Molecular Psychiatry
https://doi.org/10.1038/s41380-019-0399-z

CORRESPONDENCE

Uncovering the complex genetics of human personality: response from authors on the PGMRA Model

Igor Zwir[1,2] · Pashupati Mishra[3] · Coral Del-Val[2] · C. Charles Gu[4] · Gabriel A. de Erausquin[5] · Terho Lehtimäki[3] · C. Robert Cloninger[1,6]

Received: 10 February 2019 / Accepted: 14 February 2019

Following publication of our two articles [1, 2], a critique of the methodology of Phenotype-Genotype Many-to-Many Relations Analysis (PGMRA) [1, 3, 4] questioned the validity of our results from the perspective of polygenic risk scores (PRS) [5]. We appreciate the importance of these questions, and here provide a concise discussion of the assumptions and mathematical constraints of both approaches. We thank this commentator and others who have discussed our articles with us for their thoughtful questions and critiques.

Complex phenotypes present several challenges for genome-wide association studies including the presence of epistasis, pleiotropy, and heterogeneity. We approached these problems in a data-driven fashion to test the hypothesis that the heritability expected from twin studies but unexplained by genetic studies is distributed in heterogeneous partitions of a complex trait, each with distinct genotypic-phenotypic associations. We designed a machine learning algorithm termed PGMRA [1, 3, 4] to identify naturally occurring partitions in the data in an unsupervised fashion. PGMRA first dissects genome-wide data and uncovers a genotypic architecture composed of sets of SNPs shared by subsets of individuals (i.e., SNP sets [3, 6]). Next, phenotypic data are independently organized into natural sets of features such as clinical manifestations [4], voxels of neuroimages [7], or personality traits [1, 2] in a phenomic-like approach [8]. Cross-matching of the two types of sets reveals multiple associations restricted to subgroups of individuals, thereby uncovering the genotypic-phenotypic architecture of a trait and accounting for its distributed genetic risk or propensity.

Both approaches, PRS and PGMRA, rely on genome-wide markers (Fig. 1). However, PRS treats these markers as independent variables with additive effects, whereas PGMRA searches for sets of structurally connected markers, which may have interactive effects (epistasis). PRS assumes a global linear association model and relies on increasing sample size to improve performance [9, 10]. In contrast, PGMRA uncovers a family of models (i.e., SNP sets), each of which computes in a local partition of the data. Each model can be represented as either a linear combination of data (as in regression trees) or as a non-linear combination (as in some neural networks) [11]. Therefore, PGMRA uses a more complex model than PRS, focusing on incorporating more phenotypic variables rather than more individuals, but allows the use of smaller samples by reducing multiple comparisons.

PRS algorithms must reduce phenotypes to a single dependent variable because they use a linear supervised model [12]. In contrast, PGMRA uses an unbiased and unsupervised model to consider all possible phenotypic patterns common to a subset of individuals, regardless of their trait status (i.e., does not assign cases and controls a priori). Distinct patterns of phenotypic features can thus be

✉ C. Robert Cloninger
crcloninger44@gmail.com

1 Washington University School of Medicine, Department of Psychiatry, St. Louis, MO, USA

2 University of Granada, Department of Computer Science, Granada, Spain

3 Department of Clinical Chemistry, Fimlab Laboratories, and Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

4 Washington University, School of Medicine, Division of Biostatistics, St. Louis, MO, USA

5 University of Texas Rio-Grande Valley, School of Medicine, Department of Psychiatry and Neurology, and Institute of Neurosciences, Harlingen, TX, USA

6 Washington University, School of Arts and Sciences, Department of Psychological and Brain Sciences, and School of Medicine, Department of Genetics, St. Louis, MO, USA

Published online: 18 March 2019

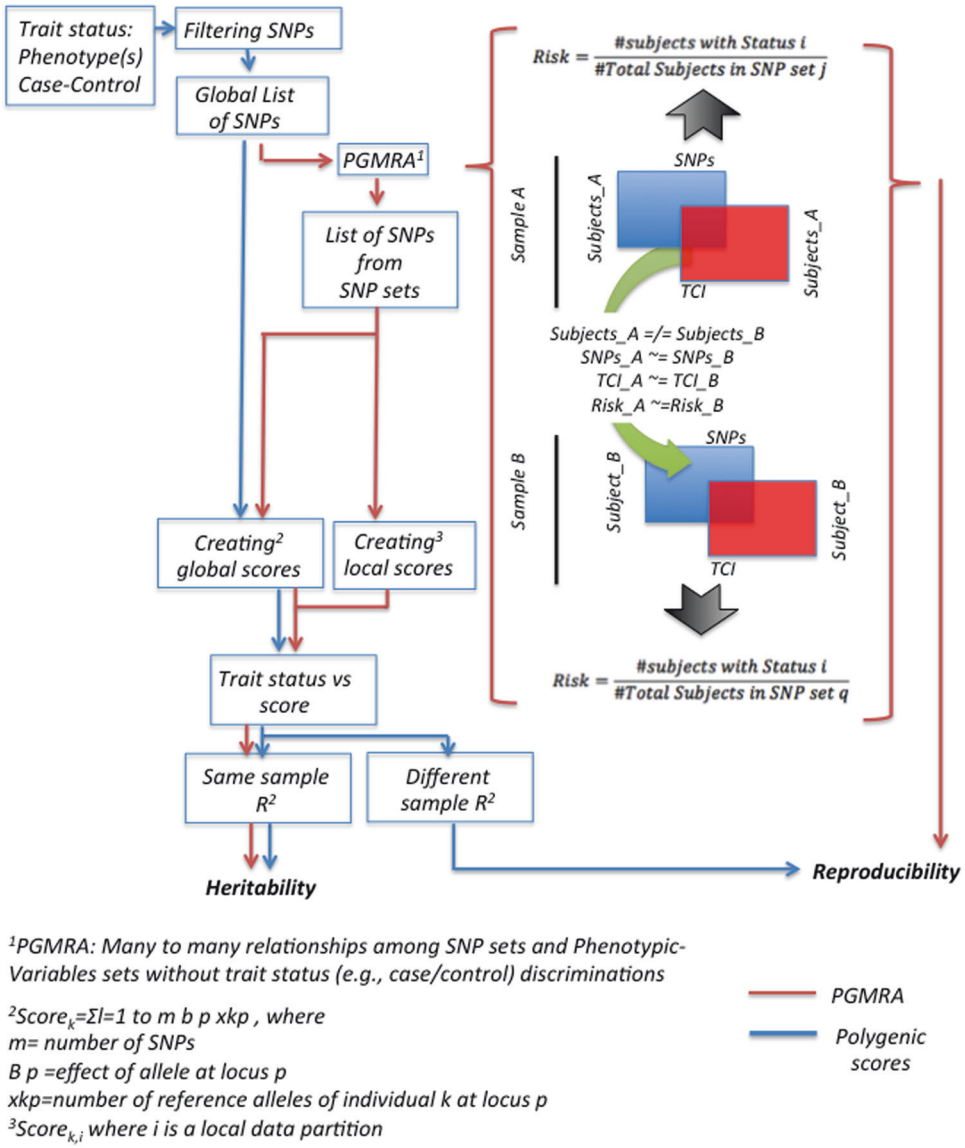

Fig. 1 Flow chart describing the common, as well as the different, roads followed by methods developed to build polygenic scores and the PGMRA method

associated with different SNP sets, thereby uncovering heterogeneous subtypes of the trait [1, 2, 4]. Finally, PGMRA incorporates trait status a posteriori to calculate the risk of such associations, and then independently tests the significance of the associations by a SNP-set Kernel Association Test [6, 13].

The validity of the replication procedure used by PGMRA was questioned too [5]. The "gold standard" approach used by PRS evaluates the reproducibility of an association by building a linear classifier trained in a *discovery* sample and testing it in a new sample assuming sample homogeneity [9, 10]. Homogeneity is a strong assumption that should be supported. By contrast, PGMRA uncovers genotypic-phenotypic associations for sample partitions and computes their corresponding risk or propensity post hoc; this process is blindly repeated independently for each new sample without assuming homogeneity within or across samples (Fig. 1). Then, similar genotypic-phenotypic associations across samples with comparable risk/propensity are uncovered using parsimonious models that balance accuracy with model complexity, thereby avoiding overfitting [11, 14, 15].

Inconsistent results obtained from applying PRS to heterogeneous samples [16, 17] has led to the suggestion of averaging scores from multiple samples [18] ignoring, at least in part, the phenotypic heterogeneity of the samples. When there is complexity derived from genetic, cultural, ethnic and environmental heterogeneity, the same global linear model is unlikely to predict across samples, especially when markers have relatively small effect [12, 16, 17]. Models learned independently in diverse samples allow analysis of replication across potentially heterogeneous samples, thereby providing a more stringent test of reproducibility [19, 20].

PRS calculates heritability as an adjusted R^2 from a global linear regression, which additively estimates variance explained by the markers. In the absence of a validated estimator of variance for "sets" of markers [6, 13], PGMRA used a similar approach (Fig. 1). For example, the estimated heritability of character, without controlling for outliers and jackknife resampling, in the Finns sample [1] was 45.67%. A criticism [5] questioned the lack of application of another sampling technique such as cross-validation. As suggested, we applied cross-validation within and across samples (e.g., R^2 of 10 k-fold is 45.05% with SD 0.049) and confirmed the observed results by bootstrapping (1,000 iterations, SE < 1.6%). We also found that the estimates of heritability for character in our paper [1] are conservative: the aggregation of the local variances explained by all SNP sets delivers a higher estimation of heritability ($R^2 > 15\%$) than the 45.67% described above (Fig. 1, unpublished results).

Some suggest that our sample size ($2126 + 972 + 902$ individuals from 3 cohorts, respectively [1, 2]) has insufficient power, even though others have calculated 80% power at nominal significance to detect heritability with the same sample size [12]. PGMRA computes genotypic-phenotypic associations based on "sets" of genotypes and "sets" of phenotypes, so the number of multiple comparisons are significantly reduced, making PGMRA less greedy of observations than PRS.

The nature of human beings embraces complex functions where every expressed gene may affect the function of any cell and their derived traits of our body in many different ways (many-to-many relationships). Complex traits are expected and known to be influenced by multiple genes acting in concert, not independently [21]. Most of the heritability in gene expression is determined by many genes far apart on the same or different chromosomes [21–23], whose effects are difficult to detect due to their small magnitude (e.g., trans eQTLs effects), as well as co-expressed genes that are vulnerable to decoherence in response to environmental perturbations [24]. PGMRA opens the door to develop new methods to explain complex genotypic-phenotypic relationships, including epistasis, pleiotropy and heterogeneous phenotypes, which present problems for PRS due to its restrictive linear model and doubtful assumption of homogeneity. Use of PGMRA would allow more thorough study of moderate-sized samples by efficient data-driven methods, which can help to bring methods of precision medicine into practice [1–3, 7, 20, 25].

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Zwir I, Arnedo J, Del-Val C, Pulkki-Raback L, Konte B, Yang SS et al. Uncovering the complex genetics of human character. Mol Psychiatry. (2018). https://doi.org/10.1038/s41380-018-0263-6. [Epub ahead of print].
2. Zwir I, Arnedo J, Del-Val C, Pulkki-Raback L, Konte B, Yang SS et al. Uncovering the complex genetics of human temperament. Mol Psychiatry. (2018). https://doi.org/10.1038/s41380-018-0264-5. [Epub ahead of print].
3. Arnedo J, del Val C, de Erausquin GA, Romero-Zaliz R, Svrakic D, Cloninger CR, et al. PGMRA: A web server for (Phenotype X Genotype) many-to-many relation analysis in GWAS. Nucleic Acids Res. 2013;41(Web Server issue):W142–9.
4. Arnedo J, Svrakic DM, del Val C, Romero-Zaliz R, Hernández-Cuervo H, Molecular Genetics of Schizophrenia Consortium. et al. Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome--wide association studies. Am J Psychiatry. 2015;172:139–53.
5. Derringer J. Explaining heritable variance in human character. bioRxiv. 2018:446518. https://doi.org/10.1101/446518.
6. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010;86:929–42.
7. Arnedo J, Mamah D, Baranger DA, Harms MP, Barch DM, Svrakic DM, et al. Decomposition of brain diffusion imaging data uncovers latent schizophrenias with distinct patterns of white matter anisotropy. Neuroimage. 2015;120:43–54.
8. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet. 2011;11:855–66.
9. International Schizophrenia C, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748–52.
10. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467:832–8.
11. Russell SJ, Norvig P. Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River, N.J.: Prentice Hall; 2010. p.pp xviii, 1,132.
12. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9:e1003348.
13. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93.
14. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc Natl Acad Sci USA. 2016;113:3932–7.

15. Deb K. Multi-objective optimization using evolutionary algorithms. 1st ed. Chichester, New York, John Wiley & Sons; 2001. pp. xix, 497.
16. Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol. 2011;35:506–14.
17. Feldman MW, Ramachandran S. Missing compared to what? Revisiting heritability, genes and culture. Philos Trans R Soc Lond B Biol Sci. 2018;373:pii. 20170064. https://doi.org/10.1098/rstb.2017.0064.
18. Krapohl E, Patel H, Newhouse S, Curtis CJ, von Stumm S, Dale PS, et al. Multi-polygenic score approach to trait prediction. Mol Psychiatry. 2018;23:1368–74.
19. Selzam S, Krapohl E, von Stumm S, O'Reilly PF, Rimfeld K, Kovas Y, et al. Predicting educational achievement from DNA. Mol Psychiatry. 2017;22:267–72.
20. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19:581–90.
21. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169:1177–86.
22. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. bioRxiv. 2018: 447367.
23. Boyle EA, Li YI, Pritchard JK. The omnigenic model: response from the authors. J Psychiatry Brain Sci. 2017;2:s8.
24. Lea A, Subramaniam M, Ko A, Lehtimaki T, Raitoharju E, Kahonen M et al. Genetic and environmental perturbations lead to regulatory decoherence. elife 2019;8:e40538.
25. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14:507–15.