An Examination of Polygenic Score Risk Prediction in Individuals With First Episode Psychosis

Supplemental Information

Supplemental Methods & Materials

Quality Control of GAP Genotyping

As the GAP sample included blood (80%) and buccal (20%) DNA, all the genotypes underwent extensive manual QC using GenomeStudio. Only swabs when DNA met quality standards were included in the analysis. The genotyping rate of the buccal DNA was similar to the blood. On 7 individuals we genotyped DNA from both blood and saliva and found identical genotype calling (pi-hat > 0.9995).

Quality control (QC) included exclusion of SNPs with minor allele frequency (MAF) <1%, SNPs and individuals with genotypic failure >1% (stepwise process excluding genotyping failure at 10%, 5% and 1% levels), SNPs with Hardy Weinberg equilibrium $p<10^{-5}$ in controls, mismatch between recorded and genotypic sex, and related individuals. Cryptic relatedness and duplicated samples were identified with pairwise identity by descent method (pi-hat>0.1875). From each pair of related individuals we selected the one to include according to the following criteria: 1) existence of case/sex info, 2) proband preferred to relative, 3) number of SNPs with missing genotyping, 4) blood DNA preferred to buccal.

Imputation was performed with IMPUTE2 (1) based on the 1000 Genomes phase 3 reference panel (2), using haplotypes from all the ancestral populations (3). The imputed markers underwent a second stage of QC to exclude SNPs that were missing in more than 5% of individuals or had imputation information score (INFO) <0.8. QC was performed with PLINK 1.9 (4).

Sample Description for the African Ancestry Cohort (SLESS)

African ancestry controls (n=887) were included from the South London Ethnicity and Stroke Study (SLESS) (5). Inclusion criteria comprised being of Black Caribbean or Black African ethnicity, and free of clinical cerebrovascular disease. SLESS control recruitment was by a) random selection from primary care lists in St George's, Guy's and St Thomas', and King's College Hospital catchment areas between 1999 and 2012; b) emailing St George's University of London/St George's Hospital staff; c) affixing posters publicizing the study in local leisure centers, primary care surgeries, churches and communities centers. Using population controls from the same catchment area as cases reduced selection bias risk.

Quality Control Procedures to Combine the GAP African with the SLESS Sample

As the two samples were genotyped with different Illumina arrays (HumanCore Exome BeadChip was used in GAP and the MEGA array in SLESS), we selected only SNPs that existed in both arrays. After applying the same QC procedures in SLESS as in GAP (exclusion of SNPs with MAF <1%, SNPs and individuals with genotypic failure >1%, SNPs with HWE $p<10^{-5}$), we merged the two samples using only the markers that had been genotyped in both arrays. We excluded any related individuals between the two datasets and repeated the above QC in the merged sample. To avoid any systematic error due to the different genotyping arrays, we performed a GWAS of controls, looking for differences between GAP and SLESS. The QQ plot is presented in Suppl. Figure S1. The top 14 markers that deviated from the QQ plot were excluded from the PRS analyses. Finally, after constructing PRS for the total African sample, we tested for polygenic prediction of sample in controls only and we did not find any difference.

Statistical Analysis

To correct for population stratification we performed principal component analysis (PCA) using EIGENSTRAT (6). After LD pruning and frequency filtering of the genotyped SNPs, we performed PCA in the total sample and retained the eigenvectors for the first 10 principal

components (PCs). The scatterplot of the first two components was compared with the ethnicity as recorded in the demographic data (Suppl. Figure S2). Some mismatch between reported ethnicity and observed based on the PCA was noticed for individuals in the "mixed" or "other" ethnic groups. Only one person recorded as "European" was found to correspond to African and one recorded as "African" to Asian based on the PCA. Since our sample was ethnically heterogeneous, we selected two subsamples, based on the loadings on the first 2 PCs, one with European only ancestry and one with African ancestry (combining African and African-Caribbean origin). A second PCA was performed in each sample to ensure that the derived PCs captured more precisely the population stratification within each ethnic subsample. We checked PRS prediction of case-control status in the GAP European ancestry sample using either the PCs from the PCA of the total GAP sample or the PCs from the PCA on Europeans only and got very similar results (Suppl. Figure S3).

Association of PRS with case-control status was performed with logistic regression. We first fitted a model predicting case-control status only from the ten PCs and DNA origin (blood or buccal) and then the full model, including the polygenic score. The model was fitted for PRS at 10 different p-value thresholds and the proportion of variance explained was calculated by subtracting the Nagelkerke's R^2 of the baseline from the full model. This procedure was performed for the total sample and the two larger ethnic groups (Europeans and Africans) separately. To estimate heritability (i.e. variance explained at the liability scale) assuming a liability-threshold model, a lifetime risk of 1% for psychosis and 0.72% for schizophrenia (7), independent SNP effects, and adjusting for case-control ascertainment, we used the GEnetic Analysis Repository software (http://sourceforge.net/p/gbchen/wiki/GEAR/; Supplemental Table S1). The discriminative power of PRS in our sample was estimated with the area under the ROC curve (AUC) (8).

As a validation of our positive findings in the European FEP group, we repeated the analysis with a "negative control" PRS using the GWAS of height as training dataset. As expected, this PRS did not predict case-control status in our FEP sample (Suppl. Figure S4).

For each analysis we estimated and analyzed PRS at 10 different levels of significance at the discovery sample. To correct for multiple hypothesis testing in each sample (European FEP, African FEP, European chronic psychosis), we estimated the equivalent number of effective tests using the correlation matrix (http://gump.qimr.edu.au/general/daleN/matSpD/) and performed Bonferroni correction on those. The effective number of independent variables were 5, 6.4, and 5.3 and the significance threshold required to keep Type I error rate at 5% were 0.01, 0.008 and 0.009 respectively.

To evaluate the specificity of PRS to schizophrenia, we divided cases according to each diagnostic approach (consensus, OPCRIT/DSM, OPCRIT/ICD, clinical) into two diagnostic categories (schizophrenia, and other psychoses) of European (two samples; FEP and chronic patients) and African ancestry. Diagnosis with at least one diagnostic approach was available in 151 FEP, 132 chronic European, and 177 FEP African cases. In an exploratory analysis we compared the adjusted PRS between each of the two diagnostic groups, those who had met criteria for schizophrenia with at least one of the four diagnostic approaches and those who had not met criteria for schizophrenia and were denoted as "any other psychosis" with controls (Suppl. Figure S5). As PRS and principal components were prepared separately for Europeans and Africans and direct comparison was not relevant, we performed a linear regression of PRS on the 10 principal components and we saved the standardized residuals. We then estimated standardized mean difference (Cohen's d) between the standardized residuals of schizophrenia or other psychosis and corresponding controls in each sample.

Using logistic regression we estimated variance explained at 10 predefined thresholds (P_T) for case-control analysis of all psychoses, schizophrenia only, any other psychosis only and case-only analysis comparing schizophrenia with other psychoses. In a secondary analysis,

using the “high resolution” function of (PRSice) (9) we also identified the P_T that maximizes PRS discriminative ability. Barplots and high resolution plots for each sample are presented in Suppl. Figures S6-S8.

To better visualize the effect of PRS on the risk of psychosis, we estimated case-control odds ratio (OR) at various levels of PRS. To obtain adjusted PRS, corrected for population stratification, first we regressed the most predictive PRS scores on the 10 PCs and saved the residuals. We then ranked all the individuals by the adjusted PRS, divided the sample in percentiles, measured the case-control ratio in each and estimated OR versus the baseline group of the median PRS. As our sample was relatively small, to have enough cases in each group for a better estimation of OR, we divided the sample in 5 quintiles and we computed 95% confidence intervals (CI) with Woolf’s method (10).

To be able to compare our estimates with the outcomes of the PGC2 schizophrenia study (11), we employed a simulation method to extract deciles from our observed data. In brief, we measured the mean and standard deviation (SD) of the adjusted PRS in cases and controls separately, and we constructed similar distributions with the same characteristics, but larger numbers, assuming an underlying normal distribution for PRS. It should be noted that the estimates of R^2 and OR reflect the ascertainment in case-control studies, where the ratio of cases and controls do not reflect the underlying disease prevalence in the population (12, 13). To address this problem, we simulated data by taking 1 case for 99 controls, to approximate a prevalence of psychosis of about 1% (7), and we followed the previous steps to estimate ORs. This gives a more representative picture of the risk of psychosis at different levels of PRS in an unselected sample. The overall distribution of simulated data was divided in deciles as previously described and ORs were calculated for each decile compared to the baseline (lowest PRS). The above procedure was repeated 100 times and we retained the mean of the estimated OR for each decile of PRS (Suppl. Figure S9).

Supplemental Tables

Table S1. Variance explained of case-control status for the total psychosis and schizophrenia-only at the observed data and liability scale. For the estimation of R^2 on the liability scale we took a lifetime risk of 1% for psychosis 1% and 0.72% for schizophrenia and adjusted for case-control ascertainment.

Sample	Cases	N cases	N controls	R^2 (%)	R^2 (%) on liability
FEP Europeans	All psychoses	160	167	9.4	5.19
	Schizophrenia	86	167	16.3	9.3
FEP Africans	All psychoses	188	881	1.1	1.05
	Schizophrenia	135	881	2.4	2.67
Chronic European	All psychoses	135	165	6	3.34
	Schizophrenia	75	165	10	5.96

Table S2. Results of PRS association with case-control status in the total GAP sample, the European FEP subsample, IMPACT chronic European sample and the African sample combining GAP with SLESS.

	GAP all			GAP EUR			IMPACT EUR			GAP SLESS AFR		
P_T	p-value	R^2	N SNPs	p-value	R^2	N SNPs	p-value	R^2	N SNPs	p-value	R^2	N SNPs
5E-08	0.0247	0.009	116	0.0044	0.034	107	0.0879	0.012	102	0.3307	0.001	100
1E-05	0.0176	0.010	387	0.014	0.025	368	0.0583	0.015	354	0.2435	0.002	377
0.0001	0.0019	0.017	860	0.0009	0.046	804	0.0118	0.027	801	0.0123	0.009	842
0.001	1E-04	0.028	2099	7E-05	0.068	1945	0.043	0.017	1906	0.0077	0.010	2096
0.01	5E-05	0.030	6170	7E-05	0.068	5640	0.0028	0.039	5595	0.0072	0.010	6296
0.05	8E-05	0.028	14172	2E-05	0.079	12749	0.0016	0.043	12857	0.0048	0.011	15093
0.1	7E-05	0.028	20654	4E-06	0.094	18413	0.0002	0.060	18488	0.0045	0.011	22322
0.2	5E-05	0.030	30281	1E-05	0.084	26760	0.0009	0.048	26813	0.0073	0.010	33538
0.5	1E-05	0.035	49386	1E-05	0.081	42781	0.0019	0.042	43152	0.0041	0.011	57161
1	1E-05	0.036	65479	3E-05	0.075	56059	0.0017	0.043	57566	0.0046	0.011	81816

P_T the P value threshold in the original PGC2 training sample. We report the p-value, variance explained (R^2) and the N of SNPs after QC and clumping in each P value threshold.

Supplemental Figures

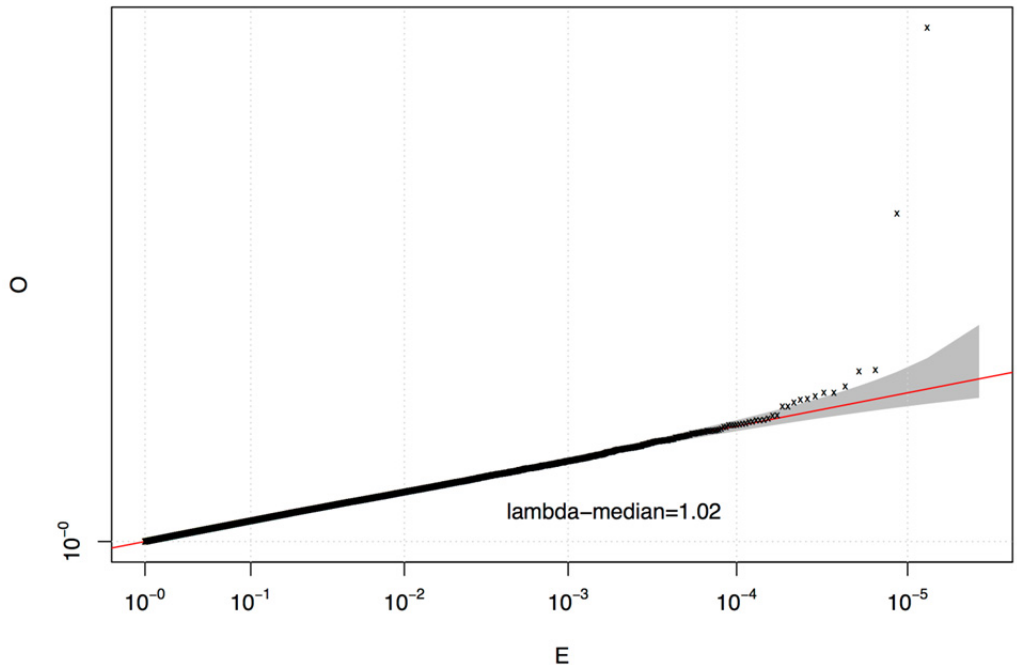

Figure S1. QQ plot of GWAS of GAP African controls with SLESS controls.

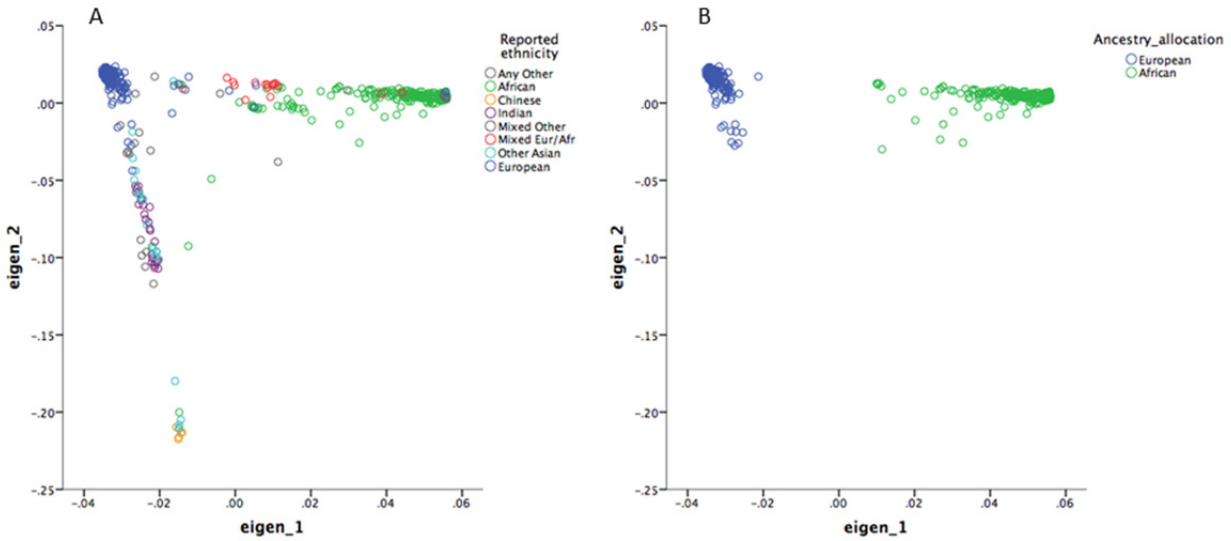

Figure S2. (A) Scatter plot of the first two Principal Components in the total GAP sample. The colors correspond to the reported ethnicity of each individual as recorded in the demographic data. Some mismatch was noticed for individuals with mixed or other ethnicity. **(B)** Scatter plot of individuals allocated in the European or African ancestry groups. Following the PCA in the total sample we selected a subsample with the 1st PC < -0.2 and the 2nd PC > -0.035 designated as Europeans and a subsample with 1st PC > 0.01 designated as Africans. In each subsample we performed separate PCA and excluded outliers.

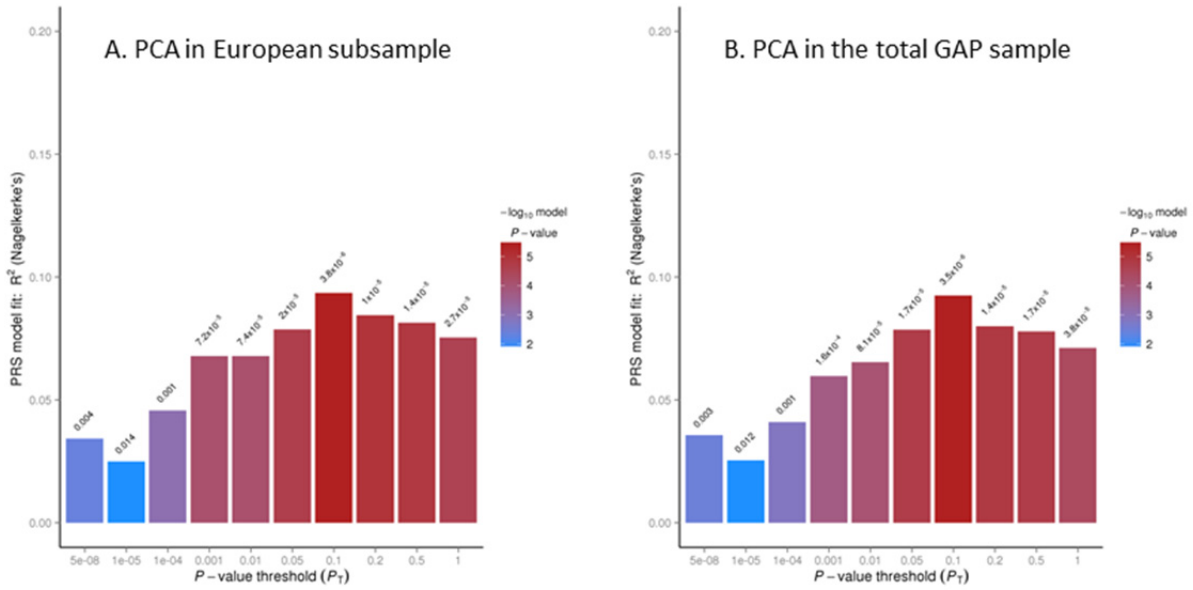

Figure S3. PRS prediction of case-control status in the GAP sample using as covariates the principal components of the PCA in the Europeans only **(A)** or in the total GAP sample **(B)**.

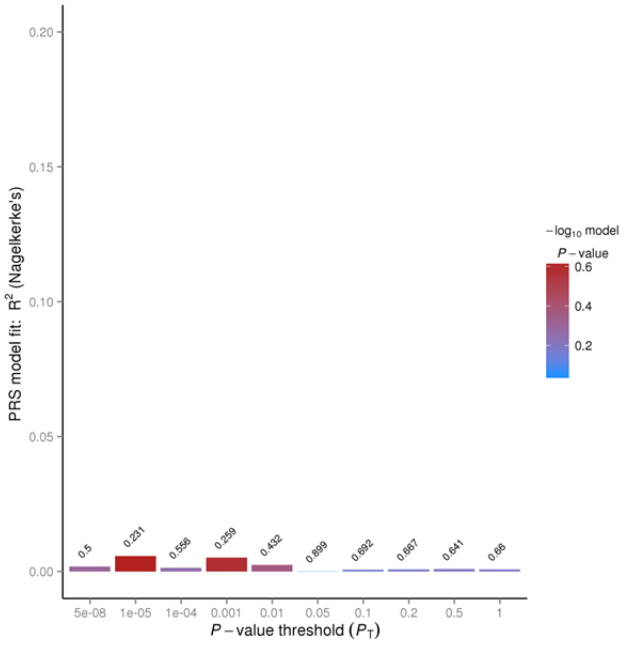

Figure S4. Prediction of case-control status in the European FEP sample using PRS from GWAS of height (as negative control).

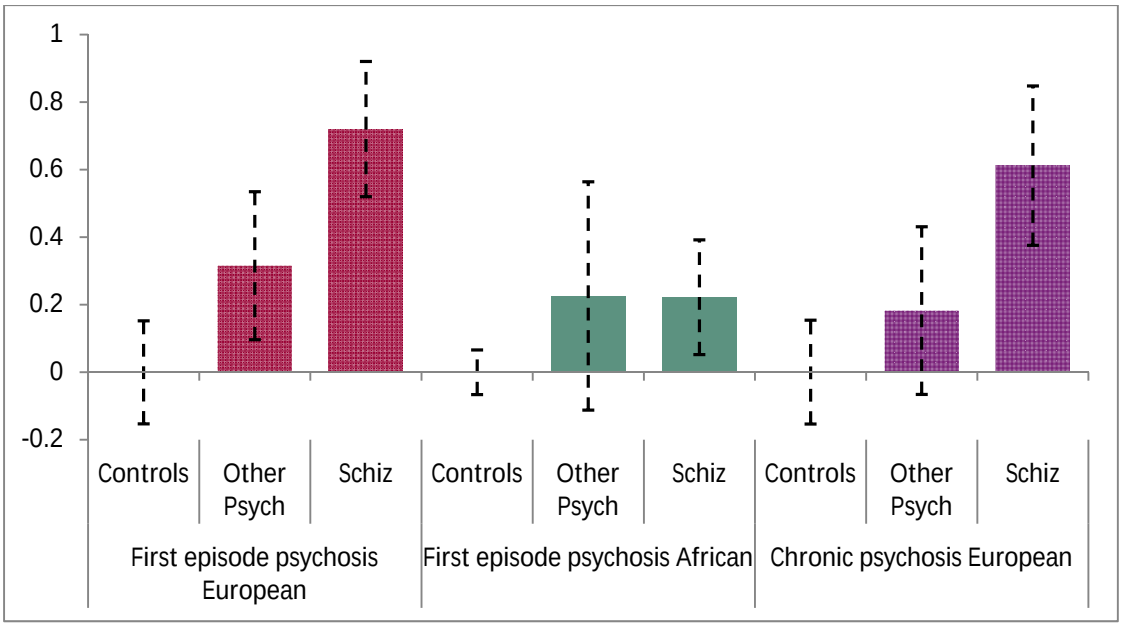

Figure S5. PRS at P_T<0.1 of controls and cases stratified by diagnosis in the 3 samples. Residuals of PRS in each sample after linear regression of the 10 principal components from the PCA on the PRS score at P_T<0.1. The y axis represents residuals of PRS standardized in the corresponding control group. The vertical bars represent 95% CI. Other, for all other psychoses and Schiz, for schizophrenia.

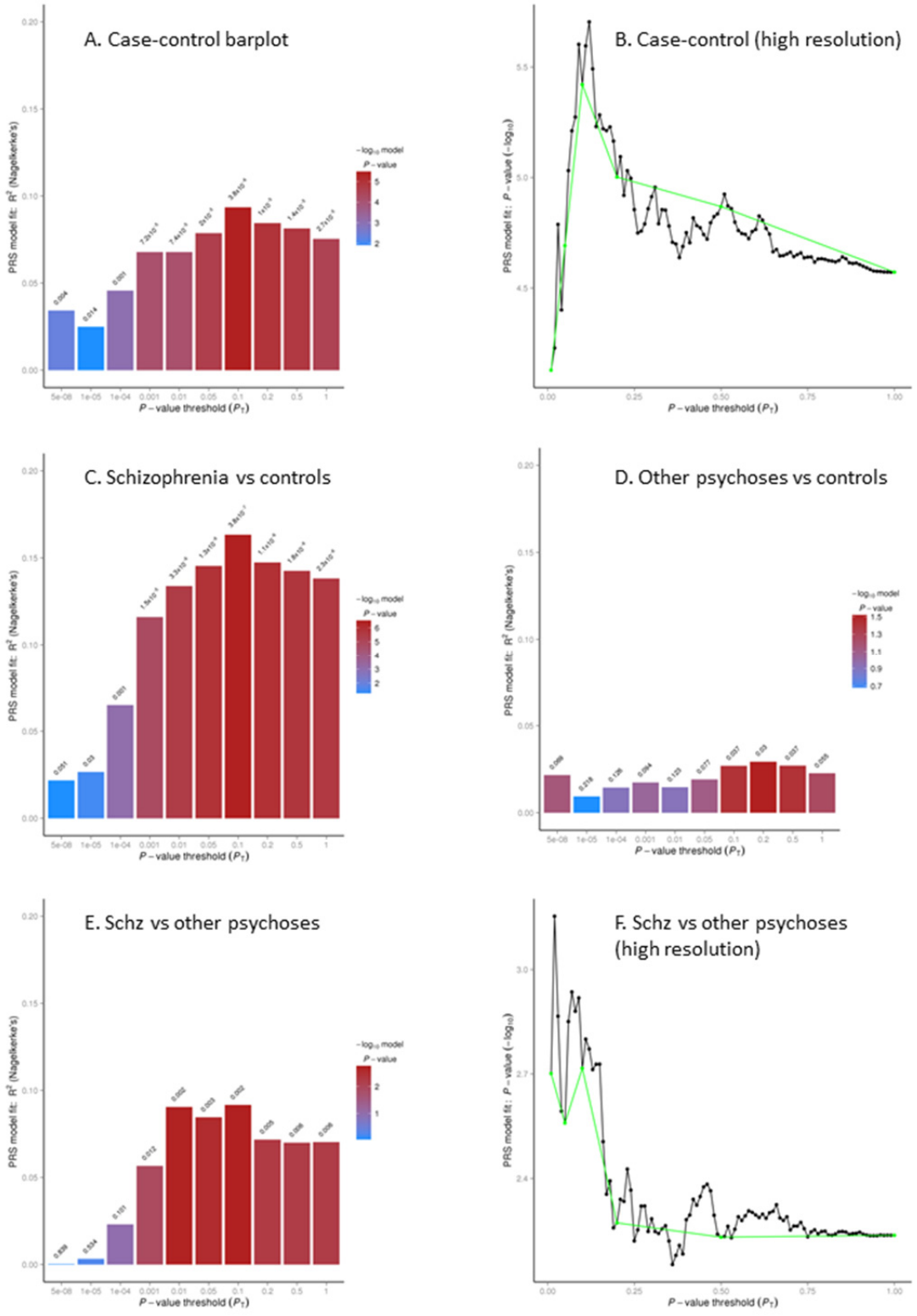

Figure S6. Barplots and high resolution plots of case-control of any psychosis **(A, B)**, schizophrenia only **(C)**, other psychoses **(D)** and case only of schizophrenia vs other psychoses **(E, F)** in the FEP European sample.

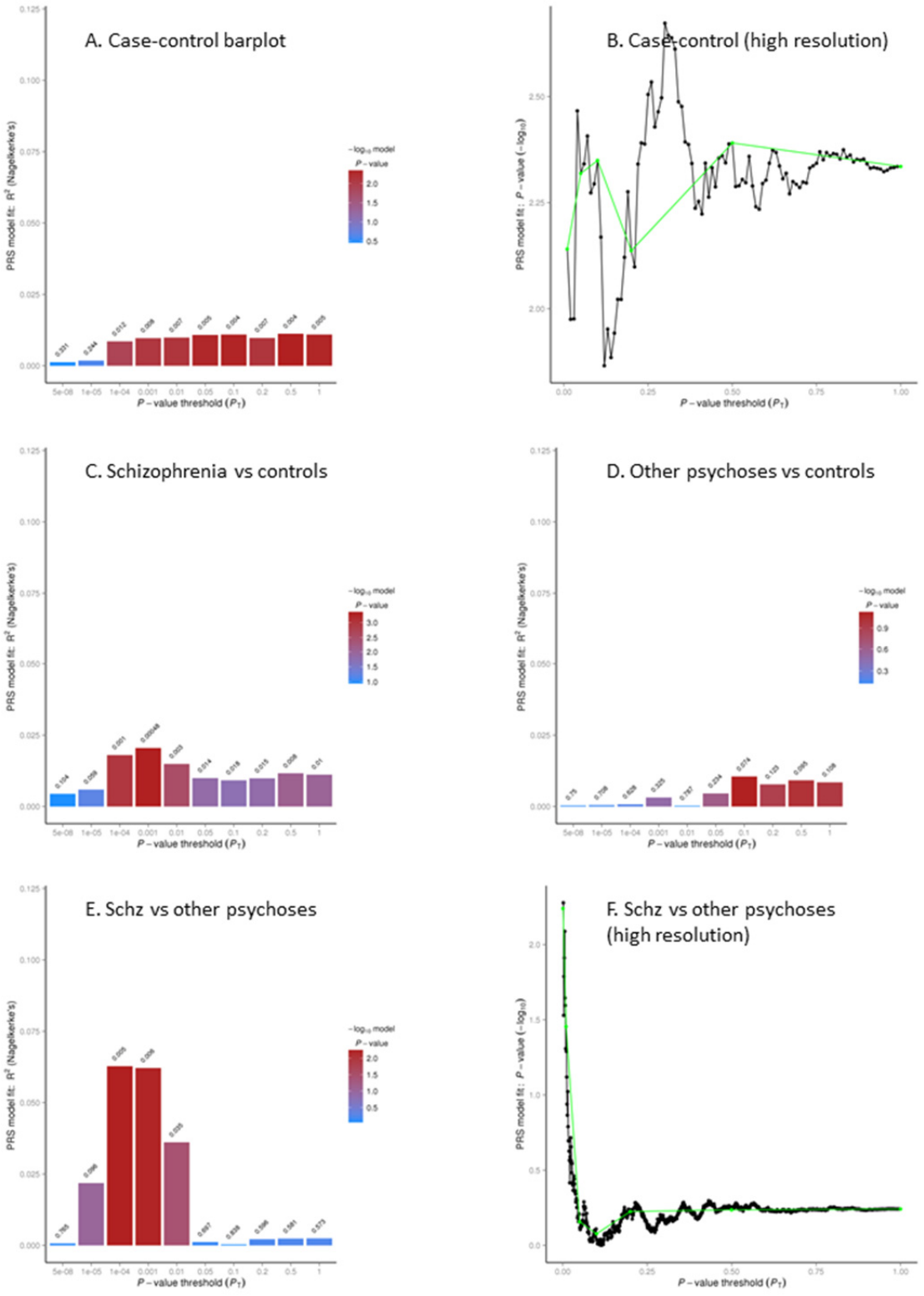

Figure S7. Barplots and high resolution plots of case-control of any psychosis **(A, B)**, schizophrenia only **(C)**, other psychoses **(D)** and case only of schizophrenia vs other psychoses **(E, F)** in the FEP African sample.

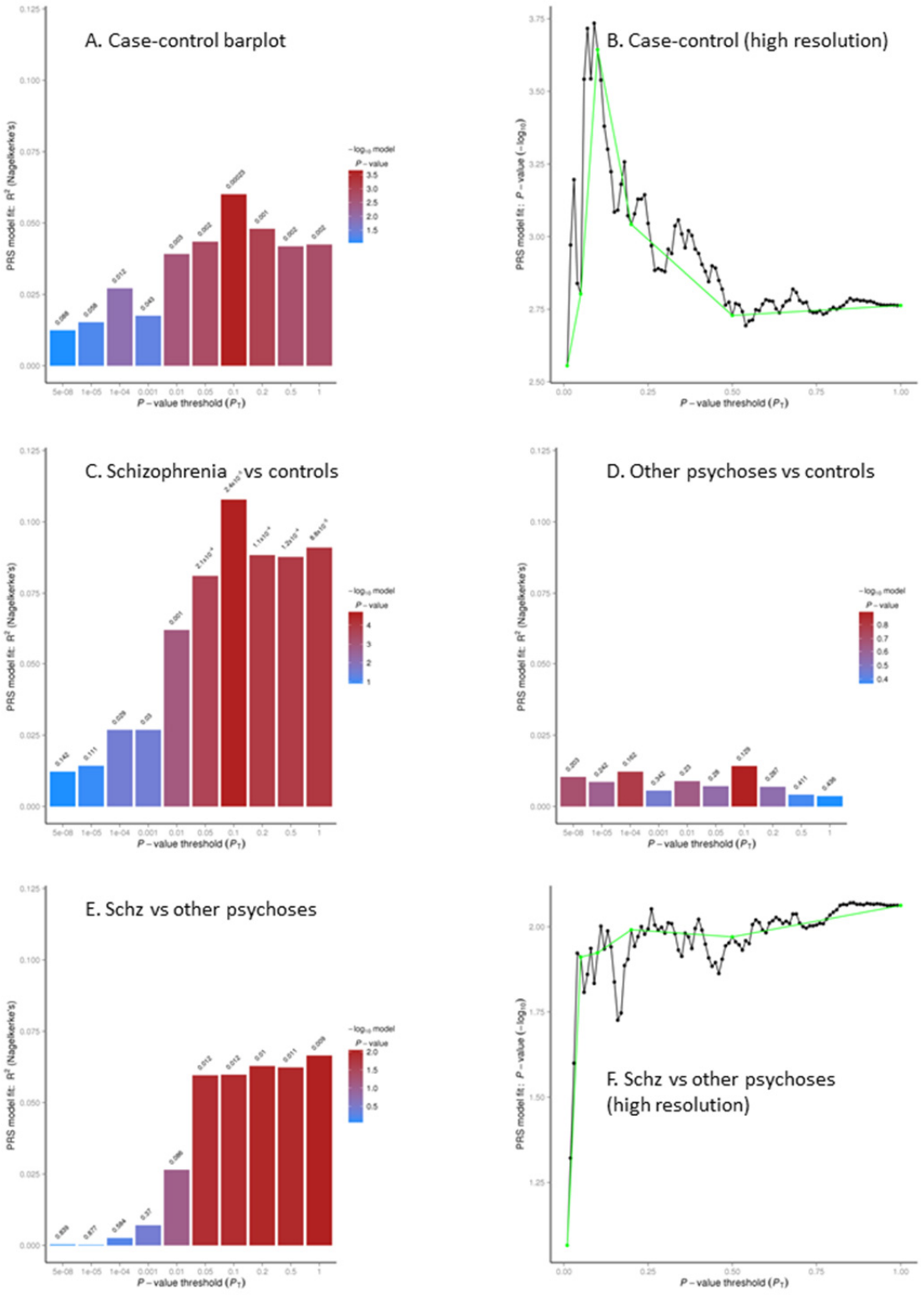

Figure S8. Barplots and high resolution plots of case-control of any psychosis **(A, B)**, schizophrenia only **(C)**, other psychoses **(D)** and case only of schizophrenia vs other psychoses **(E, F)** in the chronic psychosis European sample.

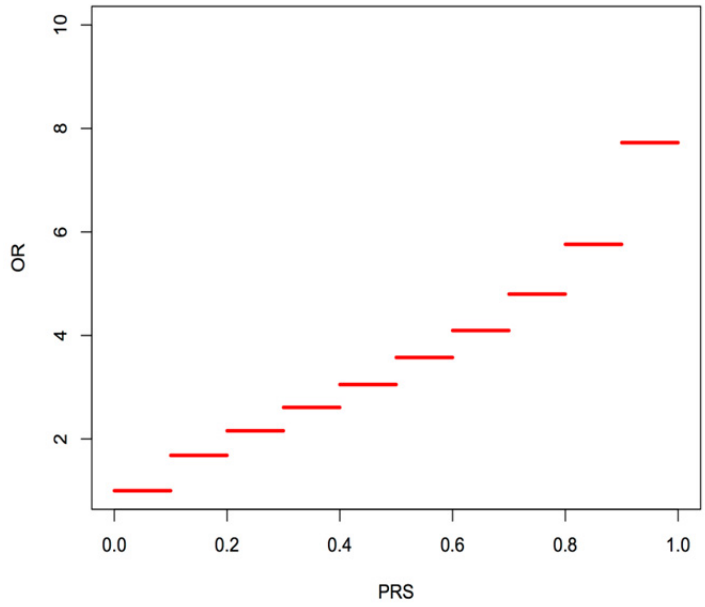

Figure S9. OR of simulated data of FEP Europeans divided in deciles assuming that the prevalence of psychosis is 1% with the lowest PRS group as baseline.

Supplemental References

1. Howie BN, Donnelly P, Marchini J (2009): A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *Plos Genet*. 5.
2. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. (2012): An integrated map of genetic variation from 1,092 human genomes. *Nature*. 491:56-65.
3. Howie B, Marchini J, Stephens M (2011): Genotype imputation with thousands of genomes. *G3*. 1:457-470.
4. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015): Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 4:7.
5. Khan U, Crossley C, Kalra L, Rudd A, Wolfe CD, Collinson P, et al. (2008): Homocysteine and its relationship to stroke subtypes in a UK black population: the south London ethnicity and stroke study. *Stroke; a journal of cerebral circulation*. 39:2943-2949.
6. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006): Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 38:904-909.
7. Saha S, Chant D, Welham J, McGrath J (2005): A systematic review of the prevalence of schizophrenia. *PLoS Med*. 2:e141.
8. Egan JP (1975): *Signal detection theory and ROC analysis*. New York ; London: Academic Press.
9. Euesden J, Lewis CM, O'Reilly PF (2015): PRSice: Polygenic Risk Score software. *Bioinformatics*. 31:1466-1468.
10. Woolf B (1955): On estimating the relation between blood group and disease. *Annals of human genetics*. 19:251-253.
11. Psychiatric Genomics Consortium Schizophrenia Working Group (2014): Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 511:421-427.
12. Witte JS, Visscher PM, Wray NR (2014): The contribution of genetic variants to disease depends on the ruler. *Nature reviews Genetics*. 15:765-776.
13. Lee SH, Goddard ME, Wray NR, Visscher PM (2012): A better coefficient of determination for genetic profile analysis. *Genetic epidemiology*. 36:214-224.