



Genetics and population analysis

GenRisk: a tool for comprehensive genetic risk modeling

Rana Aldisi ^{1,*}, Emadeldin Hassanin¹, Sugirthan Sivalingam^{1,2,3},
Andreas Bunes^{1,2,3}, Hannah Klinkhammer^{1,3}, Andreas Mayr³, Holger Fröhlich^{4,5},
Peter Krawitz ¹ and Carlo Maj^{1,6}

¹Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn 53127, Germany, ²Core Unit for Bioinformatics Analysis, University Hospital Bonn, Bonn 53127, Germany, ³Institute of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn 53127, Germany, ⁴Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, 53757 Sankt Augustin, Germany, ⁵Bonn-Aachen International Center for IT (b-it), University of Bonn, Bonn 53115, Germany and ⁶Centre for Human Genetics, University of Marburg, Marburg 35033, Germany

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on September 24, 2021; revised on February 4, 2022; editorial decision on March 4, 2022; accepted on March 9, 2022

Abstract

Summary: The genetic architecture of complex traits can be influenced by both many common regulatory variants with small effect sizes and rare deleterious variants in coding regions with larger effect sizes. However, the two kinds of genetic contributions are typically analyzed independently. Here, we present GenRisk, a python package for the computation and the integration of gene scores based on the burden of rare deleterious variants and common-variants-based polygenic risk scores. The derived scores can be analyzed within GenRisk to perform association tests or to derive phenotype prediction models by testing multiple classification and regression approaches. GenRisk is compatible with VCF input file formats.

Availability and implementation: GenRisk is an open source publicly available python package that can be downloaded or installed from Github (<https://github.com/AldisiRana/GenRisk>).

Contact: s0raaldi@uni-bonn.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

In the past decade, genome-wide association studies (GWAS) have been used extensively to investigate the genetic architecture of complex traits and diseases (Uffelmann *et al.*, 2021). However, despite the identification of many disease-associated common variants which also led to the development of several accurate polygenic risk score (PRS) models, a substantial part of the genetic architecture of common traits remains unknown (Lee *et al.*, 2014). This is known as missing heritability, which is the difference between the heritability observed in twins studies and the measured heritability explained by common variants (Génin, 2020).

Different studies suggested that the missing heritability is mainly attributable to rare variants (Young, 2019). In line with this hypothesis, many studies have observed that rare variants play a role in complex phenotypes, such as hypertension (Russo *et al.*, 2018), schizophrenia (John *et al.*, 2019) and autism (Havdahl *et al.*, 2021). Burden tests are among the most applied methods to investigate rare variant effects starting from sequencing data. These methods typically collapse rare variants in a genetic region (e.g. gene) into a single burden variable and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants (Bomba *et al.*,

2017). On the other hand, the genetic contribution of common variants is typically analyzed by mean of PRS, which is usually computed as the weighted sum of risk alleles with respect to a phenotype, where the risk alleles and the corresponding weights are derived from a reference GWAS (Choi *et al.*, 2020).

Generally, gene-based burden tests are applied on exome/target sequencing data while GWAS is performed on post-imputed chip-array data for the genotyping of high-frequent variants. In the light of the increasing availability of whole genome sequencing data, there is a need of bioinformatics solutions integrating different methodological approaches into a unique framework. With this aim in mind, we developed GenRisk, a python package that seamlessly combines different tools and libraries to analyze genotype-phenotype associations by considering both polygenic effects and the enrichment of rare deleterious variants at gene-based level.

2 Implementation

The GenRisk pipeline contains multiple modules, which can be run using a commandline interface or within a python environment. The modules can be run sequentially, so that the input of a module is the

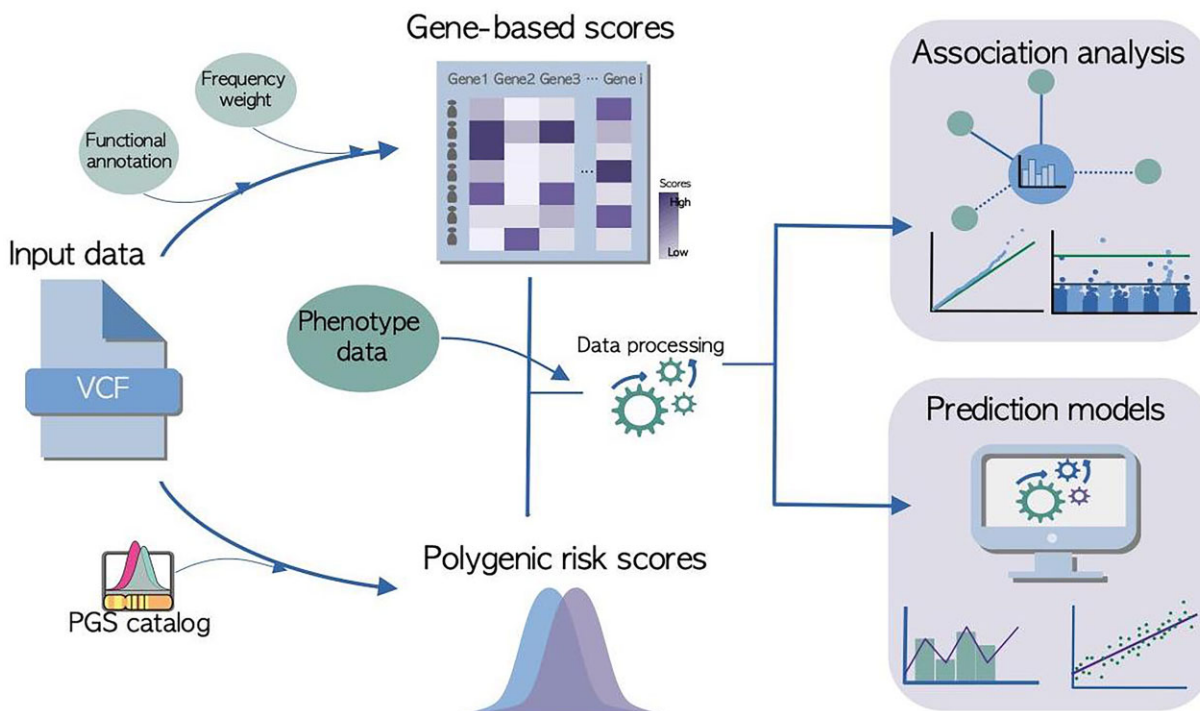


Fig. 1. GenRisk pipeline workflow. A VCF file with functional annotations and frequencies can be used to calculate gene-based scores, alternatively a VCF can be used to extract and calculate PRS. The scores can then be used with phenotypic data for association analysis or to develop prediction models

output of the previous module. In addition, each module can also be used independently with data provided by the user to increase flexibility of the tool for custom-analyses. Starting from a VCF, GenRisk computes gene scores based on variant annotations. Given a phenotype and potential covariates (possibly including PRS), the individual gene scores can be used to perform association analyses and to build phenotype prediction models. Furthermore, an interactive command implements PRS computation, the PRS model can be either provided by the user or available in *pgscatalog* (<https://www.pgscatalog.org/>).

The workflow of the pipeline is summarized in Figure 1. In the following sections the main features of GenRisk are described.

2.1 Gene-based scoring system

The gene scores are derived by the weighted sum of the variants in a gene. Each allele count is weighted according to the product of a deleteriousness score and a coefficient based on the allele frequency. Namely, a weighting function is applied to the variant frequency to potentially up-weight the biological importance of rare variants. Two weighting functions are implemented, $-\log_{10}$ as already applied in another gene-based score tool (Mossotto et al., 2019) and the beta density function, which contains two parameters α and β that can be adjusted for more flexible weight calculation as implemented in the sequence kernel association test (Lee et al., 2012). An adjustable threshold parameter for the minor allele frequency (MAF) can be also considered to filter only for rare variants.

2.2 Genetic risk scores analysis

According to the distribution of the scores, different statistical tests can be applied to analyze gene-phenotype associations starting from the derived individual-based gene scores. The association analysis results are generated as summary statistics and can be visualized via QQ-plots and Manhattan plots.

Prediction models are computed using the open-source Pycaret, a machine learning python library (Ali, 2020). The models can be generated for both quantitative and binary traits. The gene-based scores, as well as PRS and covariates, such as sex and age, can be used as features. The data given by the user can be divided into

training and testing sets (with flexible size). Cross-validation is applied on different models and the best performing model is selected, tuned and finalized. The model is then saved and can be further evaluated with external testing sets. Model evaluation reports and testing set labels are exported. Graphs like, feature importance, confusion matrix and prediction error, are also generated to visualize the model performance.

3 Usage case

We applied the pipeline on $\approx 160\,000$ samples from UK Biobank (application number 81202), the gene-based scores were calculated by applying the beta weighting function ($\alpha = 1$, $\beta = 25$) to up-weight rare variants while the CADD (Rentzsch et al., 2019) raw scores were used as deleteriousness weight and only variants with MAF $< 1\%$ were included. The derived scores were used for association test and prediction model with respect to alkaline phosphatase measurements (Field 30610) including also the first four genotyping principle components, sex, BMI and age as covariates. The association analysis based on a linear regression model detected significance in ALPL, GPLD1 and ASGR1 genes, all of which have been previously associated with alkaline phosphatase (Nioi et al., 2016; Yuan et al., 2008). In addition, a stochastic gradient boosted decision tree algorithm was identified as the best prediction model once both gene scores and PRS (from Sinnott-Armstrong et al., 2021) are taken into account and it showed an improved prediction performance compared with PRS-only model.

Detailed results, as well as comparisons with other methods, can be found in Supplementary Material.

4 Conclusion

GenRisk is a python package that processes input VCF files to generate both gene-based burden scores and PRS for association tests and development of prediction models. GenRisk provides a framework to model the effects of rare functional variants while considering the polygenic background. Thus, it is suitable for the analysis of phenotypes characterized by a complex genetic architecture.

Funding

C.M. and E.H. were supported by the BONFOR-program of the Medical Faculty, University of Bonn (O-147.0002).

Conflict of Interest: none declared.

Data availability

Genome-wide genotyping data, exome-sequencing data, and phenotypic data from the UK Biobank are available upon successful project application (<http://www.ukbiobank.ac.uk/about-biobank-uk/>). Restrictions apply to the availability of these data, which were used under license for the current study (Project ID: 81202).

References

- Ali,M. (2020) *PyCaret: An Open Source, Low-Code Machine Learning Library in Python. PyCaret Version 1.0.* <https://pycaret.gitbook.io/docs/#citation>.
- Bomba,L. *et al.* (2017) The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.*, **18**, 77.
- Choi,S.W. *et al.* (2020) Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.*, **15**, 2759–2772.
- Génin,E. (2020) Missing heritability of complex diseases: case solved? *Hum. Genet.*, **139**, 103–113.
- Havdahl,A. *et al.* (2021) Genetic contributions to autism spectrum disorder. *Psychol. Med.*, **51**, 2260.
- John,J. *et al.* (2019) Rare variant based evidence for oligogenic contribution of neurodevelopmental pathway genes to schizophrenia. *Schizophrenia Res.*, **210**, 296–298.
- Lee,S. *et al.*; NHLBI GO Exome Sequencing Project—ESP Lung Project Team. (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, **91**, 224–237.
- Lee,S. *et al.* (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.
- Mossotto,E. *et al.* (2019) GenePy – a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics*, **20**, 254.
- Nioi,P. *et al.* (2016) VariantASGR1 associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.*, **374**, 2131–2141.
- Rentzsch,P. *et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
- Russo,A. *et al.* (2018) Advances in the genetics of hypertension: the effect of rare variants. *Int. J. Mol. Sci.*, **19**, 688.
- Sinnott-Armstrong,N. *et al.*; FinnGen. (2021) Genetics of 35 blood and urine biomarkers in the UK biobank. *Nat. Genet.*, **53**, 185–194.
- Uffelmann,E. *et al.* (2021) Genome-wide association studies. *Nat. Rev. Methods Primers*, **1**, 59.
- Young,A.I. (2019) Solving the missing heritability problem. *PLoS Genet.*, **15**, e1008222.
- Yuan,X. *et al.* (2008) Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am. J. Hum. Genet.*, **83**, 520–528.