

# Supplementary Materials to "Identifying genetic variants for brain connectivity using Ball Covariance Ranking and Aggregation"

## S1 Additional non-linearity simulation

To ensure a comprehensive comparison of effects and cover diverse genetic architectures, we incorporated various nonlinear models in  $f_g(\mathbf{X}_g)$  for model (1), each designed to capture different SNP interactions both within and across SNP-sets. Specifically, we introduced the following five settings:

1. Linear SNP Effects (Setting 1): Maintains the original linear framework for basic comparison.
2. Interactive SNP Effects (Setting 2): Introduces interactions within SNP-sets to explore non-linear dependencies.
3. Exponential SNP Transformations (Setting 3): Applies exponential transformations to assess the impact of more complex nonlinear transformations.
4. Block-Diagonal Interaction Model (Setting 4): Uses a block-diagonal framework to simulate localized SNP interactions.
5. Promiscuous Interaction Model (Setting 5): Models extensive interactions across two SNP-sets, capturing both inter- and intra-set dynamics.

- Setting 1: Linear SNP effects within a SNP-set

$$\begin{aligned} Y_i &= \sum_{g=1}^G f_g(X_g) + E_i \\ &= \mathbf{B} \sum_{j=1}^{500} X_i^j \beta_j \gamma_j + E_i, \end{aligned}$$

where  $X_i^j$  represents the  $j$ -th SNP for the  $i$ -th individual,  $\beta_j \sim U(0.4, 0.8)$  and  $\gamma_j \sim \text{Bernoulli}(0.05)$ ,  $\mathbf{E}_i \sim N(0, \sigma^2 \mathbf{I})$  with  $\sigma = 1$  and  $\mathbf{B}$  has a smile face pattern (Fig.2c). This setting focuses on linear effects among SNPs within a SNP-set.

- Setting 2: Interactive effects among some SNPs within a SNP-set where 25 SNPs have true signals and divided into 5 groups. Each group contains 5 SNPs that influence

the phenotype through both individual and interactive effects.

$$Y_i = \sum_{g=1}^G f_g(X_g) + E_i$$

$$= \mathbf{B} \left( \sum_{k=1}^5 \beta_k \left( \sum_{j=1}^5 X_i^{jk} + X_i^{1k} X_i^{2k} + X_i^{1k} X_i^{3k} + X_i^{3k} X_i^{4k} X_i^{5k} \right) \right) + E_i,$$

where  $X_i^{jk}$  represents the  $j$ -th SNP for the  $i$ -th individual in the  $k$ -th group. Besides,  $\beta_k \sim U(0.4, 0.8)$ . The error term is denoted by  $\mathbf{E}_i \sim N(0, \sigma^2 \mathbf{I})$  with  $\sigma = 1$ . Additionally,  $\mathbf{B}$  follows a smile face pattern (Fig.2c). This setting accounts for interactive effects among some SNPs within a SNP-set.

- Setting 3: Exponential transformation of some SNPs within a SNP-set where 25 SNPs have true signals and divided into 5 groups. Each group contains 5 SNPs that influence the phenotype through exponentially transformed effects. Hunter et al. (2023)

$$Y_i = \sum_{g=1}^G f_g(X_g) + E_i$$

$$= \sum_{k=1}^5 \left( \sum_{j=1}^3 \exp \left( -\frac{\psi_1 j}{p} \right) + \sum_{j=4}^5 \exp \left( -\frac{\psi_2 j}{p} \right) \right.$$

$$+ \sum_{j=1}^2 \exp \left( -\frac{\psi_3 j}{p(X_i^{1k} - X_i^{2k})^2} \right) + \sum_{j \in \{1,3\}} \exp \left( -\frac{\psi_3 j}{p(X_i^{1k} - X_i^{3k})^2} \right)$$

$$\left. + \sum_{j=3}^4 \exp \left( -\psi_4 \frac{j}{p(X_i^{1k} + X_i^{2k} - X_i^{3k} - X_i^{4k})^2} - \psi_3 (X_i^{3k} \cdot X_i^{4k} \cdot X_i^{5k}) \right) \right) + E_i,$$

where  $X_i^{jk}$  represents the  $j$ -th SNP for the  $i$ -th individual in the  $k$ -th group. Besides,  $\psi_1 = 50$ ,  $\psi_2 = 25$ ,  $\psi_3 = 60$ ,  $\psi_4 = 45$ ,  $p = 500$ ,  $\mathbf{E}_i \sim N(0, \sigma^2 \mathbf{I})$  with  $\sigma = 1$ , and  $\mathbf{B}$  has a smile face pattern (Fig.2c). This setting involves exponential transformations of some SNPs within a SNP-set, capturing complex interactions and transformations.

- Setting 4: Block-diagonal (BD) interaction model within a SNP-set where 25 SNPs have true signals that influence the phenotype through localized interactive effects.

Ho & Hsu (2015)

$$Y_i = \sum_{g=1}^G f_g(X_g) + E_i$$

$$= \mathbf{B} \left( \sum_{j=1}^s \alpha_j X_i^j + \sum_{j=1}^s \beta_j (X_i^j)^2 + \sum_{j=1}^{s-1} \gamma_j X_i^j X_i^{(j+1)} \right) + E_i,$$

where  $X_i^j$  represents the  $j$ -th SNP for the  $i$ -th individual. Besides,  $s=25$ ,  $\alpha_j \sim N(1.5, 0.5)$ ,  $\beta_j \sim N(1, 0.2)$ ,  $\gamma_j \sim N(0.5, 0.1)$ ,  $p=500$ ,  $\mathbf{E}_i \sim N(0, \sigma^2 \mathbf{I})$  with  $\sigma = 1$  and  $\mathbf{B}$  has a smile face pattern (Fig.2c). Here,  $s$  causal loci interact within the same block, with linear, quadratic, and mixed terms representing their effects on the phenotype.

- Setting 5: Promiscuous (PS) interaction model within two SNP-sets where 25 SNPs in each SNP-set have true signals that influence the phenotype.

$$Y_i = \sum_{g=1}^2 f_g(X_g) + E_i$$

$$= \mathbf{B} \left( \sum_{j=1}^s \alpha_j X_i^{j1} + \sum_{j=1}^{s'} \beta_j (X_i^{j2})^2 + \sum_{j=1}^{s'/2} \gamma_j X_i^{j1} X_i^{(j2)} \right) + E_i,$$

where  $X_i^{j1}$  represents the  $j$ -th SNP for the  $i$ -th individual in the 1st SNP-set and  $X_i^{j2}$  represents the  $j$ -th SNP for the  $i$ -th individual in the 2nd SNP-set. Besides,  $s = s' = 25$ ,  $\mathbf{E}_i \sim N(0, \sigma^2 \mathbf{I})$  with  $\sigma = 1$  and  $\mathbf{B}$  has a smile face pattern (Fig.2c). Furthermore,  $\alpha_j \in \{-1, 0, 1\}$ ,  $\beta_j$  and  $\gamma_j$  are randomly chosen from normal distributions and are typically of order unity. The model has  $s$  loci which have linear but no quadratic effect on the phenotype, and  $s'$  loci have quadratic but no linear effect on the phenotype.  $s'/2$  of the latter type interact with counterparts of the former type. In biological terms, this model has subsets of loci which are entirely linear in effect, some which are entirely nonlinear, and interactions between these subsets.

Our results show that BCRA outperforms other methods in detecting nonlinear patterns within a SNP-set as displayed in Table S1, but its efficacy diminishes when assessing

interactions between SNP-sets. Conversely, subsample-BCRA shows modest power across all models, with around 70% power, except for the PS model setting, which is within expectation as the developed model is designed to capture interactions within a SNP-set. The single SNP approach, GWAS, shows a decrease in effectiveness when applied to nonlinear scenarios. Overall, BCRA and subsample-BCRA have proven to be highly effective in both straightforward linear configurations and more complex situations involving nonlinearity within single SNP-set.

## S2 Additional simulation with different distance measures

Recognizing the value of testing our results' stability against varied norms, we broadened our analysis to incorporate additional distance measures, including non-Euclidean ones such as geodesic distance. We modeled our data as follows:

$$Y_i = \sum_{g=1}^G f_g(X_g) + E_i, f_g(\mathbf{X}_g) = \mathbf{I}(\sum_{j=1}^{p_g} \mathbf{X}_i^{jg} \beta^{jg} \gamma^{jg} > 0) \cdot \mathbf{B}, \quad (1)$$

where  $\mathbf{E}_i \sim N(0, \sigma^2 \mathbf{I})$ ,  $\sigma = 1$ ,  $\gamma^{jg} \in \text{Bernoulli}(0.05)$ ,  $\beta^{jg} \sim U(0.4, 0.8)$ ,  $p_g = 500$ ,  $G = 1$  and  $\mathbf{B}$  has a smile face pattern (Fig 2c).

We choose three different distance metrics: Euclidean distance, geodesic distance and Pearson's correlation.

For Pearson's correlation, it unrolls the matrix into a vector and compute the Pearson correlation between the matrices themselves. Although it neglects matrix structure, it has yielded impressive results in identifying a participant out of a large group of participants based on FC matrix similarity, a process dubbed fingerprinting Finn et al. (2015, 2017), Amico & Goñi (2018). Another widely adopted approach is to compute the Euclidean dis-

tance between the vectorized matrices Ponsoda et al. (2017). However, since the geometry of functional connectivity is non-Euclidean, some papers proposed to use the geodesic distance Venkatesh et al. (2020), a non-Euclidean distance metric that considers the manifold on which the data lies and, demonstrates the higher participant identification compared to a similarity measure based on Pearson correlation and Euclidean distance.

Despite Pearson’s correlation omitting the matrix structure and geodesic distance acknowledging the data’s manifold nature, our simulations reveal BCRA’s high detection consistency across all three metrics as shown in Fig. S1. However, for subsample-BCRA, we noted a dip in the average detection rate, most pronounced with geodesic distance, possibly due to subsampling distorting the data’s inherent structure, thereby reducing sensitivity to this metric. Given the computational intensity of the geodesic distance, we advocate for BCRA in small samples, and for larger samples, subsample-BCRA augmented with Euclidean or Pearson’s correlation.

### S3 Additional simulation for distributions in the SPD space

We want to enrich our simulation studies with settings that generate the connectivity matrix from the semi-positive definite (SPD) space, using the Wishart distribution for error terms. Specifically, for single SNP-set simulations, we utilized the following model for our simulation:

$$Y_i = \sum_{g=1}^G f_g(X_g) + E_i, f_g(\mathbf{X}_g) = \mathbf{I}(\sum_{j=1}^{p_g} \mathbf{X}_i^{jg} \beta^{jg} \gamma^{jg} > 0) \cdot \mathbf{B}, \quad (2)$$

where  $\mathbf{E} = (E_1, \dots, E_n) \sim \text{Wishart}_q(V, n)$ ,  $\gamma^{jg} \in \text{Bernoulli}(0.05)$ ,  $\beta^{jg} \sim U(0.4, 0.8)$ ,  $p_g = 500$ ,  $G = 1$  and  $\mathbf{B}$  has a smile face pattern (Fig 2c). We chose  $q$  as the dimension of  $\mathbf{B}$  plus

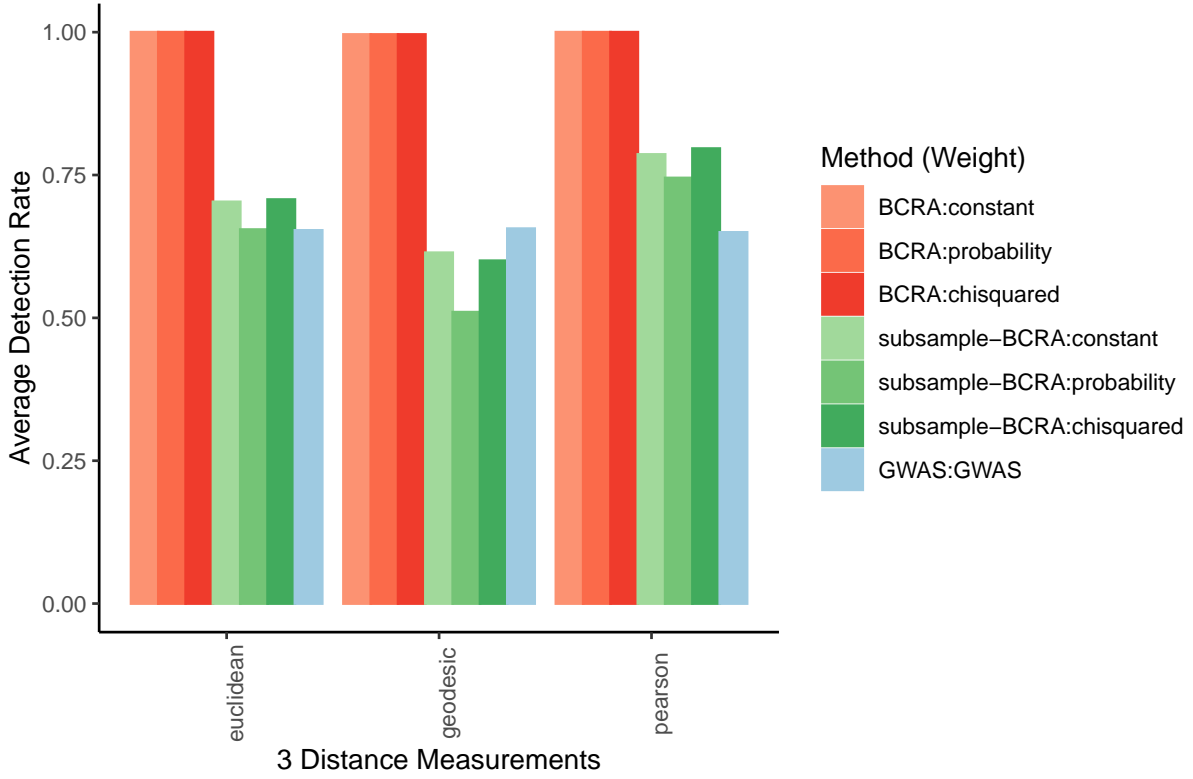


Figure S1: Sensitivity results to three choices of the distance metric, Pearson’s correlation, Euclidean distance and Geodesic distance. Detection rate, defined as the frequency of detecting this SNP-set over iterations. The BCRA performs virtually well under all three distance measurements, indicating that results of subsample-BCRA are reliable to the choice of distance metrics. While for subsample-BCRA, the average detection rate across choices of the distance metric are well for Pearson’s correlation and Euclidean distance but decreased for Geodesic distance. Different colors represented different approaches (orange: BCRA; green: subsample-BCRA). Different shapes represented different weight choices (circle: constant weight; diamond: probability weight; triangle: chisquared weight).

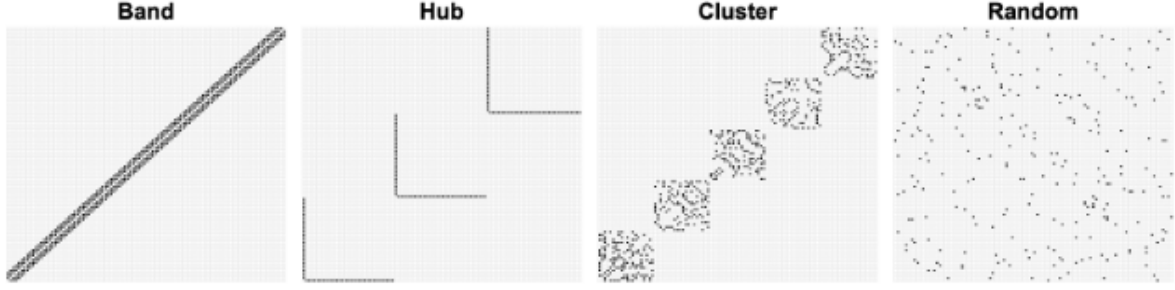


Figure S2: Four correlation structures of error terms for Wishart distribution in the simulation studies.

10 and  $V$  was structured to reflect four distinct correlation patterns: band, hub, cluster, and random shown in Fig. S2.

Our results revealed that BCRA and subsample-BCRA both maintained robust performances in Table S2, whether under independent or SPD-derived correlated error structures. However, when the connectivity matrix was generated within the SPD space exhibiting certain correlated error structures, the GWAS approach showed a slight decline in power for detecting true signals.

## S4 Additional simulation for choosing best subsample size

We perform simulation studies, utilizing various reduced sample sizes, designated as  $n_{\text{subset}}$  to select the optimal subsample size. We modeled our data as follows:

$$Y_i = \sum_{g=1}^G f_g(X_g) + E_i, f_g(\mathbf{X}_g) = \mathbf{I}(\sum_{j=1}^{p_g} \mathbf{X}_i^{jg} \beta^{jg} \gamma^{jg} > 0) \cdot \mathbf{B}, \quad (3)$$

where  $\mathbf{E}_i \sim N(0, \sigma^2 \mathbf{I})$ ,  $\sigma = 1$ ,  $\gamma^{jg} \in \text{Bernoulli}(0.05)$ ,  $\beta^{jg} \sim U(0.4, 0.8)$ ,  $G = 1$  and  $\mathbf{B}$  has a smile face pattern (Fig 2c). We vary different number of SNPs within a set  $p_g$ , total sample

size  $n$  and proportion of samples to form into subset  $p_{\text{subset}}$ . This  $n_{\text{subset}} = n \times p_{\text{subset}}$  selection was informed by a two-step process: firstly, ranking individuals by SNP call rate to prioritize data completeness, and secondly, choosing the top percentile samples with the highest rates, with this selection being further adjusted by the allele frequency of each SNP to ensure a representative subsample.

The results showed that maintaining  $n_{\text{subset}}$  at over 10% of the subjects preserves detection power above 50% as in Fig. S3. In practical terms, considering our dataset of around 30,000 subjects and SNP-sets ranging from 2 to 15,330 SNPs (mean = 3,330; median = 3,335), only a small number of SNP-sets exceeded 10,000 subjects. Given this distribution, the ratio of SNPs to subjects was generally less than 1:3, aligning with our simulation findings that a  $n_{\text{subset}}$  comprising 10% of total subjects ensures a power greater than 50%. This decision strikes a balance between computational efficiency and the statistical power required for subsample-BCRA.

## S5 Supplementary Tables and Figures

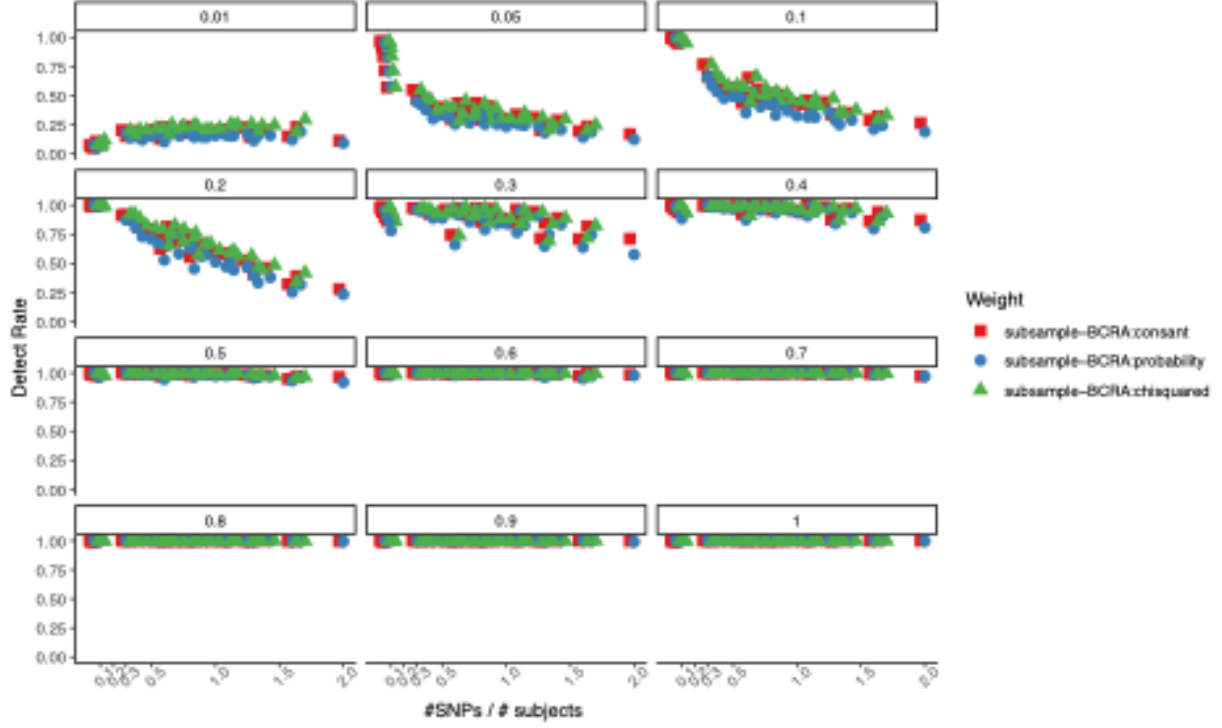
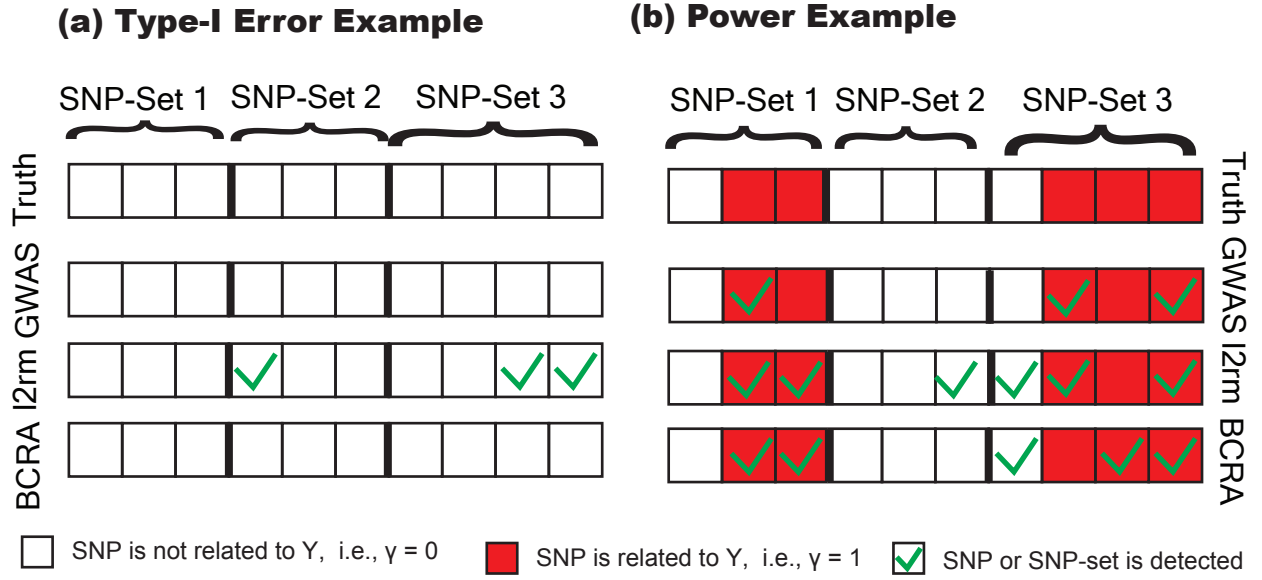


Figure S3: Simulation results to 12 choices of proportion of samples to form into subset  $p_{\text{subset}} \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ . Detection rate, defined as the frequency of detecting this SNP-set over iterations is shown in y-axis. We vary the total sample size  $n \in \{500, 600, 700, 800, 900, 1000\}$  and number of SNPs per set  $p_g \in \{50, 300, 500, 650, 800, 1000\}$ . The x-axis is the SNPs-to-sample ratio, defined as number of SNPs over number of total samples i.e.,  $p_g/n$ . The detection rate achieved over 75% regardless of the SNPs-to-sample ratio when the proportion of subset samples was larger than 30%. While subsample-BCRA failed to detect true signals when the proportion of subset samples was smaller than 10%. The subsample-BCRA keeps desirable power under the scenario where number of SNPs is smaller than one third of the number of subjects when the proportion of subset samples is 10% or 20%. Different colors and shapes represented different weight (red square: constant weight; blue circle: probability weight; green triangle: chisquared weight).



Metric (One Iteration)	GWAS	l2rm	BCRA/Subsample-BCRA
Type-I Error	0	100%	0
Detection Rate	100% for SNPs 2, 8, and 10; 0 for SNPs 3 and 9; Average = 60%	100% for SNPs 2, 3, 8, and 10; 0 for SNP 9; Average = 80%	100% for SNP-sets 1 and 2; Average = 100%
True Positive (TP)	3 SNPs	4 SNPs	2 SNP-sets
False Positive (FP)	0 SNP	2 SNPs	0 SNP-set
True Negative (TN)	5 SNPs	3 SNPs	1 SNP-set
False Negative (FN)	2 SNPs	1 SNP	0 SNP-set
$SEN = TP / (TP + FN)$	$3/5 = 60\%$	$4/5 = 80\%$	$2/2 = 100\%$
$SPE = TN / (TN + FP)$	$5/5 = 100\%$	$3/5 = 60\%$	$1/1 = 100\%$
$PREC = TP / (TP + FP)$	$3/3 = 100\%$	$4/6 = 67\%$	$2/2 = 100\%$
$NPV = TN / (TN + FN)$	$5/7 = 71\%$	$3/4 = 75\%$	$1/1 = 100\%$

Figure S4: An illustration of calculating five metrics for different approaches with 10 SNPs divided into 3 SNP-sets. The white color indicates no effects, red color means this SNP influences the response and the green frame highlights the detected SNPs or SNP-sets under different methods. (a) The situation to evaluate Type-I error, where there are no signals. (b) The situation to evaluate power, where individual SNPs 2, 3, 8, 9, 10 and SNP-set 1 and 2 have signals. The bottom table shows how different metrics are derived at one iteration.

Table S1: The detection rate of true signals for the single SNP set with non-linear effects under the smile coefficient structure. The value of Setting 5 is the average rate that both SNPs are detected across iterations. The detection rate for BCRA or subsample- BCRA is defined at the SNP-set level as the number of iterations detecting a SNP-set divided by the total number of iterations. For GWAS approach, the detection rate is defined at the SNP level as the frequency of detecting an individual SNP over iterations. The average detection rate at the SNP-level of GWAS is the average values across all true SNPs. The weight column represents various choices of weights in calculating the statistic BCov in equation (1).

Method	Weight	Setting 1:	Setting 2:	Setting3:	Setting 4:	Setting 5:
		Linear	Interaction	Exponential	BD	PS
<b>BCRA</b>	constant	1.0000	1.0000	1.0000	1.0000	0.5950
	probability	1.0000	1.0000	1.0000	1.0000	0.5880
	chisquared	1.0000	1.0000	1.0000	1.0000	0.5930
<b>subsample-BCRA</b>	constant	0.7460	0.8280	0.7555	0.6466	0.3470
	probability	0.7200	0.8040	0.7054	0.5964	0.2900
	chisquared	0.7580	0.8380	0.7615	0.6627	0.3690
<b>GWAS</b>	GWAS	0.5874	0.5343	0.5226	0.5044	0.2473

Table S2: The detection rate of true signals for the single SNP set under different error correlation structure from SPD space. The detection rate for BCRA or subsample- BCRA is defined at the SNP-set level as the number of iterations detecting a SNP-set divided by the total number of iterations. For GWAS approach, the detection rate is defined at the SNP level as the frequency of detecting an individual SNP over iterations. The average detection rate at the SNP-level of GWAS is the average values across all true SNPs. The weight column represents various choices of weights in calculating the statistic BCov in equation (1).

Method	Weight	Correlation Structure in Wishart Distribution				
		Independent	Random	Band	Hub	Cluster
<b>BCRA</b>	constant	1.0000	1.0000	1.0000	1.0000	1.0000
	probability	1.0000	1.0000	1.0000	1.0000	1.0000
	chisquared	1.0000	1.0000	1.0000	1.0000	1.0000
<b>subsample-BCRA</b>	constant	0.7823	0.8000	0.7691	0.7500	0.7992
	probability	0.7480	0.7480	0.7229	0.7120	0.7570
	chisquared	0.7702	0.8000	0.7651	0.7580	0.7992
<b>GWAS</b>	GWAS	0.5945	0.5902	0.5769	0.5794	0.5805

Table S4: The detection rate of true signals for the first three SNP sets containing true signals under the butterfly coefficient structure. The detection rate for BCRA or subsample-BCRA is defined at SNP-set level as the number of iterations detecting a SNP-set divided by total number of iterations. For GWAS approach, the detection rate is defined as SNP level as the frequency of detecting an individual SNP over iterations. The average detection rate at SNP-level of GWAS is the average values across all true SNPs. Different columns represent various choices of weights in calculating the statistic BCov in equation (1).

Method	Set	constant	probability	chis-squared
BCRA	Set1	1	1	1
	Set2	1	1	1
	Set3	1	1	1
subsample-BCRA	Set1	0.942	0.942	0.942
	Set2	0.994	0.994	0.994
	Set3	0.842	0.842	0.842
GWAS	Set1	0.523	NA	NA
	Set2	0.607	NA	NA
	Set3	0	NA	NA

Table S5: The detection rate of true signals for the first three SNP sets containing true signals under the smile coefficient structure. The detection rate for BCRA or subsample-BCRA is defined at SNP-set level as the number of iterations detecting a SNP-set divided by total number of iterations. For GWAS approach, the detection rate is defined as SNP level as the frequency of detecting an individual SNP over iterations. The average detection rate at SNP-level of GWAS is the average values across all true SNPs. Different columns represent various choices of weights in calculating the statistic BCov in equation (1).

<b>Method</b>	<b>Set</b>	<b>constant</b>	<b>probability</b>	<b>chis-squared</b>
BCRA	Set1	1	0.996	1
	Set2	1	1	1
	Set3	0.986	0.97	0.99
subsample-BCRA	Set1	0.938	0.938	0.938
	Set2	0.99	0.99	0.99
	Set3	0.808	0.808	0.808
GWAS	Set1	0.489	NA	NA
	Set2	0.549	NA	NA
	Set3	0	NA	NA

Table S6: The detection rate of true signals for the first three SNP sets containing true signals under the wink coefficient structure. The detection rate for BCRA or subsample-BCRA is defined at SNP-set level as the number of iterations detecting a SNP-set divided by total number of iterations. For GWAS approach, the detection rate is defined as SNP level as the frequency of detecting an individual SNP over iterations. The average detection rate at SNP-level of GWAS is the average values across all true SNPs. Different columns represent various choices of weights in calculating the statistic BCov in equation (1).

<b>Method</b>	<b>Set</b>	<b>constant</b>	<b>probability</b>	<b>chis-squared</b>
BCRA	Set1	0.998	0.998	0.996
	Set2	1	1	1
	Set3	0.982	0.962	0.99
subsample-BCRA	Set1	0.96	0.954	0.844
	Set2	0.99	0.99	0.99
	Set3	0.792	0.792	0.788
GWAS	Set1	0.484	NA	NA
	Set2	0.543	NA	NA
	Set3	0	NA	NA

Table S7: SNP-set (top) and SNP (bottom) level power results: average sensitivities, specificities, precisions and NPVs for BCRA, subsample-BCRA and GWAS under the butterfly scenario. For SNP-level results, because some SNPs are in LD with others, if any SNP are within the 50kb windows size with the true signal SNP is selected, this signal SNP is also considered as being identified.

Method	Weight	SEN	SPE	PREC	NPV
<b>SNP-set</b>					
<b>BCRA</b>	constant	1(0)	0.9983(0.0078)	0.9885(0.0524)	1(0)
	probability	1(0)	0.9984(0.0076)	0.989(0.0513)	1(0)
	chis-quared	1(0)	0.9987(0.0067)	0.9915(0.0453)	1(0)
<b>subsample-BCRA</b>	constant	0.9887(0.0604)	0.9994(0.0046)	0.996(0.0314)	0.9988(0.0065)
	probability	0.9867(0.0653)	0.9994(0.0046)	0.9958(0.0329)	0.9986(0.007)
	chi-squared	0.998(0.0257)	0.9995(0.0044)	0.9965(0.0294)	0.9998(0.0028)
<b>GWAS</b>	NA	0.5467(0.0367)	0.9859(5e-04)	0.0515(0.0032)	0.9994(1e-04)
<b>SNP</b>					
<b>BCRA</b>	constant	0.2698(0.0665)	0.9979(0.0035)	0.9081(0.0842)	0.9518(0.0042)
	probability	0.2279(0.0493)	0.9981(0.0032)	0.9043(0.0885)	0.9491(0.0031)
	chi-squared	0.3296(0.052)	0.998(0.0021)	0.9236(0.0513)	0.9555(0.0033)
<b>subsample-BCRA</b>	constant	0.6762(0.1713)	0.9789(0.01)	0.6988(0.0818)	0.9778(0.0114)
	probability	0.6788(0.1708)	0.9789(0.0099)	0.6989(0.0807)	0.978(0.0114)
	chi-squared	0.6824(0.1725)	0.9786(0.0101)	0.6976(0.0806)	0.9782(0.0115)
<b>GWAS</b>	NA	0.5467(0.0367)	0.9859(5e-04)	0.0515(0.0032)	0.9994(1e-04)

Table S8: SNP-set (top) and SNP (bottom) level power results: average sensitivities, specificities, precisions and NPVs for BCRA, subsample-BCRA and GWAS under the smile scenario. For SNP-level results, because some SNPs are in LD with others, if any SNP are within the 50kb windows size with the true signal SNP is selected, this signal SNP is also considered as being identified.

Method	Weight	SEN	SPE	PREC	NPV
<b>SNP-set</b>					
<b>BCRA</b>	constant	0.9953(0.0392)	0.9984(0.0076)	0.989(0.0513)	0.9995(0.0042)
	probability	0.9887(0.0604)	0.9984(0.0074)	0.9895(0.0502)	0.9988(0.0065)
	chis-quared	0.9967(0.0332)	0.9981(0.0081)	0.9873(0.0554)	0.9996(0.0036)
<b>subsample-BCRA</b>	constant	0.9043(0.1609)	0.9993(0.0049)	0.9948(0.0397)	0.9898(0.0171)
	probability	0.8735(0.185)	0.9993(0.0049)	0.9948(0.0386)	0.9865(0.0196)
	chi-squared	0.9029(0.1628)	0.9992(0.0054)	0.9936(0.0437)	0.9896(0.0173)
<b>GWAS</b>	NA	0.5064(0.0359)	0.987(4e-04)	0.052(0.0034)	0.9993(1e-04)
<b>SNP</b>					
<b>BCRA</b>	constant	0.3061(0.0699)	0.9979(0.0027)	0.9174(0.0748)	0.9541(0.0044)
	probability	0.2244(0.0581)	0.9983(0.0026)	0.9085(0.0899)	0.9489(0.0036)
	chi-squared	0.346(0.0617)	0.9975(0.0034)	0.9154(0.0729)	0.9566(0.0039)
<b>subsample-BCRA</b>	constant	0.6856(0.1784)	0.9787(0.0091)	0.6997(0.0757)	0.9784(0.0118)
	probability	0.6891(0.1796)	0.9785(0.0091)	0.698(0.0765)	0.9787(0.0119)
	chi-squared	0.6901(0.1769)	0.9784(0.0088)	0.6966(0.0748)	0.9787(0.0118)
<b>GWAS</b>	NA	0.5064(0.0359)	0.987(4e-04)	0.052(0.0034)	0.9993(1e-04)

Table S9: SNP-set (top) and SNP (bottom) level power results: average sensitivities, specificities, precisions and NPVs for BCRA, subsample-BCRA and GWAS under the wink scenario. For SNP-level results, because some SNPs are in LD with others, if any SNP are within the 50kb windows size with the true signal SNP is selected, this signal SNP is also considered as being identified.

Method	Weight	SEN	SPE	PREC	NPV
<b>SNP-set</b>					
<b>BCRA</b>	constant	0.9933(0.0467)	0.9979(0.0085)	0.986(0.0575)	0.9993(0.005)
	probability	0.9867(0.0653)	0.9979(0.0087)	0.9852(0.06)	0.9986(0.007)
	chis-quared	0.9953(0.0392)	0.998(0.0084)	0.9865(0.0565)	0.9995(0.0042)
<b>subsample-BCRA</b>	constant	0.9279(0.1452)	0.9993(0.0049)	0.9951(0.0362)	0.9923(0.0155)
	probability	0.8509(0.2015)	0.999(0.0059)	0.9915(0.0541)	0.9842(0.0213)
	chi-squared	0.897(0.1694)	0.9996(0.0041)	0.9968(0.0291)	0.989(0.018)
<b>GWAS</b>	NA	0.5012(0.0365)	0.9872(4e-04)	0.0519(0.0035)	0.9993(1e-04)
<b>SNP</b>					
<b>BCRA</b>	constant	0.3086(0.072)	0.9976(0.0038)	0.9105(0.0869)	0.9542(0.0045)
	probability	0.2266(0.0604)	0.998(0.0033)	0.8991(0.1032)	0.9491(0.0038)
	chi-squared	0.3486(0.064)	0.9974(0.0034)	0.9115(0.0727)	0.9567(0.0041)
<b>subsample-BCRA</b>	constant	0.6873(0.1668)	0.9782(0.0104)	0.6978(0.0808)	0.9785(0.0111)
	probability	0.6874(0.1674)	0.9783(0.0103)	0.6982(0.0798)	0.9785(0.0111)
	chi-squared	0.6885(0.1661)	0.9781(0.0104)	0.6964(0.0807)	0.9786(0.011)
<b>GWAS</b>	NA	0.5012(0.0365)	0.9872(4e-04)	0.0519(0.0035)	0.9993(1e-04)

Table S10: Selected SNPs for each super-variant. Each SNP is annotated with its cytogenic region and nearest gene.

Super-variant	CHR	POS	SNP	A1	A2	MAF	REGION	GENE	TYPE
chr1_144	1	143670851	rs11582530	T	C	0.0360	1q21.1	RP6-206I17.1	non-coding intronic
	1	143767646	rs148974023	A	G	0.0240	1q21.1	PPIAL4G	coding nonsyn
chr1_149	1	148544983	rs10158015	G	A	0.0131	1q21.1	LOC105371211	intronic
	1	148547345	rs58312111	A	G	0.0125	1q21.2	RP11-666A1.4	Nearest Upstream
	1	148549271	1:148549271_CT_C	C	CT	0.0122	1q21.2	RP11-666A1.4	Nearest Upstream
	1	148565416	rs9286338	G	A	0.0109	1q21.2	NBPF15	intronic
chr1_160	1	159357634	rs75276010	C	T	0.0101	1q23.2	RP11-550P17.5	intronic
chr1_223	1	222398640	rs75141700	C	T	0.0182	1q41	RP11-400N13.1	intronic
chr4_48	4	47476361	rs10002676	T	C	0.0329	4p12	COMMD8	Nearest Upstream
	4	47581697	rs76814271	T	C	0.0196	4p12	ATP10D	5upstream, intronic
	4	47715392	rs115442203	G	A	0.0102	4p12	CORIN	intronic
chr5_22	5	21476729	rs189661811	G	A	0.0130	5p14.3	GUSBP1	intronic
chr5_23	5	22430827	rs66485180	C	T	0.1242	5p14.3	CDH12	intronic
	5	22447483	rs12660009	C	G	0.1250	5p14.3	CDH12	intronic
	5	22485972	rs13169464	C	T	0.1260	5p14.3	CDH12	intronic
	5	22486235	rs34441400	C	A	0.1260	5p14.3	CDH12	intronic
	5	22486469	rs72744877	G	T	0.1261	5p14.3	CDH12	intronic
	5	22489161	rs7719756	T	C	0.1261	5p14.3	CDH12	intronic
	5	22499504	rs7727099	G	A	0.1241	5p14.3	CDH12	intronic
	5	22506398	rs1417188445	T	TC	0.1260	5p14.3	CDH12	intronic
	5	22507308	rs10040210	C	T	0.1261	5p14.3	CDH12	intronic
	5	22516890	rs4320220	T	C	0.1257	5p14.3	CDH12	intronic
	5	22556156	rs35857048	G	T	0.1122	5p14.3	CDH12	intronic
chr8_145	8	144094058	rs557243129	C	A	0.0106	8q24.3	RP11-273G15.2	5upstream, non-coding intronic
chr21_11	21	10862846	rs3916645	A	T	0.0108	21p11.2	IGHV1OR21-1	coding nonsyn
	21	10863087	rs28521368	T	G	0.0109	21p11.2	IGHV1OR21-1	Nearest Upstream
chr22_39	22	38006356	rs565047646	CT	C	0.0232	22q13.1	GGA1	intronic
	22	38805399	rs575324928	T	A	0.0116	22q13.1	RP3-449O17.1	Nearest Upstream
	22	38954208	rs145125237	T	C	0.0112	22q13.1	DMC1	intronic

Table S11: Association lookups related to psychiatric measurements for selected SNPs within verified super-variants in the NHGRI-EBI GWAS catalog.

Super-variant	SNP	CHR	POS	REGION	GENE	Phenotype	PubMed
chr5_23	rs10040210	5	22507308	5p14.3	CDH12	Bipolar Disorder	19567891
	rs12660009	5	22447483				
	rs13169464	5	22485972				
	rs34441400	5	22486235				
	rs7719756	5	22489161				
	rs7727099	5	22499504				
	rs4320220	5	22516890				
	rs66485180	5	22430827				
	rs72744877	5	22486469				
chr22_39	rs145125237	22	38954208	22q13.1	DMC1	Attention Deficit Disorder with Hyperactivity	18821565
	rs565047646	22	38006356		GGA1		
	rs575324928	22	38805399		RP3-449O17.1		

Table S12: Enriched gene pathways results.

Pathway ID	Description	Parent(s)	p-value	Genes Involved	SNPs
R-HSA-5578768	Physiological factors	Muscle contraction	0.005643	CORIN	rs115442203
R-HSA-418990	Adherens junctions interactions	Cell-Cell communication	0.015456	CDH12	rs10040210,rs12660009,rs13169464,rs34441400,rs35857048,rs4320220,rs66485180,rs72744877,rs7719756,rs7727099
R-HSA-8854214	TBC/RABGAPs	Vesicle-mediated transport	0.019638	GGA1	rs565047646
R-HSA-936837	Ion transport by P-type ATPases	Transport of small molecules	0.025654	ATP10D	rs76814271
R-HSA-912446	Meiotic recombination	Reproduction;Cell Cycle	0.026116	DMC1	rs145125237
R-HSA-421270	Cell-cell junction organization	Cell-Cell communication	0.030261	CDH12	rs10040210,rs12660009,rs13169464,rs34441400,rs35857048,rs4320220,rs66485180,rs72744877,rs7719756,rs7727099
R-HSA-977225	Amyloid fiber formation	Metabolism of proteins	0.036225	GGA1	rs565047646
R-HSA-1500620	Meiosis	Cell Cycle;Reproduction	0.040792	DMC1	rs145125237
R-HSA-446728	Cell junction organization	Cell-Cell communication	0.042614	CDH12	rs10040210,rs12660009,rs13169464,rs34441400,rs35857048,rs4320220,rs66485180,rs72744877,rs7719756,rs7727099

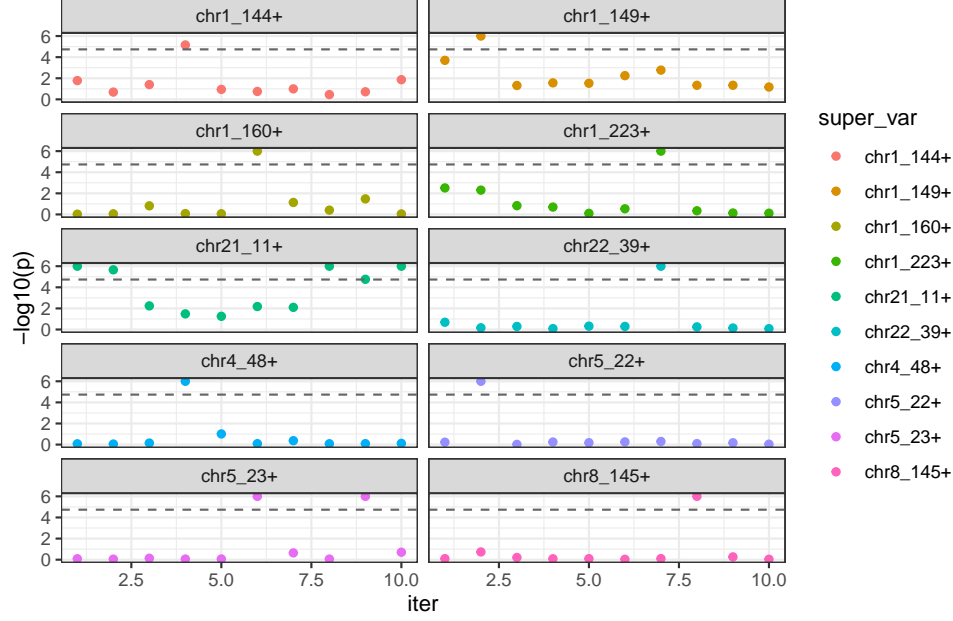


Figure S5: Discovery Phase P-Value Distributions Across 10 Iterations. This figure illustrates the reproducibility of the discovery of super-variants in our UKB White British ancestry data analysis. Each plot represents a different super-variant, indicated by chromosome number and SNP-set identifier (e.g., chr1\_144+ is for SNPs on chromosome 1 with BP ranging from 143MB to 144MB). The y-axis shows the negative log10-transformed p-values, emphasizing the significance levels across 10 iterative validation processes. Super-variants that consistently show significant values (above the horizontal threshold line representing  $1.84 \times 10^{-5}$ ) demonstrate robustness in our validation strategy.

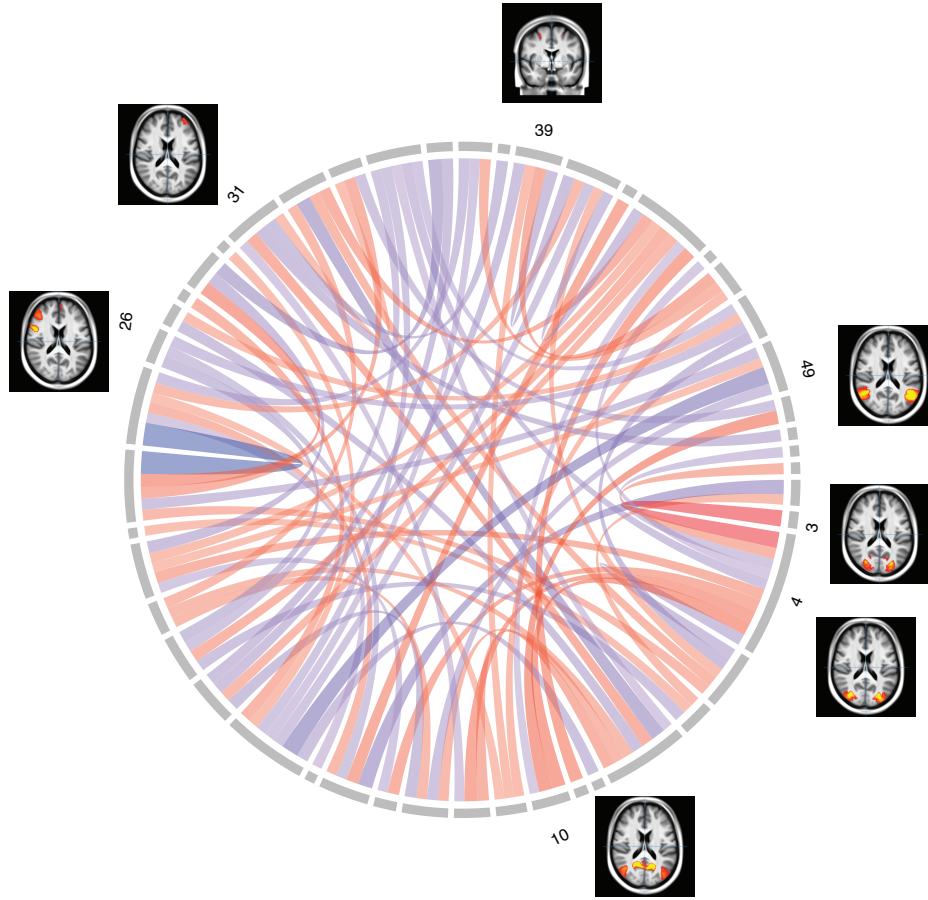


Figure S6: The influence of the super-variant on Chromosome 1 set 144 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with stronger differences in connectivity, including regions indexed as 3, 4, 10, 26, 31, 39 and 49

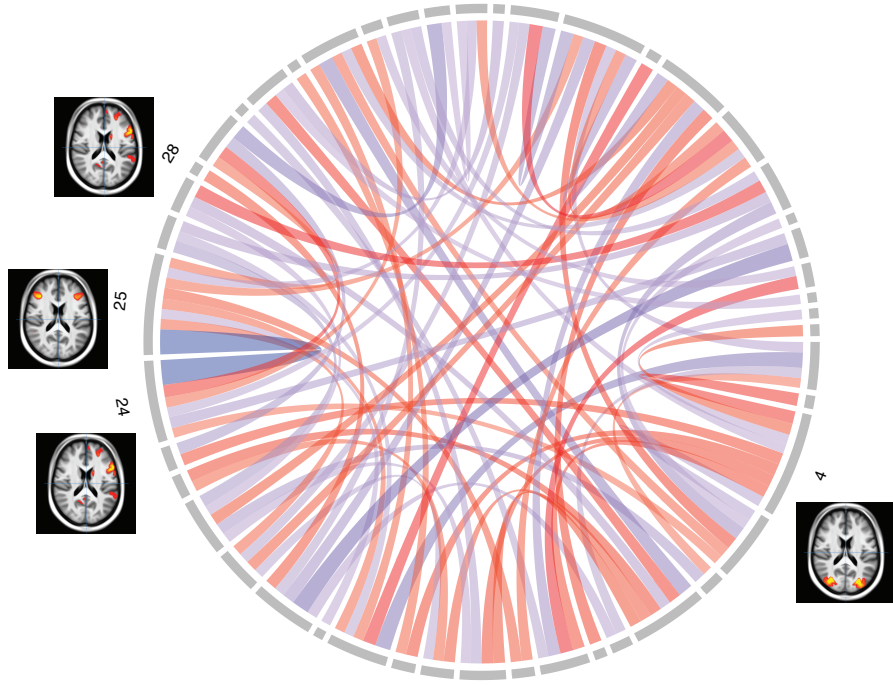


Figure S7: The influence of the super-variant on Chromosome 1 set 169 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with stronger differences in connectivity, including regions indexed as 4, 24, 25 and 28.

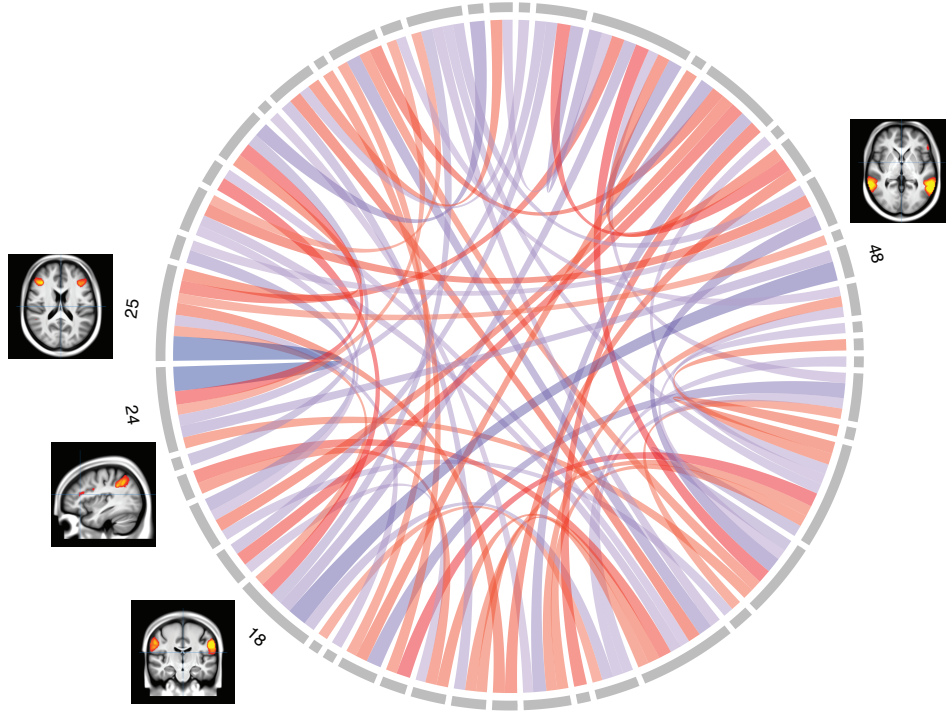


Figure S8: The influence of the super-variant on Chromosome 1 set 223 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with stronger differences in connectivity, including regions indexed as 18, 24, 25 and 48.

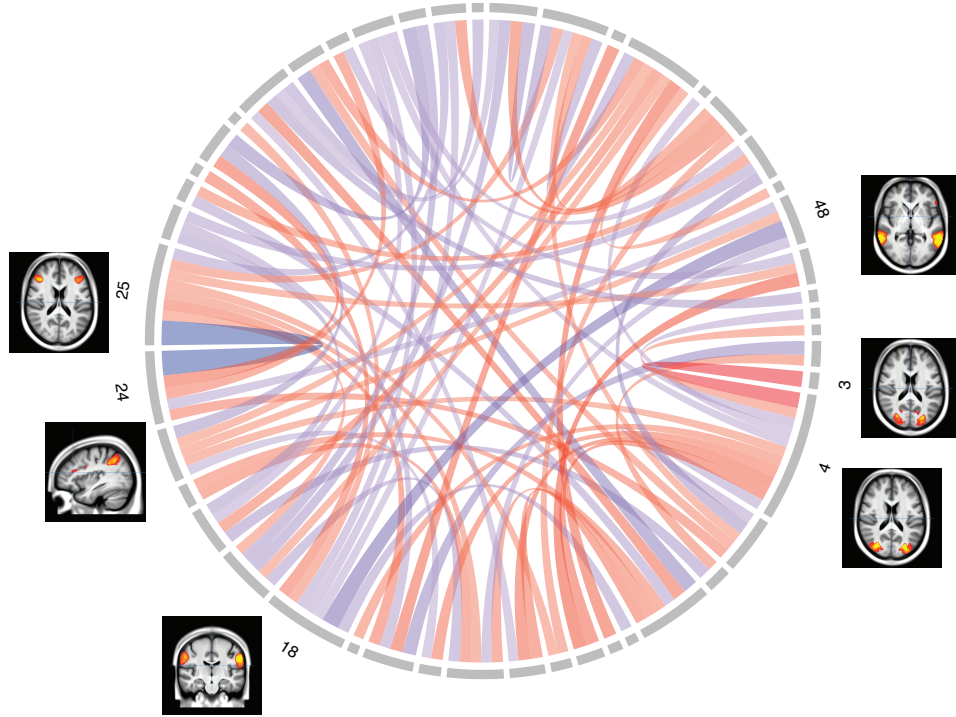


Figure S9: The influence of the super-variant on Chromosome 4 set 48 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with stronger differences in connectivity, including regions indexed as 3, 4, 18, 24, 25 and 48.

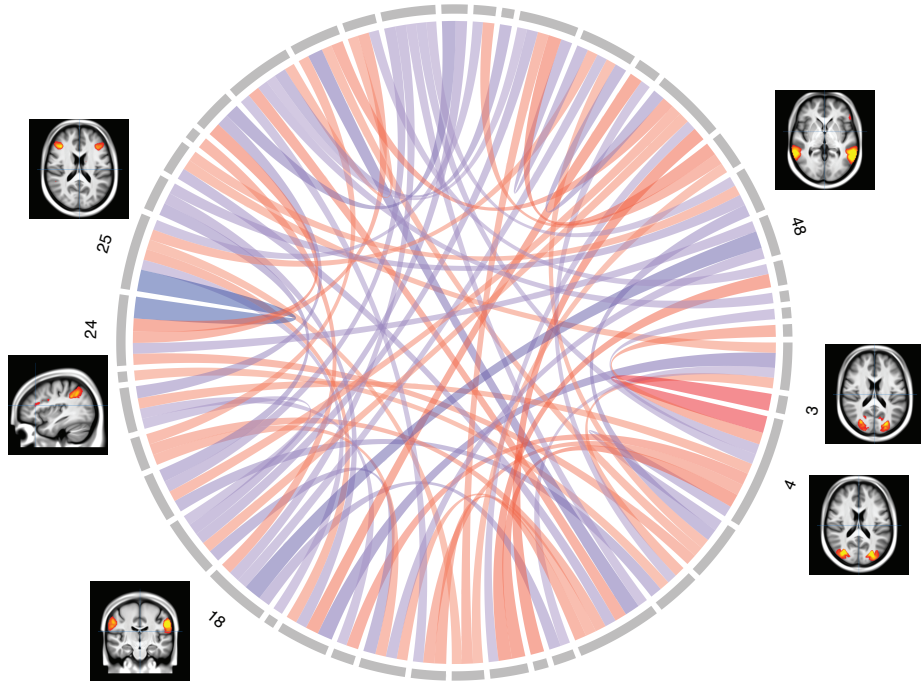


Figure S10: The influence of the super-variant on Chromosome 5 set 22 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with have stronger differences in connectivity, including regions indexed as 3, 4, 18, 24, 25 and 48.

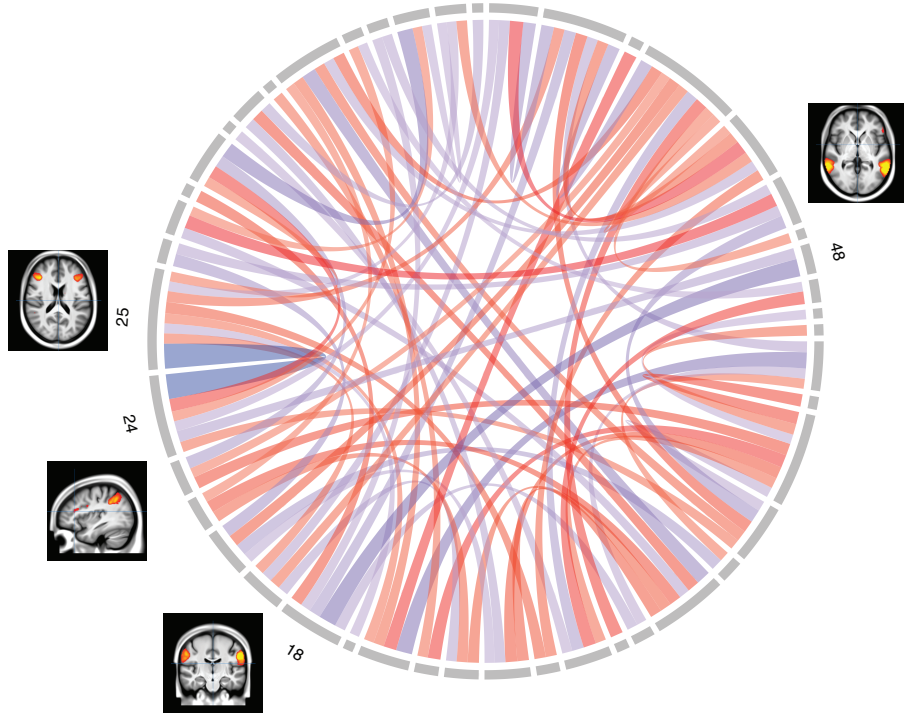


Figure S11: The influence of the super-variant on Chromosome 5 set 23 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with stronger differences in connectivity, including regions indexed as 18, 24, 25 and 48.

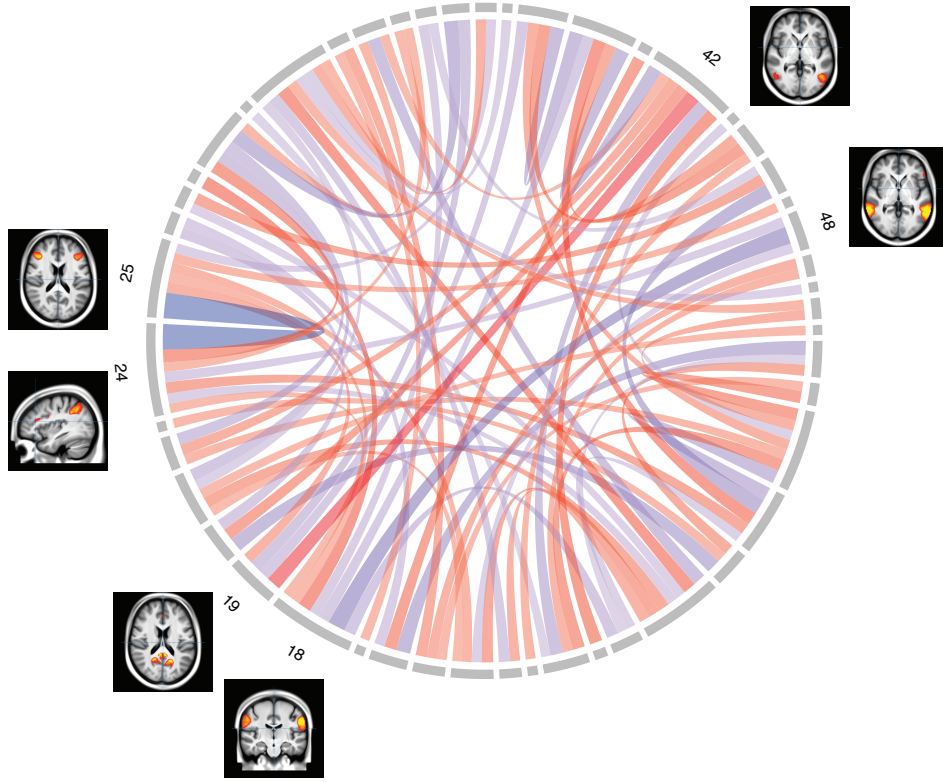


Figure S12: The influence of the super-variant on Chromosome 8 set 145 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with stronger differences in connectivity, including regions indexed as 18, 19, 24, 25, 42 and 48.

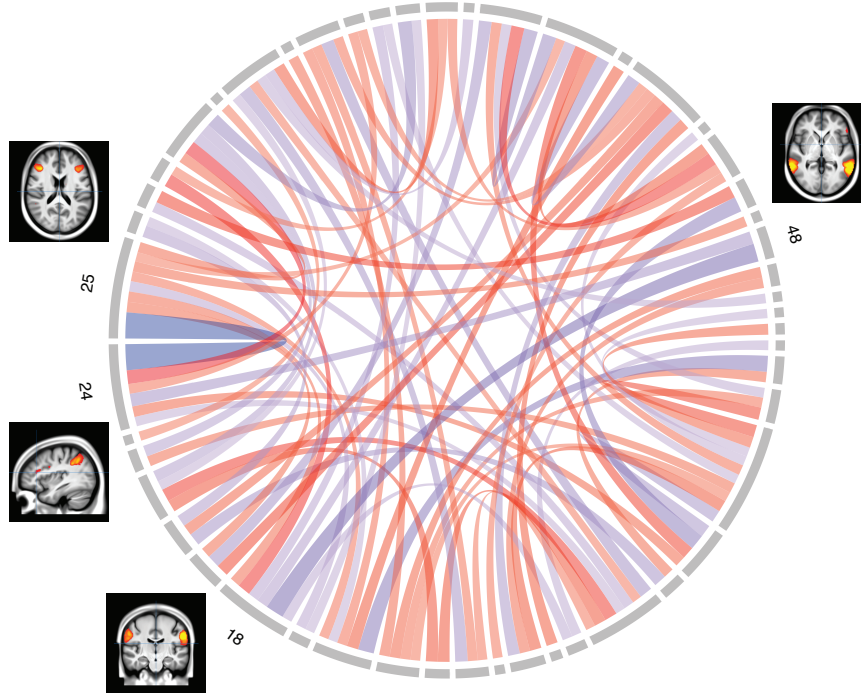


Figure S13: The influence of the super-variant on Chromosome 21 set 11 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with stronger differences in connectivity, including regions indexed as 18, 24, 25 and 48.

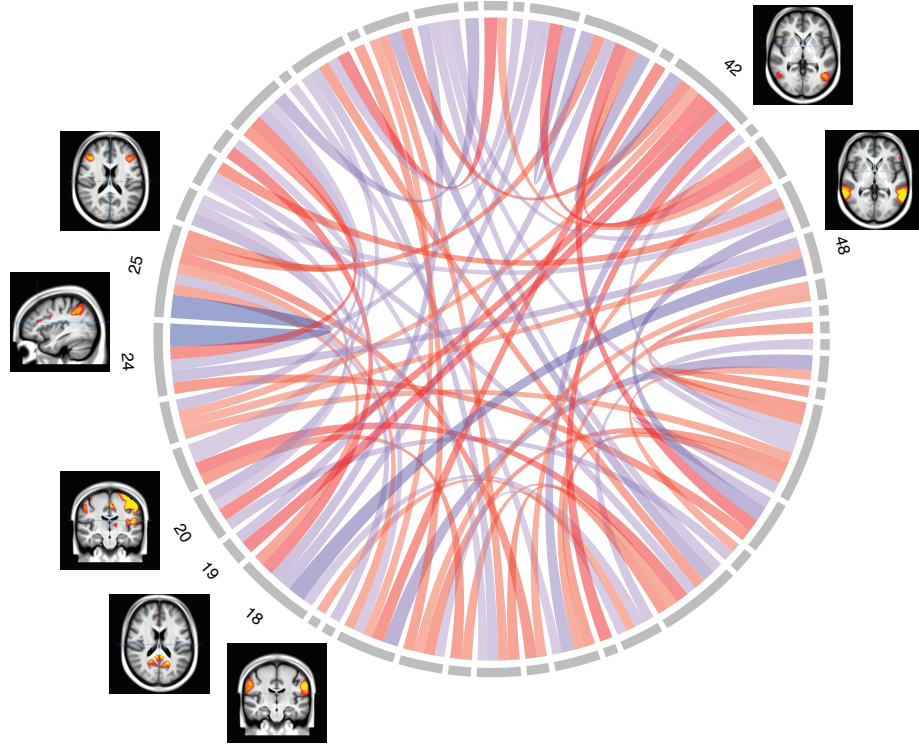


Figure S14: The influence of the super-variant on Chromosome 22 set 39 on brain connectivity. We standardize the elements of the connectivity matrices to have mean 0 and variance 1. Individuals in the combined set are separated into two groups according to the minor and major variants of the super-variant on Chromosome 1 set 149. The difference matrix is calculated by subtracting the average connectivity matrix of the group with the major variant from the average connectivity matrix of the group with the minor variant. For visualization, only differences with absolute values in top 5% are plotted in the chord diagram. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. The numbers in the outer circle indicate specific regions in the brain. We provide the axial/sagittal/coronal view of the brain regions with stronger differences in connectivity, including regions indexed as 18, 19, 20, 24, 25, 42 and 48.

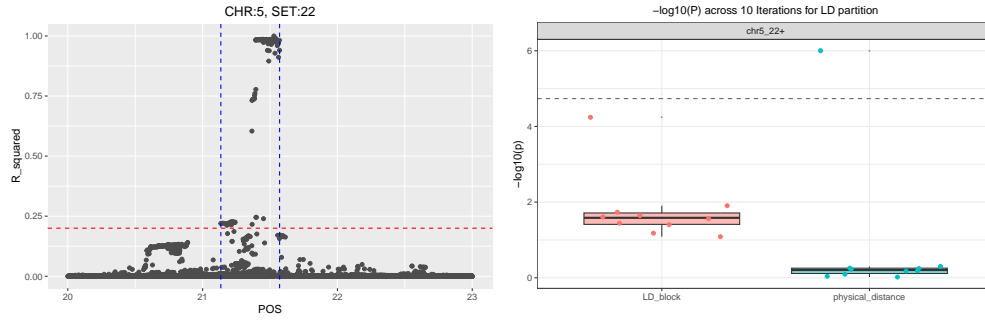


Figure S15: (Left) LD  $r^2$  values around the central SNP located between 21MB and 22MB on chromosome 5. The LD block, indicated by blue dashed lines, includes SNPs with  $r^2$  values exceeding 0.2. (Right) The negative log10-transformed p-values from 10 iterations are shown for both LD-based (red) and physical distance-based partitions (blue). Although none of the p-values achieved statistical significance, the smallest p-value ( $5.7 \times 10^{-5}$ ) was close to the significance threshold ( $1.83 \times 10^{-5}$ ), indicating a consistent trend across both partitioning methods.

Method	SNP Set	Detection Rate ( $G = 15$ )	Detection Rate ( $G = 30$ )	Detection Rate ( $G = 150$ )
BCRA	1	0.941	1.000	1
BCRA	2	1.000	1.000	0.985
BCRA	3	0.690	0.995	0.94
subsample-BCRA	1	0.586	0.944	0.872
subsample-BCRA	2	0.537	0.992	0.882
subsample-BCRA	3	0.118	0.812	0.390

Table S13: Detection rates for SNP-sets across different numbers of SNP groups ( $G$ ) in multi-set simulations with chi-squared weight and  $\mathbf{B}$  in Fig.2a. For example, when  $G=30$ , SNP Sets 1, 2, and 3 are the true sets containing signals (6 true SNPs each in Sets 1 and 2, and 9 in Set 3). However, when  $G=15$ , the true SNPs in Sets 1 and 2 are combined into a single set. In this case, the detection rate for the original Sets 1 and 2 is calculated based on whether any of the true SNPs from the original sets were selected across iterations, maintaining consistency in evaluating detection rates across different partitioning strategies.

# References

- Amico, E. & Goñi, J. (2018), ‘The quest for identifiability in human functional connectomes’, *Scientific reports* **8**(1), 8254.
- Finn, C., Abbeel, P. & Levine, S. (2017), Model-agnostic meta-learning for fast adaptation of deep networks, *in* ‘International conference on machine learning’, PMLR, pp. 1126–1135.
- Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X. & Constable, R. T. (2015), ‘Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity’, *Nature neuroscience* **18**(11), 1664–1671.
- Ho, C. M. & Hsu, S. D. (2015), ‘Determination of nonlinear genetic architecture using compressed sensing’, *GigaScience* **4**(1), s13742–015.
- Hunter, M. D., McKee, K. L. & Turkheimer, E. (2023), ‘Simulated nonlinear genetic and environmental dynamics of complex traits’, *Development and psychopathology* **35**(2), 662–677.
- Ponsoda, V., Martínez, K., Pineda-Pardo, J. A., Abad, F. J., Olea, J., Román, F. J., Barbey, A. K. & Colom, R. (2017), ‘Structural brain connectivity and cognitive ability differences: A multivariate distance matrix regression analysis’, *Human brain mapping* **38**(2), 803–816.
- Venkatesh, M., Jaja, J. & Pessoa, L. (2020), ‘Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification’, *NeuroImage* **207**, 116398.