Supplementary Note

A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases

Buhm Han[1-4,25], Jennie G Pouget[1,5-7,25], Kamil Slowikowski[1,3,4,8], Eli Stahl[9], Cue Hyunkyu Lee[10], Dorothee Diogo[1-3], Xinli Hu[1,3,4,11], Yu Rang Park[10,12], Eunji Kim[10,13], Peter K Gregersen[14], Solbritt Rantapää Dahlqvist[15], Jane Worthington[16,17], Javier Martin[18], Steve Eyre[16,17], Lars Klareskog[19], Tom Huizinga[20], Wei-Min Chen[21], Suna Onengut-Gumuscu[21], Stephen S Rich[21], Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium[22], Naomi R Wray[23], Soumya Raychaudhuri[1,3,4,19,24]

[1]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, USA
[2]Department of Convergence Medicine, University of Ulsan College of Medicine & Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea
[3]Partners Center for Personalized Genetic Medicine, Boston, USA
[4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, USA
[5]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada
[6]Institute of Medical Sciences, University of Toronto, Toronto, Canada
[7]Department of Psychiatry, University of Toronto, Toronto, Canada
[8]Bioinformatics and Integrative Genomics, Harvard University, Cambridge, USA
[9]Department of Psychiatry, Mount Sinai School of Medicine, New York, USA
[10]Asan Institute for Life Sciences, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[11]Harvard-MIT Division of Health Sciences and Technology, Boston, USA
[12]Department of Biomedical Informatics, Asan Medical Center, Seoul, Republic of Korea
[13]Department of Chemistry, Seoul National University, Seoul, Republic of Korea
[14]Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, Manhasset, USA
[15]Department of Public Health and Clinical Medicine, Rheumatology, Umeå University, Umeå, Sweden
[16]Arthritis Research UK Centre for Genetics and Genomics, Musculoskeletal Research Centre, Institute for Inflammation and Repair, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK
[17]National Institute for Health Research, Manchester Musculoskeletal Biomedical Research Unit, Central Manchester University Hospitals National Health Service Foundation Trust, Manchester Academic Health Sciences Centre, Manchester, UK
[18]Instituto de Parasitologíay Biomedicina López-Neyra, Consejo Superior de Investigaciones Científicas, Granada, Spain
[19]Rheumatology Unit, Department of Medicine, Karolinska Institutet and Karolinska University Hospital Solna, Stockholm, Sweden
[20]Department of Rheumatology, Leiden University Medical Centre, Leiden, the Netherlands
[21]Center for Public Health Genomics, University of Virginia, Charlottesville, USA
[22]A full list of members and affiliations appears in the **Supplementary Note**
[23]The University of Queensland, Queensland Brain Institute, Brisbane, Australia
[24]Institute of Inflammation and Repair, University of Manchester, Manchester, UK
[25]These authors contributed equally to this work

Correspondence to:
Soumya Raychaudhuri
77 Avenue Louis Pasteur, Harvard New Research Building, Suite 250D, Boston, MA 02446, USA. soumya@broadinstitute.org; 617-525-4484 (tel); 617-525-4488 (fax)

Buhm Han
Asan Institute for Life Sciences, Asan Medical Center, 88, Olympic-ro 43-gil, Songpa-gu, Seoul 138-736, Korea. buhm.han@amc.seoul.kr; 82-2-3010-2648 (tel); 82-2-3010-2619 (fax)

TABLE OF CONTENTS

MOTIVATING INTUITION FOR BUHMBOX

Allele dosages are uncorrelated if there is no subgroup heterogeneity

Our method is built upon the intuition that if there is no subgroup heterogeneity in case individuals, then for loci that are independent in control individuals, under the additive model assumption, the risk alleles at those loci will be independent in the case individuals as well. Although this seems to be intuitive, analytically proving this will set up the reasoning for our null hypothesis of no correlations.

We will assume haploid model and consider two loci that are biallelic. If we consider two loci together, there will be 4 possible pairs of alleles. Thus, we can define a new “virtual locus” that consists of the two loci, which has 4 alleles. To deal with 4 alleles, in the following, it will be useful to define the multiallelic odds ratio. For a single biallelic locus, if we let p^+ and p^- be the risk allele frequencies (RAFs) in cases and controls respectively, the biallelic odds ratio (OR) is

$$\gamma = \frac{p^+/(1-p^+)}{p^-/(1-p^-)}$$

Thus, case RAF is a function of OR and control RAF,

$$p^+ = \frac{\gamma p^-}{(\gamma-1)p^-+1} \tag{1}$$

Then we generalize biallelic OR to multiple alleles. Suppose that we have M different risk alleles in addition to a reference allele (total M+1 alleles). Let p_R^+ and p_R^- be the case and control reference allele frequencies. Let p_{i}^+ and p_{i}^- be the case and control RAFs of allele *i* (*i=1,…,M*). The constraints are, $p_1^+ + p_2^+ + \cdots + p_M^+ + p_R^+ = 1$ and $p_1^- + p_2^- + \cdots + p_M^- + p_R^- = 1$. The multiallelic odds ratios are defined as

$$\begin{array}{ll} \gamma_1 & = \frac{p_1^+/p_R^+}{p_1^-/p_R^-} \\ \cdots & \cdots \\ \gamma_M & = \frac{p_M^+/p_R^+}{p_M^-/p_R^-} \end{array}$$

Again, case RAF is a function of OR and control RAF. For reference allele,

$$p_R^+ = \frac{1}{1+\frac{1}{p_R^-}\sum_{i=1}^{M} p_i^- \gamma_i} \tag{2}$$

and for allele i=1,…,M,

$$p_i^+ = p_i^- \gamma_i \frac{p_R^+}{p_R^-} \tag{3}$$

Now recall that we consider two independent biallelic risk loci. Let γ_1 and γ_2 be their respective ORs and p_1^- and p_2^- be their control RAFs. Let p_1^+ and p_2^+ be the case RAFs, which can be derived from equation (1). We assume the standard additive model that, if an individual has risk alleles at both loci, the resulting OR is $\gamma_1 \times \gamma_2$. If we consider "virtual locus" spanning the two loci, there will be 4 possible alleles. We can build the following multi-allelic frequency table,

Alleles at locus 1 and 2	Case	Control	OR
Risk at both	p_{12}^+	$p_{12}^- = p_1^- p_2^-$	$\gamma_1\gamma_2$
Reference at 1, Risk at 2	$p_{\bar{1}2}^+$	$p_{\bar{1}2}^- = (1-p_1^-)p_2^-$	γ_2
Risk at 1, Reference at 2	$p_{1\bar{2}}^+$	$p_{1\bar{2}}^- = p_1^-(1-p_2^-)$	γ_1
Reference at both	$p_{\bar{1}\bar{2}}^+$	$p_{\bar{1}\bar{2}}^- = (1-p_1^-)(1-p_2^-)$	1

Note that the control allele frequency is simply multplication of the frequencies at the two loci ("Control" column in the table above), because these loci are uncorrelated in control individuals. Our goal is to prove that the case allele frequency can be similarly decomposed into the multiplication of the frequencies of the two loci. If that is the case, that will show that these loci are independent in the case individuals as well. Consider the reference allele frequency, $p_{\bar{1}\bar{2}}^+$. By equation (2), we have

$$p^{+}_{\bar{1}\bar{2}} = \frac{1}{1+\frac{p^{-}_{\bar{1}2}}{p^{-}_{\bar{1}\bar{2}}}\gamma_1+\frac{p^{-}_{1\bar{2}}}{p^{-}_{\bar{1}\bar{2}}}\gamma_2+\frac{p^{-}_{12}}{p^{-}_{\bar{1}\bar{2}}}\gamma_1\gamma_2}$$

$$= \frac{1}{1+\frac{p^{-}_{1}(1-p^{-}_{2})}{(1-p^{-}_{1})(1-p^{-}_{2})}\gamma_1+\frac{(1-p^{-}_{1})p^{-}_{2}}{(1-p^{-}_{1})(1-p^{-}_{2})}\gamma_2+\frac{p^{-}_{1}p^{-}_{2}}{(1-p^{-}_{1})(1-p^{-}_{2})}\gamma_1\gamma_2}$$

$$= \frac{(1-p^{-}_{1})(1-p^{-}_{2})}{(1-p^{-}_{1})(1-p^{-}_{2})+p^{-}_{1}(1-p^{-}_{2})\gamma_1+(1-p^{-}_{1})p^{-}_{2}\gamma_2+p^{-}_{1}p^{-}_{2}\gamma_1\gamma_2}$$

$$= \frac{(1-p^{-}_{1})(1-p^{-}_{2})}{((1-p^{-}_{1})+\gamma_1 p^{-}_{1})((1-p^{-}_{2})+\gamma_2 p^{-}_{2})}$$

$$= (1-p^{+}_{1})(1-p^{+}_{2}) \tag{4}$$

Thus, the case allele frequency is also a multiplication of the frequencies of the two loci. The similar decomposition can be done for the other three alleles. For example, by equations (3) and (4),

$$p^{+}_{12} = p^{-}_{12}\gamma_1\gamma_2\frac{p^{+}_{\bar{1}\bar{2}}}{p^{-}_{\bar{1}\bar{2}}}$$

$$= p^{-}_{1}p^{-}_{2}\gamma_1\gamma_2\frac{(1-p^{+}_{1})(1-p^{+}_{2})}{(1-p^{-}_{1})(1-p^{-}_{2})}$$

$$= \gamma_1 p^{-}_{1}\frac{(1-p^{+}_{1})}{(1-p^{-}_{1})}\gamma_2 p^{-}_{2}\frac{(1-p^{+}_{2})}{(1-p^{-}_{2})}$$

$$= p^{+}_{1}p^{+}_{2}$$

Since the frequency is the product of frequencies of each allele, the two loci are independent. Thus, we conclude that the loci that are independent in controls will also be independent in case individuals under the standard additive model.

Subgroup heterogeneity induces positive correlations

An equally important intuition is that if there is subgroup heterogeneity, the risk alleles at loci that are independent in control individuals will show positive correlations in case individuals. This fact sets the foundation stone for our alternative hypothesis of positive correlations.

Suppose that disease A (D_A) case individuals consist of two groups: one group genetically similar to a second trait (disease B, D_B) and the rest not similar to D_B. Say (π x 100)% of case individuals are in the D_B-similar group. We will call π the *heterogeneity proportion*. Consider two independent SNPs that are associated to the second trait. Let their risk allele frequencies be p_1^+ and p_2^+ in the D_B-similar group and p_1^- and p_2^- in the rest. Note that the two loci are uncorrelated within each subgroup (the D_B-similar group comprises individuals that are genetically cases for D_B, and we have already shown that independent risk loci will be uncorrelated in case individuals under the standard additive model).

If we consider frequencies of haplotypes spanning the two loci in the D_A case individuals consisting of both subgroups,

Alleles at locus 1 and 2	**Haplotype**	**Frequency**
Risk at both	p_{12}	$\pi p_1^+ p_2^+ + (1-\pi)p_1^- p_2^-$
Reference at 1, Risk at 2	$p_{\bar{1}2}$	$\pi(1-p_1^+)p_2^+ + (1-\pi)(1-p_1^-)p_2^-$
Risk at 1, Reference at 2	$p_{1\bar{2}}$	$\pi p_1^+(1-p_2^+) + (1-\pi)p_1^-(1-p_2^-)$
Reference at both	$p_{\bar{1}\bar{2}}$	$\pi(1-p_1^+)(1-p_2^+) + (1-\pi)(1-p_1^-)(1-p_2^-)$

The expected value of Pearson correlation is therefore

$$r_{12} = \frac{p_{12}p_{\bar{1}\bar{2}} - p_{\bar{1}2}p_{1\bar{2}}}{\sqrt{p_{1\cdot}p_{\bar{1}\cdot}p_{\cdot 2}p_{\cdot\bar{2}}}} \tag{5}$$

where dot ($\cdot$) in the subscript denotes marginal frequency, for example $p_{1\cdot} = p_{12} + p_{1\bar{2}}$. A few interesting characteristics are, (1) r_{12} is always positive or zero because we considered risk allele dosage at both loci. (2) $r_{12} = 0$ if $\pi = 0$ or $\pi = 1$. (3) $r_{12} = 0$ if risk is zero ($p_1^+ = p_1$ and $p_2^+ = p_2$). (4) r_{12} is a function of RAF, OR, and π (but not of sample size). r_{12} is typically a very small value. **Supplementary Figure 2** shows the value of r_{12}

as a function of OR at the two loci (when we fix p_1 and p_2 to 0.5) and the value of r_{12} as a function of p_1 and p_2 (when we fix OR to 1.5 at both loci).

DERIVATION OF BUHMBOX

Null and alternative hypotheses

Building upon the aforementioned intuitions, we can build the BUHMBOX statistic to detect positive correlations between independent loci, which will be evidence of subgroup heterogeneity. Suppose that we examine D_A case individuals at M independent D_B-associated loci. Between these loci, we calculate correlations of risk allele dosages to obtain an $M\times M$ correlation matrix, $\mathbf{R}$. The null hypothesis of our method is that the non-diagonal elements of **R** are zero. The alternative hypothesis of our method is that the non-diagonal elements of **R** are positive. We build our method in the following steps.

Combining correlations into one statistic

The first challenge is to combine $M(M-1)/2$ non-diagonal elements of $\mathbf{R}$ into one statistic. To this end, we show that under the null hypothesis of no correlations, the non-diagonal elements of the observed correlation matrix will be independent of each other. We employ the framework of Jennrich[1]. Jennrich describes a framework for testing deviance of a correlation matrix from a specified null matrix. To describe the framework briefly, let $\mathbf{P}=(\rho_{ij})$ be a specific $M\times M$ correlation matrix that defines the null hypothesis. The goal is to test if the observed sample correlation matrix $\mathbf{R}$ deviates from $\mathbf{P}$. To define a statistic, what we need is the inverse of asymptotic covariance matrix for the maximum likelihood estimates of ρ_{ij}, which we call $\mathbf{\Gamma}^{-1}$. Note that $\mathbf{\Gamma}^{-1}$ is a $q \times q$ matrix where $q = M(M-1)/2$.

Let $\mathbf{P}^{-1} = (\rho^{ij})$. Let δ_{ij} be the Kronecker delta, that is, 1 if $i = j$ and zero otherwise. We define $\mathbf{T} = (t_{ij})$ as the following,

$$t_{ij} = \delta_{ij} + \rho_{ij}\rho^{ij} \tag{6}$$

Then, $\mathbf{\Gamma}^{-1}$ is given by

$$\mathbf{\Gamma}^{-1}(i,j;k,l) = \rho^{ik}\rho^{jl} + \rho^{il}\rho^{jk} - \rho^{ij}(t^{ik} + t^{jk} + t^{il} + t^{jl})\rho^{kl} \tag{7}$$

Given these, if we define

$$\mathbf{Y} = \sqrt{N}(\mathbf{R} - \mathbf{P}) = (y_{ij}) \tag{8}$$

where N is the number of samples used to calculate R, the test statistic is

$$S_{Jennrich} = \sum_{i<j} \sum_{k<l} y_{ij}\mathbf{\Gamma}^{-1}(i,j;k,l)y_{kl} \tag{9}$$

which follows χ^2 distribution with q degrees of freedom under the null. The computation is challenging if p is large because of the time complexity $O(q^2) = O(p^4)$. Jennrich applies an optimization technique to simplify the formula to

$$S_{Jennrich} = \frac{1}{2}tr(\mathbf{Y}\mathbf{P}^{-1}\mathbf{Y}\mathbf{P}^{-1}) - dg' \quad (\mathbf{P}^{-1}\mathbf{Y})\mathbf{T}^{-1}dg(\mathbf{P}^{-1}\mathbf{Y}) \tag{10}$$

which only involves operations between $M \times M$ matrices requiring only $O(M^3)$.

In our situation, this statistic simplifies further. Our null hypothesis is no correlation. Thus, the identity matrix $\mathbf{I}$ is our null correlation matrix ($\mathbf{P} = \mathbf{I}$). Substituting $\mathbf{P}$ with $\mathbf{I}$, the statistic simplifies to

$$S_{Jennrich}|_{\mathbf{P}=\mathbf{I}} = \frac{1}{2}tr(\mathbf{YY}) = \sum_{i<j} y_{ij}^2 \tag{11}$$

Note that each y_{ij} asymptotically follows a normal distribution (thus, a z-score). The statistic can be interpreted as the following; under the specific situation $\mathbf{P} = \mathbf{I}$, the z-scores become asymptotically independent. Thus, we can combine information by simply summing up their squares which will follow χ^2 distribution with q degrees of freedom.

OPTIMIZATION OF BUHMBOX

Accounting for directions to increase power

A straightforward application of Jennrich's approach in equation (11) is not optimal for our situation, because Jennrich's test is a general test that does not account for the direction of correlations. As we have described, subgroup heterogeneity only results in positive expected correlations between risk loci. Thus, accounting for this fact may give us better power.

To this end, we employ the meta-analytic framework Wei[2,3] and Lin and Sullivan[4]. This framework combines multiple estimates whose asymptotic covariance is known while accounting for their directions. Applying this approach, we obtain a new statistic that is alternative to equation (9),

$$S_{Directional} = \frac{\sum_{i<j}\sum_{k<l} y_{ij}\Gamma^{-1}(i,j;k,l)}{\sqrt{\sum_{i<j}\sum_{k<l}\Gamma^{-1}(i,j;k,l)}}$$

which follows $N(0,1)$ under the null hypothesis of $R = P$. To reduce computational complexity, we can apply the same optimization technique of Jennrich to simplify the statistic to

$$S_{Directional} = \frac{\frac{1}{2}tr(\mathbf{YP}^{-1}\mathbf{EP}^{-1}) - dg' \;\; (\mathbf{P}^{-1}\mathbf{Y})\mathbf{T}^{-1}dg(\mathbf{P}^{-1}\mathbf{E})}{\sqrt{\frac{1}{2}tr(\mathbf{EP}^{-1}\mathbf{EP}^{-1}) - dg' \;\; (\mathbf{P}^{-1}\mathbf{E})\mathbf{T}^{-1}dg(\mathbf{P}^{-1}\mathbf{E})}}$$

where $\mathbf{E}$ is an $M \times M$ matrix whose elements are all ones.

In our specific situation that $\mathbf{P} = \mathbf{I}$, this statistic further simplifies to

$$S_{Directional}|_{\mathbf{P}=\mathbf{I}} = \frac{\sum_{i<j} y_{ij}}{\sqrt{q}} \qquad (12)$$

which follows the standard normal distribution under the null hypothesis. We calculate significance of this statistic using the normal distribution in a positive one-sided test.

Optimizing weights based on effect sizes and allele frequencies

The directional test in equation (12) accounts for the directions of correlations, but does not account for the effect size and frequency differences between loci. As we have shown, expected correlation of two loci is a function of not only π, heterogeneity proportion, but also the effect sizes and RAFs of the two loci. Thus, each pair of loci will have different expected correlations. When we combine correlations to one statistic, we can have better power by giving higher weights to the pairs of high expected correlations.

Recall that the expected correlation is, as given in equation (5),

$$r_{12} = \frac{p_{12}p_{\overline{1}\overline{2}} - p_{\overline{1}2}p_{1\overline{2}}}{\sqrt{p_{1\cdot}p_{\overline{1}\cdot}p_{\cdot 2}p_{\cdot\overline{2}}}}$$

We examine what is the increase in r_{12} given an increase in π at the local region around the null hypothesis $\pi = 0$.

$$w_{12} = \frac{\partial r_{12}}{\partial \pi}|_{\pi=0} = \frac{\sqrt{p_1(1-p_1)p_2(1-p_2)}(\gamma_1-1)(\gamma_2-1)}{((\gamma_1-1)p_1+1)((\gamma_2-1)p_2+1)}$$

The value w_{12} can be thought of as the slope of the curves evaluated at $\pi = 0$. For any loci pair i and j, we can calculate w_{ij}.

Then what we need is an optimal strategy to incorporate w_{ij} into our testing, so that our method can have the local optimum property at around $\pi = 0$. That is, should we weight y_{ij} by w_{ij} or $\sqrt{w_{ij}}$? We note that our situation is analogous to a situation where in a meta-analysis, the effect sizes are the same for all participating studies but their units or scales are different, thus requiring different weights. We extended the traditional meta-analysis method, fixed effects model (FE), to a new model which can deal with the situation that the scales are different between studies. We present the details in the end of **Supplmentary Note**. Briefly speaking of the conclusion, the optimal strategy is multiplying the scaling parameters directly into the weights of the sum of weighted z-scores.

Based on this reasoning, our statistic becomes

$$S_{BUHMBOX} = \frac{\sum_{i<j} w_{ij} y_{ij}}{\sqrt{\sum_{i<j} w_{ij}^2}} \tag{13}$$

which follows $N(0,1)$ under the null hypothesis. We calculate the significance of this statistic uising the normal distribution assuming a positive one-sided test.

Controlling for LD and utilizing control samples

Because linkage disequilibrium (LD) can induce unexpected correlations between loci, one should prune the loci before applying BUHMBOX. We suggest a harsh criterion (e.g. removing SNPs that are $r^2 < 0.1$ or that are nearby (within $\pm$1Mb) to other SNPs). However, even after harsh pruning, there can be residual LD that can affect the results. To minimize the effect of residual LD, BUHMBOX uses control samples.

Recall that when we defined the z-scores based on correlations in case samples, we used the formula in equation (8),

$$\mathbf{Y} = \sqrt{N}(\mathbf{R} - \mathbf{P}) = (y_{ij})$$

which means that we should multiply the correlation elements by square root of sample size to obtain z-scores. Now if we use control samples, and let $\mathbf{R}'$ be the correlation matrix of control individuals, we can define a new $\mathbf{Y}'$ as

$$\mathbf{Y}' = \sqrt{\frac{NN'}{N+N'}}(\mathbf{R} - \mathbf{R}')$$

where N' is the control sample size.

Although the basic form of our approach is case-only statistic, using only cases requires a strong assumption of no residual LD. By subtracting control correlations from case correlations to use "*delta-correlations*", we obtain a more robust statistic against the effect of LD.

Controlling for population stratification

We correct for population stratification by regressing out PCs from the vector of case/control allele dosage of each locus. Note that we regress out PCs from each locus one by one, not simultaneously from the whole dosage matrix including multiple loci. This way, our approach can be thought of as obtaining partial correlations.

Meta-analysis of BUHMBOX results

The BUHMBOX statistic is a z-score. Therefore, we can meta-analyze BUHMBOX results using the standard weighted sum of z-score approach, where z-scores are weighted by the square root of the total sample size.

POLYGENIC MODELING AND BUHMBOX

We assessed by simulations whether our method can benefit by taking advantage of a polygenic modeling approach. In GWAS, the use of a stringent threshold (P-value threshold t=5x10^{-8}) minimizes false positives, but likely misses true positives due to imperfect power. Therefore, investigators often use a polygenic modeling approach that applies a more liberal threshold to define a larger set of variants[5]. We simulated GWASs based on the Bayesian polygenic model[6] and employed more liberal values of *t* ranging from 5x10^{-8} up to 0.01, to obtain larger sets of variants.

Specifically, We adapted Stahl *et al.*'s Bayesian polygenic model[6], which predicted 2,231 causal loci among 84,000 independent genome-wide loci for RA. To simulate 2,231 causal variants, we combined 71 independent known loci of RA[7] to an additional 2,160 loci sampled from the joint posterior distribution of RAF and OR presented in Stahl *et al.* For null loci, we also used the null RAF distribution presented in Stahl *et al.* Given this disease model, we simulated a GWAS with 3,964 cases and 12,052 controls (sample sizes from Stahl *et al.*), assuming prevalence of 0.01. Given GWAS results, we used only the top *k* GWAS loci defined by p-value threshold *t* and their observed odds ratios for BUHMBOX power simulations. We assumed $N=5,000$ and $\pi=0.5$ for power evaluation and tried different p-value threshold *t* from 5x10^{-8} to 0.01.

We observed that the statistical power of BUHMBOX increased when we included variants with moderately significant p-values (**Supplementary Figure 4**). In this simulation, 29.5% power at t=5.0x10^{-8} increased up to 88.1% at t=3.6x10^{-4} and then gradually dropped as we used even more liberal *t* values. This shows that BUHMBOX can benefit from polygenic modeling.

INTERPRETATION OF BUHMBOX RESULTS:

Here we sought to describe in detail what can specifically cause subgroup heterogeneity.

Misclassifications can cause subgroup heterogeneity

Phenotypic misclassifications can cause subgroup heterogeneity. Suppose that a proportion of diagnosed case individuals were actually case individuals for a different disease. In this situation, the heterogeneity proportion π corresponds to the misclassification proportion.

Molecular subtypes can cause subgroup heterogeneity

Suppose that disease A can occur because of multiple different molecular pathways, and one of the pathways is shared with disease B. Then, the subgroup that shares a molecular pathway with disease B will show genetic characteristics that are similar to disease B patients.

Phenotypic causality can cause subgroup heterogeneity

This is a situation that is often called "mediated pleiotropy"[8]. Assume that there are two conditions A and B whose population prevalences are K_A and K_B. First, consider the null situation that A and B are not causal to each other. Obviously, B-associated loci will be uncorrelated within A cases. Within A cases, the prevalence of B will be K_B.

Now consider the situation that B causes A (having condition B increases the chance of acquiring A). Let $K_{A|B}$ and $K_{A|\bar{B}}$ be the frequency of A among B patients and among non-B-patients. The population attributable risk of A to B is

$$\mathrm{PAR} = K_{A|B} - K_{A|\bar{B}} > 0$$

The population prevalence of A can be written

$$K_A = K_B K_{A|B} + (1 - K_B) K_{A|\bar{B}}$$

Thus, within A cases, the proportion of individuals having B will be

$$K_{B|A} = \frac{K_B K_{A|B}}{K_A} = \frac{K_B K_{A|B}}{K_B K_{A|B} + (1 - K_B) K_{A|\bar{B}}} > F_B$$

In this situation, the heterogeneity proportion π corresponds to the excessive proportion of individuals having B among A cases, that would not have occurred without the causal relationship;

$$\pi = K_{B|A} - K_B$$

which has the following relationship to PAR

$$\pi = \frac{K_B(1 - K_B)}{K_A}\ \mathrm{PAR}$$

Whole-group pleiotropy <u>cannot</u> cause subgroup heterogeneity

Common genetic basis between all patients of A and all patients of B (whole-group pleiotropy) does not cause subgroup heterogeneity. If disease A and B share risk alleles, within A cases, the frequencies of risk alleles for B may increase. However, there will not be a subgroup who has excessivie numbers of risk alleles. Instead, the risk alleles will occur independently and homogeneously across all A cases even if only a subset of variants have pleiotropic effects. Thus, there will not be correlations among B-associated-risk alleles among A cases. Note that the prevalence of B may increase

$(K_{B|A} > K_B)$, but we can think of A cases being sampled from a new homogeneous population which has a larger prevalence of B.

Inverse causal relationship cannot cause subgroup heterogeneity

We previously considered the causal relationship of disease B causing A, but now consider the inverse situation that A causes B, while B does not cause A. Again we examine the correlations between B-associated loci in A cases. This is a similar situation to pleiotropy in that we can also think of A cases being sampled from a new population which has a larger prevalence of B. Since there will not be a subgroup, there will not be correlations.

META-ANALYSIS WITH SCALE DIFFERENCES

In deriving the strategy to incorporate weights in equation (13), our BUHMBOX method utilized a meta-analytic framework that accounts for the scale or unit differences between studies. Because the description of the framework is long, we pushed the description back to here. We will first introduce the well-known meta-analytic method, fixed effects model, and extend it to account for scale differences.

Fixed effects model

We first review the fixed effects model meta-analysis method. Let X_i be the observed effect size of study i and V_i be the variance of it. By definition, z-score is defined $Z_i = \frac{X_i}{\sqrt{V_i}}$. Let $W_i = V_i^{-1}$ be the inverse variance. Under the fixed effects model, we assume that X_i has mean μ that is constant (fixed) across the studies. There are two common approaches under the fixed effects model: the inverse variance weighted average and the weighted sum of z-scores. Two approaches are related as shown below.

Inverse variance weighted average In the inverse variance weighted average approach, the goal is to obtain the best estimate of μ. A commonly used estimate is weighted average, $\bar{X}_i = \frac{\sum_i c_i X_i}{\sum_i c_i}$, which is an unbiased estimate of μ for any $c_i > 0$. The variance of $\bar{X}_i$ is $\bar{V}_i = \frac{\sum_i c_i^2 V_i}{(\sum_i c_i)^2}$. To obtain the best estimate, we choose c_i that minimizes the variance. By the Cauchy-Schwarz inequality,

$$\bar{V}_i = \frac{\sum_i c_i^2 V_i}{(\sum_i c_i)^2} \geq \left(\sum_i \frac{c_i\sqrt{V_i}}{\sum_j c_j} \times \frac{1}{\sqrt{V_i}}\right) / \sum_i \frac{1}{V_i} = 1/\sum_i \frac{1}{V_i}$$

The equality is achieved when

$$\frac{c_i\sqrt{V_i}}{\sum_j c_j} = k \cdot \frac{1}{\sqrt{V_i}}$$

for a constant $k > 0$. Without losing generality, we can assume $\sum_j c_j$ is a constant. Thus, we can achieve the equality by choosing $c_i = W_i$, which is why the method is called inverse variance weighted average. Given these weights, the average estimate and its variance are

$$\bar{X} = \frac{\sum_i W_i X_i}{\sum_i W_i} \quad (15)$$

$$\bar{V} = \frac{1}{\sum_i W_i} \quad (16)$$

Weighted sum of z-scores In the weighted sum of z-scores approach, the goal is to maximize the power of statistic. Given z-scores $Z_i = \frac{X_i}{\sqrt{V_i}}$, we can define a weighted sum of z-scores statistic $\bar{Z} = \frac{\sum_i b_i Z_i}{\sqrt{\sum_i b_i^2}}$. $\bar{Z}$ is also a z-score (normally distributed and of variance 1) for any weights $b_i > 0$. To maximize power, we want to maximize the non-centrality parameter of $\bar{Z}$, $E[\bar{Z}]$. Since in each study $E[Z_i] = \frac{\mu}{\sqrt{V_i}}$, we have

$$E[\bar{Z}] = \frac{\sum_i b_i \mu / \sqrt{V_i}}{\sqrt{\sum_i b_i^2}}$$

Again, by the Cauchy-Schwarz inequality,

$$E[\bar{Z}] = \mu \cdot \sum_i \frac{b_i}{\sqrt{\sum_j b_j^2}} \cdot \frac{1}{\sqrt{V_i}} \leq \mu \sqrt{\left(\sum_i \frac{b_i^2}{\sum_j b_j^2}\right)\left(\sum_i \frac{1}{V_i}\right)}$$

The equality is achieved when

$$\frac{b_i}{\sqrt{\sum_j b_j^2}} = k \cdot \frac{1}{\sqrt{V_i}}$$

for a constant $k > 0$. Without losing generality, we can assume $\sum_j b_j^2$ is a constant. Thus, we can achieve the equality by choosing $b_i = 1/\sqrt{V_i} = \sqrt{W_i}$. Given these weights, the weighted sum of z-scores statistic is

$$\bar{Z} = \frac{\sum_i \sqrt{W_i} Z_i}{\sqrt{\sum_i W_i}} \tag{17}$$

Note that in many applications, we can approximate $\sqrt{W_i} \propto \sqrt{N_i p_i (1 - p_i)}$ where N_i is the sample size of study i and p_i is the allele frequency in study i. If we can assume the allele frequencies are the same for all studies, the weights b_i approximates to $\sqrt{N_i}$, which is the widely used sample-size-based weight for this approach.

Relation of two approaches The two approaches, the inverse weighted average and the weighted sum of z-scores, are closely related. Given the inverse variance weighted average and its variance in equations (15) and (16), one can construct a z-score for statistical testing. The z-score is

$$\bar{Z}^* = \bar{X}/\sqrt{\bar{V}} = \frac{\sum_i W_i X_i}{\sum_i W_i} / \frac{1}{\sqrt{\sum_i W_i}} = \frac{\sum_i \sqrt{W_i} Z_i}{\sqrt{\sum_i W_i}} = \bar{Z}$$

, exactly resulting in the same z-score in the weighted sum of z-scores approach in equation (17). Thus, although the goals of the two approaches were different (obtaining the best estimate and maximizing power), the results of statistical test will be exactly the same for the two approaches, or at least similar if we use the approximation $\sqrt{W_i} \propto \sqrt{N_i p_i (1 - p_i)}$ or $\sqrt{W_i} \propto \sqrt{N_i}$.

Fixed effects model with scale differences

We extend the fixed effects model (FE) to a new model accounting for scale differences. We use the similar notations; let X_i and $V_i = W_i^{-1}$ be the observed effect size and its variance in study i. In the new model, we assume that there is a baseline effect μ which is manifested in different scales for each study. We assume that in study i, X_i has mean $\mu_i = \rho_i \mu$, where ρ_i is a scaling factor that is known a priori. Under this model, we can also propose the inverse variance weighted average and the weighted sum of z-score approaches.

Inverse variance weighted average In this approach, our goal is to obtain the best estimate of μ. In each study, we can define $Y_{i=}\frac{X_i}{\rho_i}$, an estimator of μ. The variance of Y_i is $Var(Y_i) = V_i/\rho_i^2$. We can define the weighted average estimate for μ,

$$\bar{Y} = \frac{\sum_i c_i Y_i}{\sum_i c_i}$$

Using Cauchy-Schwarz inequality, we can show that the variance of $\bar{Y}$ is minimum when $c_i = \frac{1}{Var(Y_i)} = \frac{\rho_i^2}{V_i}$. Given these weights, the average estimate and its variance are

$$\bar{Y} = \frac{\sum_i \rho_i W_i X_i}{\sum_i \rho_i^2 W_i} \tag{18}$$

$$Var(\bar{Y}) = \frac{1}{\sum_i \rho_i^2 W_i} \tag{19}$$

Weighted sum of z-scores Given z-scores $Z_i = \frac{Y_i}{\sqrt{Var(Y_i)}} = \frac{X_i}{\sqrt{V_i}}$, we can construct a weighted sum of z-scores statistic, $\hat{Z} = \frac{\sum_i b_i Z_i}{\sqrt{\sum_i b_i^2}}$. Since in each study $E[Z_i] = \frac{\rho_i \mu}{\sqrt{V_i}}$, we have

$$E[\hat{Z}] = \frac{\sum_i b_i \rho_i \mu / \sqrt{V_i}}{\sqrt{\sum_i b_i^2}}$$

Again, by the Cauchy-Schwarz inequality,

$$E[\hat{Z}] = \mu \cdot \sum_i \frac{b_i}{\sqrt{\sum_j b_j^2}} \cdot \frac{\rho_i}{\sqrt{V_i}} \leq \mu \sqrt{\left(\sum_i \frac{b_i^2}{\sum_j b_j^2}\right)\left(\sum_i \frac{\rho_i^2}{V_i}\right)}$$

The equality is achieved when

$$\frac{b_i}{\sqrt{\sum_j b_j^2}} = k \cdot \frac{\rho_i}{\sqrt{V_i}}$$

for a constant $k > 0$. Thus, we can achieve the equality by choosing $b_i = \rho_i / \sqrt{V_i} = \rho_i \sqrt{W_i}$. In other words, the scaling factor ρ_i is directly multiplied to the original z-score weights used in FE. Given these weights, the weighted sum of z-scores statistic is

$$\hat{Z} = \frac{\sum_i \rho_i \sqrt{W_i} Z_i}{\sqrt{\sum_i \rho_i^2 W_i}} \tag{20}$$

Relation of two approaches The two approaches, the inverse weighted average and the weighted sum of z-scores, have close relation in the new model, as they have close relation in FE. Given the inverse variance weighted average and its variance in equations (18) and (19), one can construct a z-score for statistical testing. The z-score is

$$\hat{Z}^* = \bar{Y} / \sqrt{Var(\bar{Y})} = \frac{\sum_i \rho_i W_i X_i}{\sum_i \rho_i^2 W_i} / \frac{1}{\sqrt{\sum_i \rho_i^2 W_i}} = \frac{\sum_i \rho_i \sqrt{W_i} Z_i}{\sqrt{\sum_i \rho_i^2 W_i}} = \hat{Z}$$

, exactly resulting in the same z-score in the weighted sum of z-scores approach in equation (20). Thus, the results of statistical test will be the same for the two approaches, or similar if we use the approximation $\sqrt{W_i} \propto \sqrt{N_i p_i (1 - p_i)}$ or $\sqrt{W_i} \propto \sqrt{N_i}$.

REFERENCES

1. Jennrich, R.I. An asymptotic χ^2 test for the equality of two correlation matrices. *J Am Statist Assoc* **65,** 904-912 (1970).
2. Wei, L.J., Lin, D.Y. & Weissfeld, L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Statist Assoc* **84,** 1065-1073 (1989).
3. Wei, L.J. & Johnson, W.E. Combining dependent tests with incomplete repeated measurements. *Biometrika* **72,** 359-364 (1985).
4. Lin, D.Y. & Sullivan, P.F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* **85,** 862-872 (2009).
5. International Schizophrenia Consortium*.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748-752 (2009).
6. Stahl, E.A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet* **44,** 483-489 (2012).
7. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet* **44,** 1336-1340 (2012).
8. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14,** 483-495 (2013).

Supplementary Tables

Supplementary Table 1. Null and alternative hypotheses of GRS and BUHMBOX

GRS approach		Subgroup Heterogeneity	
		No	Yes
Whole-group Pleiotropy	No	Null hypothesis	Alternative hypothesis
	Yes	Alternative hypothesis	Alternative hypothesis

BUHMBOX		Subgroup Heterogeneity	
		No	Yes
Whole-group Pleiotropy	No	Null hypothesis	Alternative hypothesis
	Yes	Null hypothesis	Alternative hypothesis

Supplementary Table 2. False positive rate of BUHMBOX

We simulated a million null studies assuming sample size *N*=2,000 and number of risk loci *M*=50. Then given threshold α, the false positive rate was estimated as the proportion of simulated studies with p-value ≤ α.

True threshold α	False positive rate
0.05	0.051
0.01	0.011
0.005	0.0056
0.001	0.0012
0.0005	0.00060

Supplementary Table 3. Detailed SNP information used for GRS and BUHMBOX analyses

Please refer to separate Excel file.

Supplementary Table 4. GRS and BUHMBOX results
Please refer to separate Excel file.

Supplementary Table 5. MDD sample description

Full dataset (used for BUHMBOX analysis)

Study	Tag	N Total	N Cases
Genetic Association Information Network (GAIN)-MDD[1]	gain	3,461	1,694
Genetics of Recurrent Early-Onset Depression[2]	genred	2,283	1,030
Glaxo-Smith-Kline (GSK)[3]	gsk	1,751	887
MDD2000[4]	mdd2kb	1,184	433
	mdd2ks	1,977	1,017
Max Planck Institute of Psychiatry, Munich[5]	munich	913	376
RADIANT-GERMANY[6] and Bonn/Mannheim[7]	radbon	2,225	935
RADIANT-UK[6]	raduk	3,213	1,625
Sequenced Treatment Alternatives to Relieve Depression (STARD)[8]	stardfull	1,752	1,241
	Total	**16,759**	**9,238**

Schizophrenia-GWAS-independent dataset (used for GRS analysis)

Study	Tag	N Controls	N Cases
Genetic Association Information Network (GAIN)-MDD[1]	gain	1,682	1,693
MDD2000[4]	mdd2kb	751	433
	mdd2ks	960	1,016
Max Planck Institute of Psychiatry, Munich[5]	munich	537	375
RADIANT-UK[6]	raduk	1,583	1,624
Sequenced Treatment Alternatives to Relieve Depression (STARD)[8]	stardfull	101	1,241
	Total	**5,614**	**6,382**

1. Sullivan, P.F. *et al*. Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol Psychiatry* **14**, 359-375 (2009).
2. Shi, J. *et al*. Genome-wide association study of recurrent early-onset major depressive disorder. *Mol Psychiatry* **16**, 193-201 (2011).
3. Muglia, P. *et al*. Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts. *Mol Psychiatry* **15**, 589-601 (2010).
4. Wray, N.R. *et al*. Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol Psychiatry* **17**, 36-48 (2012).
5. Kohli, M.A. *et al*. The neuronal transporter gene SLC6A15 confers risk to major depression. *Neuron* **70**, 252-265 (2011).
6. Lewis, C.M. *et al*. Genome-wide association study of major recurrent depression in the U.K. population. *Am J Psychiatry* **167**, 949-957 (2010).
7. Rietschel, M. *et al*. Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. *Biol Psychiatry* **68**, 578-585 (2010).
8. Shyn, S.I. *et al*. Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol Psychiatry* **16**, 202-215 (2011).
9. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371-1379 (2013).
10. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-994 (2013).
11. Bulik-Sullivan, B. *et al*. An atlas of genetic correlations across human diseases and traits. *Nat Genet* (2015).

Supplementary Table 6. Summary of selected previous estimates of genetic overlap between MDD and SCZ

Study	**Method**	**MDD results**	**BPD Results**
Cross-Disorder Group of the PGC[9]	Polygenic risk scoring, p_T=1 (all SNPs)	R^2=0.009, p<1x10^{-16}	R^2=0.025, p<1x10^{-50}
Cross-Disorder Group of the PGC[10]	REML	r_g=0.43, SE=0.06, p<1.0x10^{-16}	r_g=0.68, SE=0.04, p=6.0 x10^{-15}
Bulik-Sullivan *et al.*[11]	LDSR	r_g=0.51, SE= 0.08, p=1.32x10^{-11}	r_g=0.79, SE=0.04, p=7.45x10^{-94}

1. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371-1379 (2013).
2. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-994 (2013).
3. Bulik-Sullivan B. *et al*. An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**,1236-1241 (2015).

Members of the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium

Subgroup	Published Name	Institution	Department	City	State	Country
	Patrick F Sullivan	Karolinska Institutet	Medical Epidemiology and Biostatistics	Stockholm		SE
		University of North Carolina at Chapel Hill	Genetics	Chapel Hill	NC	US
		University of North Carolina at Chapel Hill	Psychiatry	Chapel Hill	NC	US
	Stephan Ripke	Charite Universitatsmedizin Berlin Campus Benjamin Franklin	Department of Psychiatry	Berlin	Berlin	DE
		Broad Institute	Medical and Population Genetics	Cambridge	MA	US
		Massachusetts General Hospital	Analytic and Translational Genetics Unit	Boston	MA	US
	Danielle Posthuma	VU Medical Center	Clinical Genetics	Amsterdam	Noord-Holland	NL
		VU University Amsterdam	Complex Trait Genetics	Amsterdam		NL
Rotterdam Study	Henning Tiemeier	Erasmus MC	Psychiatry	Rotterdam	Zuid-Holland	NL
Rotterdam Study	Henning Tiemeier	Erasmus MC	Epidemiology	Rotterdam	Zuid-Holland	NL
Rotterdam Study	Henning Tiemeier	Erasmus MC	Child and Adolescent Psychiatry	Rotterdam	Zuid-Holland	NL
Rotterdam Study	André G Uitterlinden	Erasmus MC	Internal Medicine	Rotterdam	Zuid-Holland	NL
Rotterdam Study	Nese Direk	Erasmus MC	Epidemiology	Rotterdam	South-Holland	NL
Rotterdam Study	Saira Saeed Mirza	Erasmus MC	Epidemiology	Rotterdam	Zuid-Holland	NL
Rotterdam Study	Albert Hofman	Erasmus MC	Epidemiology	Rotterdam	Zuid-Holland	NL
MARS	Susanne Lucae	Max Planck Institute of Psychiatry		Munich		DE
MARS	Stefan Kloiber	Max Planck Institute of Psychiatry		Munich		DE
MARS	Klaus Berger	University of Muenster	Institute of Epidemiology and Social Medicine	Muenster		DE
MARS	Jürgen Wellmann	University of Muenster	Institute of Epidemiology and Social Medicine	Muenster		DE
MARS	Bertram Müller-Myhsok	Munich Cluster for Systems Neurology (SyNergy)		Munich		DE
MARS	Bertram Müller-Myhsok	University of Liverpool		Liverpool		GB
MARS	Bertram Müller-Myhsok	Max Planck Institute of Psychiatry	Department of Translational Research in Psychiatry	Munich		DE
JANSSEN	Qingqin S Li	Janssen Research and Development, LLC	Neuroscience Therapeutic Area	Titusville	NJ	US
GSK_MUNICH	Bertram Müller-Myhsok	Munich Cluster for Systems Neurology (SyNergy)		Munich		DE
GSK_MUNICH	Bertram Müller-Myhsok	University of Liverpool		Liverpool		GB
GSK_MUNICH	Bertram Müller-Myhsok	Max Planck Institute of Psychiatry	Department of Translational Research in Psychiatry	Munich		DE
GSK_MUNICH	Marcus Ising	Max Planck Institute of Psychiatry		Munich		DE
GSK_MUNICH	Till F M Andlauer	Max Planck Institute of Psychiatry	Department of Translational Research in Psychiatry	Munich		DE
GSK_MUNICH	Till F M Andlauer	Munich Cluster for Systems Neurology (SyNergy)		Munich		DE
GSK_MUNICH	Susanne Lucae	Max Planck Institute of Psychiatry		Munich		DE
GSK_MUNICH	Stefan Kloiber	Max Planck Institute of Psychiatry		Munich		DE
BoMa	Marcella Rietschel	Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University	Department of Genetic Epidemiology in Psychiatry	Mannheim	Baden-Württemberg	DE
BoMa	Andreas J Forstner	University of Bonn	Institute of Human Genetics	Bonn		DE

BoMa	Andreas J Forstner	University of Bonn	Life&Brain Center, Department of Genomics	Bonn		DE
BoMa	Fabian Streit	Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University	Department of Genetic Epidemiology in Psychiatry	Mannheim	Baden Württemberg	DE
BoMa	Jana Strohmaier	Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University	Department of Genetic Epidemiology in Psychiatry	Mannheim		DE
BoMa	Maren Lang	Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University	Department of Genetic Epidemiology in Psychiatry	Mannheim	Baden-Württemberg	DE
BoMa	Josef Frank	Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University	Department of Genetic Epidemiology in Psychiatry	Mannheim	Baden-Württemberg	DE
BoMa	Stefan Herms	University of Basel	Division of Medical Genetics and Department of Biomedicine	Basel		CH
BoMa	Stefan Herms	University of Bonn	Institute of Human Genetics	Bonn	NRW	DE
BoMa	Stefan Herms	University of Bonn	Life&Brain Center, Department of Genomics	Bonn	NRW	DE
BoMa	Stephanie H Witt	Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University	Department of Genetic Epidemiology in Psychiatry	Mannheim		DE
BoMa	Jens Treutlein	Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University	Department of Genetic Epidemiology in Psychiatry	Mannheim		DE
BoMa	Markus M Nöthen	University of Bonn	Institute of Human Genetics	Bonn		DE
BoMa	Markus M Nöthen	University of Bonn	Life&Brain Center, Department of Genomics	Bonn		DE
BoMa	Sven Cichon	University of Basel	Department of Biomedicine	Basel		CH
BoMa	Sven Cichon	University of Basel	Division of Medical Genetics	Basel		CH
BoMa	Sven Cichon	Research Center Juelich	Institute of Neuroscience and Medicine (INM-1)	Juelich		DE
BoMa	Sven Cichon	University of Bonn	Institute of Human Genetics	Bonn		DE
BoMa	Franziska Degenhardt	University of Bonn	Institute of Human Genetics	Bonn		DE
BoMa	Franziska Degenhardt	University of Bonn	Life&Brain Center, Department of Genomics	Bonn		DE
BoMa	Per Hoffmann	University of Basel	Human Genomics Research Group, Department of Biomedicine	Basel		CH
BoMa	Per Hoffmann	University of Bonn	Institute of Human Genetics	Bonn		DE
BoMa	Per Hoffmann	University of Bonn	Life&Brain Center, Department of Genomics	Bonn		DE
BoMa	Thomas G Schulze	Medical Center of the University of Munich, Campus Innenstadt	Institute of Psychiatric Phenomics and Genomics (IPPG)	Munich	Bayern	DE
BoMa	Thomas G Schulze	University Medical Center Göttingen	Department of Psychiatry and Psychotherapy	Goettingen	Niedersachsen	DE
BoMa	Thomas G Schulze	Johns Hopkins University	Department of Psychiatry and Behavioral Sciences	Baltimore	MD	US
BoMa	Thomas G Schulze	NIMH Division of Intramural Research Programs	Human Genetics Branch	Bethesda	MD	US
BoMa	Thomas G Schulze	Central Institute of Mental Health	Department of Genetic Epidemiology in Psychiatry	Mannheim		DE
CoFaMS - Adelaide	Bernhard T Baune	University of Adelaide	Discipline of Psychiatry	Adelaide	SA	AU
CoFaMS - Adelaide	Udo Dannlowski	University of Marburg	Department of Psychiatry	Marburg	Hessen	DE
CoFaMS - Adelaide	Udo Dannlowski	University of Münster	Department of Psychiatry	Münster		DE
CoFaMS - Adelaide	Tracy M Air	University of Adelaide	Discipline of Psychiatry	Adelaide	SA	AU
CoFaMS - Adelaide	Grant C B Sinnamon	James Cook University	School of Medicine and Dentistry	Townsville	QLD	AU
CoFaMS - Adelaide	Naomi R Wray	The University of Queensland	Queensland Brain Institute	Brisbane	QLD	AU
EDINBURGH	Andrew M McIntosh	University of Edinburgh	Division of Psychiatry	Edinburgh	Edinburgh	GB
EDINBURGH	Douglas H R Blackwood	University of Edinburgh	Division of Psychiatry	Edinburgh		GB
EDINBURGH	Toni-Kim Clarke	University of Edinburgh	Division of Psychiatry	Edinburgh		GB
EDINBURGH	Donald J MacIntyre	University of Edinburgh	Division of Psychiatry	Edinburgh	Edinburgh	GB
GENSCOT	Andrew M McIntosh	University of Edinburgh	Division of Psychiatry	Edinburgh	Edinburgh	GB

GENSCOT	David J Porteous	University of Edinburgh	Medical Genetics Section, CGEM, IGMM	Edinburgh	Edinburgh	GB
GENSCOT	Toni-Kim Clarke	University of Edinburgh	Division of Psychiatry	Edinburgh		GB
GENSCOT	Donald J MacIntyre	University of Edinburgh	Division of Psychiatry	Edinburgh	Edinburgh	GB
GENSCOT	Caroline Hayward	University of Edinburgh	Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine	Edinburgh	Edinburgh	GB
EGCUT	Tõnu Esko	University of Tartu	Estonian Genome Center	Tartu		EE
EGCUT	Tõnu Esko	Broad Institute	Program in Medical and Population Genetics	Cambridge	MA	US
EGCUT	Tõnu Esko	Children's Hospital Boston	Division of Endocrinology	Boston	MA	US
EGCUT	Tõnu Esko	Harvard Medical School	Department of Genetics	Boston	MA	US
EGCUT	Evelin Mihailov	University of Tartu	Estonian Genome Center	Tartu		EE
EGCUT	Evelin Mihailov	Estonian Biocentre		Tartu	Tartumaa	EE
EGCUT	Lili Milani	University of Tartu	Estonian Genome Center	Tartu		EE
EGCUT	Andres Metspalu	University of Tartu	Estonian Genome Center	Tartu		EE
EGCUT	Andres Metspalu	University of Tartu	Institute of Molecular and Cell Biology	Tartu		EE
SHIP-LEGEND/TREND	Hans J Grabe	University Medicine Greifswald	Department of Psychiatry and Psychotherapy	Greifswald		DE
SHIP-LEGEND/TREND	Henry Völzke	University Medicine Greifswald	Institute for Community Medicine	Greifswald		DE
SHIP-LEGEND/TREND	Alexander Teumer	University Medicine Greifswald	Institute for Community Medicine	Greifswald		DE
SHIP-LEGEND/TREND	Sandra Van der Auwera	University Medicine Greifswald	Department of Psychiatry and Psychotherapy	Greifswald	Mecklenburg-Vorpommern	DE
SHIP-LEGEND/TREND	Georg Homuth	University Medicine and Ernst Moritz Arndt University Greifswald	Interfaculty Institute for Genetics and Functional Genomics, Department of Functional Genomics	Greifswald	Mecklenburg-Vorpommern	DE
SHIP-LEGEND/TREND	Matthias Nauck	University Medicine Greifswald	Institute of Clinical Chemistry and Laboratory Medicine	Greifswald	Mecklenburg-Vorpommern	DE
RADIANT	Cathryn M Lewis	King's College London	Department of Medical & Molecular Genetics	London	London	GB
RADIANT	Cathryn M Lewis	King's College London	MRC Social Genetic and Developmental Psychiatry Centre	London	London	GB
RADIANT	Gerome Breen	King's College London	NIHR BRC for Mental Health	London	London	GB
RADIANT	Gerome Breen	King's College London	MRC Social Genetic and Developmental Psychiatry Centre	London	London	GB
RADIANT	Margarita Rivera	University of Granada	Instituto de Investigación Biosanitaria ibs.Granada and CIBER en Salud Mental (CIBERSAM)	Granada	Granada	ES
RADIANT	Margarita Rivera	King´s College London	MRC Social Genetic and Developmental Psychiatry Centre	London	London	GB
RADIANT	Michael Gill	Trinity College Dublin	Department of Psychiatry	Dublin		IE
RADIANT	Nick Craddock	Cardiff University	Psychological Medicine	Cardiff	Cardiff	GB
RADIANT	John P Rice	Washington University in Saint Louis	Department of Psychiatry	Saint Louis	MO	US
RADIANT	Michael J Owen	Cardiff University School of Medicine	MRC Centre for Neuropsychiatric Genetics and Genomics	Cardiff	Cardiff	GB
RADIANT	Peter McGuffin	King's College London	MRC Social Genetic and Developmental Psychiatry Centre	London	London	GB
Danish Radiant	Henriette N Buttenschøn	Aarhus University	Department of Clinical Medicine, Translational Neuropsychiatry Unit	Aarhus		DK
Danish Radiant	Ole Mors	Aarhus University Hospital, Risskov	Research Department P	Aarhus	Central Denmark Region	DK
Danish replication	Henriette N Buttenschøn	Aarhus University	Department of Clinical Medicine, Translational Neuropsychiatry Unit	Aarhus		DK
Danish replication	Ole Mors	Aarhus University Hospital, Risskov	Research Department P	Aarhus	Central Denmark	DK

					Region	
Danish replication	Anders D Børglum	iPSYCH, Lundbeck Foundation Initiative for Integrative Psychiatric Research		Aarhus		DK
Danish replication	Anders D Børglum	Aarhus University	iSEQ, Centre for Integrative Sequencing, Department of Biomedicine	Aarhus		DK
Danish replication	Jakob Grove	iPSYCH, Lundbeck Foundation Initiative for Integrative Psychiatric Research		Aarhus		DK
Danish replication	Jakob Grove	Aarhus University	iSEQ, Centre for Integrative Sequencing	Aarhus		DK
Danish replication	Jakob Grove	Aarhus University	Bioinformatics Research Centre (BiRC)	Aarhus		DK
Danish replication	Jakob Grove	Aarhus University	Department of Biomedicine	Aarhus		DK
Danish replication	Jesper Krogh	University of Copenhagen	Deparment of Endocrinology at Herlev University Hospital	Copenhagen		DK
Roche MDD	Enrico Domenici	F. Hoffmann-La Roche Ltd	Roche Pharmaceutical Research and Early Development, Neuroscience, Ophthalmology and Rare Diseases Discovery & Translational Medicine Area, Roche Innovation Center Basel	Basel		CH
Roche MDD	Daniel Umbricht	F. Hoffmann-La Roche Ltd	Roche Pharmaceutical Research and Early Development, Neuroscience, Ophthalmology and Rare Diseases Discovery & Translational Medicine Area, Roche Innovation Center Basel	Basel		CH
Roche MDD	Jorge A Quiroz	F. Hoffmann-La Roche Ltd	Roche Pharmaceutical Research and Early Development, Neuroscience, Ophthalmology and Rare Diseases Discovery & Translational Medicine Area, Roche Innovation Center Basel	Basel		CH
Roche MDD	Carsten Horn	F. Hoffmann-La Roche Ltd	Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel	Basel		CH
QIMR	Enda M Byrne	The University of Queensland	Queensland Brain Institute	Brisbane	QLD	AU
QIMR	Baptiste Couvy-Duchesne	QIMR Berghofer Medical Research Institute	Genetics and Computational Biology	Herston	QLD	AU
QIMR	Baptiste Couvy-Duchesne	The University of Queensland	Centre for Advanced Imaging	Saint Lucia	QLD	AU
QIMR	Baptiste Couvy-Duchesne	The University of Queensland	School of Psychology	Saint Lucia	QLD	AU
QIMR	Scott D Gordon	QIMR Berghofer Medical Research Institute	Genetics and Computational Biology	Brisbane	Queensland	AU
QIMR	Andrew C Heath	Washington University in Saint Louis School of Medicine	Psychiatry	Saint Louis	MO	US
QIMR	Anjali K Henders	The University of Queensland	Queensland Brain Institute	Brisbane	QLD	AU
QIMR	IB Hickie	University of Sydney	Brain and Mind Research Institute	Sydney	NSW	AU
QIMR	Pamela AF Madden	Washington University in Saint Louis School of Medicine	Department of Psychiatry	Saint Louis	MO	US
QIMR	Nicholas G Martin	The University of Queensland	School of Psychology	Brisbane	QLD	AU
QIMR	Nicholas G Martin	QIMR Berghofer Medical Research Institute	Genetics and Computational Biology	Brisbane	Queensland	AU
QIMR	Sarah Elizabeth Medland	QIMR Berghofer Medical Research Institute	Genetics and Computational Biology	Herston	QLD	AU
QIMR	Grant W Montgomery	QIMR Berghofer Medical Research Institute	Genetics and Computational Biology	Brisbane	QLD	AU
QIMR	Dale R Nyholt	Queensland University of Technology	Institute of Health and Biomedical Innovation	Brisbane	QLD	AU
QIMR	Michele L Pergadia	Florida Atlantic University	Charles E. Schmidt College of Medicine	Boca Raton	FL	US
QIMR	Divya Mehta	The University of Queensland	Queensland Brain Institute	Brisbane	QLD	AU
QIMR	Naomi R Wray	The University of Queensland	Queensland Brain Institute	Brisbane	QLD	AU
PsyColaus	Martin Preisig	University Hospital of Lausanne	Department of Psychiatry	Prilly	Vaud	CH
PsyColaus	Enrique Castelao	University Hospital of Lausanne	Department of Psychiatry	Prilly	Vaud	CH
PsyColaus	Zoltán Kutalik	University Hospital of Lausanne	Institute of Social and Preventive Medicine (IUMSP)	Lausanne	VD	CH

PsyColaus	Zoltán Kutalik	Swiss Institute of Bioinformatics		Lausanne	VD	CH
STAR*D	Steven P Hamilton	Kaiser Permanente Northern California	Psychiatry	San Francisco	CA	US
GenPod/Newmeds	Katherine E Tansey	University of Bristol	MRC Integrative Epidemiology Unit	Bristol	Bristol	GB
GenPod/Newmeds	Rudolf Uher	Dalhousie University	Psychiatry	Halifax	NS	CA
GenPod/Newmeds	Glyn Lewis	University College London	Division of Psychiatry	London		GB
GenPod/Newmeds	Michael C O'Donovan	Cardiff University	MRC Centre for Neuropsychiatric Genetics and Genomics	Cardiff	Cardiff	GB
NESDA	Brenda WJH Penninx	VU University Medical Center and GGZ inGeest	Department of Psychiatry	Amsterdam		NL
NESDA	Yuri Milaneschi	VU University Medical Center and GGZ inGeest	Department of Psychiatry	Amsterdam		NL
NESDA	Wouter J Peyrot	VU University Medical Center and GGZ inGeest	Department of Psychiatry	Amsterdam		NL
NESDA	Johannes H Smit	VU University Medical Center and GGZ inGeest	Department of Psychiatry	Amsterdam		NL
NESDA	Rick Jansen	VU University Medical Center and GGZ inGeest	Department of Psychiatry	Amsterdam		NL
NESDA	Aartjan TF Beekman	VU University Medical Center and GGZ inGeest	Department of Psychiatry	Amsterdam		NL
NESDA	Robert Schoevers	University of Groningen, University Medical Center Groningen	Department of Psychiatry	Groningen	Groningen	NL
NESDA	Albert M van Hemert	Leiden University Medical Center	Department of Psychiatry	Leiden		NL
NESDA	Gerard van Grootheest	VU University Medical Center and GGZ inGeest	Department of Psychiatry	Amsterdam	Noord-Holland	NL
NTR	Dorret I Boomsma	VU University Amsterdam	Dept of Biological Psychology	Amsterdam		NL
NTR	Jouke- Jan Hottenga	VU University Amsterdam	Dept of Biological Psychology	Amsterdam		NL
NTR	Christel M Middeldorp	VU University Amsterdam	Dept of Biological Psychology	Amsterdam		NL
NTR	EJC de Geus	VU University Medical Center	EMGO+ Institute	Amsterdam		NL
NTR	EJC de Geus	VU University Amsterdam	Dept of Biological Psychology	Amsterdam		NL
NTR	Abdel Abdellaoui	VU University Amsterdam	Dept of Biological Psychology	Amsterdam		NL
NTR	Gonneke Willemsen	VU University Amsterdam	Dept of Biological Psychology	Amsterdam		NL
Harvard	Erin C Dunn	Broad Institute	Stanley Center for Psychiatric Research	Cambridge	MA	US
Harvard	Erin C Dunn	Massachusetts General Hospital	Department of Psychiatry	Boston	MA	US
Harvard	Erin C Dunn	Massachusetts General Hospital	Psychiatric and Neurodevelopmental Genetics Unit (PNGU)	Boston	MA	US
Harvard	Roy H Perlis	Harvard Medical School	Psychiatry	Boston	MA	US
Harvard	Roy H Perlis	Massachusetts General Hospital	Psychiatry	Boston	MA	US
Harvard	Jordan W Smoller	Massachusetts General Hospital	Department of Psychiatry	Boston	MA	US
Harvard	Jordan W Smoller	Massachusetts General Hospital	Psychiatric and Neurodevelopmental Genetics Unit (PNGU)	Boston	MA	US
Harvard	Jordan W Smoller	Broad Institute	Stanley Center for Psychiatric Research	Cambridge	MA	US
TwinGene	Patrik K Magnusson	Karolinska Institutet	Department of Medical Epidemiology and Biostatistics	Stockholm		SE
TwinGene	Nancy L Pedersen	Karolinska Institutet	Department of Medical Epidemiology and Biostatistics	Stockholm		SE
TwinGene	Alexander Viktorin	Karolinska Institutet	The Department of Medical Epidemiology and Biostatistics	Stockholm	Stockholm	SE
TwinGene	Erik Pettersson	Karolinska Institutet	Medical Epidemiology and Biostatistics	Stockholm		SE
DK Control	Thomas Werge	The Lundbeck Foundation Initiative for Psychiatric Research	iPSYCH	Copenhagen		DK
DK Control	Thomas Werge	University of Copenhagen	Institute of Clinical Medicine	Copenhagen		DK
DK Control	Thomas Werge	Mental Health Services Capital Region of Denmark	Institute of Biological Psychiatry, Mental Health Center Sct. Hans	Copenhagen		DK
DK Control	Thomas Hansen	Mental Health Services Capital Region of Denmark	Institute of Biological Psychiatry, Mental Health Center Sct. Hans	Copenhagen		DK
DK Control	Thomas Hansen	The Lundbeck Foundation Initiative for Psychiatric Research	iPSYCH	Copenhagen		DK

Pfizer	Sara A Paciga	Pfizer Global Research and Development	Human Genetics and Computational Biomedicine	Groton	CT	US
Pfizer	Hualin S Xi	Pfizer Global Research and Development	Computational Sciences Center of Emphasis	Cambridge	MA	US
Pfizer	Ashley R Winslow	Pfizer Global Research and Development	Human Genetics and Computational Medicine	Cambridge	MA	US
GenRED, GenRED2, DGN	Douglas F Levinson	Stanford University	Psychiatry & Behavioral Sciences	Stanford	CA	US
GenRED, GenRED2, DGN	Myrna M Weissman	New York State Psychiatric Institute	Division of Epidemiology	New York	NY	US
GenRED, GenRED2, DGN	Myrna M Weissman	Columbia University College of Physicians and Surgeons	Psychiatry	New York	NY	US
GenRED, GenRED2, DGN	James B Potash	University of Iowa	Psychiatry	Iowa City	IA	US
GenRED, GenRED2, DGN	Jianxin Shi	National Cancer Institute	Division of Cancer Epidemiology and Genetics	Bethesda	MD	US
GenRED, GenRED2	James A Knowles	University of Southern California	Psychiatry & The Behavioral Sciences	Los Angeles	CA	US