

CNV GDRC Data

Curated Data

selected_phenotypes.csv

This is a plain text file containing the phenotypes analyzed in the pre-print. This is a subset of all the phenotypes present in the data set. This maps between UKB field IDs and field names. The field IDs are used to label files in the rest of the data set.

Column	Description
field_id	UKB field ID
field_title	UKB field title
category_title	UKB category title
class_1	Always LoF
class_2	Always Dup
selection	Always All
gamma2_1_hat	Mean squared effect of LoF burden tests
gamma2_1_hat_se	SE of mean squared effect of LoF burden tests
gamma2_2_hat	Mean squared effect of duplication burden tests
gamma2_2_hat_se	SE of mean squared effect of duplication burden tests
M	Number of overlapping genes available for the trait

Summary Statistics

LoF_Burden_Tests.tar.gz

This is a tarball containing loss-of-function (LoF) variant burden test results. Each individual file is a gzip-format plain text file that contains the summary statistics for one trait. The file name is labelled using the UKB field ID of the trait that was analyzed.

The five masks are defined as follows:

1. M1.0.01 - All LoF variants with MAF < 1%
2. M2.0.01 - All LoF variants with MAF < 1% and misannotation proability < 10%
3. M3.0.01 - All LoF variants with MAF < 1% and misannotation probability < 5%
4. M4.0.01 - All LoF variants with MAF < 1% and misannotation probability < 1%
5. M5.0.01 - All synonymous variants with MAF < 1%

Column	Description
CHROM	Chromosome
GENPOS	Midpoint of the gene in GRCh38 coordinates
ID	Unique identifier for the association test using the gene name and mask that was used
ALLELE0	Always ref, standing for the functional copy
ALLELE1	Always the mask that was used to aggregate LoF variants
A1FREQ	Burden genotype frequency
N	Number of individuals used to perform the association test
TEST	Type of test, always additive
BETA	Effect size
SE	Standard error of effect size
CHISQ	Chi-square statistic for the effect size
LOG10P	Negative logarithm base 10 p-value
EXTRA	Always NA

LoF_Burden_Tests_Proteins.tar.gz

This is a tarball containing loss-of-function (LoF) variant burden test results for ~3000 proteins. Each individual file is a gzip-format plain text file that contains the summary statistics for one trait. The file name is labelled using the protein name of the trait that was analyzed.

The five masks are defined as follows:

- 1. M1.0.01 - All LoF variants with MAF < 1%
- 2. M2.0.01 - All LoF variants with MAF < 1% and misannotation proability < 10%
- 3. M3.0.01 - All LoF variants with MAF < 1% and misannotation probability < 5%
- 4. M4.0.01 - All LoF variants with MAF < 1% and misannotation probability < 1%
- 5. M5.0.01 - All synonymous variants with MAF < 1%

Column	Description
CHROM	Chromosome
GENPOS	Midpoint of the gene in GRCh38 coordinates
ID	Unique identifier for the association test using the gene name and mask that was used
ALLELE0	Always ref, standing for the functional copy
ALLELE1	Always the mask that was used to aggregate LoF variants
A1FREQ	Burden genotype frequency
N	Number of individuals used to perform the association test
TEST	Type of test, always additive
BETA	Effect size
SE	Standard error of effect size
CHISQ	Chi-square statistic for the effect size
LOG10P	Negative logarithm base 10 p-value
EXTRA	Always NA

LoF_Burden_Tests_Confounding.tar.gz

This is a tarball containing loss-of-function (LoF) variant burden test results in different subsets of individuals in the UKB. These subsets include unrelated individuals (unrelated/), individuals with genetic similarity to the EUR “superpopulation” (EUR/), and the set of WES individuals with different principal components (different_PCs/). Each individual file is a gzip-format plain text file that contains the summary statistics for one trait. The file name is labelled using the UKB field ID of the trait that was analyzed.

The five masks are defined as follows:

- 1. M1.0.01 - All LoF variants with MAF < 1%
- 2. M2.0.01 - All LoF variants with MAF < 1% and misannotation proability < 10%
- 3. M3.0.01 - All LoF variants with MAF < 1% and misannotation probability < 5%
- 4. M4.0.01 - All LoF variants with MAF < 1% and misannotation probability < 1%
- 5. M5.0.01 - All synonymous variants with MAF < 1%

Column	Description
CHROM	Chromosome
GENPOS	Midpoint of the gene in GRCh38 coordinates
ID	Unique identifier for the association test using the gene name and mask that was used
ALLELE0	Always ref, standing for the functional copy
ALLELE1	Always the mask that was used to aggregate LoF variants
A1FREQ	Burden genotype frequency
N	Number of individuals used to perform the association test
TEST	Type of test, always additive
BETA	Effect size
SE	Standard error of effect size

Column	Description
CHISQ	Chi-square statistic for the effect size
LOG10P	Negative logarithm base 10 p-value
EXTRA	Always NA

LoF_pQTL.tar.gz

This is a tarball containing loss-of-function (LoF) targeted protein quantitative trait locus (pQTL) mapping results. Each individual file is a gzip-format plain text file that contains the summary statistics for one trait. The file name is labelled using the protein name of the trait that was analyzed.

Column	Description
CHROM	Chromosome
GENPOS	Midpoint of the gene in GRCh38 coordinates
ID	Unique identifier for the association test using the gene name and mask that was used
ALLELE0	The non-LoF allele
ALLELE1	The LoF allele
A1FREQ	Genotype frequency
N	Number of individuals used to perform the association test
TEST	Type of test, always additive
BETA	Effect size
SE	Standard error of effect size
CHISQ	Chi-square statistic for the effect size
LOG10P	Negative logarithm base 10 p-value
EXTRA	Always NA

CNV_Burden_Tests.tar.gz

This is a tarball containing deletion, duplication, and potential loss-of-function (pLoF) variant burden test results. For descriptions of the variant classes, please refer to the pre-print. Summary statistics for each variant class are stored in separate files for each trait. The file name is labelled using the UKB field ID of the trait that was analyzed.

The duplication and deletion genotypes use the following masks:

1. Pad0KB.0.01 - Burden genotype frequency < 1% and 0 Kbp of flanking region surrounding the gene body
2. Pad1KB.0.01 - Burden genotype frequency < 1% and 1 Kbp of flanking region surrounding the gene body
3. Pad10KB.0.01 - Burden genotype frequency < 1% and 10 Kbp of flanking region surrounding the gene body

Column	Description
CHROM	Chromosome
GENPOS	Midpoint of the gene in GRCh38 coordinates
ID	Unique identifier for the association test using the gene name and mask that was used
ALLELE0	Always ref, standing for the functional copy
ALLELE1	Always the mask that was used to aggregate LoF variants
A1FREQ	Burden genotype frequency
N	Number of individuals used to perform the association test
TEST	Type of test, always additive
BETA	Effect size
SE	Standard error of effect size
CHISQ	Chi-square statistic for the effect size
LOG10P	Negative logarithm base 10 p-value
EXTRA	Always NA

CNV_Burden_Tests_Proteins.tar.gz

This is a tarball containing deletion, duplication, and potential loss-of-function (pLoF) variant burden test results for ~3000 proteins. For descriptions of the variant classes, please refer to the pre-print. Summary statistics for each variant class are stored in separate files for each trait. The file name is labelled using the UKB field ID of the trait that was analyzed.

The duplication and deletion genotypes use the following masks:

- 1. Pad0KB.0.01 - Burden genotype frequency < 1% and 0 Kbp of flanking region surrounding the gene body
- 2. Pad1KB.0.01 - Burden genotype frequency < 1% and 1 Kbp of flanking region surrounding the gene body
- 3. Pad10KB.0.01 - Burden genotype frequency < 1% and 10 Kbp of flanking region surrounding the gene body

Column	Description
CHROM	Chromosome
GENPOS	Midpoint of the gene in GRCh38 coordinates
ID	Unique identifier for the association test using the gene name and mask that was used
ALLELE0	Always ref, standing for the functional copy
ALLELE1	Always the mask that was used to aggregate LoF variants
A1FREQ	Burden genotype frequency
N	Number of individuals used to perform the association test
TEST	Type of test, always additive
BETA	Effect size
SE	Standard error of effect size
CHISQ	Chi-square statistic for the effect size
LOG10P	Negative logarithm base 10 p-value
EXTRA	Always NA

CNV_LD.tar.gz

This is a tarball containing signed R values (linkage disequilibrium) between duplication and deletion burden genotypes with the Pad1KB.0.01 mask.

A plain text file called `UKB_CNV_duplications_1kb_chr[chromosome].signed_R.csv` will contain the LD matrix of the chromosome for duplication burden genotypes.

A plain text file called `UKB_CNV_duplications_1kb_chr[chromosome].signed_R.genes.txt` will contain the genes that map to the LD matrix of the chromosome for the duplication burden genotypes.

A plain text file called `UKB_CNV_deletionss_1kb_chr[chromosome].signed_R.csv` will contain the LD matrix of the chromosome for deletion burden genotypes.

A plain text file called `UKB_CNV_deletions_1kb_chr[chromosome].signed_R.genes.txt` will contain the genes that map to the LD matrix of the chromosome for the deletion burden genotypes.

CNV_LD_Uncorrected.tar.gz

This is a tarball containing signed R values (linkage disequilibrium) between duplication burden genotypes with the Pad1KB.0.01 mask. This LD is not corrected for covariates. This was used for the simulations.

A plain text file called `UKB_CNV_duplications_1kb_chr[chromosome].signed_R.tsv` will contain the LD matrix of the chromosome for duplication burden genotypes.

A plain text file called `UKB_CNV_duplications_1kb_chr[chromosome].signed_R.vars.txt` will contain the genes that map to the LD matrix of the chromosome for the duplication burden genotypes.

UKB_CNV_duplications_1kb_chr3.signed_R.WES.csv.gz

This is a gzip-format file that contains the LD matrix for chromosome 3, estimated for all individuals with whole-exome sequencing data.

GWAS_Top_Hits.txt.gz

This is gzip-format file containing the conditionally independent top hits from genome-wide association studies.

Column	Description
TRT	Neale lab GWAS ID
chrombpID	SNP ID
rsID	refSNP ID
EA	Effect allele
EAF	Effect allele frequency
EABeta	Allele effect size
DA	Derived allele
DAF	Derived allele frequency
DABeta	Derived allele effect size

total_del_length_ecdf.csv

The empirical CDF of deletion lengths in the UKB DRAGEN copy number data.

total_dup_length_ecdf.csv

The empirical CDF of duplication lengths in the UKB DRAGEN copy number data.

RSS MASH

RSS_MASH_Samples.tar.gz

This is a tarball containing the inferred prior mixture weights from stochastic approximation expectation maximization. It also contains posterior samples from the empirical Bayes inference for burden effect sizes.

A plain text file called `p[UKB field ID].rss_mash_mcem.csv` contains the inferred mixture weights for the 25 mixture components used in the model for the specified trait.

A plain text file called `p[UKB field ID].rss_mash_samples.csv` contains 2000 samples for the LoF and duplication burden effects for the specified trait.

Analyses

mom_simulations.csv

This is a plain text file containing inference results from our fixed-effect and random-effect simulations. Inference was performed using our unbiased and method-of-moments estimators.

Column	Description
experiment	The value being perturbed in the simulation
iteration	The simulation iteration number
gamma_bar_1	The mean LoF burden effect size
gamma_bar_2	The mean duplication burden effect size
sigma2_11	The variance of the LoF burden effect size
sigma_12	The covariance of the burden effect sizes
sigma2_22	The variance of the duplication burden effect size
rho	The correlation between effect sizes
phi	The monotonicity
phi_hat	The estimated monotonicity using a method-of-moments estimator
phi_hat_se	The standard error

Column	Description
<code>gamma_bar_1_hat</code>	Estimated average burden effect
<code>gamma_bar_1_hat_se</code>	The standard error
<code>gamma_bar_2_hat</code>	Estimated average burden effect
<code>gamma_bar_2_hat_se</code>	The standard error
<code>sigma2_11_hat</code>	Estimated variance
<code>sigma2_11_hat_se</code>	SE of estimated variance
<code>sigma_12_hat</code>	Estimated covariance
<code>sigma2_22_hat</code>	Estimated variance
<code>sigma2_22_hat_se</code>	SE of estimated variance

mse_simulations.csv

This is a plain text file containing inference results from our fixed-effect simulations. Inference was performed using our unbiased mean squared error estimator.

Column	Description
<code>experiment</code>	The value being perturbed in the simulation
<code>iteration</code>	The simulation iteration number
<code>gamma_bar_sq_1</code>	The mean squared effect size for LoF variants
<code>gamma_bar_sq_2</code>	The mean squared effect size for duplication variants
<code>gamma2_1_hat</code>	The estimated mean sqquared effect for LoF variants
<code>gamma2_1_hat_se</code>	The standard error
<code>gamma2_2_hat</code>	The estimated mean sqquared effect for duplication variants
<code>gamma2_2_hat_se</code>	The standard error

ngd_simulations.csv

This is a plain text file containing inference results from our fixed-effect and random-effect simulations. Inference was performed using our natural gradient ascent approach for maximum likelihood estimation.

Column	Description
<code>experiment</code>	The value being perturbed in the simulation
<code>iteration</code>	The simulation iteration number
<code>gamma_bar_1</code>	The mean LoF burden effect size
<code>gamma_bar_2</code>	The mean duplication burden effect size
<code>sigma2_11</code>	The variance of the LoF burden effect size
<code>sigma_12</code>	The covariance of the burden effect sizes
<code>sigma2_22</code>	The variance of the duplication burden effect size
<code>rho</code>	The correlation between effect sizes
<code>phi</code>	The monotonicity
<code>phi_hat</code>	MLE estimator for monotonicity
<code>phi_prime_hat</code>	MLE estimator in transformed space
<code>phi_prime_hat_se</code>	The standard error
<code>gamma_bar_1_hat</code>	Estimated average burden effect
<code>gamma_bar_1_hat_se</code>	The standard error
<code>gamma_bar_2_hat</code>	Estimated average burden effect
<code>gamma_bar_2_hat_se</code>	The standard error
<code>sigma2_11_hat</code>	Estimated variance
<code>sigma_12_hat</code>	Estimated covariance

Column	Description
sigma2_22_hat	Estimated variance
ll	Log-likelihood at the MLE
llr	Log-likelihood ratio of the last and penultimate iteration

mom_estimates.csv

This is a plain text file containing unbiased estimates of the average burden effect and method-of-moments estimates of the monotonicity for traits.

Column	Description
field_id	UKB field ID
class_1	Type of variants labelled 1 in the rest of the columns
class_2	Type of variants labelled 2 in teh rest of the columns
selection	Set of genes used to estimate the quantity (All = all genes, Extreme/Strong/Weak = selection buckets)
phi_hat	Estimated monotonicity
phi_hat_se	Standard error
gamma_bar_1_hat	Estimated average burden effect
gamma_bar_1_hat_se	Standard error
gamma_bar_2_hat	Estimated average burden effect
gamma_bar_2_hat_se	Standard error
sigma2_11_hat	Estimated variance
sigma2_11_hat_se	SE of estimated variance
sigma_12_hat	Estimated covariance
sigma2_22_hat	Estimated variance
sigma2_22_hat_se	SE of estimated variance

sq_eff_estimates.csv

This is a plain text file containing unbiased estimates of the average squared burden effects.

Column	Description
field_id	UKB field ID
class_1	Type of variants labelled 1 in the rest of the columns
class_2	Type of variants labelled 2 in teh rest of the columns
selection	Always All
gamma2_1_hat	Estimated average squared burden effect
gamma2_1_hat_se	Standard error
gamma2_2_hat	Estimated average squared burden effect
gamma2_2_hat) se	Standard error

monotonicity.csv

This is a plain text file containing estimates of the monotonicity for traits.

Column	Description
field_id	UKB field ID
class_1	Type of variants labelled 1 in the rest of the columns
class_2	Type of variants labelled 2 in the rest of the columns
phi_hat	Estimated monotonicity
phi_prime_hat	Estimated monotonicity in transformed space

Column	Description
phi_prime_hat_se	Standard error
gamma_bar_1_hat	Estimated average burden effect
gamma_bar_1_hat_se	Standard error
gamma_bar_2_hat	Estimated average burden effect
gamma_bar_2_hat_se	Standard error
sigma2_11_hat	Estimated variance
sigma_12_hat	Estimated covariance
sigma2_22_hat	Estimated variance
ll	Log likelihood at the end of natural gradient ascent
llr	Log likelihood ratio between the last and penultimate iteration of natural gradient ascent

xi_mash.csv

This is a plain text file containing estimates of trait buffering for traits.

Column	Description
field	UKB field ID
tau	LFSR < tau bin
xi_mean	Posterior mean trait buffering
xi_low	Lower bound of 95% credible interval
xi_high	Upper bound of 95% credible interval
xi_lfsr	Local false sign rate
xi_nm_comp_mean	Posterior component of trait buffering from non-monotone GDRCs
xi_nm_comp_low	Lower bound of 95% credible interval
xi_nm_comp_high	Upper bound of 95% credible interval
xi_nm_comp_lfsr	Local false sign rate
xi_m_comp_mean	Posterior component of trait buffering from monotone GDRCs
xi_m_comp_low	Lower bound of 95% credible interval
xi_m_comp_high	Upper bound of 95% credible interval
xi_m_comp_lfsr	Local false sign rate
m_enrichment_mean	Posterior enrichment of monotone curves
m_enrichment_low	
m_enrichment_high	Upper bound of 95% credible interval

genetic_correlation.csv

This is a plain text file containing estimates of genetic correlation between loss-of-function burden tests and deletion burden tests.

Column	Description
field	UKB field ID
class_1	Type of variants labelled 1 in the rest of the columns
class_2	Type of variants labelled 2 in the rest of the columns
corr_hat	Estimated genetic correlation
corr_prime_hat	Estimated genetic correlation in transformed space
corr_prime_hat_se	Standard error
sigma2_11_hat	Estimated variance
sigma_12_hat	Estimated covariance
sigma2_22_hat	Estimated variance
ll	Log likelihood at the MLE

Column	Description
llr	Log likelihood ratio between the last and penultimate iteration of natural gradient ascent