https://doi.org/10.1186/s11689-024-09571-8

Journal of
Neurodevelopmental Disorders

RESEARCH **Open Access**

Commonly used genomic arrays may lose information due to imperfect coverage of discovered variants for autism spectrum disorder

Michael Yao[1†], Jason Daniels[1,2†], Luke Grosvenor[1,3], Valerie Morrill[1,2], Jason I. Feinberg[1,3], Kelly M. Bakulski[4], Joseph Piven[5,6], Heather C. Hazlett[5,6], Mark D. Shen[5,6], Craig Newschaffer[7,8], Kristen Lyall[7], Rebecca J. Schmidt[9,10], Irva Hertz-Picciotto[9,10], Lisa A. Croen[11], M. Daniele Fallin[1,2,3,12], Christine Ladd-Acosta[1,2,3], Heather Volk[1,3] and Kelly Benke[1,3*]

Abstract

Background Common genetic variation has been shown to account for a large proportion of ASD heritability. Polygenic scores generated for autism spectrum disorder (ASD-PGS) using the most recent discovery data, however, explain less variance than expected, despite reporting significant associations with ASD and other ASD-related traits. Here, we investigate the extent to which information loss on the target study genome-wide microarray weakens the predictive power of the ASD-PGS.

Methods We studied genotype data from three cohorts of individuals with high familial liability for ASD: The Early Autism Risk Longitudinal Investigation (EARLI), Markers of Autism Risk in Babies-Learning Early Signs (MARBLES), and the Infant Brain Imaging Study (IBIS), and one population-based sample, Study to Explore Early Development Phase I (SEED I). Individuals were genotyped on different microarrays ranging from 1 to 5 million sites. Coverage of the top 88 genome-wide suggestive variants implicated in the discovery was evaluated in all four studies before quality control (QC), after QC, and after imputation. We then created a novel method to assess coverage on the resulting ASD-PGS by correlating a PGS informed by a comprehensive list of variants to a PGS informed with only the available variants.

Results Prior to imputations, None of the four cohorts directly or indirectly covered all 88 variants among the measured genotype data. After imputation, the two cohorts genotyped on 5-million arrays reached full coverage. Analysis of our novel metric showed generally high genome-wide coverage across all four studies, but a greater number of SNPs informing the ASD-PGS did not result in improved coverage according to our metric.

Limitations.

The studies we analyzed contained modest sample sizes. Our analyses included microarrays with more than 1-million sites, so smaller arrays such as Global Diversity and the PsychArray were not included. Our PGS metric for ASD

[†] Michael Yao and Jason Daniels contributed equally to this work.

*Correspondence:
Kelly Benke
kbenke1@jhu.edu
Full list of author information is available at the end of the article

is only generalizable to samples of European ancestries, though the coverage metric can be computed for traits that have sufficiently large-sized discovery findings in other ancestries.

Conclusions We show that commonly used genotyping microarrays have incomplete coverage for common ASD variants, and imputation cannot always recover lost information. Our novel metric provides an intuitive approach to reporting information loss in PGS and an alternative to reporting the total number of SNPs included in the PGS. While applied only to ASD here, this metric can easily be used with other traits.

Keywords Polygenic scores (PGS), Autism spectrum disorder (ASD), Information Loss

Background

PolyGenic Scores (PGS) are potentially useful tools for research and prediction [1, 2], but require additional development before their utility can be fully realized [3]. PGS are weighted sums of risk alleles that are computed using genotype data from individuals in a select target sample. The list of genetic variants to be summed and attendant effect sizes that serve as weights are identified in large-scale, genome-wide discovery studies. PGS represent an individual's genetic loading for a given trait, and although they are not expected to be sufficiently predictive for psychiatric and mental health outcomes on their own [4], they are likely to serve in the future as an essential component of risk modeling that guides clinical decision making for preventive strategies or post-diagnostic treatment.

Several factors limit the potential utility of the PGS by either directly or indirectly influencing the extent to which the PGS captures the true genetic susceptibility. Some of these factors are not easily addressed. Lack of generalizability between discovery and target samples, for example, will require data collection in diverse ancestries around the world [5, 6]. Also, the sample size or phenotypic measurement employed in current discovery studies may result in power loss reflecting incompleteness of variant identification and imprecision of effect sizes to serve as weights in the PGS [7]. Regardless of the completeness of discovery, when applying discovery results to target genotypes, the incomplete overlap of the discovery variant list and target genotyping array data can also lead to information loss when computing PGS. Specifically, the genotyping array used in the target sample is likely to lack some portion of the index discovery variants, and this will differ by array. Even if correlated, proxy variants are available as substitutes for the index variant, there is still some expected loss of PGS signal. It is often assumed that genotype imputation can recoup the lost information from unrepresented polymorphisms [8]. However, if no index or proxy genetic variants can be imputed with high quality, information loss will occur. Collectively, this lack of discovery representation in the genetic variants in the target sample could lead to a lower observed variance explained compared to the initial PGS R^2 metric often reported as part of the large discovery GWAS effort [9].

PGS for numerous psychiatric traits have been developed and have both global and unique considerations for their use and development. This project focuses on a PGS for Autism Spectrum Disorder (ASD). ASD is a neurodevelopmental disability characterized by social interaction and social communication deficits, and restrictive, repetitive patterns in behaviors, interests, or activities [10]. The heritability for common genetic variation is substantial and was estimated to be about 50% [11]. A PGS for ASD (ASD-PGS) could assist in the early detection of infants or toddlers displaying early symptoms. ASD-PGS, if accurate in the prediction of later abilities and challenges among diagnosed children, could also be used to inform parents and guide decisions about more specific and individualized supports and services.

To inform the ASD-PGS, the most recent discovery effort by Grove et al. (2019) identified common genetic susceptibility variants predisposing to ASD and was accomplished via genome-wide association study (GWAS) in combined samples from the Psychiatric Genomic Consortia (PGC) and the Danish Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) studies [12]. This scan successfully identified three independent loci and then attempted to replicate the findings for the top 88 variants using meta-analysis of five follow up studies conducted in Northern European populations. In addition to the top 88 GWAS hits, the Grove et al. study evaluated the predictive ability of ASD-PGS. While confirming that common genetic prediction is not currently clinically useful [1], the ASD-PGS was observed to explain 2.8% of variance in the trait [12], the SNP-based heritability, representing the potential of polygenic prediction with sufficient discovery sample size, was estimated at 11% [13].

A number of recent studies have used the Grove et al. discovery information to derive ASD-PGS and found statistically significant associations with ASD [14–17] and other ASD-related traits [18–20]. However, when reported, variance explained is well below the value estimated in Grove et al. (2019) and measures of the strength of association are modest or null. The failure to achieve the full predictive power of the ASD-PGS can be attributed to small target sample size and lack of generalizability of the target population to the discovery study. The incomplete coverage of independent loci that were discovered due to lack of adequate representation on the target study genome-wide microarray, however, is a specific cause that can be empirically characterized.

In this project, we evaluate the presence or absence of the 88 previously identified variants that were carried forward for replication in Grove et al. on standard microarrays before quality control (QC), after QC and post-imputation. To accomplish this, we use genome-wide data from three studies of familial ASD and one population-based case–control study. Collectively, these studies used a diversity of microarrays spanning approximately one million to more than five million sites. We then expand our evaluation of coverage in each study to a genome-wide PGS approach using a novel method and report on our findings.

Methods

Familial ASD studies

Individuals included in our analyses were participants in one of three studies of high familial likelihood for ASD. These studies were designed to investigate the early life factors involved in ASD and neurodevelopmental outcomes by enrolling mothers who had already given birth to a previous child with a clinical ASD diagnosis. The Early Autism Risk Longitudinal Investigation (EARLI) Study [21] and Markers of Autism Risk in Babies-Learning Early Signs (MARBLES) [22] both enrolled mothers who already had a child with ASD to participate when they became pregnant with another child and followed them throughout pregnancy until 36 months of age. We also included participants from the Infant Brain Imaging Study (IBIS), [23–25] which enrolled these high likelihood mothers and infants at 6 months and were then re-evaluated at 12 and 24 months of age.

Population-based study to explore early development phase I

A total of 3,769 children were enrolled in the Study to Explore Early Development, Phase 1 (SEED 1), a multisite cohort initiative designed to obtain a representative sample of ASD and typically developing preschool-aged children in the US. Children between 2 and 5 years old, born between September 1, 2003, and August 31, 2005, and living in one of six study site vicinities (San Francisco Bay Area, Philadelphia metropolitan area, northeast Maryland, central North Carolina, and the Atlanta metropolitan area) were ascertained through a variety of methods, including diagnostic clinics, organizations providing evaluation or services for children with developmental problems, educational departments, and population vital records. Detailed recruitment procedures are described elsewhere [26].

Genetic data

Cleaning and imputation

Samples from familial ASD cohorts were genotyped using the MEGA 1-Million (M)(MARBLES) and 5 M Illumina (EARLI and IBIS) chips. The Multi-Ethnic Global Array (MEGA) is a high-density array consisting of more than 1.7 million single nucleotide polymorphisms (SNPs) and is designed to represent diverse ancestries. The 5 M array is more comprehensive, consisting of around 5-million high-density SNPs, and has high overlap with the 1 M array. The SEED samples were genotyped on the more recently available Global Screening Array (GSA), which contains 640 K variants and represents diverse ancestry. For all studies, whole blood or buccal tissue was collected and samples were processed and stored. Genotyping was performed at the Johns Hopkins Genetic Resources Core Facility (GRCF).

The resulting genotypes were then subject to quality control filters according to standard criteria [27]. Briefly, using PLINK v1.9 [28], individuals failing QC checks for sex, relatedness, sample call rate of 0.03 (using –missing), and divergent ancestry were removed. SNPs were also removed based on the following criteria: MAF ≥ 0.05 among European samples, missing call rates exceeding 0.05, and Hardy–Weinberg equilibrium exact test p-value below 0.00001 among European samples.

Following cleaning, studies were imputed on the Michigan Imputation Server using the Minimac4 pipeline provided by the University of Michigan. We specified the 1000 genomes project (1000G) Phase 3 v5 reference panel, hg19 array build, Eagle v2.4 phasing [29], and the quality control and imputation mode. We imputed each target study separately and included 2504 1000G samples along with the target samples surviving QC. Post-imputation, we required imputed SNPs to correlate with the true unobserved genotypes at an r-squared (R^2) > 0.80.

Genetic ancestry classification

After applying both the variant and sample level filters, measured genotypes were used to compute genetic ancestry variables from principal components analysis in Eigensoft [30] according to a recommended procedure [27], which includes pooling the target samples with the 2504 1000G samples from diverse ancestral groups [31].

Classification to a defined ancestral group was carried out using K-means (R v4.0.3 with "kmeans" function) based on the first 2 principal components (PC1 and PC2) resultant from this procedure. Only the 661 African, 504 East Asian, and 503 European samples from 1000G were used as anchors to define three ancestral clusters. The minimum, maximum and standard deviation of PC1 and PC2 for each of the three 1000G ancestry groups were computed. Target sample principal components were then compared to these values to classify into an ancestral group. European ancestry classification required that the target PC1 and PC2 value fell within 1.96 standard deviations of the minimum and maximum values for the 1000G European corresponding principal components. All other ancestries were classified as non-European.

Top ranking ASD discovery SNPs

This analysis uses the top 88 discovery variants by *p*-value that were implicated in the ASD discovery GWAS [12]. The discovery GWAS combined samples from the Danish iPSYCH population as well as samples from the Psychiatric Genomics Consortium (PGC), totaling 18,381 cases of ASD and 27,969 controls reflecting European ancestry. Following initial identification of 88 top loci, a replication analysis was conducted with another 2,119 cases of ASD and 142,379 controls pooled across five different populations of Northern European ancestry. The three identified loci were found to be significant, while two additional loci became significant when meta-analyzed with the discovery sample. Although the majority of the single variant tests did not achieve statistical significance, a test to replicate the direction of effects was significant. For each identified top variant, the minor allele frequency, *p*-value, and odds ratios were provided in Supplementary Table 1.

Identifying correlated SNPs

In addition to the original list of 88 top variants, nearby SNPs that were highly correlated with the discovery index variants were identified as proxies. We accomplished this by accessing the GRCH37 REST API database (http://grch37.rest.ensembl.org) via R software. The Ensemble database uses the 1000G phase 3 reference to perform searches for correlated SNPs in specific windows. Proxy searches were performed specifying a reference panel made up of samples with European ancestry. Proxies were kept if they were found to be in linkage disequilibrium (LD) with the index SNP at $R^2 > = 0.80$. When multiple proxies were available for a missing SNP, the proxy was selected based on highest R^2 with the index SNP and closest physical distance.

Calculating single variant coverage among top ASD hits

Representation of the original index or proxy SNP was determined for cleaned and imputed target datasets by evaluating the overlap using chromosome, base pair, and variant identifier (rs number). An identical process was applied to obtain coverage for the Global Diversity Array-8 (GDA) and the Infinium PsychArray. Because we did not have target samples typed on these arrays, we were only able to evaluate coverage for variants on the pre-cleaned manifest files. The manifest files we downloaded for evaluation are publicly available on the Illumina website ([1, 2]).

Literature search for published reports using an ASD polygenic score

To characterize the methods for deriving and describing ASD-PGS that are commonly employed by researchers to date, we conducted a literature search in PubMed to identify manuscripts published through October 2022 that reported on ASD-PGS associations, where the target sample was in children and the outcome was either ASD or an ASD-related trait. We searched on the terms “ASD” and “Polygenic Risk Score” to obtain 109 potential hits. We also supplemented our search with the same terms in Google, evaluated the suggested literature from each identified manuscript, and considered references returned by the PGS Catalog [32] when searching for “autism”. Two researchers evaluated each abstract to rule out those studies that did not report on an ASD-PGS, did not perform analyses in children, or did not report on child ASD status or child ASD trait. We did not consider randomized trials. After a review of abstracts, 36 potential manuscripts met our criteria, and of these, 24 were selected for extraction. Extraction included the method and software used to derive ASD-PGS, parameters specified for the method, and whether the number of SNPs informing the ASD-PGS was reported.

Creation of an information metric for polygenic scores

While there exist several methods to derive PGS, the majority of researchers employ the clumping and thresholding (C + T) method [33], where redundant SNPs due to linkage disequilibrium (LD) are removed (i.e. clumping), and only the SNPs that fall below an established discovery *p*-value threshold level are included in the score (i.e. thresholding). Scores are derived using PLINK software [28, 34], which can also be implemented via PRSice version 1 [35] or version 2 [36]. To derive a score, the researcher must specify: 1) a reference panel, 2) a target panel and 3) a target population. The reference panel will be used to determine the amount of LD between nearby SNPs, supervised by discovery effect sizes or *p*-values, to make decisions about which SNPs to prune in the “clump” procedure. The target panel will limit the choice of SNPs to those available after cleaning and/or imputing the target array data. Thus, together, the reference and target panels will incorporate the discovery information to select the list of SNPs that inform the PGS. Finally, using the clumped SNP list, the “score” command will sum and weight each genotype for each sample in the target population.

With this C+T method, if a SNP, or any proxy SNP, is not available to represent a genomic region, its failure to be incorporated into the PGS calculation is likely to result in information loss. Even in the case when a proxy SNP is selected from the target panel to represent the index discovery SNP, some information loss will occur. Our goal was to assess the impact of these missing representations of discovery loci due to lack of coverage when using a C+T method to derive ASD-PGS, even after high quality imputation recovered a number of SNPs in the target data that were not present on the cleaned, measured GWAS array.

To accomplish this, we made use of the 1000G phase 3 v5 reference panel [31], which contains a comprehensive set of genetic variants from whole genome sequencing across 26 different populations. We limited the reference panel to the 498 unrelated individuals of European ancestry (eur1kg). The eur1kg panel can serve as the reference panel when clumping, and can also act as a high coverage, comprehensive target panel, in addition to serving as the target population. Alternatively, the high quality post-imputed data from a target cohort can also serve as a target panel, and the samples can be used as the target population in whom scoring takes place. We compute two different PGS as outlined in Fig. 1. All PGS are informed by the Grove et al. discovery results, and the eur1kg panel is always specified as the reference panel of choice. The first score represents the score computed from a comprehensive panel and specifies the eur1kg as both the target panel and the target population (full-eur1kg PGS); the process to create full-eur1kg PGS is shown in Fig. 1 on the left. The second score represents the PGS computed using the cohort target panel, which is not as comprehensive as eur1kg. This second PGS

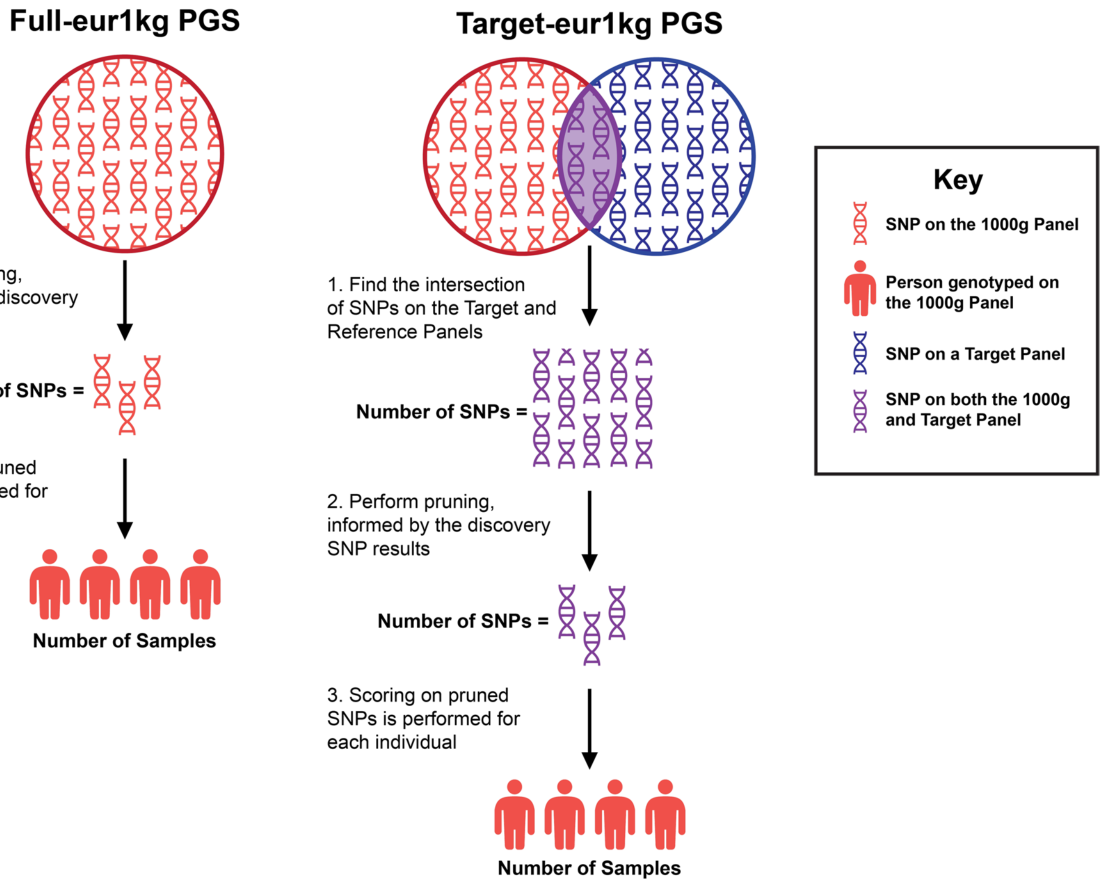
Full-eur1kg PGS
1. Perform pruning, informed by the discovery SNP results
Number of SNPs =
2. Scoring on pruned SNPs is performed for each individual
Number of Samples
Target-eur1kg PGS
1. Find the intersection of SNPs on the Target and Reference Panels
Number of SNPs =
2. Perform pruning, informed by the discovery SNP results
Number of SNPs =
3. Scoring on pruned SNPs is performed for each individual
Number of Samples
Key
SNP on the 1000g Panel
Person genotyped on the 1000g Panel
SNP on a Target Panel
SNP on both the 1000g and Target Panel

Fig. 1 PGS Computation Workflow. Legend: Our genome-wide information metric is simply the direct correlation between full-eur1kg and target-eur1kg PGS, which reflects information loss genome-wide. A correlation of 1.0 indicates that no information loss occurred, whereas a low correlation suggests a substantial loss of information. We derive each ASD-PGS using a C+T method, limiting to biallelic, high quality (info > 0.80) SNPs, for a suite of *p*-value discovery thresholds (5×10^{-8}, 1×10^{-6}, 10^{-4},10^{-3}, .01, 0.05, 0.10, 0.20, 0.50 and 1.0) scored in PLINK software. Because target-eur1kg-ASD-PGS were clumped and scored for each cohort separately, a different SNP selection informed each of the target-eur1kg PGS, leading to cohort-specific correlations with the full-eur1kg PGS

is scored in the eur1kg sample to serve as an “apples to apples” comparison to the full-eur1kg PGS, as shown on the right of Fig. 1 (referred to as target-eur1kg PGS). Our coverage metric is the correlation between the full-eur1kg and target-eur1kg PGS, which will range from 0.0 to 1.0.

Results

Characteristics of the target cohort samples

The total number of samples, as well as sex and ancestry distributions per target cohort, are provided in Table 1. The samples include ASD children, parents and siblings. A large majority of the samples reflect European ancestry in both the IBIS (88.5%) and SEED (69.6%) studies, while EARLI (56.4%) and MARBLES (54.5%) are slight majority European. All cohorts had < 10% of the sample East Asian and African, but admixed individuals make up a sizeable minority (ranging from about 10% to 30%). EARLI and IBIS have a more balanced sex proportion than MARBLES and SEED, which have predominantly female samples. The total number of samples across all cohorts ranges from 633 to 914 individuals.

Characteristics of top 88 variants

Grove et al. (2019) provided metrics for the 88 variants selected for follow up from their discovery efforts, and we provided the ranges reported on allele frequency and effect sizes in Supplementary Table 1. All minor allele frequencies for each of the top 88 variants were > 1%. The odds ratios for the variants reflected moderate to modest effect sizes, ranging from 0.658 to 1.342. Of the total number of SNPs, 70 (79.55%) were bi-allelic. As expected, all 88 variants exhibited suggestive statistical significance (p-values $< 1 \times 10^{-5}$) in the original analysis, with 53 of the 88 variants also achieving significance in the follow-up study [12].

Table 1 Characteristics of target study participants

Characteristic	Type	EARLI	IBIS	MARBLES	SEED
N	-	827	914	633	863
Sex (%)	Male	461 (55.7)	529 (57.9)	177 (27.9)	295 (34.1)
	Female	366 (44.2)	385 (42.1)	456 (72.0)	568 (65.8)
	European	467 (56.5)	809 (88.5)	345 (54.5)	601 (69.6)
	Non-European	360 (43.5)	105 (11.5)	288 (45.5)	262 (30.4)

Coverage of top ASD discovery variants

The surviving number of SNPs and coverage among the 88 variants identified via discovery are provided in Table 2 for clean, measured genotypes and for high quality ($R^2 > 0.80$) post-imputation genotypes. Considering the clean, measured genotypes, no study panel we examined directly covered all 88 of the top variants. For the IBIS and EARLI studies, which were both genotyped using the 5 M arrays, 31 (35%) and 32 (36%) variants were represented, respectively. MARBLES (1 M array) and SEED (GSA array) only contained 11 of the index variants. Including proxy SNPs improved coverage for measured genotypes across all studies, with 52 of the index discovery variants represented in IBIS, 54 represented in EARLI, 28 in MARBLES and 31 in SEED. Once we considered high quality imputed variants, coverage for IBIS and EARLI cohorts reached 100%, but MARBLES was still missing 6 of the 88 variants and SEED 5 of the 88 variants.

We also analyzed coverage of the top 88 variants and their proxies on the GDA and iPSYCH arrays using publicly available manifest files. Initial analysis of the GDA array revealed representation of 16 (18%) of the top 88 variants, and improved to 38 (43%) when including proxy variants (see Supplementary Table 2 for full proxy list). Analysis of iPSYCH yielded 11 of the original 88 discovery variants on the array, and the inclusion of proxy SNPs improved coverage to 25 (28% of all 88 variants).

Table 2 Sample size, variant numbers, and coverage statistics for top ASD hits per cohort

Cohort (Array)	N measured Variants surviving QC	N Variants surviving post-imputation QC	N Top ASD Variants on GWA Array[a]	N Top ASD Variants surviving post-imputation QC	N SNPs in Score[b]	ASD-PGS coverage metric[c]
IBIS (5 M)	2,400,509	38,343,801	52 (31)/88	88/88	249,698	0.9569
EARLI (5 M)	2,525,262	38,123,095	54 (32)/88	88/88	242,199	0.9445
MARBLES (1 M)	578,578	33,317,727	28 (11)/88	82/88	229,119	0.9295
SEED (GSA)	877,115	32,275,019	31 (11)/88	83/88	268,035	0.9849

[a] Coverage reflects the presence on the panel of either the discovery variants itself or at least one proxy SNP. Number in parentheses refers to the non-proxy number

[b] Full-eur1kg containes 281,593 SNPs after clumping

[c] Coverage metric reflects the correlation of full-eur1kg and target-eur1kg scores as explained in Fig. 1

Findings from a literature search for ASD-PGS associations

Our literature search identified 24 published manuscripts with reports of ASD-PGS with ASD or ASD-related traits in children, and extracted information from these manuscripts is in Supplementary Table 3. Almost all microarrays were from Illumina and ranged from the 550 Quad chip to the more recent GSA chips. In all verifiable studies, PGS were made from imputed genotypes. Four studies employed a method other than C+T to derive ASD-PGS. Among those using C+T, we observed a range of specified r^2 for LD correlation and window sizing, and about half employed a version of PRSice to carry out scoring. Only one study employed both long- and short-range pruning. Importantly, among the 24 studies, 16 reported the number of SNPs that informed the ASD-PGS, however, no alternate metric to reflect information loss due to SNPs that weren't directly represented or indirectly represented by proxy in the score was reported.

Genomic coverage using ASD-PGS

To determine potential loss of information in the ASD-PGS from unrepresented or indirectly represented variants across the genome, we correlated two different PGS, both derived in the 1000G sample, using complete SNP data and again using only the available SNP target data, as explained in the Methods and depicted in Fig. 1. Figure 2 shows a scatter plot between the eur1kg-full-PGS and the eur1kg-target-PGS, color coded for each high ASD likelihood target cohort and the SEED 1 target cohort. The two different PGS informing our metric exhibited a linear relationship with each other for all target cohorts, confirming that a correlation is an appropriate comparison statistic. In general, correlation was high for all study platforms, ranging from 0.9295 to 0.9849, using the most liberal discovery *p*-value threshold of 1.0 (see last column Table 2).

Figure 3 shows correlations between the full and target PGS by select discovery *p*-value thresholds. We consistently saw a drop in our PGS coverage metric for the suggestive discovery *p*-value threshold of 1×10^{-6}, but coverage improved thereafter, suggesting that coverage is particularly low for ASD in the statistically suggestive range. Results for all *p*-value threshold levels are provided in Supplementary Table 4. We did not observe clear patterns between PGS coverage and platform array density or for the number of SNPs informing the ASD-PGS.

Discussion

We evaluated coverage of top GWAS hits for ASD for a number of Illumina microarrays. We used available genotype data from the Illumina 5 M, MEGA and GSA arrays, plus two arrays using publicly available manifest files. The modern GWA arrays increasingly used in genetic research include the Illumina PsychArray (iPSYCH), Global Diversity Array (GDA) and Global Screening Array (GSA). These modern arrays are less costly

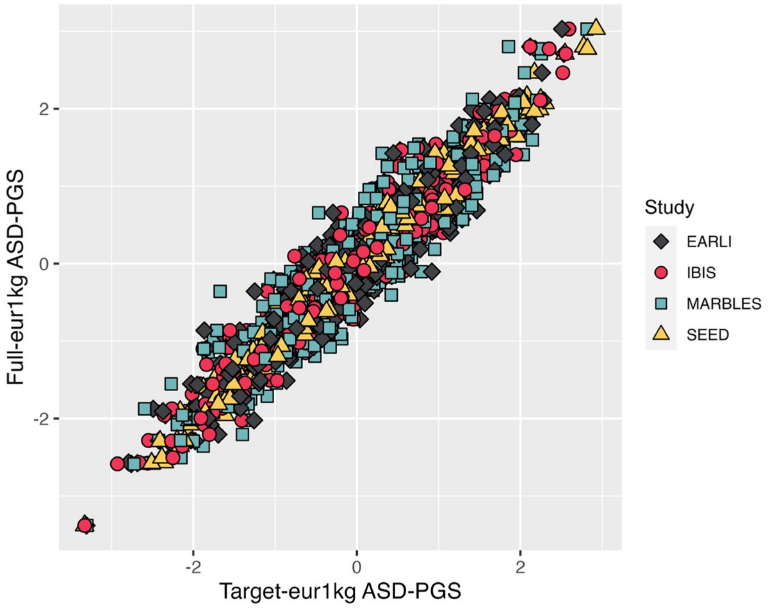

Fig. 2 Correlation between Full-eur1kg ASD-PGS and Target-eur1kg ASD-PGS for each of the 498 eur1kg samples. Legend: A separate Target-eur1kg ASD-PGS was made in each of the four target cohorts, reflecting the different SNP selection that emerged from the pruning process."The correlation coefficient for each study using this data reflect the ASD-PGS coverage metric reported in the last column of Table 2."The Full-eur1kg is the same for each cohort. All scores represent a discovery *p*-value threshold of 1.0

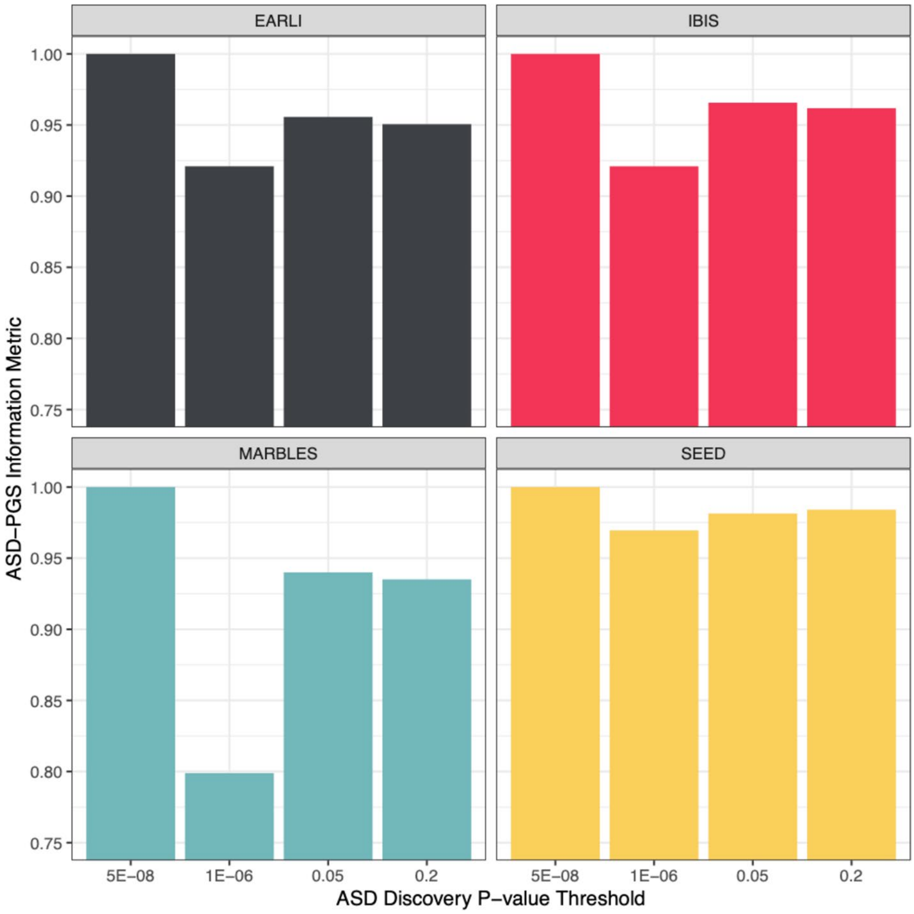

Fig. 3 The correlation metric (Full-eur1kg ASD-PGS versus Target-eur1kg ASD-PGS) representing loss of information in each of the four target cohorts. Legend: The metric represents information loss or genome-wide coverage from using a pruning and thresholding derivation method in each of the four target cohorts. Thresholds for discovery *p*-values are genome-wide significant (5×10^{-8}), genome-wide suggestion (1×10^{-6}), 0.05 and 0.20

compared to the denser (5 M or 2.5 M) arrays. They are also designed to expand on the intent of the Multi-ethnic arrays (MEGA or 1 M), which included variants meant to tag diverse ancestries and improve imputation performance. The modern arrays continue to represent diverse ancestry while providing content for an increasing number of clinical traits or disease outcomes. Our evaluation of coverage determined that many variants are not directly represented. For GSA, 5 M and MEGA chips, less than half of the 88 variants identified in large-scale GWA for ASD were directly represented. Using publicly available manifest files, we observed similar low coverage for the iPSYCH and GDA arrays as well. Further, when considering proxy SNPs that are highly correlated with the index variant, we still observe incomplete coverage for all arrays.

We then re-calculated coverage using available study genotype data after standard QC procedures for GSA, MEGA and 5 M arrays, and confirmed that coverage for measured genotyping is incomplete. Perhaps most importantly, we identified that for the dense 5 M array, high quality imputed data can bring the coverage to 100%, but for less dense arrays, while providing high coverage, high quality imputation can still fall short of completion. These findings suggest that attempts to recover lost SNPs are not guaranteed to be successful, and that different microarrays can yield varying levels of coverage.

The results of our literature search provide a gestalt for the state of current research practices when deriving ASD-PGS in target study data. We found that the vast majority of studies used a C + T method rather than using alternative methods, such as LDPred [37], that may somewhat improve variance explained [38]. Papers published more recently, however, are beginning to employ these newer methods. The C + T approach, however, is still likely to continue to be a popular approach due to its intuitive appeal, lower computational requirements, and the fact that it can be scored in PLINK software, which is familiar to many researchers. We also observed that though software and parameter specification for C+T varied somewhat, the general procedures and software were quite similar across studies. Importantly, the only metric reported across these studies that reveal insight into the coverage of the target study ASD-PGS was the total number of SNPs incorporated into the score, highlighting the gap in current research practice for evaluating information loss. It is also notable that one-third of the studies did not report any metric yielding some insight into SNP coverage for the ASD-PGS. Interestingly, the University of Michigan Imputation Server is offering a beta version to derive PGS for hundreds of different traits, using discovery findings published in the PGS Catalog [32], but appears to only offer the total number of index SNPs contained in the score as a measure of PGS coverage.

We then explored a simple and novel approach that would intuitively provide an indication of information loss. Our metric, in addition to addressing the selection of SNPs, reflects the extent to which this imperfect or missing information is influenced by observed allele frequencies and weighted by the discovery effect size, so that the joint impact of all these factors on the PGS is incorporated into the measure. Thus, we provide an intuitive, scalar measure that reflects the true loss of information genome-wide from multiple sources that is not readily intuitive by reporting the total number of SNPs.

To compute our PGS coverage metric for ASD, we use the eur1kg reference panel. For the ASD trait, our choice to compute the full-eur1kg and target-eur1kg scores in a European panel is appropriate because reference and discovery ancestry should reflect each other [39], and the ASD discovery is comprised almost entirely of individuals of European descent. There exist many alternative reference panels that offer a comprehensive set of variants via sequencing to calculate a PGS coverage metric. For example, the Haplotype Reference Consortium (HRC) [40] may provide better imputation accuracy, particularly for rare variants, due to its larger sample size compared to the 1000G European populations [41]. Several panels are also available to researchers representing African ancestry including the Consortium on Asthma among African ancestry Populations in the Americas (CAAPA) [42], the African Genome Resources (AGR) (https://www.apcdr.org/) and African Genome Variation Project (AGVP) [43]. The selection of reference panel to compute a coverage metric will depend on the ancestry of the discovery population as well as consideration for the ancestry of the target population.

Our PGS metric can serve as a useful tool for researchers. Current published studies for ASD-PGS focus on the association with ASD outcomes for the purpose of gaining insight into the disorder's genetic etiology. With the establishment of the PGS Catalog and the opportunity to compute PGS via the Michigan Imputation Server based on the PGS Catalog discovery input, focus may shift to creating genetic scores in target samples that are derived from pre-prepared, filtered and pruned SNP lists from a standard reference population. In this case, the target sample PGS value can be placed directly along the distribution of PGS in the reference population to determine if a target individual has a high, average, or low genetic load, rather than relying on the relative ordering of samples within a target study. If index SNPs are not directly genotyped and/or not surviving pre-imputation QC procedures, then knowledge about how this potentially influences coverage would be important to calculate and report. As PGS methods and discovery findings develop for ASD and other traits, we argue that the reporting of an intuitive coverage metric that captures lost discovery information should become an essential part of any future PGS effort.

Limitations

There were several limitations to our study. First, although we present coverage prior to cleaning and imputation, using Illumina manifest files for several arrays, we did not measure our own genotypes using either of the GDA or iPSYCH arrays. A lack of in-hand data for these studies prevented us from assessing top hits coverage after cleaning and imputation procedures were applied as well as from computing our PGS coverage metric. These arrays, however, have high overlap with the GSA, so our findings using the SEED data, which was genotyped on the GSA, may provide a good estimate of the GDA and iPSYCH coverage. Second, our analysis is in studies with modest sample sizes, and additional evaluation with studies ranging from a small to large number of samples may offer the opportunity to better explore trends in coverage. Third, our computation of top variant coverage as well as PGS coverage could have been influenced by the presence in our target cohort samples of non-Europeans via any influence these samples may have had on the QC and imputation process. To acknowledge this influence, we computed minor allele frequency and Hardy Weinberg Equilibrium p-values in separate ancestries before applying filtering criteria. When imputing, we included all 1000G ancestries along with our target samples and impute to the full multi-ethnic 1000G panel. Despite differing LD pattern between ancestries, previous research has suggested that this strategy of imputing to a diverse ancestry panel can result in higher imputation accuracy [44, 45], and thus, the presence of non-European target samples should have little to no influence on our coverage metric. Fourth, the PGS coverage metric was computed in Europeans, and is not generalizable to non-European ancestries. Although beyond the scope we have defined here, large-scale GWAS in non-European ancestries are available for some traits and can be used to explore PGS coverage. Finally, we limit our approach to the clumping and thresholding method of PGS derivation, but extensions to LDpred2 [37, 39] and other derivation methods including SBayesR [46], SDPR [47], PRS-CS [48] and others may be possible.

Conclusions

In summary, we provide insight into the coverage of top ASD GWAS variants for a number of commonly used genome-wide microarrays. We also create and apply a genome-wide coverage metric to assess how well the ASD-PGS in a particular target sample is incorporating the available information from the discovery GWA results. While applied only to ASD here, our approach can be used for any trait with available discovery results and offers a more intuitive and satisfying alternative to reporting the total number of SNPs included in the PGS. There may be natural extensions of our metric for other PGS derivation methods beyond the clumping and thresholding approach that can be explored in future research.

Abbreviations

ASD	Autism Spectrum Disorder
PGC	Psychiatric Genomics Consortium
MAF	Minor Allele Frequency
PGS	Polygenic Score
ASD-PGS	Polygenic Score for Autism Spectrum Disorder
SNP	Single Nucleotide Polymorphism
1000G	1000 Genomes
EARLI	Early Autism Risk Longitudinal Investigation
IBIS	Infant Brain Imaging Study
MARBLES	Markers of Autism Risk – Learning Early Signs
SEED	Study to Explore Early Development
HRC	Haplotype Reference Consortium
GDA	Global Diversity Array
GSA	Global Screening Array
iPSYCH	Infinium PsychArray
QC	Quality Control
MEGA	Multi-Ethnic Global Array

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s11689-024-09571-8.

Additional file 1: Supplementary Table 1. Characteristics of ASD Discovery GWA Top 88 Variants. Contains information about the MAF, odds ratio, p values, and number of bi-allelic SNPs in the original 88 variants identified by Grove et al.

Additional file 2: Supplementary Table 2. List of Proxy SNPs to the Top 88 Variants from ASD Discovery GWAS. Contains the list of proxy SNPs for the top 88 variants ordered by chromosome with corresponding rs ID, R^2, average MAF, and distance in base pairs.

Additional file 3: Supplementary Table 3. Characteristics of ASD-PGS Derivation in Reported Studies. Contains the results of the literature search on characteristics of ASD-PGS derivation in 24 studies, including information from the following categories: ASD Discovery GWAS, Target GWA Chip, Imputed/Reference Panel, Post-imputation filters, PRS software, Clumping reference panel, Clump r^2, clump window size, presence of a 2nd clump round, specification of a PRS threshold, and whether # of SNPs was reported.

Additional file 4: Supplementary Table 4. PGS coverage metric by Discovery *P*-value threshold. Contains values of our novel PGS coverage metric for IBIS, EARLI, MARBLES, and SEED for 9 selected *P*-values.

Acknowledgements

Thank you to all the EARLI, IBIS, MARBLES, and SEED study participants and families for their contributions and involvement in this research.

Authors' contributions

This study was conceived and designed by KB with input from CL-A, MDF, and VM. Analyses were performed by JD and MY. MY prepared Figs. 1 and 2 and JD prepared Fig. 3. VM performed reproducibility analyses. Writing was shared between KB, JD, and MY. All authors reviewed the manuscript.

Funding

This work was supported by Centers for Disease Control and Prevention (CDC) Cooperative Agreements announced under the following RFAs: 01086, 02199, DD11-002, DD06-003, DD04-001, and DD09-002. IBIS was supported by grants from NIH R01 HD055741 (ACE), Autism Speaks, and Simons Foundation (140209) The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Availability of data and materials

The SEED 1 data in this study are not publicly available due to lack of explicit consent for such sharing in the written informed consents for SEED sites, per the CDC IRB that governs the SEED network. Discovery GWAS results are available in the study by Grove et. Al. The datasets generated during the EARLI, IBIS, and MARBLES studies are not publicly available but are available from the corresponding author on reasonable request. 1000 Genome Data is publicly available.

Declarations

Ethics approval and consent to participate

This study was approved by the institutional review boards (IRBs) at each SEED site. SEED 1 recruitment was approved by the IRB of each recruitment site: IRB-C, CDC Human Research Protection Office; Kaiser Foundation Research Institute (KFRI) Kaiser Permanente Northern California IRB, Colorado Multiple IRB, Emory University IRB, Georgia Department of Public Health IRB, Maryland Department of Health and Mental Hygiene IRB, Johns Hopkins Bloomberg School of Public Health IRB, University of North Carolina IRB and Office of Human Research Ethics, IRB of The Children's Hospital of Philadelphia, and IRB of the University of Pennsylvania. All enrolled families provided written consent for participation.

Infants were enrolled in IBIS at one of four clinical sites (University of North Carolina, University of Washington, Washington University, and Children's Hospital of Philadelphia). Parents provided informed consent, and the institutional review board at each site approved the research protocol.

Consent for publication

Not applicable.

Competing Interests

The authors declare no competing interests.

Author details

[1]Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [2]Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [3]Wendy Klag Center for Autism and Developmental Disabilities, JHSPH, Baltimore, MD, USA. [4]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [5]Department of Psychiatry, University of North Carolina, North Carolina, Chapel Hill 27599,

USA. [6]Carolina Institute for Developmental Disabilities, Chapel Hill, NC 27599, USA. [7]7AJ Drexel Autism Institute, Drexel University, 3020 Market St, Suite 560, Philadelphia, PA 19104, USA. [8]College of Health and Human Development, Penn State, University Park, PA 16802, USA. [9]Department of Public Health Sciences, University of California, Davis, CA 95616, USA. [10]UC Davis MIND (Medical Investigations of Neurodevelopmental Disorders) Institute, Sacramento, CA 95817, USA. [11]Autism Research Program, Kaiser Permanente Division of Research, 2000 Broadway, Oakland, CA 94612, USA. [12]Rollins School of Public Health, Emory University, 1518 Clifton Rd, Suite 8011, Atlanta, GA 30355, USA.

Received: 26 September 2023 Accepted: 29 August 2024
Published online: 12 September 2024

References

1. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12:44.
2. Martin AR, Daly MJ, Robinson EB, Hyman SE, Neale BM. Predicting polygenic risk of psychiatric disorders. Biol Psychiatry. 1969;2019(86):97–109.
3. Janssens ACJW. Validity of polygenic risk scores: are we measuring what we think we are? Hum Mol Genet. 2019;28:R143–50.
4. Wray NR, Trzaskowski M, Byrne EM, Abdellaoui A, Adams MJ, Agerbo E, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50:668–81.
5. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51:584–91.
6. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nat Lond. 2019;570:514–8.
7. Lam M, Lencz T, Consortium (COGENT) CG. SU101 - identification of key snps and pathways underlying differential genetic correlations between education and cognition on schizophrenia. Eur Neuropsychopharmacol. 2019;29:S943-4.
8. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499–511.
9. Nguyen DT, Tran TTH, Tran MH, Tran K, Pham D, Duong NT, et al. A comprehensive evaluation of polygenic score and genotype imputation performances of human SNP arrays in diverse populations. Sci Rep. 2022;12:17556.
10. Dr M-CL, PhD MVL, Prof SB-C. Autism. Lancet. 2014;383:896–910.
11. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. Nat Genet. 2014;46:881–5.
12. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. 2019; Available from: https://research.vumc.nl/en/publications/a4919ac9-a15d-4b8b-892a-6ed1a324754f.
13. Baselmans BML, Yengo L, van Rheenen W, Wray NR. Risk in relatives, heritability, snp-based heritability, and genetic correlations in psychiatric disorders: a review. Biol Psychiatry. 1969;2021(89):11–9.
14. Klei L, McClain LL, Mahjani B, Panayidou K, Rubeis SD, Grahnat ACS, et al. How rare and common risk variation jointly affect liability for autism spectrum disorder. Mol Autism. 2021;12:66.
15. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet. 2017;49:978–85.
16. Torske T, Nærland T, Bettella F, Bjella T, Malt E, Høyland AL, et al. Autism spectrum disorder polygenic scores are associated with every day executive function in children admitted for clinical assessment. Autism Res. 2020;13:207–20.
17. Jansen A, Dieleman G, Jansen P, Verhulst F, Posthuma D, Polderman TJ. Psychiatric polygenic risk scores as predictor for attention deficit/hyperactivity disorder and autism spectrum disorder in a clinical child and adolescent sample. Behav Genet. 2020;50:203–12.
18. Takahashi N, Harada T, Nishimura T, Okumura A, Choi D, Iwabuchi T, et al. Association of genetic risks with autism spectrum disorder and early neurodevelopmental delays among children without intellectual disability. JAMA Netw Open. 2020;3: e1921644.
19. Serdarevic F, Tiemeier H, Jansen PR, Alemany S, Xerxa Y, Neumann A, et al. Polygenic risk scores for developmental disorders, neuromotor functioning during infancy, and autistic traits in childhood. Biol Psychiatry. 1969;2020(87):132–8.
20. Clarke T-K, Lupton MK, Fernandez-Pujals AM, Starr J, Davies G, Cox S, et al. Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. Mol Psychiatry. 2016;21:419–25.
21. Newschaffer CJ, Croen LA, Fallin MD, Hertz-Picciotto I, Nguyen DV, Lee NL, et al. Infant siblings and the investigation of autism risk factors. J Neurodev Disord. 2012;4: 7.
22. Hertz-Picciotto I, Schmidt RJ, Walker CK, Bennett DH, Oliver M, Shedd-Wise KM, et al. A prospective study of environmental exposures and early biomarkers in autism spectrum disorder: design, protocols, and preliminary data from the MARBLES study. Environ Health Perspect. 2018;126:117004.
23. Hazlett HC, Gu H, Munsell BC, Kim SH, Styner M, Wolff JJ, et al. Early brain development in infants at high risk for autism spectrum disorder. Nature. 2017;542:348–51.
24. Shen MD, Swanson MR, Wolff JJ, Elison JT, Girault JB, Kim SH, et al. Subcortical brain development in autism and fragile X syndrome: evidence for dynamic, age- and disorder-specific trajectories in infancy. Am J Psychiatry. 2022;179:562–72.
25. Wolff JJ, Gu H, Gerig G, Elison JT, Styner M, Gouttard S, et al. Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. Am J Psychiatry. 2012;169:589–600.
26. Schendel DE, DiGuiseppi C, Croen LA, Fallin MD, Reed PL, Schieve LA, et al. The Study to Explore Early Development (SEED): a multisite epidemiologic study of autism by the Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) Network. J Autism Dev Disord. 2012;42:2121–40.
27. Anderson CA, Zondervan KT, Pettersson FH, Clarke GM, Cardon LR, Morris AP. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5:1564–73.
28. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.
29. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the haplotype reference consortium panel. Nat Genet. 2016;48:1443–8.
30. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
31. Altshuler DM, Albers CA, Abecasis GR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
32. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021;53:420–5.
33. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nat Lond. 2009;460:748–52.
34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
35. Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. Bioinformatics. 2015;31:1466–8.
36. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience. 2019;8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31307061.
37. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2020;36:5424–31.
38. Ni G, Wang Y, Ge T, Smoller JW, Ripke S, Farh K-H, et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. Biol Psychiatry. 1969;2021(90):611–20.
39. Gusev A, Ripke S, Walters JTR, Agartz I, Albus M, Bene J, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet. 2015;97:576–92.

40. Mccarthy S, Das S, Kretzschmar W, Luo Y, Timpson N, Zhang H, et al. A reference panel of 64,976 haplotypes for genotype imputation. 2016. Available from: https://explore.openaire.eu/search/publication?articleId=dedup_wf_001::a754d81bb6b6cd0c831e119802af6cc3.
41. Vergara C, Parker MM, Franco L, Cho MH, Valencia-Duarte AV, Beaty TH, et al. Genotype imputation performance of three reference panels using African ancestry individuals. Hum Genet. 2018;137:281–92.
42. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, et al. A continuum of admixture in the western hemisphere revealed by the African diaspora genome. Nat Commun. 2016;7:12522.
43. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African genome variation project shapes medical genetics in Africa. Nature. 2015;517:327–32.
44. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, et al. Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet. 2009;84:235–50.
45. Jostins L, Morley KI, Barrett JC. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. Eur J Hum Genet EJHG. 2011;19:662–6.
46. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun. 2019;10:5086–111.
47. Zhou G, Zhao H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. PLoS Genet. 2021;17: e1009697.
48. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10:1776.

Publisher's Note