SUPPLEMENTAL MATERIAL

for

Joint contributions of rare CNVs and common SNPs to risk for schizophrenia

Bergen et al. 2018

Supplementary Table S1. Information on contributing studies, including country of origin, study PI, presence of family trios in the study, CNV- and GWAS genotyping arrays used as well as the numbers of subjects passing CNV quality control (QC).

Collection country	Data ID	PI	Trio_fam	CNV genotyping array	GWAS genotyping array	QC final dataset size	QC final CNV cases	QC final CNV controls
Scotland	aber	St Clair	no	A5.0	A6.0	1375	690	685
Portugal	port	Pato	no	A5.0	A6.0	472	281	191
Sweden	swe1	Sullivan	no	A5.0	A5.0	265	140	125
UK	uclo	McQuillan	no	A5.0	A6.0	461	461	0
USA	cati	Sullivan	no	A500	A500	515	314	201
7 countries	pewb	Bramon	no	A6.0	I1M	1630	365	1265
Spain	pews	Bramon	no	A6.0	I1M	60	42	18
USA	mgs2	Gejman	no	A6.0	A6.0	4913	2537	2376
Bulgaria	buls	Kirov	no	A6.0	A6.0	767	182	585
Ireland	dubl	Corvin	no	A6.0	A6.0	1058	252	806
Scotland	edin	Blackwood	no	A6.0	A6.0	617	341	276
Sweden	s234	Sullivan	no	A6.0	A6.0	3223	1314	1909
Norway	top8	Andreassen	no	A6.0	A6.0	431	166	265
Ireland	irwt	Corvin	no	A6.0	A6.0	1886	1089	797
Bulgaria	butr	Kirov	yes	A6.0	A6.0	293	293	0
Canada/USA	lktu	Knight	no	A6.0	A6.0	172	172	0
USA	msaf	Buxbaum	no	A6.0	A6.0	399	274	125
Germany	munc	Rujescu	no	I300	I317	695	410	285
Germany	boco	Rietschel/Rujescu	no	I550	I550	1711	458	1253
USA	ucla	Ophoff	no	I550	I550	1124	672	452
Australia	asrb	Mowry	no	I610	I650	611	367	244
Denmark	denm	Werge	no	I610	I650	830	451	379
USA	cims	Buxbaum	no	omni_express	ill	89	35	54
UK	clo3	O'Donovan	no	omni_express	omni	3165	2096	1069
Estonia	egcu	Esko	no	omni_express	omni	1347	229	1118
Sweden	swe5	Sullivan	no	omni_express	omni	4238	1729	2509
Sweden	swe6	Sullivan	no	omni_express	omni	2077	952	1125
UK	uktr	Kirov	yes	omni_express	omni	39	39	0
Sweden	umeb	Adolfsson	no	omni_express	omni	850	325	525
Sweden	umes	Adolfsson	no	omni_express	omni	848	186	662
UK	clm2	O'Donovan	no	omni_express_plus	I1M	3418	3418	0
UK	cou3	Walters	no	omni_express_plus	omni	1127	526	601
Sweden	ersw	Jönsson	no	omni_express_plus	omni	577	260	317
USA	cims	Petryshen	no	omni_2.5	ill	38	28	10

Supplementary Table S2. Previously associated CNVs and reported effect sizes from: Rees E, Walters JT, Georgieva L, Isles AR, Chambert KD, Richards AL, et al. Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry.* 2014;204(2):108-14. Number of carriers by affection status and study odds ratio are for the current material.

CNV Type	Implicated Locus	Lit. OR[1]	Cases	Controls	Study OR[2]
Deletions	22q11.2	28.3	56	0	∞
	15q11.2	2.2	97	50	1.9
	1q21.1	8.4	30	6	4.8
	3q29	57.7	14	0	∞
	15q13.3	7.5	23	2	11.0
	NRXN1 chr2	9.0	36	3	11.5
Duplications	16p11.2	11.5	53	4	12.7
	7q11.23	11.4	10	0	∞
	1q21.1	3.5	20	4	4.8
	16p13.11	2.3	57	46	1.2
	15q11.2-15q13.1	13.2	11	0	∞

[1] Lit. OR = odds ratio as reported in the literature

[2] Study OR = crude odds ratio estimated from our material

Supplementary Table S3. Average PRS for carriers and non-carriers, separately for cases and controls, by individual CNV. Ref = reference

Status	CNV	N	Mean	Difference	P-Value
Cases	Non-carrier	20681	0.97 (0.017)	ref	ref
OR <5	15q11.2_del	97	1.20 (0.222)	0.22 (0.250)	0.370
	16p13.11_dup	57	0.74 (0.308)	-0.23 (0.326)	0.480
	1q21.1_dup	20	0.90 (0.529)	-0.07 (0.551)	0.900
5 <= OR < 15	15q13.3_del	23	1.37 (0.450)	0.40 (0.514)	0.440
	1q21.1_del	30	0.45 (0.410)	-0.52 (0.450)	0.250
	NRXN1	36	0.78 (0.371)	-0.19 (0.411)	0.640
	7q11.23_dup	10	2.17 (0.826)	1.20 (0.779)	0.120
	16p11.2_dup	53	0.76 (0.352)	-0.22 (0.339)	0.520
	15q11.2_dup	11	0.80 (0.831)	-0.17 (0.742)	0.820
15 < OR	22q11_del	56	-0.59 (0.350)	-1.56 (0.329)	**2.2e-06**
	3q29_del	14	-0.15 (0.649)	-1.12 (0.658)	0.089
Controls	Non-carrier	20107	-1.01 (0.017)	ref	ref
OR <5	15q11.2_del	50	-0.90 (0.267)	0.11 (0.341)	0.74
	16p13.11_dup	46	-1.38 (0.379)	-0.37 (0.356)	0.30
	1q21.1_dup	4	-1.81 (1.562)	-0.80 (1.207)	0.51
5 <= OR < 15	15q13.3_del	2	0.88 (0.945)	1.89 (1.706)	0.27
	1q21.1_del	6	0.43 (0.256)	1.44 (0.985)	0.14
	NRXN1	3	-1.20 (0.795)	-0.19 (1.393)	0.89
	7q11.23_dup	0	-	-	-
	16p11.2_dup	4	0.91 (1.393)	1.92 (1.207)	0.11
	15q11.2_dup	0	-	-	-
15 < OR	22q11_del	0	-	-	-
	3q29_del	0	-	-	-

Supplementary Table S4a. Logistic liability models for schizophrenia as a function of PRS and CNV status in non-carriers of specific CNVs (n = 40,732). *Large deletions* codes the presences of deletions ≥500kb as a binary 0/1 variable, and *Total CNV burden* includes the sum of measured CNVs (in kb) as a continuous linear predictor. "+" indicates an additive main effect, "x" a full interaction model with both main effects and an interaction term.

All models are adjusted for collection site, sex, population substructure and CNV-metric. P-values are based on a likelihood-ratio test of the null hypothesis that the model of interest does not perform better than the reference model.

Model	**Genetic Exposure**	**df**	**AIC**	**R^2**	**Ref Model**	**Delta AIC**	**Delta R^2**	**p-value**
0	None	40	5787.0	30.94	-	-	-	-
1	PRS only	41	1072.0	42.13	0	4715.0	11.19	1E-99
2	Large delonly	41	5778.2	30.96	0	8.8	0.02	0.00098
3	Total CNV burden only	41	5772.6	30.98	0	14.4	0.04	0.000051
4	PRS + large deletion	42	1062.9	42.16	1	9.1	0.03	0.00089
5	PRS + total CNV burden	42	1057.7	42.17	1	14.3	0.04	0.000056
6	PRS x large deletion	43	1064.9	42.16	4	-2.0	0.00	0.91
7	PRS x total CNV burden	43	1058.2	42.17	5	-0.5	0.00	0.21

Supplementary Table S4b. Logistic liability models for schizophrenia as a function of GRS and CNV status in carriers of specific CNVs (n = 522). *Specific CNV-OR* includes the previously reported log(OR) for the specific CNV as predictor. *Large deletions* codes the presences of deletions ≥500kb as a binary 0/1 variable, and *Total CNV burden* includes the sum of measured CNVs (in kb) as a continuous linear predictor. "+" indicates an additive main effect, "x" a full interaction model with both main effects and an interaction term.

All models are adjusted for collection site, sex, population substructure and CNV-metric. P-values are based on a likelihood-ratio test of the null hypothesis that the model of interest does not perform better than the reference model.

Model	Genetic Exposure	df	AIC	R^2	Ref Model	Delta AIC	Delta R^2	p-value
0	None	37	146.0	33.5	-	-	-	-
1	PRS only	38	117.6	40.3	0	28.4	6.8	3.5E-08
2	CNV-OR only	38	65.8	51.0	0	80.2	17.5	1.2E-19
3	Large del only	38	141.4	35.0	0	4.6	1.5	0.01
4	Total CNV burden only	38	147.8	33.5	0	-1.8	0.1	0.63
5	PRS + CNV-OR	39	34.1	57.4	1	83.5	17.1	2.3E-20
6	PRS + Large del	39	110.6	42.2	1	7.0	1.9	0.0027
7	PRS x CNV-OR	40	28.3	58.8	5	5.8	1.4	0.0052
8	PRS x Large del	40	91.8	46.6	6	18.8	4.3	5.1E-06
9	PRS x CNV-OR + large del	41	30.2	58.8	7	-1.9	0.0	0.77

Supplementary Methods

Polygenic risk scores Polygenic risk scores (PRS) were calculated for the same p-value thresholds as previously used[1]. In a multi-variable logistic regression model for disease status, adjusted for site, sex, CNV quality and five ancestral components, the strength of the association between the different PRS and case-control status in our material was comparable to what has been reported previously[1] (Extended data figures 5, 6a), see Supplementary Table 6 and Supplementary Figure 1A.

Definition of PRS1 Given that the PRS at different p-value thresholds are by construction correlated, we attempted to concentrate the information contained in the ten original scores S1-S10 by considering a suitable subset of the ten principal components as weighted indices for polygenic risk. We found that the first principal component of the PRS explained indeed 65% of the overall variability across the ten underlying scores S1-S10, and contributed 11.1% to the R^2 of the multivariable model, which is more than any of the original scores (maximum R^2 10.2% for S6/S7, Supplementary Table 6). However, eight out of the nine remaining principal components were still highly significantly associated with schizophrenia in a multivariable model (range of p-values: 4E-5 to 3E-273), offering little opportunity for conceptual simplification.

We found however that normalization of the PRS across sites, as discussed below, and subsequent principal component analysis of the normalized scores was extremely effective in concentrating polygenic risk information: the first principal component of the normalized scores, referred to as PRS1, explains 69% of the variability across the normalized scores and contributes 11.2% to the R^2 of the multivariable model. All other principal components have an R^2 of less than 0.05% and are not statistically significantly associated with schizophrenia, with the exception of component PRS8 with p-value p=0.04 (Supplementary Table 7 and Supplementary Figure 1B-D). Consequently, we used PRS1 throughout as summary measure of polygenic risk for schizophrenia in the analyses presented in the paper.

Normalization of the original PRS During initial quality control, we had previously found extensive variability in PRS levels between sites, with between-site differences accounting for between 16% and 66% of total PRS variability across different p-value thresholds (Supplementary Figure 2A). For e.g. score S6, which exhibits intermediate between-site variability, this already translates into dramatic shifts in distribution between sites which is difficult to explain in terms of differential disease risk (Supplementary Figure 2B).

Visual inspection shows that almost all of the between-site variability is shared between cases and controls: for all thresholds, the average PRS among cases follows the average PRS among controls extremely closely (Supplementary Figure 3). We can model this relationship by fitting a linear regression for the mean PRS among cases ($\overline{\mathrm{PRS}}_{\mathrm{Cases}}$) as a function of the mean PRS among controls ($\overline{\mathrm{PRS}}_{\mathrm{Controls}}$) as

$$\overline{\mathrm{PRS}}_{\mathrm{Cases}} = \beta_0 + \beta_1 \overline{\mathrm{PRS}}_{\mathrm{Controls}} \qquad (SuppEq.1)$$

for all 28 sites that contribute both cases and controls; these are the regression lines shown in Supplementary Figure 3. The corresponding R^2 for these ten models varies from 92% to 99% (median: 98%), suggesting that more than 90% of the between-site variability is indeed shared between cases and controls.

Furthermore, the corresponding estimates $\hat{\beta}_1$ for the slopes of these models are all close to 1 (range: 0.90-1.05, median: 1.00), and their 95% confidence intervals all cover 1. We can therefore simplify the model by fixing the slope β_1 at 1 without loss of generality:

$$\overline{\mathrm{GRS}}_{\mathrm{Cases}} = \beta_0 + \overline{\mathrm{GRS}}_{\mathrm{Controls}} \qquad (SuppEq.2)$$

This means that for our collection of sites with both cases and controls, the average PRS among cases and controls differs by a fixed constant; in other words, the excess polygenic risk among cases compared to controls is (on average) constant across sites. We can therefore eliminate more than 90% of the between-site variability simply by subtracting from all PRS values measured at one site the mean PRS among controls at that site: in other words, we align the site distributions of PSR so that all controls are centered at zero.

For sites that only contributed cases (n = 5), this does not work, as we cannot estimate the mean PRS among cases. Instead, we use the estimated intercept $\hat{\beta}_o$ to align the case means directly.

We re-scaled all PRS normalized in this manner to have the same mean and standard deviation as the original scores S1-S10, which produced the normalized scores nS1-nS10 used in the principal component analysis above. As expected, these normalized scores show considerably less variability between sites (Supplementary Figure 1C).

Crucially, this normalization procedure does not affect the association between PRS and schizophrenia: when comparing the normalized scores nS1-nS10 in Supplementary Table 8 with the original scores S1-S10 in Supplementary Table 6, we find that the unscaled odds ratios and their p-value as well as the R^2 contributions are identical. This is due to the fact that the multivariable logistic regression model underlying these estimates is by necessity adjusted for site to allow for site-specific baseline risks of schizophrenia; replacing an original PRS in the multivariable model with its normalized version, which only differs by a term that is constant within site, is equivalent to re-parametrizing the site-specific effect. Neither the overall model fit nor the regression parameter for the PRS is affected by such a re-parametrization. The normalization does not affect statistical inference for individual scores at different p-value thresholds, but has the beneficial effect of aligning the correlation structure between scores in such a way that almost all of the polygenic risk is concentrated in the first principal component PRS1, as demonstrated above.

Modelling strategy. We fit separate models for carriers and non-carriers of specific CNVs in order to quantify and test the predictive power of different model terms involving PRS1 and CNV status by comparing models with and without the terms of interest. We contrasted a series of nested models:

a. a baseline model, including neither PRS1 nor CNV status, but all covariates;
b. individual effect models, including either PRS1 or CNV to these, in order to quantify the separate effects of these predictors;
c. additive models, including PRS1 and CNV status, to quantify improvement in predictive power by adding CNV to a PRS1 model;
d. non-additive effect models adding a PRS1 x CNV interaction term to the additive models to test for non-linear interactions between predictors; and
e. combination models, including PRS1 and multiple CNV predictors, to test for non-overlapping effects between different categories of CNVs.

These models are compared using Nagelkerke's pseudo-R^2 for predictive power, and Akaike's information criterion for model fit. We use likelihood ratio tests to calculate one-sided p-values for the hypothesis that a pair of nested models perform equally well.

1. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421-7.

Supplementary Table S4. Original polygenic risk scores calculated as in[1], definition and association with schizophrenia.

PRS	Threshold	OR	P-value	Scaled OR	Scaled 95% CI	ΔR^2
S1	5x10^{-8}	3.44	1.0E-164	1.60	1.54 - 1.65	1.8
S2	1x10^{-6}	2.85	1.2E-243	1.68	1.62 - 1.73	2.7
S3	1x10^{-4}	2.14	0.0E+00	2.60	2.50 - 2.71	5.5
S4	0.001	1.79	0.0E+00	2.30	2.23 - 2.37	7.3
S5	0.01	1.53	0.0E+00	2.31	2.24 - 2.37	9.4
S6	0.05	1.38	0.0E+00	2.82	2.73 - 2.92	10.2
S7	0.1	1.33	0.0E+00	3.15	3.04 - 3.27	10.2
S8	0.2	1.29	0.0E+00	3.31	3.18 - 3.44	9.9
S9	0.5	1.26	0.0E+00	3.74	3.59 - 3.91	9.9
S10	1.0	1.26	0.0E+00	3.69	3.54 - 3.86	9.8

Threshold is the p-value threshold for including SNPs for calculating the PRS.
OR is the odds ratio for a +1 increase of the PRS in a multivariable logistic regression model for schizophrenia adjusted for sex, site, CNV quality and five ancestral components; *P-value* is the Likelihood ratio test p-value for the PRS.
Scaled OR is the odds ratio for an increase by +1 standard deviation of the PRS in the same multivariable model, and *Scaled 95% CI* is the corresponding 95% confidence interval. *ΔR2* is the change in Nagelkerke's R^2 when removing the PRS from the multivariable model.

Supplementary Table S5. Principal components of the normalized polygenic risk scores as described in the Supplementary Methods and their association with schizophrenia. PRS1 is the first principal component which is used as summary index for polygenic risk of schizophrenia for the logistic regression models presented in Table 2 of the paper.

Component	Variance	OR	P-value	Scaled OR	Scaled 95% CI	ΔR^2
PRS1	69.2	1.39	0.0E+00	2.40	2.33 - 2.47	11.2
PRS2	17.9	1.00	0.70	1.00	0.98 - 1.03	0.0
PRS3	6.1	1.00	1.00	1.00	0.98 - 1.02	0.0
PRS4	2.3	1.03	0.24	1.01	0.99 - 1.04	0.0
PRS5	1.9	0.99	0.69	1.00	0.97 - 1.02	0.0
PRS6	1.3	1.02	0.47	1.01	0.99 - 1.03	0.0
PRS7	0.8	0.99	0.75	1.00	0.98 - 1.02	0.0
PRS8	0.3	1.14	0.04	1.02	1.00 - 1.05	0.0
PRS9	0.2	0.88	0.16	0.98	0.96 - 1.01	0.0
PRS10	0.0	0.71	0.14	0.98	0.96 - 1.01	0.0

Variance is the percentage of variance in the normalized scores nS1-nS10 that is explained by the principal component.
OR is the odds ratio for a +1 increase of the principal component in a multivariable logistic regression model for schizophrenia adjusted for sex, site, CNV quality and five ancestral components; *P-value* is the corresponding Wald test p-value.
Scaled OR is the odds ratio for an increase by +1 standard deviation of the principal component in the same multivariable model, and *Scaled 95% CI* is the corresponding 95% confidence interval.
ΔR2 is the change in Nagelkerke's R^2 when removing the principal component from the multivariable model.

Supplementary Table S6. Normalized polygenic risk scores as described in the Supplementary Methods and their association with schizophrenia.

Normalized PRS	OR	P-value	Scaled OR	Scaled 95% CI	ΔR^2
nS1	3.44	1.0E-164	1.36	1.33 - 1.39	1.8
nS2	2.85	1.2E-243	1.46	1.43 - 1.49	2.7
nS3	2.14	0.0E+00	1.75	1.71 - 1.79	5.5
nS4	1.79	0.0E+00	1.94	1.89 - 1.98	7.3
nS5	1.53	0.0E+00	2.17	2.12 - 2.23	9.4
nS6	1.38	0.0E+00	2.30	2.23 - 2.36	10.2
nS7	1.33	0.0E+00	2.29	2.23 - 2.35	10.2
nS8	1.29	0.0E+00	2.27	2.21 - 2.33	9.9
nS9	1.26	0.0E+00	2.26	2.20 - 2.33	9.9
nS10	1.26	0.0E+00	2.26	2.20 - 2.32	9.8

OR is the odds ratio for a +1 increase of the normalized PRS in a multivariable logistic regression model for schizophrenia adjusted for sex, site, CNV quality and five ancestral components; *P-value* is the Likelihood ratio test p-value for the normalized PRS.

Scaled OR is the odds ratio for an increase by +1 standard deviation of the normalized PRS in the same multivariable model, and *Scaled 95% CI* is the corresponding 95% confidence interval.

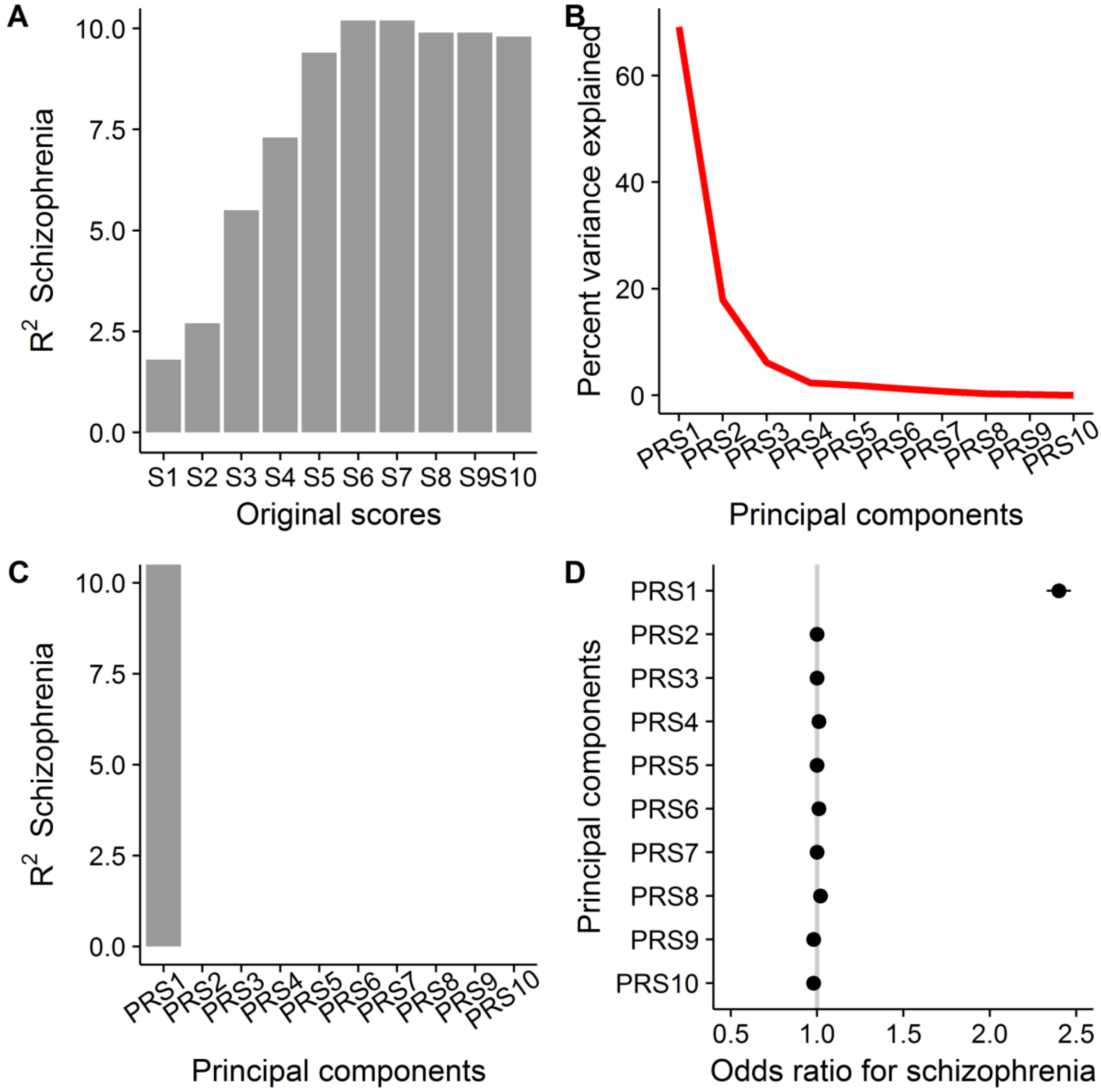
ΔR2 is the change in Nagelkerke's R^2 when removing the normalized PRS from the multivariable model.

Supplementary Figure S1. A: Increase in R^2 when adding an original score S1-S10 to the multivariable logistic regression model

B: Percentage of variance explained by principal components PRS1-PRS10 of the normalized scores nS1-nS10

C: Increase in R^2 when adding a principal component PRS1-PRS10 to the multivariable model

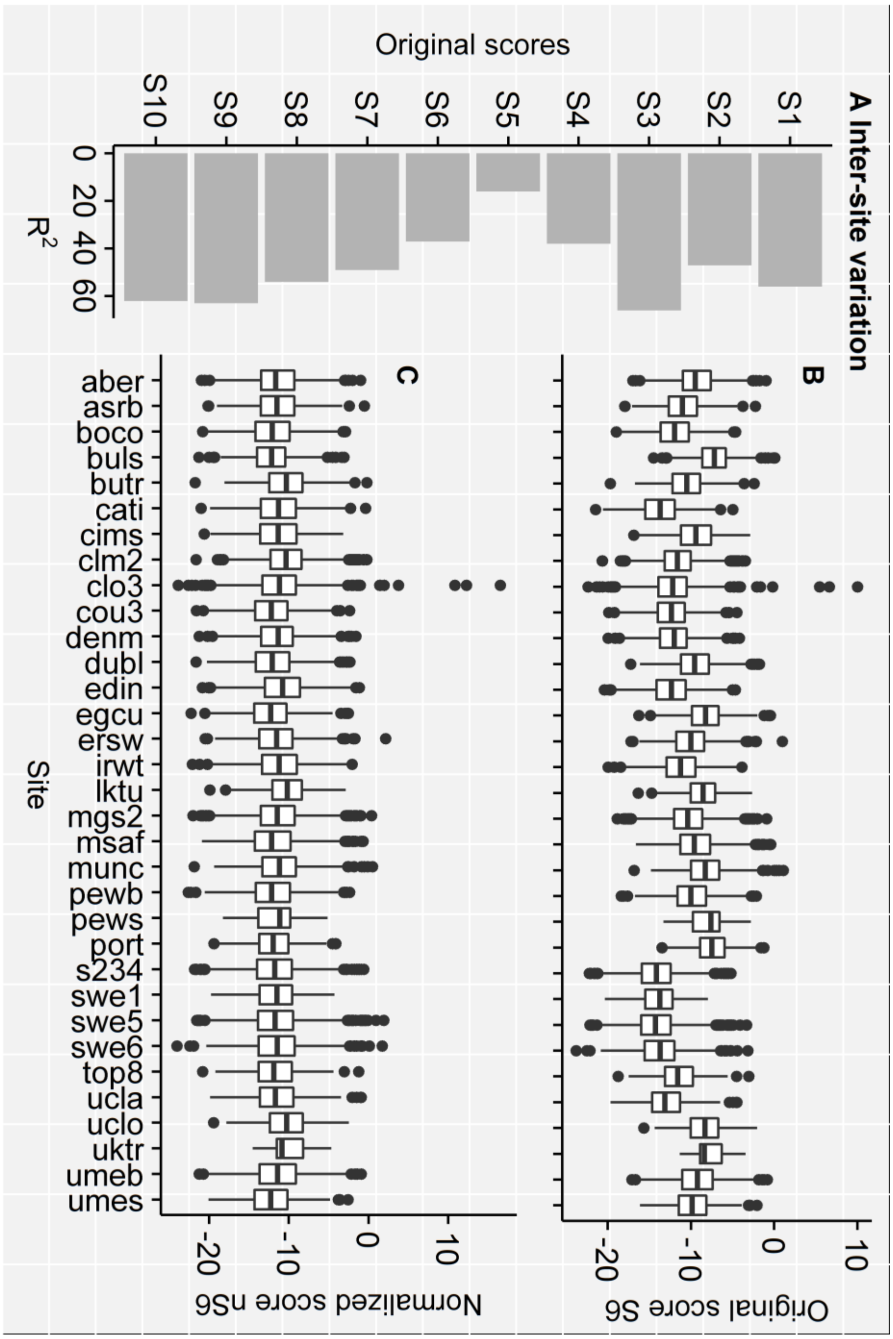
D: Odds ratios and 95% confidence intervals for PRS1-10 in a multivariable model



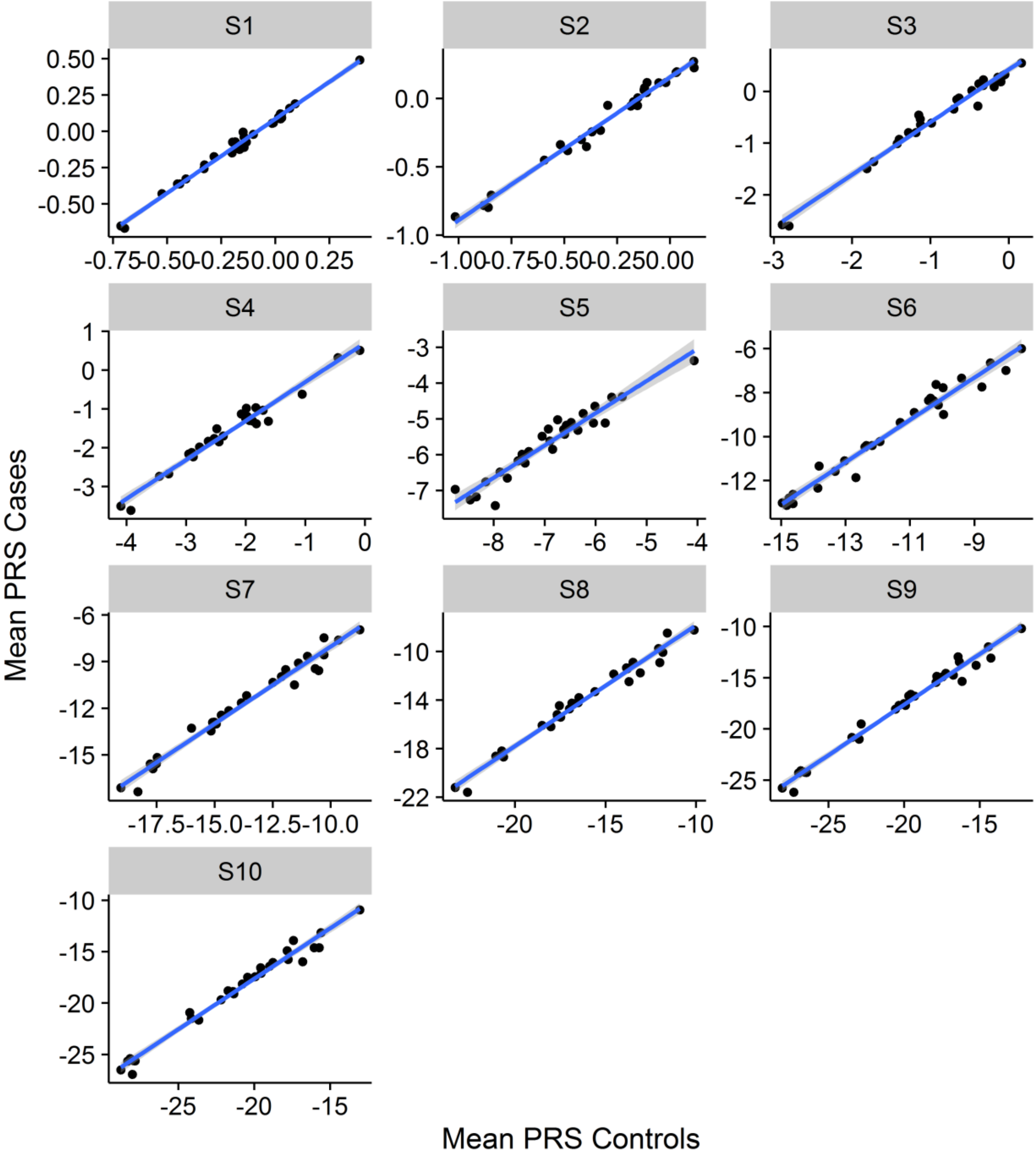
Supplemental Figure S2. A: Percent variance of original scores S1-S10 explained by between-site variation (expressed as R^2 from linear regression model)

B: Distribution of original score S6 across sites

C: Distribution of normalized score nS6 across sites



Supplemental Figure 3. Scatter plots of mean PRS among cases (vertical axes) against mean PRS among controls (horizontal) axes for all n=28 sites contributing with both cases and controls. A linear regression fit is shown as a blue line, with the pointwise 95% confidence area for the expected value shown in grey.


S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
Mean PRS Cases
Mean PRS Controls