






Genome analysis

PRScalc, a privacy-preserving calculation of raw polygenic risk scores from direct-to-consumer genomics data

Lorena Sandoval ^{1,2,*}, Saleet Jafri¹, Jeya Balaji Balasubramanian², Praphulla Bhawsar², Jacob L. Edelson², Yasmmin Martins ³, Wolfgang Maass⁴, Stephen J. Chanock ², Montserrat Garcia-Closas ², Jonas S. Almeida ²

¹Department of Biomedical Informatics, George Mason University, Fairfax, VA 22030, United States

²Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute, Rockville, MD 20850, United States

³Bioinformatics Laboratory, National Laboratory for Scientific Computing, Petropolis 25651, Brazil

⁴Saarland University, 66123 Saarbrücken, Germany

*Corresponding author. Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute, 9609 Medical Center Dr, Rockville, MD 20850, United States. E-mail: lorena.sandoval@nih.gov

Associate Editor: Thomas Lengauer

Abstract

Motivation: Currently, the Polygenic Score (PGS) Catalog curates over 400 publications on over 500 traits corresponding to over 3000 polygenic risk scores (PRSs). To assess the feasibility of privately calculating the underlying multivariate relative risk for individuals with consumer genomics data, we developed an in-browser PRS calculator for genomic data that does not circulate any data or engage in any computation outside of the user's personal device.

Results: A prototype personal risk score calculator, created for research purposes, was developed to demonstrate how the PGS Catalog can be privately and readily applied to readily available direct-to-consumer genetic testing services, such as 23andMe. No software download, installation, or configuration is needed. The PRS web calculator matches individual PGS catalog entries with an individual's 23andMe genome data composed of 600k to 1.4 M single-nucleotide polymorphisms (SNPs). Beta coefficients provide researchers with a convenient assessment of risk associated with matched SNPs. This in-browser application was tested in a variety of personal devices, including smartphones, establishing the feasibility of privately calculating personal risk scores with up to a few thousand reference genetic variations and from the full 23andMe SNP data file (compressed or not).

Availability and implementation: The PRScalc web application is developed in JavaScript, HTML, and CSS and is available at GitHub repository (<https://episphere.github.io/prs>) under an MIT license. The datasets were derived from sources in the public domain: [PGS Catalog, Personal Genome Project].

1 Introduction

The increasing identification of risk loci from large-scale 'genome-wide association studies' (GWAS) holds promise for risk prediction to inform both individualized and population-level preventative strategies (e.g. screening, precision prevention) (Chatterjee *et al.* 2016). GWAS-derived 'polygenic risk scores' (PRSs) are collections of genetic variants weighted by their effect size (association with trait of interest) that inform the cumulative effect of an individual's genetic risk for some trait (Loos 2020). The Polygenic Score (PGS) Catalog curates publications of PRSs for multiple traits and corresponding metadata, variants associated with a trait, effect alleles and its associated effect size, and outcome odds ratios, all served through a public REST API and FTP site (Lambert *et al.* 2021).

'Direct-to-consumer' products that calculate the risk of disease/traits based on genotype data, often through server-side methods, have become widely available (Horton *et al.* 2019). One such product, 23andMe (<https://www.23andme.com>), is used here to assess the privacy-preserving approach. 23andMe

relies on genotyping technology to identify genetic markers or 'single-nucleotide polymorphisms' (SNPs) associated with a disease or trait. Additionally, 23andMe provides users with their raw data (see 'Your 23andMe data download' in the in-browser application for the direct link). Impute.me was an open-source PRS calculation tool (Folkersen *et al.* 2020); however, it required genotype data to be uploaded to a server for processing, prompting data breach prevention strategies. Furthermore, providing users with access to their raw genetic data can allow for a sense of empowerment which may offset the balance between privacy and informed consent. To address this limitation in our application, personal PRSs are calculated on 23andMe genotype data, using pure Javascript, completely on the client-side, on all modern browsers, without the need for opaque software downloads or installations. This serves as a privacy-preserving exercise in scalability of portable risk score calculators running in personal mobile devices under full user oversight.

Specifically, no personal data circulate external computational environments, while making full use of the public reference offered by hundreds of published PRS in the PGS catalog.

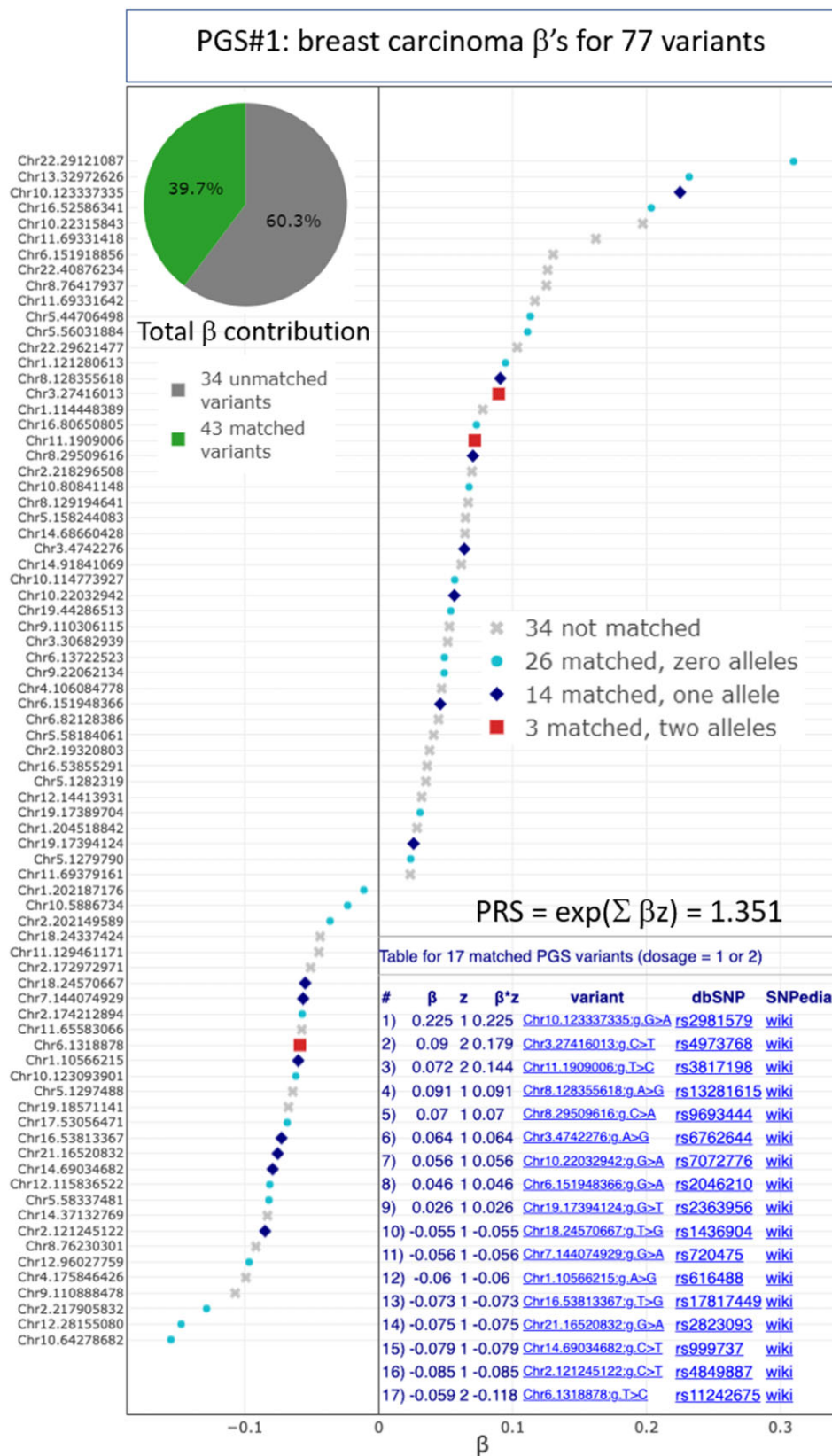


Figure 1. Composition of graphic elements produced by the privacy-preserving PRS calculator for the public 23andme datafile at Harvard Personal Genome Project ([personalgenomes.org](#)), with ref #hu54EEB2, under the name ‘Dorothy Wolf’, based on PGS Catalog ([pgscatalog.org](#)) entry PGS000001, for breast carcinoma. The three graphic elements in this composition are, in order of appearance in the web calculator result: (i) the table listing only the variants that match both the PGS catalog and for which the individual 23andme file reports a variant (where z value is 1 or 2), placed in the lower right of the pot; (ii) the pie chart accessing what proportion of risk, $\sum \beta$, reported in the PGS catalog is actually covered by 23andme panel, displayed at the top left of the figure; and (iii) the background plot sorting all the SNPs assessed in the catalog by β value with different symbols and colors for unmatched. The report provides links to additional references, including the publication, phenotype ontology, and dbSNP.

2 Methods

For the general user, 23andMe does not currently support an authentication API to pull SNP data directly from their account. As an alternative, a 23andMe raw data file containing up to 610 527 variants from an individual, is locally loaded onto the application. The calculation is processed entirely in the browser on the 23andMe file in memory in the user's local system, i.e. without ever circulating any data by a server. The calculation consists of matching PGS variants by chromosome and position to 23andMe variants (build37). The PRS is calculated from the summation of the products of the PGS beta (effect size) and the individual's allele dosage (total number of effect alleles) according to the following equation:

$$\text{PRS} = \exp\left(\sum_i^n (\beta_i z_i)\right)$$

The above equation is a standard equation to calculate the underlying multivariate relative risk as weighted (β) PRS multiplied by dosage (z , number of alleles with variation).

In this equation, N is the number of SNPs in PGS catalog entry, β_i is the effect size (or beta) of variant i , and dosage, z_i , is the number of copies of the effect allele of SNP i for that 23andme individual (Collister *et al.* 2022). The 'exp(betas)' component can also be defined as odds ratios measuring the relative risk for a variant. Additional information about the user's SNPs matched against the chosen PGS catalog entry scoring file is retrieved from the Ensembl database using an API request to <https://rest.ensembl.org>

3 Results

PGS scoring files are standardized compressed text files, typically composed of a few hundred rows, one per variant. The prototype application displays the scoring and genomic file as a decompressed text for direct inspection and an odds ratio plot of associated variants. To illustrate the characteristics and working of this application, the PGS scoring file PGS000001, listing 77 variants associated with breast cancer, was matched for a public 23andme report data file. Figure 1 reproduces that PRS calculation, with specific reference to the sources and process detailed in that figure's legend.

In summary, in the example depicted in this figure, the underlying multivariate relative risk score reported by PGS#1 (breast carcinoma) and calculated for 'Dorothy Wolf' 23andme result is 1.351. This calculation considered 77 variants listed in the PGS catalog #1, of which 17 were found to occur amongst the 43 that matched Dorothy's 23andme data covered 40% of the risk score reported by the reference publication.

4 Discussion

As advised by the FAIR principles, research genomics tools should be open-source and either run on the client side or under its governance, in order not to raise obstacles to reusability (Wilkinson *et al.* 2016). The PRS calculator relies solely on published/validated/open-source risk scores from the PGS catalog, whereas consumer genomics companies use proprietary

algorithms that cannot be readily reproduced by others. Additionally, we observed that the inherent reusability of this design led to the use of the PRS calculator as a research tool to compare different PGS catalog entries, by reviewing the in-browser SNPedia and dbSNP analysis. It should be noted that this calculator was devised as a proof of concept, and it has a number of important limitations for research or clinical use. The current tool uses pre-imputation genotype data and thus results in a low number of SNP matches. Most important, only the relative risk model (see in Chatterjee *et al.* 2016, Box 1, eq 2) is calculated, with no effort to determine absolute risk of disease. Finally, the risk calculation will only be attempted for PGS catalog entries reporting effects as beta values under 'effect_weight'. Future work enabling the use of imputed or sequencing data (of much larger size) covering variation across the whole genome, and integration of raw PRS into risk models, is needed. Further considerations are also required prior to clinical use of PRS (Hao *et al.* 2022).

5 Conclusion

The prototype PRS calculator reported here is a web application developed as an exercise in privacy-preserving personalized risk computing. Specifically, it probes the logistic limits of what a commodity mobile device can realistically compute. We found that, up to a few thousand SNPs, web computing (in-browser) provides an effective avenue for researchers to assess their contributions to PGS catalog in a manner that addresses the FAIR principles (Findable, Accessible, Interoperable, Reusable) (Wilkinson *et al.* 2016, García-Closas *et al.* 2023). Given the rapid advancement of web technologies, the software architecture implemented may also inform future logistics of computer-assisted Precision Prevention.

Author contributions

Jonas Silva Almeida (Conceptualization [equal], Software [equal], Supervision [lead], Writing—review & editing [equal])

Conflict of interest

None declared.

References

- Chatterjee N, Shi J, García-Closas M *et al.* Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 2016;17:392–406.
- Collister JA, Liu X, Clifton L *et al.* Calculating polygenic risk scores (PRS) in UK Biobank: a practical guide for epidemiologists. *Front Genet* 2022;13:818574.
- García-Closas M, Ahearn TU, Gaudet MM *et al.* Moving toward findable, accessible, interoperable, reusable practices in epidemiologic research. *Am J Epidemiol* 2023;192:995–1005.
- Folkersen L, Pain O, Ingason A *et al.* Impute.me: an open-source, non-profit tool for using data from direct-to-consumer genetic testing to calculate and interpret polygenic risk scores. *Front Genet* 2020; 11:578.

- Hao L, Kraft P, Berriz GF *et al.* Development of a clinical polygenic risk score assay and reporting workflow. *Nat Med* 2022;28:1006–13.
- Horton R, Crawford G, Freeman L *et al.* Direct-to-consumer genetic testing. *BMJ* 2019;367:l5688.
- Lambert SA, Gil L, Jupp S *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* 2021;53:420–5.
- Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. *Nat Commun* 2020;11:5900–19.
- Wilkinson MD, Dumontier M, Aalbersberg IJJ *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.