Novel functional genomics approaches bridging neuroscience and psychiatry

Jose M. Restrepo-Lozano, Cecilia Flores, Patricia P. Silveira

PII: S2667-1743(22)00092-1
DOI: https://doi.org/10.1016/j.bpsgos.2022.07.005
Reference: BPSGOS 158

To appear in: Biological Psychiatry Global Open Science

Received Date: 27 January 2022
Revised Date: 15 July 2022
Accepted Date: 20 July 2022

Please cite this article as: Restrepo-Lozano J.M., Flores C. & Silveira P.P., Novel functional genomics approaches bridging neuroscience and psychiatry, Biological Psychiatry Global Open Science (2022), doi: https://doi.org/10.1016/j.bpsgos.2022.07.005.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc. on behalf of Society of Biological Psychiatry.
Novel functional genomics approaches bridging neuroscience and psychiatry

Jose M. Restrepo-Lozano\textsuperscript{1,2}, Cecilia Flores\textsuperscript{2,3,5\#}, and Patricia P. Silveira\textsuperscript{2,4,5\#}

\textsuperscript{1}Integrated Program in Neuroscience, McGill University, Montreal, Quebec, Canada, \textsuperscript{2}Douglas Mental Health University Institute, Montreal, Quebec, Canada, \textsuperscript{3}Department of Neurology and Neurosurgery, Faculty of Medicine, McGill University, Montreal, Quebec, Canada, \textsuperscript{4}Ludmer Centre for Neuroinformatics and Mental Health, McGill University, Montreal, Quebec, Canada, \textsuperscript{5}Department of Psychiatry, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.

\#Correspondence should be addressed to

Cecilia Flores, PhD, Department of Psychiatry, Douglas Mental Health University Institute, Perry Pavilion, Room 2111, 6875 LaSalle Boulevard, Montréal (Verdun), Quebec H4H 1R3, Canada; E-mail: cecilia.flores@mcgill.ca

Patricia Pelufo Silveira, MD PhD, Department of Psychiatry and Ludmer Centre for Neuroinformatics & Mental Health, Douglas Mental Health University Institute, Perry Pavilion, Room 4116, 6875 LaSalle Boulevard, Montréal (Verdun), Quebec H4H 1R3, Canada; E-mail: patricia.silveira@mcgill.ca.

Running title: Novel genomic approaches in psychiatry

Keywords: functional genomics, polygenic risk score, gene expression, psychiatry, neuroscience
Abstract

The possibility to establish a metric of individual genetic risk to a particular disease or trait has sparked the interest of the clinical and research communities, with many groups developing and validating genomic profiling methodologies for their potential application in clinical care. Current approaches for calculating genetic risk to specific psychiatric conditions consist of aggregating GWAS-derived estimates into polygenic risk scores (PRS) which broadly represents the number of inherited risk alleles by an individual. While the traditional approach for PRS calculation aggregates estimates of gene-disease associations, novel alternative approaches have started to consider functional molecular phenotypes that are closer to genetic variation and are less penalized by the multiple testing required in genome-wide associations studies. Moving the focus from genotype-disease towards genotype-gene regulation frameworks, these novel approaches incorporate prior knowledge regarding biological processes involved in disease, and aggregate estimates for the association of genotypes and phenotypes using multi-omics data modalities. In this review we discuss and list different functional genomics tools that can be used and integrated to inform researchers and clinicians for a better understanding and diagnosis of psychopathology. We suggest that these novel approaches can help generate biologically-driven hypotheses for polygenic signals that can ultimately serve the clinical community as potential biomarkers for psychiatric disease susceptibility.
Introduction

Establishing potential high-risk scenarios prior to the onset of neuropsychiatric conditions could profoundly improve mental health trajectories worldwide by presenting an opportunity for timely interventions, especially during sensitive neurodevelopmental windows. Although the well established practice of inquiring about an individual’s family history when diagnosing physical and psychiatric conditions is a useful tool to indirectly assess potential heritable risk (1–3), an individual’s genomic profile could provide information to guide overall health management. However, the true value of genomic data relies on our understanding of the complex interaction between genes, environments, and lifestyle choices over time (4–6), and efforts to elucidate this complex interplay have the potential to help developing tools to assess disease susceptibility prior to symptom onset, informing preventive and therapeutic decisions.

Current genotyping technology allows the identification of inherited DNA differences in the order of millions – mostly in the form of single nucleotide polymorphisms (SNPs) – across a given population and in a rapid and affordable manner (7). As a result, studying genotype-phenotype associations went from interrogating a few carefully selected candidate genes at a time to unbiased genome-wide surveys, with constant increases in sample sizes leading to the identification of more and more genetic loci that could modify risk for a given disease (8). Although this systematic interrogation of genomes yielded several loci reliably associated with an increased risk for psychiatric phenotypes, linking such loci to specific biological functions remains a challenge, in particular because most identified genome-wide significant associations lie in noncoding portions of the genome and require fine-mapping resolution to determine the
real causal variants implicated (8–10). Establishing a neurobiological framework underlying psychiatric risk will require a multi-omics data integration approach, with the purpose of mapping the molecular processes linking genomes and disease-relevant phenotypes (11). Such frameworks may ultimately help improve models of disease-risk prediction based on genomic profiles and provide actionable insights for clinical decision making. In this review we discuss emerging genomic risk assessment approaches in psychiatry, emphasizing methods that explore the neurobiological mechanisms by which gene networks contribute to psychiatric phenotypes.

GWAS as the basis for mapping genetic susceptibility to psychiatric phenotypes

To date, the most common population-based method to find genotype-phenotype associations is the performance of Genome-Wide Association Studies (GWAS, see (12,13)), which has successfully helped identifying genomic variants associated with increased risk to develop different psychiatric conditions (9,14–16). Essentially, GWAS studies entail the assessment of millions of variants across many individuals to detect those statistically associated with a specific phenotype. The primary outcome of GWAS typically includes a list of tested variants together with their respective effect sizes. Then, after identifying the relationship between the phenotypic variance and each genotype by means of a linear (for continuous) or logistic (for binary outcomes) regression, significant loci can be functionally annotated for post-GWAS analyses (Figure 1A).

Psychiatric genomics studies for conditions like schizophrenia (17) and depression (18) have yielded > 100 robustly associated risk loci, with ~43.7% and ~8.9% of heritability explained by common SNPs, respectively. The remarkable collaborative effort from the Psychiatric Genomics Consortium (PGC) has helped generate important discoveries in the identification of risk-
conferring variants as well as in advancing our understanding of the genetic architecture across 11 psychiatric disorders (17,18,27,19–26).

GWAS-derived quantified effects of common human variation have translated into different clinical applications. For example, using data derived from human genetics studies has improved the successful development of novel drugs (28,29). Another application, central to this review, is the calculation of polygenic risk scores (PRS), which aims to predict the contribution of an individual’s genomic profile to a given trait or disease (16,30–35). The possibility to establish a metric of individual genetic risk to a particular disease or trait has sparked the interest of the clinical community, with many researchers now investigating and exploiting the utility of PRS-profiling in clinical care (for examples see (34) or (36)).

Aggregating GWAS-derived signals into Polygenic Risk Scores- a proxy for genetic liability to psychiatric traits

For many years, studies in psychiatric genetics used a candidate gene approach, investigating the role of SNPs in particular phenotypes (see for example (37) where a specific mutation in HTR2B gene was associated with increased impulsivity). However, this approach to study the contribution of common variants to psychiatric phenotypes required a prior defined SNP target that was arbitrarily selected, albeit with very few exceptions. Indeed, conditions like Huntington’s disease (38) are caused by large effect variants, and there is a marked increase in risk for Alzheimer’s disease (although not a determinant of the disease itself) in people with the isoform e4 of the APOE gene (39). But Huntington’s and AD are neurological conditions with a more
defined clinical phenotype compared to psychiatric conditions like mood disorders, where the
degree of polygenicity is even more evident. Nowadays the candidate gene approach is outdated
as it has failed to yield useful insights for psychiatry (see (8) for a perspective on how GWAS made
candidate gene studies obsolete). Current psychiatric genetics studies use an unbiased
examination of the genome, as a continuously growing body of evidence established the highly
polygenic architecture across disorders, with many small-effect risk loci distributed across the
entire genome (40–43). As psychiatry gradually adopted a more probabilistic and risk-oriented
mindset, evidence for a concept that could explain a significant proportion of heritability in
independent target samples, based entirely on inherited DNA differences, began to emerge
(44,45).

Current methodologies for PRS calculation in psychiatry and important considerations to obtain
meaningful genetic signals
In principle, all methods for PRS calculation provide an estimate of an individual’s genetic
susceptibility to a trait, by aggregating the GWAS-derived effect size estimates into an indexed
score, as shown in Figure 1B (for detailed PRS tutorial see (32), for detailed PRS review see (46)).
The classic method for PRS calculation uses clumping/pruning and thresholding (C/P + T method),
to prune out SNPs in high LD and apply varying stringencies to p-value thresholds that can be
higher than genome-wide significance, to calibrate and maximize predictability (30,46,47).
Essentially, SNPs with p-values below an established threshold will keep the original estimate of
their effect size, while SNPs with higher p-values are excluded from the PRS, shrinking their effect
sizes to 0. This process can be carried out iteratively, using a range of p-value thresholds, with
the resulting PRSs tested for an association with the target trait in a test sample, determining the optimal p-value in a forward selection method (48,49). Other methods for PRS calculation are based on Bayesian frameworks where the shrinkage of all SNPs is based on a prior distribution specification (for more details see (50,51)). One example that seems to be particularly suitable to calculate PRSs for psychiatric disorders (52) is the Bayesian Multiple Regression summary statistic ($SBayesR$) (53), which can use publicly available GWAS summary statistics while employing prior distributions of alternative genetic effects and analyzing all SNPs together, accounting for their pattern of co-inheritance.

Ideally, a PRS can serve as a tool to stratify the population in terms of disease risk, as this can help decide on potential follow-up actionable measures such as therapeutic interventions, more in-depth screening, or life-style modifications. One of the earliest examples of a successful PRS came in 2009, when the International Schizophrenia Consortium published an aggregated polygenic signal derived from a GWAS that could predict risk for both schizophrenia and bipolar disorder (44). As the sample size for the schizophrenia GWAS increased, the phenotypic variance explained by the aggregated polygenic signal also increased. Current estimates indicate that individuals with PRS in the top 10% and top 1% of the population have an approximate 3-fold and 6-fold increase in their risk for developing schizophrenia, respectively, compared to 1% baseline risk when selecting someone randomly from the population (17,54). Another example comes from the study of Desikan et al., (2017) (55), where the researchers calculated a PRS based on a large Alzheimer’s disease GWAS meta-analysis (56) to investigate the PRS’s predictability of age-specific risk for developing the disease. By combining epidemiologic data on population-based
incidence rates and PRS scores, they found that individuals in the highest PRS quartile developed 
AD at a lower age and showed the highest yearly AD incidence rate. This finding was then 
replicated in other independent cohorts, where the PRS was associated with known 
neurodegenerative markers and with age of disease onset (55).

It is important to note that existing GWAS are predominantly performed in individuals with 
European ancestry. Missing genetic effects present in other populations and genetic variants with 
very low frequency may dramatically decrease the accuracy of a PRS. This is especially true when 
the ancestry of the target sample does not match the population of the original GWAS (57,58). 
In addition, it has been shown that PRSs work better when considered in combination with other 
clinical risk factors, with a joint model improving overall disease risk calculation, the identification 
of individuals that can benefit from early diagnosis, and predictive accuracy (55,59–62). 
Prediction is a difficult task, and most GWASs necessitate many millions of individuals to allow 
PRSs to achieve higher discriminatory power and to reach the upper bound of their predictive 
performance (i.e. heritability estimates) (33). Some groups have started to propose alternatives 
to investigate polygenic signals in psychiatry, considering phenotypes closely linked to genetic 
variation and, therefore, more directly affected by it.

From Genetics to Functional Genomics: PRS methodologies that go beyond the link between 
geneic variability and psychiatric traits by addressing biological mechanisms/functions 
The PRS methodologies described so far have been useful tools for clinicians and researchers, but 
one common characteristic is the agnosticism when it comes to the biological functions
implied in disease risk. In the classic GWAS-PRS methods, the first step consists of identifying statistically significant genetic associations such that afterwards, while conducting post-GWAS work, the biological functions implicated in those gene-disease associations can be dissected (see Figure 1A) and further explored as potential therapeutic avenues. But another way to investigate the role that genes play in disease (together with their associated transcripts, proteins, and epigenomes) is to first identify disease-relevant biological processes and functions to create PRSs that somehow capture and quantify those functions, to then test their association with disease (see Table 1 listing methods for functional PRS calculation). Moving the focus from genotype-disease towards genotype-gene regulation frameworks, we review below these novel methodologies and resources used by some groups to guide the selection of variants and phenotypes, emphasizing those that take into consideration (i) meaningful networks of genes co-regulated (or co-expressed) with spatiotemporal specificity, and (ii) highly quantifiable phenotypes, such as transcriptomic or epigenomic data. We suggest that these approaches can help generate biologically driven hypotheses for polygenic signals that can ultimately serve the clinical community as potential biomarkers for disease susceptibility.

Genotype-disease effects are small for most common genetic variation, but the fact that a large proportion of disease risk can be explained by variants that modulate gene expression levels (9,63,64) is intriguing and may provide clues for the cellular and biological mechanisms underlying disease (65,66). The transcriptome-wide association studies (TWAS) methodology was developed with the goal of detecting associations between measured or predicted levels of gene expression and particular traits (see Figure 2A) (67). For example, in the study of Girgenti et
al (2021) (68), researchers used the Million Veteran Program PTSD GWAS dataset to impute gene expression and identify genes significantly associated with PTSD risk and illness state, uncovering novel functional signals that confer genetic liability for PTSD. This method provides key advantages with respect to GWAS. First, using a gene-based approach reduces the burden of multiple-testing, prevalent in other SNP-based approaches. There are ~20,000 genes for which one can impute their transcript levels. Although still large, this number is considerably smaller compared to a typical GWAS consisting of several million SNPs that are individually tested for an association with a given trait. By incorporating functional information about the regulation of gene expression, this method can help uncover the underlying biological mechanisms affecting a trait. Another advantage of this method is to facilitate the interpretation of the direction of the effect. A gene-based signal which includes the direction of the effect is highly amenable to systems biology approaches, because if the increased (or decreased) expression of a gene is associated with a particular trait, the information can be easily incorporated into pathway or network analyses, making the interpretation of results more straight-forward, especially when compared to SNP-based signals. This approach is nonetheless limited regarding the tissue accessibility in study participants, in particular if the tissue of interest is a specific brain region. To this end, more novel computational frameworks like the probabilistic transcriptome-wide association study (PTWAS) can be of help, because it can predict gene expression from genotypes and investigate causal relationships between tissue- or cell-type-specific gene expression and complex traits (69).
Many research groups are actively developing tools to predict the transcriptional effects of genetic variation (see (70–72) for some examples), most likely driven by similar motivations: a unidirectional effect (from genes to gene expression) that ultimately narrows the gap between genetic variation and disease. One of the most prominent examples of this type of work is the PrediXcan methodology, which developed a machine learning algorithm to predict tissue-specific gene expression based on genomic profiles (73). Using genotype and gene expression data from the Genotype-Tissue Expression (GTEx) project (74), this method generates a database where, in a tissue-specific manner, transcript levels can be predicted using as input the genotypic data from any target sample (Figure 2B). PrediXcan serves to calculate an endophenotype (genetically-regulated gene expression) that is known to drive biological processes, to test for associations with a particular trait (for the entire data repository, and the PrediXcan family of methods see (75)). The more novel version, MultiXcan, can help investigate the mediating role of gene expression on many complex traits, using only summary statistics from publicly available GWASs (76).

The predicted gene expression approach has been applied to existing GWAS studies for bipolar disorder to identify novel risk-conferring genes, PTPRE and BBX, whose predicted transcript levels in whole blood and in the anterior cingulate cortex, respectively, were found to be associated with increased bipolar disorder risk (77). This study highlights the importance of gene expression to help understand the potential underlying mechanisms driving disease risk. However, this approach fails to simultaneously consider genes that are co-regulated as part of common biological processes, bypassing the established polygenicity of most psychiatric phenotypes.
Gene co-expression networks to inform polygenic metrics

As discussed previously, most identified genome-wide significant associations are devoid of a clear functional interpretation because they lie in noncoding portions of the genome, requiring fine-mapping resolution to determine the real causal variants implicated (9,10). Many of these non-coding disease-associated variants are regulatory in nature (a high proportion of these variants have been determined to be cis and/or trans-eQTLs, see (9)), suggesting they are likely affecting the expression of their associated genes, in the end placing gene expression as an imminent molecular phenotype linking genetics and disease. More crucially, however, disease-associated genes do not operate in isolation, but as part of complex networks that function with an exquisite degree of spatiotemporal specificity for precise biological processes. By operating under the assumption that functional groups of genes are co-regulated as part of specific molecular pathways, the identification of disease-relevant and tissue-specific gene networks provides a framework for mapping transcriptionally co-regulated processes into a type of polygenic score. This approach can potentially increase the likelihood of discovering psychiatrically relevant markers of disease (see (78)).

A study that aimed to determine genetic susceptibility to cognitive disability used an unsupervised genome-wide co-expression network analysis leveraging measurements of gene expression in human hippocampal tissue, with the goal of capturing modules of co-varying genes, which can ultimately provide clues for the molecular mechanisms driving the susceptibility (79). They identified a module of 150 genes with significant enrichment for 1) genes associated with
relevant cognitive phenotypes, 2) genes related to neural activity and synaptic processes, and 3) genes intolerant to mutations which, when mutated, are associated with intellectual disability (80,81). Another group carried a similar approach, but instead of using an unsupervised analysis, hypothesized that genes conferring risk to disease must translate into biological risk by acting as part of a co-regulated gene network on a measurable molecular phenotype, which could then be associated with the disease (82). They were interested in elucidating the genetic architecture of the D2 receptor molecular pathway, since genetic variation within the DRD2 gene has been linked with schizophrenia-related phenotypes, including response to treatment (17). Starting with human post-mortem tissue, the authors identified a prefrontal DRD2 co-expression network using Weighted Gene Co-expression Network analysis (WGCNA), and then defined potential SNPs in the form of eQTLs affecting the expression of the genes within the network. Combining these regulatory SNPs into a particular PRS (referred as “polygenic co-expression index”), the study captured the genetic component (eQTLs) of the expression of the network, and associated the PRS with brain activity measurements during working memory tasks. Finally, they found that individuals with a higher prefrontal cortex DRD2 co-expression PRS are predisposed to a less efficient working memory, which is a known risk-associated phenotype for schizophrenia. This study is an example of how identifying a disease/trait-relevant gene network can help generate hypotheses for novel type of PRSs based on biological frameworks.

Another innovative way to identify co-regulated biological processes underlying the genetic susceptibility to psychiatric conditions leverages data from the Genotype-Tissue Expression project (GTEx), to quantify the genotypic effect linked to gene expression across several tissues.
One such example is the method eMAGMA, which integrates both genetic and transcriptomic data to identify disease-specific risk genes, and test for their enrichment across different gene modules (83). This method can exploit a systems biology approach to generate polygenic signals that are essentially based on tissue-specific gene co-expression networks. A similar approach from our group has generated tissue-specific polygenic signals associated with traits/diseases (Figure 2C). We first identify co-expression networks using genome-wide gene expression data from a specific tissue, then map all SNPs within the co-expressed genes and eliminate those in linkage disequilibrium. We then assign to each SNP the weight of the association between alleles and gene expression estimated by GTEx (84), ultimately obtaining a set of SNPs that lie within a tissue-specific co-expression network, where each SNP is weighted by its estimated influence on gene expression. We can then identify all SNPs from the co-expressed genes in a test sample of subjects with available genotype data, and weight SNPs according to the GTEx. The derived expression-based polygenic signal (or “ePRS”) reflects variation in the expression of the gene network and can be calculated in target samples with available genotype data (85–89).

In a recent study, we investigated whether an ePRS based on corticolimbic-specific gene co-expression networks associates with impulsive phenotypes in children (90). We aimed at capturing individual variation in the molecular processes involved in the maturation of corticolimbic substrates, which are known to support inhibitory control. Similar to most studies using functional polygenic signals, we compared the predictive ability of the score against a conventional PRS derived from the latest GWAS for ADHD and found the ePRS to be a better overall predictor of impulsivity. This type of polygenic signal did not suffer from the
generalizability problems seen in other polygenic score methodologies, as the experiment was conducted in three ethnically diverse cohorts, all showing similar effects. This approach exploits the fact that genes engage within complex networks for precise biological functions, and they do so with a remarkable tissue-specificity. Based on knowledge of the neurobiological processes of brain development, this score aimed to predict psychiatric-relevant phenotypes.

Gene × environment interplay - quantifying environmental influences and their interaction with multi-omics data

One of the biggest challenges faced by researchers studying models of disease risk prediction is to develop a methodology to accurately represent an individual’s “environment” in a quantitative metric. Similar to how a functional PRS can represent a restricted set of phenotype-relevant biological processes, some studies have narrowed down the “environment” variable to a composite score made up of clearly defined constructs (see (48,49,85,86,88,89,91) for examples). By doing so, researchers can start investigating the interplay between genes and environments while also assessing potential ways in which genetic and environmental effects interact (Figure 2D). Although not in psychiatry, the study by Belsky et al., (2018) (92) is a good example of how an individual’s environment can exert a powerful influence in his/her socio-economic attainment. In this study the authors tested whether a PRS based on a GWAS for educational attainment (which is currently one of the PRSs with highest predictive value; see (35)) could predict socioeconomic mobility (i.e., any shift in a person’s social class relative to their parents). While higher PRS scores did predict more socioeconomic success compared to parents and siblings, additional analyses revealed that a mother’s polygenic score for educational attainment predicts
above and beyond the child’s own polygenic score, suggesting an environmentally-mediated genetic effect.

Some studies have started to integrate epigenetic data into genome-wide scores with the goal of identifying individuals that might be at an increased risk for psychiatric phenotypes. For example, given the association between early life stress and behavioral and psychiatric problems later in life (93), the study of Provençal et al (2020) (94) assessed differentially methylated sites following exposure to glucocorticoids in a human hippocampal progenitor cell line and in human blood cells. In addition, a subsequent glucocorticoid exposure induced important transcriptional changes. The overlapping differentially methylated sites were then used to calculate a weighted polyepigenetic score, which was proposed as a potential biomarker for conditions associated with prenatal glucocorticoid exposure in newborns. The calculated score was applied to newborns’ cord blood DNA (N=817), with the GC-responsive score significantly associated with levels of maternal anxiety and depression, suggesting that early life stress induces lasting epigenetic changes that can ultimately modify the vulnerability to stress exposure in later years.

Concluding remarks

As the field of psychiatric genomics continues to evolve, so will the models of disease-risk prediction based on strong biological foundations. Advances in big data availability and complexity (e.g. longitudinal studies like the ABCD cohort, deep phenotyping like the UK Biobank), mapping the developmental trajectories and including a wealth of data in large numbers of individuals will benefit the understanding of factors that ultimately play an important
role in determining mental health. Polygenic scores and polyepigenetic scores by themselves – like any other marker - have a limited capacity to predict with perfect accuracy the condition for which they were generated. It should be noted, however, that the optimal selection of genetic variants and other genomic markers and the aggregation of their associated weights are active areas of research (50,51,53). The continued improvement of the technology (increases in GWAS sample size and incorporation of different ancestries, higher genotyping resolution, etc.,) entails continued revision of the guidelines for their calculation and interpretation. Due to the recency for the appearance of several methods discussed in the present review, evidence for their clinical utility is still lacking, but as the technology driving functional genomics approaches continues to improve, we expect researchers and clinicians to be encouraged to investigate/test their clinical utility in psychiatry. One can assume that some of the methods highlighted here will be replaced by newer approaches. However, incorporating functional aspects rather than being informed exclusively by data-driven approaches is our core message.

Although it is highlighted that functional PRS can focus on a particular network/system, efforts should be made to maintain a genome-wide platform for unbiased querying of the relevant signals (e.g. genome wide RNA sequencing). Moreover, one of the advantages of the functional methods is to provide tissue-specific information, but this can be challenging for certain research questions (e.g. epigenetics markers collected from peripheral studies inferring brain mechanisms). Finally, brain gene expression data in humans is postmortem, and is limited in numbers, ancestry and developmental stage representation. These features can influence gene expression and therefore can bias the generation of functional PRS.
Training of the clinical workforce on how to handle and communicate genome-wide information is an issue that is becoming more pressing with time. Commercially available and direct-to-consumer genotyping services, which allow users to download their genotype data, are already reporting PRS for some traits and users can upload their data into other online PRS calculators (95). It is important to clearly communicate to the public the utility but most importantly the limitations of PRS profiling, in particular driving away the idea that genetic testing can accurately predict every aspect of a person’s health, as they have inherent limitations similar to other tests commonly used in clinical settings (96).

Future efforts in disease-risk prediction should aim at integrating data at multiple levels, aggregating genomics, epigenomics, transcriptomics, proteomics, and metabolomics data into predictive models. Some of the examples presented in this review highlight the significant contribution from each of these data to disease liability. In addition, models should be able to assess the role of the environment at multiple levels (person, family, community), since all biological processes occurring within an individual are physically contained processes functioning together as part of society, including household, neighborhood, school, and work, etc., (4). The technological advance should occur in parallel to societal progress in overcoming the menace of racism and structural inequalities, which still are an unfortunate reality that has a major impact on mental and physical health. The ability to combine multilevel biological information with the constant changes in a person’s environment for overall health risk assessment, in trusted
clinician-patient relationships with joint decision-making, can revolutionize the diagnosis and early prevention of psychopathology.

Acknowledgments

This work was funded by the National Institute on Drug Abuse (CF Grant number: R01DA037911), the Canadian Institutes of Health Research (CF Grant Number: MOP-74709; PPS Grant Number: PJT-166066, and PPS Grant Number: PJT-173237), the Natural Science and Engineering Research Council of Canada (CF Grant Number: 29822268; PPS Grant Number: RGPIN-2018-05063), a Doctoral Training fellowship from the Fonds de Recherche du Québec en Santé to JMRL, and a Healthy Brains for Healthy Lives Graduate Student Fellowship to JMRL. This publication is the work of the authors and Cecilia Flores and Patricia Pelufo Silveira will serve as guarantors for the contents of this paper.

All figures were created using BioRender.com

Conflict of interest: The authors report no biomedical financial interests or potential conflicts of interest.
References


7. Perkel J. SNP genotyping: six technologies that keyed a revolution. 2008;


1. 1068–87.


41. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to


54. Ripke S. PGC SCZ WORKGROUP: GWAS WITH OVER 70,000 CASES AND 100,000 CONTROLS. Eur Neuropsychopharmacol. 2019 Jan 1;29:S814.


Home | PredictDB [Internet]. [cited 2021 Dec 27]. Available from: https://predictdb.org/


Pergola G, Di Carlo P, D’Ambrosio E, Gelao B, Fazio L, Papalino M, et al. DRD2 co-expression network and a related polygenic index predict imaging, behavioral and clinical


88. Silveira PP, Pokhvisneva I, Parent C, Cai S, Rema ASS, Broekman BFP, et al. Cumulative prenatal exposure to adversity reveals associations with a broad range of
neurodevelopmental outcomes that are moderated by a novel, biologically informed polygenetic score based on the serotonin transporter solute carrier family C6, member 4 (. Dev Psychopathol. 2017 Dec 22;29(5):1601–17).


Figure 1. Overview of the steps involved when conducting a GWAS and creating a PRS. (A) Genotypic data from cases and controls or from a population-based sample is gathered to compare the proportion of specific alleles from each SNP among cases and controls, or to determine the linear relationship between genotypes and a continuous trait. After proper quality control of the genotype data and determining the underlying population structure in the sample, a statistical analysis is conducted to investigate whether the observed allele proportions (for case control-studies) or relationships (for continuous traits) deviate
significantly from expected values at each SNP, correcting for the number of tests applied. When an allele is found in the cases more frequently than it would be expected by chance, it is reported as a candidate SNP for the entire haplotype block, together with its estimated effect size which quantifies the increased odds of having the disease, per risk allele count. For continuous traits, the regression coefficient will determine the effect size attributed to the “effect” allele. Ideally, the observed GWAS signal should be replicated in an independent cohort, to minimize false-positives and to calibrate the effect sizes attributed to all SNPs. Genome-wide signals (shown in a Manhattan plot) that have been replicated are typically further investigated during post-GWAS work, which consists of i) fine-mapping the genomic region to find the true causal variant, ii) investigating the tissues/cell types where the variant is known to be active, iii) determining the genes that are impacted by the variant, and iv) identifying the molecular pathways implicated. (B) Using a base and a target dataset, the GWAS-derived estimated effects can be applied to a target sample for which genotype data is available. The calculated PRS is an aggregated score of the individual-level genotype weighted by the SNP effect sizes described in a discovery GWAS, resulting in a normally distributed score in the target sample. The distributions depicted in (B) reflect raw standardized values of real PRS scores, which could be associated with a particular trait of interest.

Figure 2. Novel approaches to functional genomics. (A) The transcriptome-wide association study (TWAS) consists of associating measured (or predicted) gene expression data with a disease/trait, making this a gene-based rather than a SNP-based association study, considerably reducing the number of multiple comparisons while also providing insight into the potential biological
mechanisms driving disease risk. (B) Models that predict gene expression based on genotype data. These models can be tissue or cell-type specific, allowing to test for the association of the cell/tissue-specific imputed transcriptome and a given disease/trait. (C) The expression-based polygenic risk score (ePRS) model starts with the identification of a gene co-expression network from RNA sequencing data, to identify co-regulated disease-relevant biological processes in specific tissues. Once a gene network is identified, it maps the genetic variation within the co-expression network in an independent target sample, to weigh the SNPs according to a GWAS (typically a GWAS of gene expression, e.g., GTEx). The resulting ePRS, which aims to capture individual variation in the expression of the gene network, can then be associated with a disease/trait, mapping the association between functional biological processes and the target phenotype. (D) A global model that incorporates all the potential intermediate phenotypes that can occur between genetic variation and disease, considering the feedback from the environment and lived experiences across all levels. This model, like the ones discussed above, generates a unidirectional effect that starts from individual genetic variability and provide alternative approaches to assess the effects of inherited DNA polymorphisms on particular traits. All these approaches can help generate biologically informed predictors of susceptibility to psychiatric-relevant phenotypes.
<table>
<thead>
<tr>
<th>Method name</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAGMA</td>
<td>Software tool for mapping genome-wide significant variants to genes and gene sets. Novel variations of this method (i.e., h-MAGMA and eMAGMA) are meant to refine the mapping of variants by incorporating long-range and tissue-specific interactions, and the enrichment of variants across different gene modules.</td>
<td>(1)</td>
</tr>
<tr>
<td>LDSC</td>
<td>Method that leverages GWAS summary statistics and LD scores from an external panel to distinguish between inflated effect sizes and true polygenic effects. This method is commonly used for determining genetic correlation between complex traits, partitioned heritability, and stratified heritability.</td>
<td>(2)</td>
</tr>
<tr>
<td>SMR</td>
<td>Method that integrates GWAS summary statistics and data from eQTL studies, allowing the user to identify and prioritize genes whose expression levels are associated with specific complex traits.</td>
<td>(3)</td>
</tr>
<tr>
<td>lassosum</td>
<td>Method to construct a PRS in a penalized regression framework that uses GWAS summary statistics and a LD reference panel.</td>
<td>(4)</td>
</tr>
<tr>
<td>LD-Hub</td>
<td>Centralized database of GWAS summary statistics that automates LDSC analysis pipeline, allowing the user to estimate SNP heritability and genetic correlation across complex traits.</td>
<td>(5)</td>
</tr>
<tr>
<td>ANNOpred</td>
<td>Bayesian framework for disease risk prediction that integrates genomic functional annotations using GWAS summary statistics and estimates LD from reference genotype data.</td>
<td>(6)</td>
</tr>
<tr>
<td>SBLUP</td>
<td>Method that re-scales SNP effect sizes using an external LD reference panel, converting the ordinary least-squares SNP estimates into approximate best linear unbiased predictions.</td>
<td>(7)</td>
</tr>
<tr>
<td>PRS-CS</td>
<td>Polygenic prediction method that infers posterior effect sizes of SNPs using GWAS summary statistics and an external LD panel. This model places a continuous shrinkage prior on SNP effect sizes.</td>
<td>(8)</td>
</tr>
<tr>
<td>JAMPred</td>
<td>Method for modeling polygenic risk using the joint analysis of marginal summary statistics (JAM) software, adjusting for local and for long-range LD. The computed polygenic risk predictions are</td>
<td>(9)</td>
</tr>
</tbody>
</table>
obtained through a Bayesian variable selection framework.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBayesR</td>
<td>Polygenic prediction method that adjusts SNP effect sizes based on Bayesian multiple regression model (BayesR), using GWAS summary statistics data.</td>
<td>(10)</td>
</tr>
<tr>
<td>LDpred-func</td>
<td>Probabilistic model for deriving PRS that accounts for LD and incorporates trait-specific functional priors to increase prediction accuracy. This model assumes a point-normal distribution as a prior.</td>
<td>(11)</td>
</tr>
<tr>
<td>LDpred2</td>
<td>Method for deriving polygenic scores using GWAS summary statistics and LD information from an external reference sample to infer posterior mean effect sizes of SNPs. Optimization of LD and p-value thresholds is achieved using a Bayesian framework for shrinkage of SNP effects. This model assumes a point-normal distribution as a prior.</td>
<td>(12)</td>
</tr>
<tr>
<td>PTRS</td>
<td>Method for calculating polygenic transcriptomic risk scores that can be applied as a gene-based complement to other PRS methods, as it does not outperform other current PRS technologies. This method can help improve portability across ancestries and facilitate interpretation of underlying biological effects.</td>
<td>(13)</td>
</tr>
</tbody>
</table>

1. Table 1. Overview of the different methodologies for PRS calculation

4 References


A

<table>
<thead>
<tr>
<th>Genetic variation</th>
<th>Association testing (+ QC)</th>
<th>GWAS signal</th>
<th>Post-GWAS work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls</td>
<td>Case-control studies</td>
<td>Quantitative Phenotypes</td>
<td>Fine-mapping causal variants</td>
</tr>
<tr>
<td>Cases</td>
<td></td>
<td></td>
<td>Tissue/Cell-type enrichment</td>
</tr>
<tr>
<td>OR</td>
<td></td>
<td></td>
<td>Implicated gene(s)</td>
</tr>
<tr>
<td>Population-based sample</td>
<td></td>
<td></td>
<td>Implicated pathway</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Target/Base data</th>
<th>PRS calculation (+ QC)</th>
<th>PRS results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual-level genotype</td>
<td>PRS = ( \sum \beta_i X_i )</td>
<td>N SNPs</td>
</tr>
<tr>
<td>SNPs effect size derived from GWAS</td>
<td></td>
<td>M Subjects</td>
</tr>
<tr>
<td>PRS(_1), PRS(_2), ...</td>
<td></td>
<td>1% percentile, 99% percentile</td>
</tr>
<tr>
<td>Computed PRS</td>
<td></td>
<td>SNPs</td>
</tr>
<tr>
<td>PRS(_1), PRS(_2), ...</td>
<td></td>
<td>Low risk allele, High risk allele</td>
</tr>
<tr>
<td>PRS(_1), PRS(_2), ...</td>
<td></td>
<td>Polygenic Score (PHS)</td>
</tr>
</tbody>
</table>