, stat=fit$stat)                     # compute genomic scores
   acc(yobs=y[valid], ypred=grs[valid,], typeoftrait="binary")   # assess accuracy (e.g. AUC or r2)

# Gene Set Enrichment Analysis
   res <- gsea(stat=fit$stat, sets=sets)                         # marker set association statistics
```

**Figure 1.** Overview of the simple and streamlined workflow for genetic analysis of complex traits that takes PLINK files as input. The output from each function is designed to match the input format of downstream analyses.

## 3 Usage

The different statistical genetic methods are implemented using a simple and streamlined workflow (Fig. 1). Initially, PLINK files are processed to construct a Glist-object that summarizes information about the genetic data and computes genotype and allele frequencies, genotype missingness etc. This information can be used to perform standard quality control of the genetic data and to compute sparse LD matrices and LD scores. After the initial pre-processing, genome-wide association study (GWAS) summary data can be directly computed, followed by adjustment using either clumping and thresholding (C+T) or one of the implemented BLR models (bayes A, bayes C, bayes R, bayesian Lasso and bayesian mixed model). Finally, the GWAS summary data can be used to construct PGS or to identify biologically relevant gene sets enriched for associated variants.

In the accompanying Supplementary Material we showcase the new features implemented in qgg, by performing single- and multi-trait genetic analyses of body mass index (BMI) and standing height using data from the United Kingdom Biobank (UKB) (Bycroft *et al.* 2018).

## 4 Conclusion

We have presented an update of the R package qgg. We highlight the BLR models for single and multiple trait analyses for (i) constructing PGS with utilities in genomic medicine, (ii) fine mapping of GWAS results and (iii) accurate estimation of quantitative genetic parameters.

## Supplementary data

Supplementary data are available at Bioinformatics online.

## Conflict of interest

None declared.

## Funding

P.S. has received funding from The Novo Nordisk Foundation initiative Open Discovery Innovation Network (ODIN) (NNF20SA0061466).

## Data availability

The phenotypic and genotype data underlying this article cannot be shared publicly. The data comes from the UK Biobank, which researchers can gain access to by applying for research access. See https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access for instructions for the application process.

## References

Bycroft C, Freeman C, Petkova D *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9.

Chang CC, Chow CC, Tellier LC *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**:7–16.

Ehsani A, Janss L, Pomp D *et al.* Decomposing genomic variance using information from GWA, GWE and eQTL analysis. *Anim Genet* 2016;**47**:165–73.

Ehsani A, Sørensen P, Pomp D *et al.* Inferring genetic architecture of complex traits using Bayesian integrative analysis of genome and transcriptome data. *BMC Genomics* 2012;**13**:456.

Euesden J, Lewis CM, O'Reilly PF *et al.* PRSice: polygenic risk score software. *Bioinformatics* 2015;**31**:1466–8.

Lloyd-Jones LR, Zeng J, Sidorenko J *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun* 2019;**10**:5086–11.

Moser G, Lee SH, Hayes BJ *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet* 2015;**11**:e1004969.

Privé F, Arbel J, Vilhjálmsson BJ *et al.* LDpred2: better, faster, stronger. *Bioinformatics* 2021;**36**:5424–31.

Rohde PD, Fourie Sørensen I, Sørensen P *et al.* qgg: an R package for large-scale quantitative genetic analyses. *Bioinformatics* 2020;**36**: 2614–5.

Rohde PD, Nyegaard M, Kjolby M *et al.* Multi-trait genomic risk stratification for type 2 diabetes. *Front Med (Lausanne)* 2022;**8**: 711208.

Shrestha M, Bai Z, Gholipourshahraki T *et al.* Evaluation of Bayesian linear regression models as a fine mapping tool. bioRxiv 2023, https://doi.org/10.1101/2023.09.01.555889, preprint: not peer reviewed.

Sørensen P, de los Campos G, Morgante F *et al.* Genetic control of environmental variation of two quantitative traits of *Drosophila melanogaster* revealed by whole-genome sequencing. *Genetics* 2015;**201**:487–97.

Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res* 2014;**24**:1550–7.

Speed D, Cai N, Johnson MR *et al.*; UCLEB Consortium. Reevaluation of SNP heritability in complex human traits. *Nat Genet* 2017;**49**:986–92.

Timpson NJ, Greenwood CMT, Soranzo N *et al.* Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* 2018;**19**:110–24.

Vilhjálmsson BJ, Yang J, Finucane HK *et al.*; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet* 2015;**97**:576–92.

Yang J, Lee SH, Goddard ME *et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**: 76–82.

Zhang Q, Privé F, Vilhjálmsson B *et al.* Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat Commun* 2021;**12**:4192.