Cell Genomics

CellPress
OPEN ACCESS

Preview

Polygenic score development in the era of large-scale biobanks

Vincent Plagnol[1,*]
[1]Genomics plc, King Charles House, Park End Street, Oxford OX1 1JD, UK
*Correspondence: vincent.plagnol@genomicsplc.com
https://doi.org/10.1016/j.xgen.2021.100088

In this issue of *Cell Genomics*, Xu et al. report a comprehensive analysis of the genetics of 26 blood cell traits, leveraging data from two large biobanks to construct and make available machine-learning optimized polygenic scores (PGSs). In addition to delivering insights into the biology and clinical associations of these traits, the authors evaluate and provide recommendations on methods for PGS construction.

Blood measurements are ubiquitous in healthcare and key biomarkers for many diseases. Their broad availability has enabled the construction of large cohorts combining blood traits with genetic data, which in turn have powered extensive genome-wide association studies (GWASs). These GWASs[1] have identified more than 17,000 genetic associations for a range of blood measurements. Polygenic scores (PGSs; or a polygenic risk score [PRS] when targeted to a disease trait) are constructed to estimate the shared effects of many genetic variants on a phenotype. While PRSs are typically used in a context of identifying individuals at risk for common diseases, PGSs can also be used to quantify the contribution of genetic factors to quantitative traits. Now, Mike Inouye and colleagues[2] report, in this issue of *Cell Genomics*, a comprehensive evaluation of methods for PGS construction for 26 blood cell traits. They provide insights into the genetic basis of blood cell traits, develop and evaluate a range of predictive methods for PGS construction, and provide a resource of optimized PGSs. More broadly, their work guides on optimal methods in constructing PGSs using either summary or individual-level genetic datasets.

Prior GWASs in large-scale cohorts, including UK Biobank,[3] have demonstrated that blood cell traits are heritable and have identified numerous genetic associations. The most recent large-scale meta-analysis performed by Vuckovic et al.[1] have assembled a list of ~17,000 associations for 29 blood cell phenotypes at more than 7,000 genomic loci. In the current study, in order to assess the predictive ability of these genetic associations, Xu et al.[2] analyze measurements for 26 blood traits available in the UK Biobank[3] and INTERVAL[4] datasets. They use a range of methodologies to construct PGSs using UK Biobank data for training and then evaluate these PGSs in the INTERVAL cohort, which plays the role of an out-of-training evaluation cohort.

The evaluation in the INTERVAL dataset confirms the substantial contribution of genetics to the inter-individual variability of these blood traits. The Pearson correlation coefficient in the INTERVAL evaluation cohort between PGS and blood trait ranged from ~0.17 (for basophil percentage of white cells) to ~0.6 (for mean platelet volume). Using the INTERVAL dataset and a regression analysis with a sex interaction term, the authors identified significant sex-specific differences in PGS effect sizes for 10 of the 26 blood traits. For example, one standard deviation of the PGS for hemoglobin concentration predicted a mean difference of 1.48 g/dL in men and slightly more than 2 g/dL in women.

The predictive power of genetics for blood traits matters because this genetic component has the potential to alter the interpretation of blood tests. Specifically, does only the absolute value of the biomarker matter, or is there benefit in considering the difference between the measured blood cell trait and its genetically predicted value? Relevant to this question, the inclusion of a PGS for HBA1c levels has been shown to improve type 2 diabetes (T2D) diagnosis by altering the HBA1c threshold for T2D diagnosis for each individual.[5] Hence, the larger the predictive power of PGS is for a given blood trait, the greater the opportunity to refine the clinical interpretation.

To learn about shared genetics between diseases and blood cell traits, Xu et al.[2] also evaluated correlation between their PGSs with six common disease PRSs (asthma, allergy, coronary artery disease, Crohn's disease, rheumatoid arthritis, and schizophrenia). The authors found several statistically significant correlations capturing known associations between the actual traits, such as between asthma and eosinophil count/eosinophil percentages, or white blood cell count and Crohn's disease. Among the intriguing findings, the association between the monocyte count PGS and the schizophrenia PRS supports a role of inflammation in the etiology of schizophrenia and should warrant additional explorations.

Xu et al.[2] also provide useful insights into the performances of existing methodologies for PGS construction by evaluating six PGS methodologies. These include the method of pruning and thresholding (P+T), a current standard in PGS construction, as well as five supervised learning methods: LDpred2,[6] elastic net (EN), Bayesian ridge (BR), multilayer perceptron (MLP), and convolutional neural network (CNN). To support further evaluation and benchmarking, the authors made their code to construct these PGSs available on the GitHub public repository (https://github.com/xuyu-cam/PGS-BC-Traits-Using-ML-DL). They also released the PGS models for the 26 blood traits in the PGS catalog.[7]

Three key methodological features of this evaluation should be highlighted prior to stating its results. First, the INTERVAL biobank was used as an ''out-of-training'' evaluation cohort, allowing the authors to spot when methodologies overfit to the training cohort (UK Biobank in this case). Second, the variant selection step that is performed prior to running the machine-learning toolkit includes the selection of interaction terms between variants, which are handled elegantly by the EN and BR methodologies and most likely contribute to their effectiveness. Because the concept of interaction or genetic epistasis is often confusing and hard to interpret, it should be noted that these interaction terms may well be capturing predominantly statistical concepts about missing variant or haplotype information rather than biologically meaningful interacting variants. Third, in a final analysis, the authors alter the complexity of the training space by tuning the number of variants included in the model. They then evaluate the impact of this variant space on the performance of the three methods that best scale to large variant sets (EN, LDpred2, and P+T).

What do we learn from this methodological evaluation? First, and consistent with other PGS evaluation work,[8] the predictive performance of machine-learning methodologies (LDpred2, EN, BR) exceeded the performance of the simpler P+T methodology. Second, more complex models including non-linear terms (e.g., CNN and MLP) did not provide additional predictive benefit. Hence, the incredible excitement around the development of deep-learning methodologies for multiple applications has yet to translate into predictive benefit for PGS construction.[9]

The third, and most useful, lesson of this evaluation is that under specific conditions, the performance of EN can exceed the performance of LDpred2. This performance gap becomes measurable when two things jointly happen: use of a larger genetic variant space and evaluation of the predictive model on the out-of-training INTERVAL evaluation set. A potential interpretation of this finding is that when performing training on a complex variant space, LDpred2 appears more prone to overfitting than EN. This result is intriguing because LDpred2 is one of the most commonly used methods for PGS construction. These findings provide a baseline for future evaluation and benchmarking of PGS methodologies as well as additional motivation for the development of PGS methods that can leverage individual level data.

However, some caution is required prior to concluding that EN should become a method of choice for PGS construction. It is essential to note that, unlike the other methods evaluated here, LDpred2 and P+T are designed to be trained using GWAS summary statistics and require minimum access to individual-level data. Such training data are much easier to share than individual-level data. Close inspection of the results of Xu et al.[2] shows that the benefit of individual-level data methodologies over LDpred2 is often limited and trait dependent. Therefore, the added methodological flexibility may not overcome the reduced training sample size availability if summary statistics cannot be used. To some extent, the datasets for blood cell traits investigated here are atypical and favor methods based on individual-level data because of the extremely large sample size training set provided by UK Biobank. However, for most diseases, especially low-incidence diseases, prospective cohorts with individual-level data, even as large as UK Biobank, only provide a limited number of cases. Targeted case control cohorts will often be more valuable for PRS training but have shared individual-level data less frequently, hence putting more emphasis on summary statistics methodologies. Lastly, a method like LDpred2 requires substantial tuning and optimization. Its observed overfitting may reflect sub-optimum choices in variant filtering and parameter optimization. Different choices could narrow or even reverse the performance gap observed in this study.

While keeping in mind the caveats stated above, this study opens the door to the use of flexible methodologies based on individual-level data for PGS/PRS construction, provided that sufficient training data are available. While the work of Xu et al.[2] uses two biobanks that are extremely well powered for blood cell traits, PRS training for most diseases requires considerably larger sample sizes than what UK Biobank (or any other single biobank) can currently offer to be effective.

Is the perspective of individual-level data training across multiple biobanks realistic? While the current study was able to use individual-level data for blood cell traits in two biobanks, in general, the lack of individual-level data sharing across biobanks and large cohort studies has been a major obstacle for PRS training. The perspective of using federated cross-biobank datasets for training and evaluating PRS without having to set up complex data-sharing approaches is exciting and forward looking and has the potential to impact genomic prediction and medicine significantly by enabling the derivation of more predictive models.

However, the technical challenges associated with PRS training across multiple biobanks without requiring combining these datasets within a single computing instance are vast. As an illustration, recent efforts from the Global Biobank Meta-Analysis Initiative (GBMI[10]) still rely on an initial GWAS step, basing their main analyses on summary statistics rather than individual-level data. One can imagine technical solutions where PRS training happens independently in each biobank and the resulting insights (but not the raw data) are aggregated in a central computing instance. However, this vision remains a relatively distant future.

The future of PGS methodology development and, more broadly, genomic prediction is tied to the increasing availability of biobank and large-scale cohort datasets. The appropriate software infrastructure will need to be developed to maximize the value of the data. Depending on the trait of interest, individual-level data approaches may add value to summary statistics methodologies, but the optimal decision will depend on data availability together with the methodological path taken by the research community. The work of Xu et al.[2] contributes to this journey and will inform future steps in PGS methods development, with the end goal to support future clinical use cases.

DECLARATION OF INTERESTS

Vincent Plagnol is a full-time employee of Genomics plc.

REFERENCES

1. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al.; VA Million Veteran Program (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell *182*, 1214–1231.e11.
2. Xu, Y., Vuckovic, D., Ritchie, S.C., Akbari, P., Jiang, T., Grealey, J., Butterworth, A.S., Ouwehand, W.H., Roberts, D.H., Di Angelantonio, E., et al. (2022). Machine learning optimized polygenic scores for blood cell traits stratify sex-specific trajectories and identify genetic correlations with disease. Cell Genomics *2*, 100086-1–100086-12.
3. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
4. Moore, C., Sambrook, J., Walker, M., Tolkien, Z., Kaptoge, S., Allen, D., Mehenny, S., Mant, J., Di Angelantonio, E., Thompson, S.G., et al. (2014). The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. Trials *15*, 363.
5. Dornbos, P., Koesterer, R., Ruttenburg, A., Cole, J.B., Leong, A., Meigs, J.B., Florez, J.C., Rotter, J.I., Udler, M.S., and Flannick, J.; AMP-T2D-GENES Consortia (2021). A combined polygenic score of 21,293 rare and 22 common variants significantly improves diabetes diagnosis based on hemoglobin A1C levels. medRxiv. https://doi.org/10.1101/2021.11.04.21265868.
6. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.
7. Lambert, S.A., Gil, L., Jupp, S., Ritchie, S.C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., et al. (2021). The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat. Genet. *53*, 420–425.
8. Pain, O., Glanville, K.P., Hagenaars, S.P., Selzam, S., Fürtjes, A.E., Gaspar, H.A., Coleman, J.R.I., Rimfeld, K., Breen, G., Plomin, R., et al. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. PLoS Genet. *17*, e1009021.
9. Bellot, P., de Los Campos, G., and Pérez-Enciso, M. (2018). Can Deep Learning Improve Genomic Prediction of Complex Human Traits? Genetics *210*, 809–819.
10. Wang, Y., Namba, S., Lopera, E., Kerminen, S., Tsuo, K., Läll, K., Kanai, M., Zhou, W., Wu, K.-H., Favé, M.-J., et al. (2021). Global biobank analyses provide lessons for computing polygenic risk scores across diverse cohorts. medRxiv. https://doi.org/10.1101/2021.11.18.21266545.