Supplementary Information

Trans-Ancestry Analysis of Psychosis Biotypes: Shared Polygenic Risk and Unique Genomic Associations

Table of contents

Summary of Supplementary Information

This document summarizes the ancestry information (Table S1, Figure S1), case-control variance explained by PRS (Table S2, Figure S6), inflation of TWAS results (Table S3-S5), prediction accuracy of PRS-CSx (unadjusted) versus PRS-CSx meta (unadjusted) (Figure S3), prediction accuracy of PRS (Figure S4), estimated overlapped areas of PRS (Figure S5), Biotype comparison results on PRS (Figure S2) and TAPRS (Figure S7) in each diagnostic group (SCZ and BD) in B-SNIP dataset.

Supplementary Tables

Table S1. Concordance between PCA-derived ancestry and Self-reported race questionnaire.

PCA-derived ancestry	Self-reported race							
	AA	AE	AS	CA	MR	NH	OT	UNK
AA	481	1	0	5	37	0	11	1
ASIAN	0	1	26	29	5	1	7	1
EUR	2	4	18	623	15	0	21	0

Legend: PCA = Principal Component Analysis (see Figure S1). Self-reported race: response of each person based on current cultural categories. ASIAN = Asian ancestry, EUR = European ancestry. AA = African American, AE = American Indian, AS = Asian, CA = Caucasian, MR = Multiracial/Mixed Race, NH = Native Hawaiian, OT = Other Race, UNK = Unknown/Missing. Within ancestry, the concordance rates of PCA inferred ancestry with self-reported ancestry in AA, EUR and ASIAN are 90%, 89% and 39%, respectively.

Table S2. Two-way Analysis of Variance (ANOVA) of Ancestry and Biotype effects on Schizophrenia PRS under different trans-ancestry (TAPRS) methods.

Factor	PRS-CSx meta (unadjusted)	TAPRS (adjusted-Ge)	TAPRS (adjusted-Khera)
Ancestry	39.4***	12.3***	1.8***
Biotype	1.7***	2.6***	3.1***
Interaction	0.4*	0.5*	0.7*
Residuals	58.5	84.6	94.4

Legend: ANOVA test *** $P < 0.001$ ** $P < 0.01$ * $P < 0.05$. Contribution of ancestry to PRS variance is smallest in the Khera method. Residuals include cumulative effects of individual SNPs.

Table S3. TWAS of adult brain gene expression: Inflation of results with and without genotype principal components (PCs) 1 and 2 as covariates.

Ancestry	Group	Inflation		Lambda	
		noPCs	withPCs	noPCs	withPCs
AA + EUR	case-ctrl	1.16	1.01	1.45	1.03
AA + EUR	HC-BT1	1.37	0.99	2.15	1.01
AA + EUR	HC-BT2	1.25	1.01	1.67	1.02
AA + EUR	HC-BT3	1.02	1	1.08	1.01
AA + EUR	BT1-BT2	1.01	0.98	1.05	1
AA + EUR	BT1-BT3	1.5	0.98	2.73	1
AA + EUR	BT2-BT3	1.36	1	2.12	1.04
AA	case-ctrl	0.98	0.98	1	1
AA	HC-BT1	0.98	0.98	0.97	0.98
AA	HC-BT2	0.98	0.99	1.01	1
AA	HC-BT3	0.98	0.98	1.02	1.01
AA	BT1-BT2	0.98	0.98	0.97	0.98
AA	BT1-BT3	0.98	0.98	0.99	0.98
AA	BT2-BT3	0.98	0.98	1	0.99
EUR	case-ctrl	1.02	0.99	1.1	1
EUR	HC-BT1	1	0.99	0.99	0.99
EUR	HC-BT2	1.01	0.98	1.03	0.98
EUR	HC-BT3	1.01	0.99	1.06	1.01
EUR	BT1-BT2	0.99	0.99	1.04	1.03
EUR	BT1-BT3	0.99	0.98	1.02	0.98
EUR	BT2-BT3	0.98	0.98	1.02	1

Legend: PCs = Principal Components. “noPCs” means association results without any PCs included in the logistic regression model. “withPCs” means association results with genotype PC1 and PC2 included in the logistic regression model. The metric *Inflation* is calculated using a Bayesian method based on the empirical null distribution. The metric *Lambda* is the genomic inflation factor, which is calculated by comparing the observed distribution of p-values to the expected uniform distribution under the null hypothesis of no association. A value close to 1 for both metrics indicates that the distribution of p-values closely follows the expected uniform distribution, suggesting minimal inflation. This result shows there is an inflation when genotype PCs are not adjusted for TWAS in adult brain in the combined sample (AA + EUR) but not in AA or EUR ancestries separately.

Table S4. TWAS of fetal brain gene expression: Inflation of results with and without genotype principal components (PCs) 1 and 2 as covariates.

Ancestry	Group	Inflation		Lambda	
		noPCs	withPCs	noPCs	withPCs
AA + EUR	case-ctrl	1.16	1	1.48	1.02
AA + EUR	HC-BT1	1.4	0.98	2.27	1.01
AA + EUR	HC-BT2	1.26	1.01	1.76	1.05
AA + EUR	HC-BT3	1.01	0.98	1.07	0.99
AA + EUR	BT1-BT2	1.01	0.98	1.02	0.99
AA + EUR	BT1-BT3	1.54	0.98	3.14	1.02
AA + EUR	BT2-BT3	1.42	0.99	2.33	0.97
AA	case-ctrl	0.99	0.99	1.01	1
AA	HC-BT1	0.99	0.99	1.02	1.01
AA	HC-BT2	0.99	0.99	1	1.01
AA	HC-BT3	0.99	0.98	1	0.99
AA	BT1-BT2	0.96	0.96	0.99	0.99
AA	BT1-BT3	0.98	0.98	0.98	0.95
AA	BT2-BT3	0.98	0.98	0.98	0.95
EUR	case-ctrl	1.02	0.98	1.08	1
EUR	HC-BT1	1	0.98	1.03	1.01
EUR	HC-BT2	1.01	0.99	1.09	1.04
EUR	HC-BT3	1	0.97	0.99	0.95
EUR	BT1-BT2	0.99	0.99	1.01	1.02
EUR	BT1-BT3	0.99	0.97	1.02	0.97
EUR	BT2-BT3	0.97	0.97	1	0.99

Legend: PCs = Principal Components. “noPCs” means association results without any PCs included in the logistic regression model. “withPCs” means association results with genotype PC1 and PC2 included in the logistic regression model. The metric *Inflation* is calculated using a Bayesian method based on the empirical null distribution. The metric *Lambda* is the genomic inflation factor, which is calculated by comparing the observed distribution of p-values to the expected uniform distribution under the null hypothesis of no association. A value close to 1 for both metrics indicates that the distribution of p-values closely follows the expected uniform distribution, suggesting minimal inflation. This result shows there is an inflation when genotype PCs are not adjusted for TWAS in fetal brain in the combined sample (AA + EUR) but not in AA or EUR ancestries separately.

Table S5. TWAS of adult brain isoform expression: Inflation of results with and without genotype principal components (PCs) 1 and 2 as covariates.

Ancestry	Group	Inflation		Lambda	
		noPCs	withPCs	noPCs	withPCs
AA + EUR	HC-BT1	1.611	1.064	3.125	1.143
AA + EUR	HC-BT2	1.393	0.985	2.203	1.01
AA + EUR	HC-BT3	1.152	1.046	1.4	1.117
AA + EUR	BT1-BT2	1.059	0.965	1.157	0.937
AA + EUR	BT1-BT3	1.803	0.996	4.174	1.011
AA + EUR	BT2-BT3	1.536	0.976	2.869	0.954
AA	case-ctrl	1	1	1.02	1.02
AA	HC-BT1	1	1	1.02	1.02
AA	HC-BT2	0.99	0.99	0.99	0.98
AA	HC-BT3	0.99	0.99	0.99	0.99
AA	BT1-BT2	0.97	0.98	0.98	0.97
AA	BT1-BT3	0.99	1	1	0.99
AA	BT2-BT3	0.99	0.99	1.01	1
EUR	case-ctrl	1.02	1.03	1.09	1.09
EUR	HC-BT1	1.01	1.01	1.06	1.06
EUR	HC-BT2	1.03	1.03	1.11	1.11
EUR	HC-BT3	1.01	1.01	1.06	1.05
EUR	BT1-BT2	1.01	1.02	1.07	1.07
EUR	BT1-BT3	1.01	1.01	1.08	1.08
EUR	BT2-BT3	0.99	1	1.01	1.01

Legend: PCs = Principal Components. “noPCs” means association results without any PCs included in the logistic regression model. “withPCs” means association results with genotype PC1 and PC2 included in the logistic regression model. The metric *Inflation* is calculated using a Bayesian method based on the empirical null distribution. The metric *Lambda* is the genomic inflation factor, which is calculated by comparing the observed distribution of p-values to the expected uniform distribution under the null hypothesis of no association. A value close to 1 for both metrics indicates that the distribution of p-values closely follows the expected uniform distribution, suggesting minimal inflation. This result shows there is an inflation when genotype PCs are not adjusted for Isoform-WAS (IWAS) in adult brain in the combined sample (AA + EUR) but not in AA or EUR ancestries separately.

Supplementary Figures

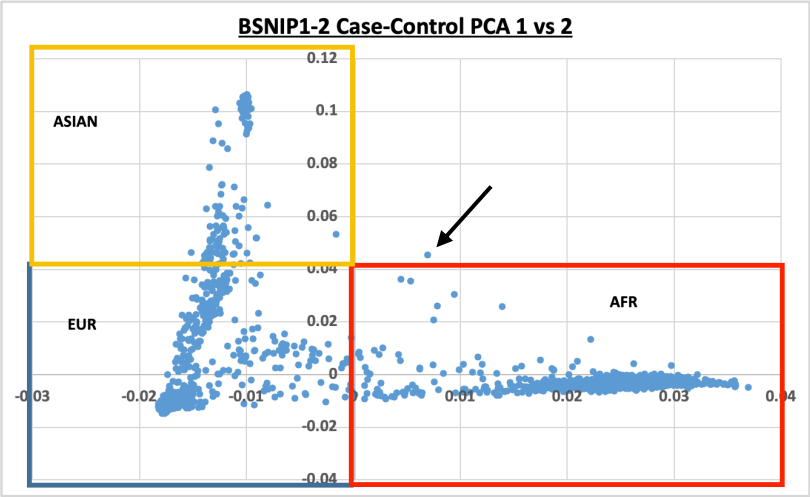

Figure S1. Principal Component Analysis (PCA) of genotypes from 2505 BSNIP individuals. Legend: There are three ancestries assigned: AA (African American ancestry, inside the red box), EUR (European ancestry, inside the blue box), and ASIAN (Asian ancestry, inside the yellow box). One individual (pointed by the black arrow, outside the three boxes) is not assigned to any of the three ancestries. The thresholds for categorizing the ancestry groups are 0 and 0.04 in the two PC dimensions.

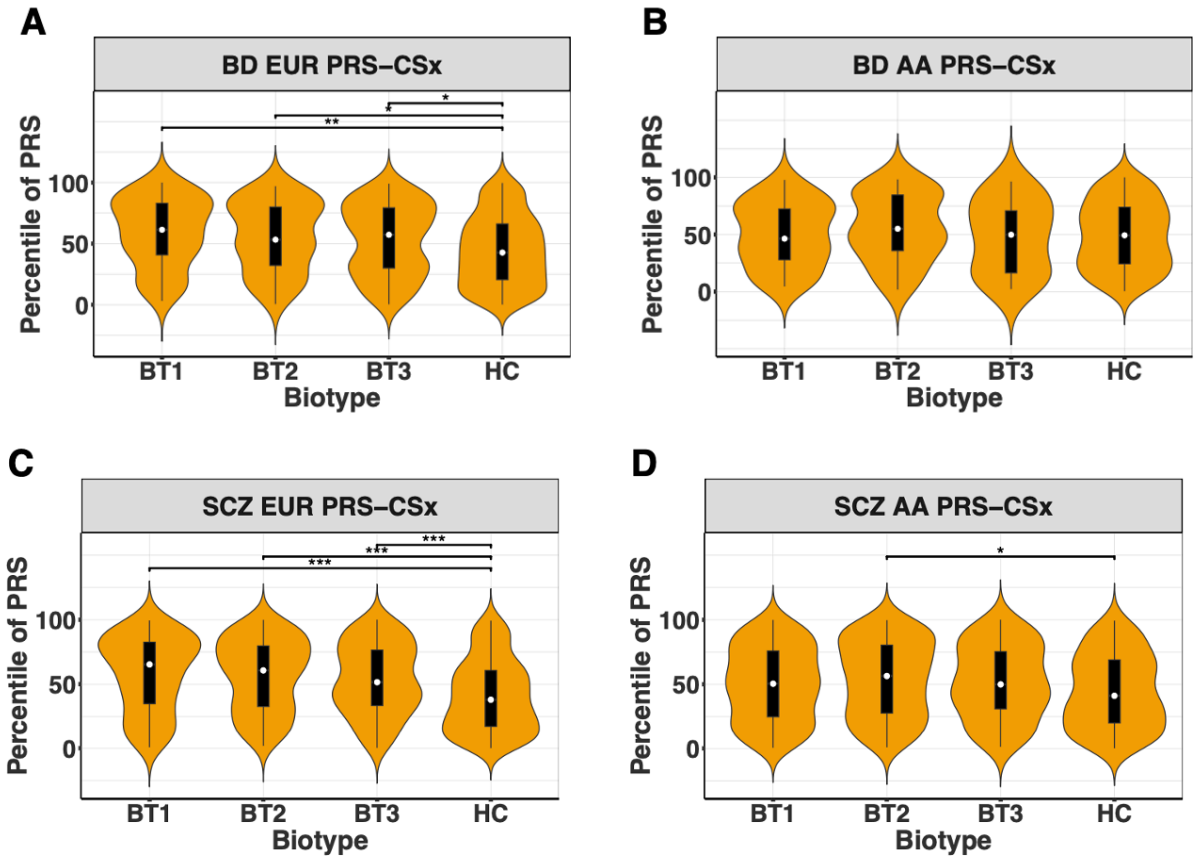

Figure S2. Ancestry-specific Biotype differences on psychosis PRS-CSx (meta), unadjusted (before any *post hoc* ancestry adjustment) in each diagnostic group. Each possible pair of Biotypes was tested for difference in PRS, and each Biotype (BT) was compared with Healthy Control (HC). (A) Violin plot of psychosis PRS-CSx (meta option) scores among Biotypes and with Healthy Control in EUR samples in BD group. (B) Same comparisons in AA samples in BD group. (C) Violin plot of psychosis PRS-CSx (meta option) scores among Biotypes and with Healthy Control in EUR samples in SCZ group. (D) Same comparisons in AA samples in SCZ group. EUR = European ancestry; AA = African American ancestry. Wilcoxon tests were used for comparisons. Bonferroni-corrected significance threshold over 6 two-sample Wilcoxon tests for each ancestry is P-value < 8.33e-03. Only significant comparison results are labeled with asterisks. *** indicates P-value < 1.67E-04, ** indicates P-value < 1.67E-03, * indicates P-value < 8.33E-03. Consistent with the Combined Diagnosis (SCZ + BD) results, the three Biotypes show similar PRS (no significant differences) in separate diagnoses within ancestries.

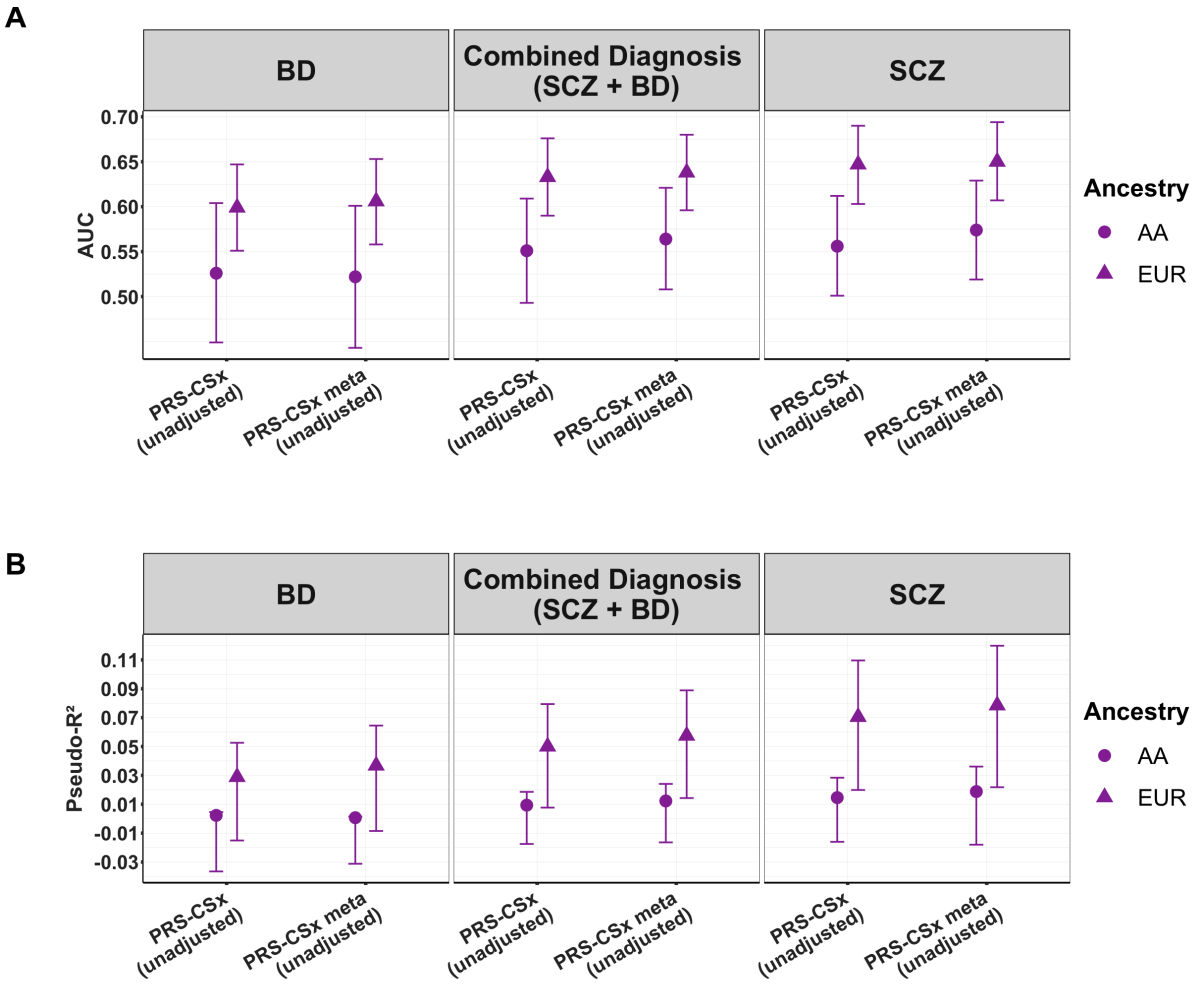

Figure S3. Prediction accuracy of PRS-CSx and PRS-CSx (meta) for case-control status within ancestry in each diagnostic group and in the combined diagnoses. (A) The area under the receiver operating characteristic (ROC) curve (AUC) of polygenic risk scores (PRSs) within ancestry. (B) Nagelkerke's pseudo-R^2 of PRSs within ancestry. Lines for Nagelkerke's pseudo-R^2 in (B) correspond to 95% confidence intervals calculated via 1000 bootstrapping. AA = African American ancestry, EUR = European ancestry. "meta" means we used the *--meta* option results for PRS-CSx. "Unadjusted" means the risk scores before any *post hoc* ancestry adjustment. The meta option of PRS-CSx generated marginally higher PRS prediction accuracy for case-control status in SCZ only, BD only and combined diagnosis (SCZ + BD).

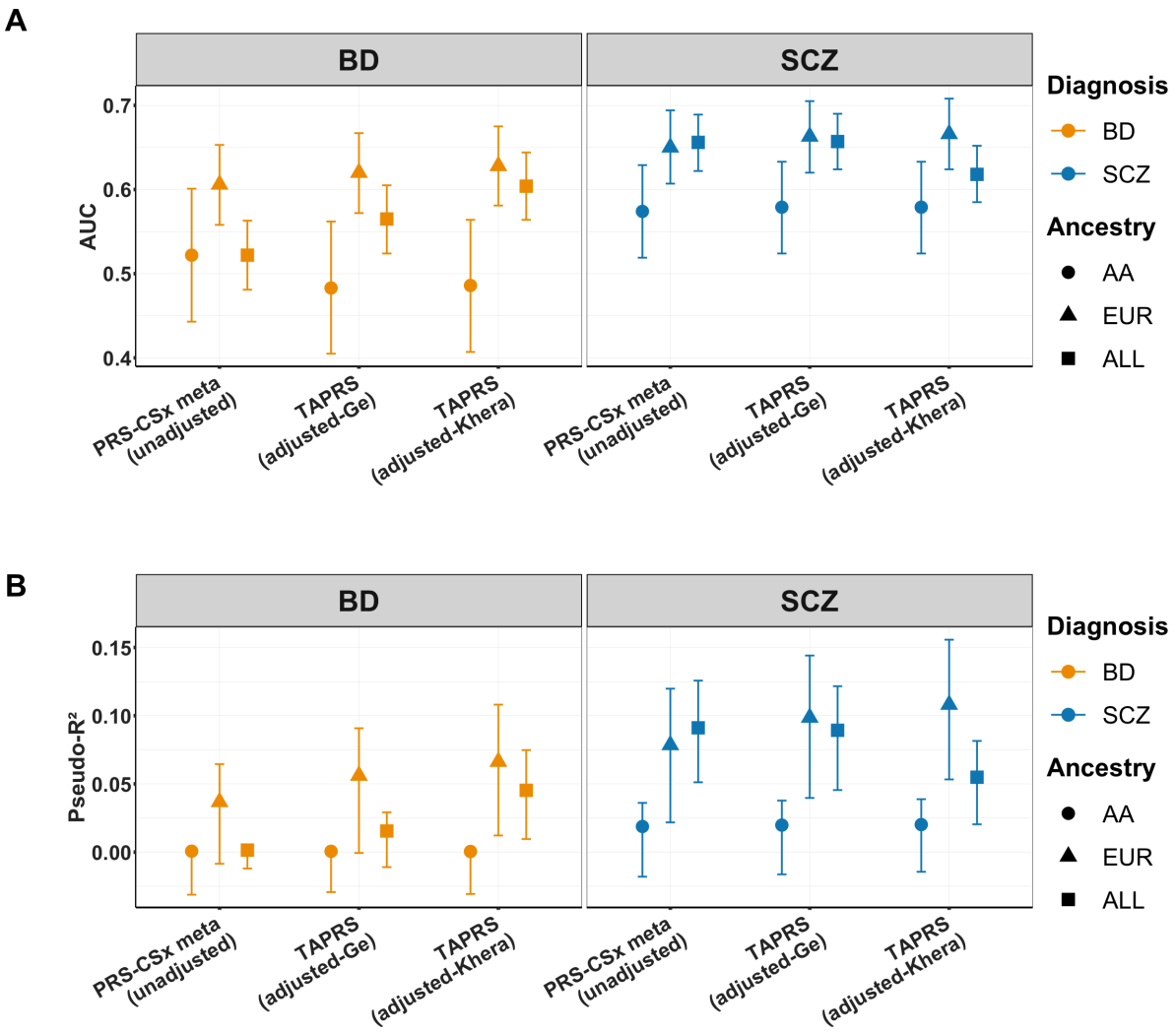

Figure S4. Prediction accuracy of PRS for case-control status before and after ancestry adjustment in each diagnostic group. (A) The area under the receiver operating characteristic (ROC) curve (AUC) of PRSs. (B) The proportion of the case-control variance (Nagelkerke's pseudo-R^2) explained by PRSs. Lines for Nagelkerke's pseudo-R^2 in (B) correspond to 95% confidence intervals calculated via 1000 bootstrapping. The combined sample (ALL) is the mixed samples of African American (AA) and European (EUR) ancestries. Terms: *--meta* option results for PRS-CSx. "Unadjusted" risk scores are prior to *post hoc* ancestry adjustment, and "adjusted" refers to TAPRS (trans-ancestry polygenic risk score) with *post hoc* ancestry adjustment of Khera or Ge. We find no overall advantage in prediction accuracy of case-control status for either adjustment method in SCZ only or BD only.

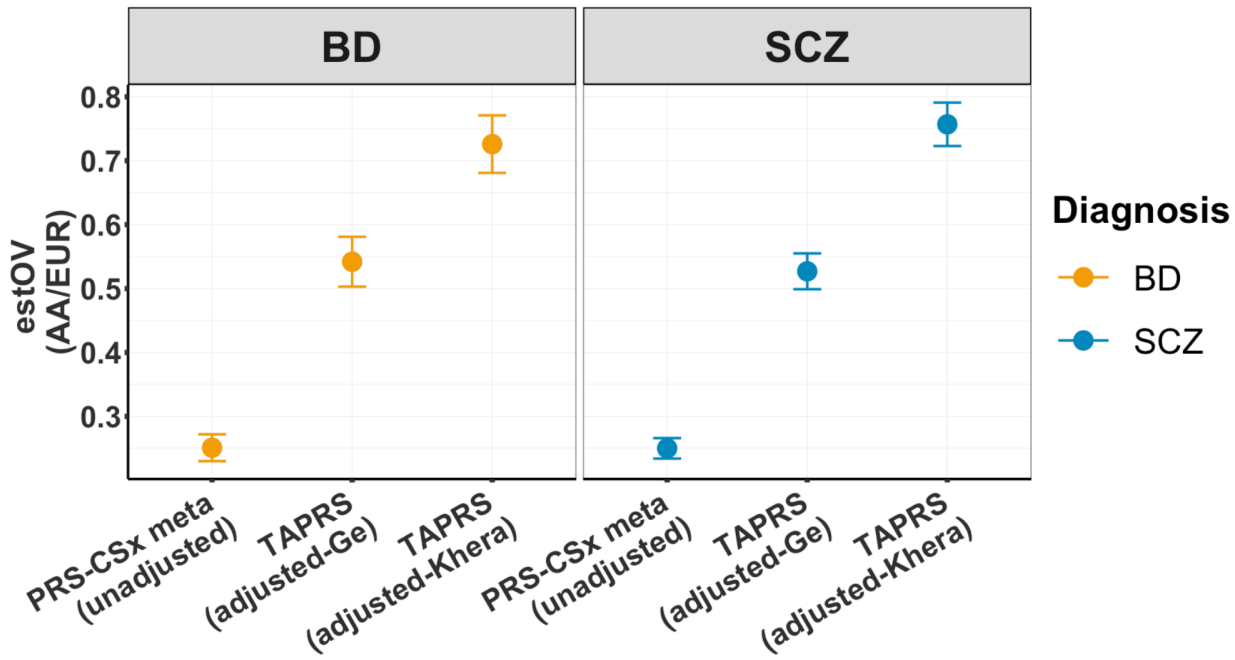

Figure S5. Effects of *post hoc* ancestry adjustment to generate TAPRS on the overlap of PRS distributions between ancestries in each diagnostic group. Figure shows the overlap of distributions between AA and EUR ancestries. EUR = European ancestry; AA = African American ancestry. 'estOV' = estimated overlapping area of risk scores between AA and EUR ancestries. Standard error of estOV is calculated by 1000 bootstrap draws (meaning 100 iterations with bootstrapping), and the labelled error bar of upper and lower values is estOV +/- SE. "Meta" is the --*meta* option for PRS-CSx. Between the two PRS *post hoc* ancestry adjustment methods, Khera adjustment gave greater PRS overlap between AA and EUR ancestries in SCZ only and BD only.

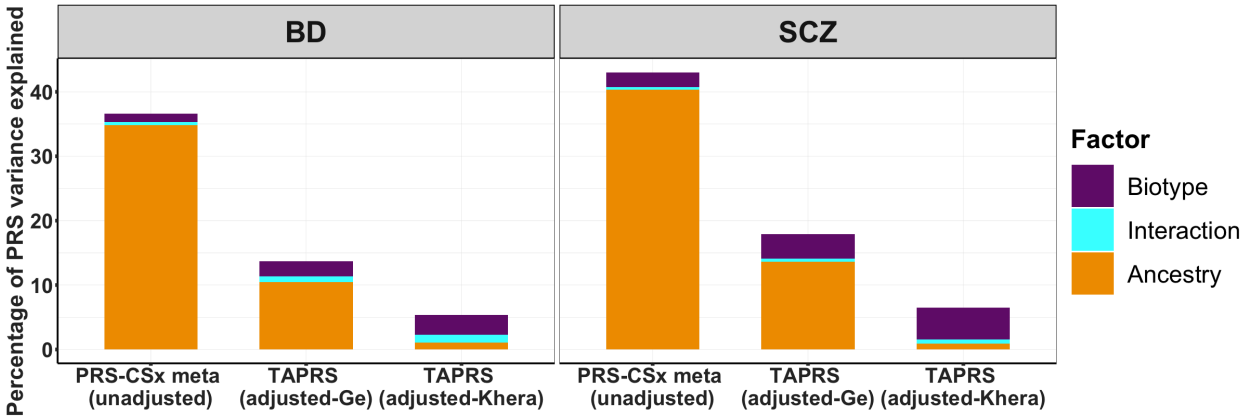

Figure S6. Percentage of PRS variance explained by each factor in two-way Analysis of Variance (ANOVA) in each diagnostic group. "meta" means we used the *--meta* option results for PRS-CSx. "Unadjusted" is PRS-CSx *meta* before *post hoc* ancestry adjustment. "Adjusted" refers to trans-ancestry risk scores (TAPRS) with *post hoc* ancestry adjustment of Khera or Ge. Minimal ancestry variance is desirable for TAPRS in a combined multi-ancestry sample. The Khera-adjusted TAPRS has the smallest ancestry variance in SCZ only and BD only. Residuals presumably include cumulative effects of individual SNPs.

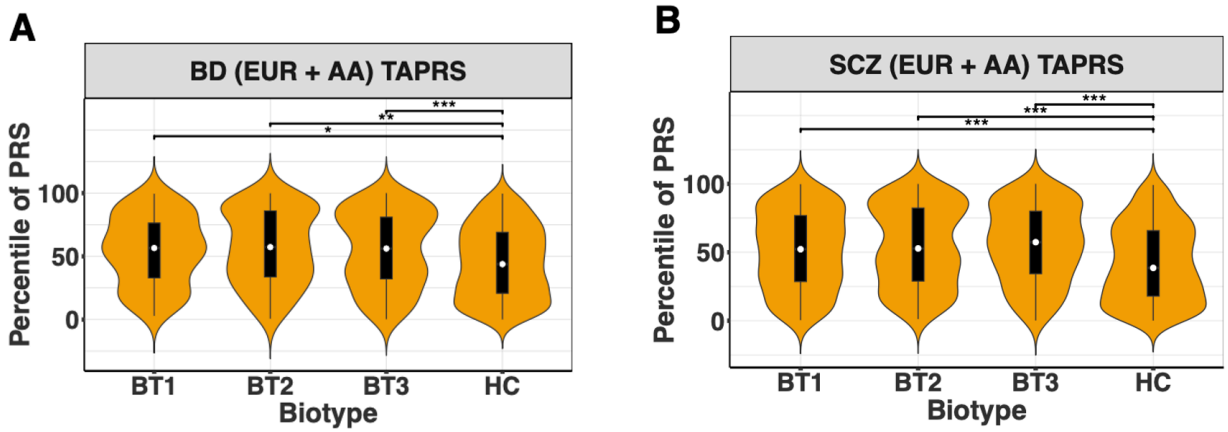

Figure S7. Trans-ancestry Biotype differences on TAPRS in each diagnostic group. (A) Violin plot of psychosis TAPRSs among Biotypes and with Healthy Control in the combined sample (EUR + AA) in BD group. (B) Violin plot of psychosis TAPRSs among Biotypes and with Healthy Control in the combined sample (EUR + AA) in SCZ group. EUR = European ancestry; AA = African American ancestry. The TAPRS is based on the Khera-adjusted results. Wilcoxon tests were used for the comparison. Bonferroni-corrected significance threshold over 6 two-sample Wilcoxon tests is P-value < 8.33E-03. Only significant comparison results are labeled with asterisks. *** indicates P-value < 1.67E-04, ** indicates P-value < 1.67E-03, * indicates P-value < 8.33E-03. Consistent with the Combined Diagnosis (SCZ + BD) results, the three Biotypes show no significant differences in TAPRSs in the separate diagnoses.