RVAT: a unified framework to discover & interpret rare variant associations in large DNA sequencing datasets

Paul J. Hop[1,2], Kees de Jong[1,2], Tessa A. Zonneveld[3], Maarten Kooyman[2], Brendan J. Kenna[2], Kevin P. Kenna[1]

1. Department of Translational Neuroscience, UMC Utrecht Brain Center, University Medical Center Utrecht, 3584 CG Utrecht, the Netherlands.
2. Department of Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht, 3584 CX Utrecht, the Netherlands.
3. Genetic Epidemiology, Department of Psychiatry, Amsterdam UMC, location University of Amsterdam, Amsterdam, the Netherlands.

Contact:
P.J.Hop-2@umcutrecht.nl
K.P.Kenna@umcutrecht.nl

Abstract

The proliferation of whole-genome sequencing has transformed our ability to study how rare variants contribute to health and disease. This creates new opportunities to map disease modifying genes, resolve variants of unknown significance and to discover the aggregate effects of hidden rare variant associations on biological pathways and cell types. With this, there is an increasing need for accessible user-friendly data infrastructures and software tools that efficiently store, query, analyze and interpret these data. We developed RVAT (Rare Variant Association Toolkit) as a one-stop solution to address these needs and perform a comprehensive and customizable range of rare variant analyses and visualizations. RVAT is embedded in the Bioconductor ecosystem and uses a compressed out-of-memory data structure based on SQLite to facilitate efficient integration of large sequencing datasets with variant and sample annotations. The file format is complemented by object types and functions that support single variant, gene level, gene partitioning and gene set analyses through both R and command-line interfaces. We demonstrate the utility of RVAT in bridging the gap between the discovery and interpretation of rare variant associations using case studies wherein we recover mutation hotspots linked to amyotrophic lateral sclerosis (ALS) and reveal biologically

relevant gene sets and cell-types associated with health-related traits in UK biobank sequencing data.

Introduction

Recent years have seen a rapid rise in the number, size and applications of whole-genome (WGS) and whole-exome (WXS) datasets. While the identification of disease-causing rare variants used to be confined to linkage analysis in Mendelian disorders, the advent of these sequencing datasets has enabled the systematic identification of rare variants in Mendelian and non-Mendelian disorders alike. These include a wide range of diseases such as bipolar disease, diabetes, breast cancer, Alzheimer's disease and Crohn's disease[1–5] as well as health-related and molecular traits such as BMI, smoking, lipid levels and protein levels[6–9]. Similarly, our group has focused on using rare variant analyses to discover susceptibility genes for neurodegenerative disorders including ALS and Parkinson's disease[10–13], work which has led us to pinpoint and address key challenges inherent in identifying rare variants in large sequencing datasets.

Though rare variant analyses share many concepts and challenges with common variant GWAS, there are notable distinctions to consider. The first is having to handle and analyze a considerably larger number of variants, as the vast majority of human variants are rare[14]. This not only presents challenges in terms of managing and processing larger volumes of data but also necessitates tailored analytical approaches. Secondly, while single variant tests conventionally performed in GWAS have also proven useful in the context of certain rare variants, they often lack the statistical power required to identify associations among the rarest variants such as singletons. Therefore, single variant tests are typically complemented or replaced by gene- or region-based tests in which variants are tested jointly across genes or other functional units of interest[15]. A key aspect of these tests is prioritizing so-called "qualifying variants"[15]. This prioritization process increases power to discover disease associations by filtering out benign genetic variants and technical artifacts through the use of an ever expanding array of variant effect predictions (VEP), quality control metrics and minor allele frequency (MAF) thresholds[15].

Together, several challenges therefore emerge, including the management and querying of large sequencing datasets (often terabyte-sized), the integration of complex variant and sample annotations, performing a variety of rare variant tests and downstream analyses, all typically necessitating the unifying of disparate data formats and software tools. Moreover, a significant challenge lies in the interpretation of rare variant signals including fine-mapping gene-based associations, resolving variants of unknown significance (VUS) as well as disentangling the contribution of rare variants beyond individual functional units, such as biological pathways and cell types.

In this manuscript we describe how RVAT was designed to mitigate these challenges and provide a low learning curve and accessible interface that supports a wide range of rare variant analyses on both compute clusters and local computers. Also central to the RVAT framework are novel features focused on the fine-mapping and interpretation of rare variant signals. These include both supervised and unsupervised methods to identify and visualize mutation hotspots and a comprehensive suite of rare variant gene set analyses. We illustrate these features through case studies in which we pinpoint mutation hotspots in amyotrophic lateral sclerosis (ALS) and uncover relevant biology in several health-related traits through rare variant gene set analyses in the UK biobank.

Methods

Case study 1

Data assembly, processing, and quality control was performed as described in Hop *et al.*[13]. Variants were annotated using snpEff[16], dbscSNV[17] and Ensembl Release 105 gene models[18]. Variants were classified as loss of function (LOF) when predicted by snpEff to have a high impact (including nonsense mutations, splice acceptor/donors and frameshift mutations) or predicted as potentially splice-altering by dbscSNV ('ada' or 'rf' score > 0.7). Variants were classified as having moderate impact when predicted as such by snpEff (including missense mutations, inframe deletions and UTR truncations).

For the domain-based analyses, protein coordinates for Interpro domains, coiled coils, transmembrane helices, low complexity regions, and cleavage sites were retrieved from Ensembl version 105 (http://dec2021.archive.ensembl.org/biomart/martview/)[18]. For each transcript, variants were annotated to domains by remapping both the domain coordinates and variant positions to coding sequence (CDS) relative coordinates using the *mapToCDS* method. Variants up to 12bp from the coding sequence border (introns and UTRs) were mapped to the respective border (*exonPadding* = 12). To generate spatial clusters, the *spatialClust* method was applied to the CDS-relative positions. We used a sliding window step of 30 variants and an overlap of 15 variants as parameters for the clustering algorithm. Gene-based variant sets were generated using the *buildVarSet* method.

Region-based burden tests (across the gene, domains, or spatial clusters) were performed using firth logistic regression, testing for an association between case-control status and the total number of minor alleles per sample per gene (burden score). Sex, ten principal components, and the total number of qualifying synonymous variants in each individual were included as covariates. For the gene-based tests, we additionally performed SKAT (robust version, 'skat_robust' in RVAT) and ACAT-v (SPA-corrected, 'acatvSPA' in RVAT)[19,20].

Case study 2

Exome test-statistics for the following phenotypes were downloaded from Genebass[21] (https://genebass.org/): bone mineral density (phenocode = bone_mineral_density_custom), white matter integrity of tapetum (phenocode = 25439), osteoporosis (phenocode = 131964), LDL (phenocode = 30780) and red blood cellcounts (RBC; phenocode = 30010). We focused on the pLoF SKAT-O results.

Gene set analysis. Ontology gene sets (C5) were downloaded from MSigDb (https://www.gsea-msigdb.org/gsea/msigdb)[22]. Gene sets were imported into the RVAT *geneSetFile* format using the *buildGeneSet* method. For each phenotype, we performed competitive GSA (one-sided tests) using the *geneSetAssoc* method, adjusting for total

CDS length and the number of variants in the gene. Results were visualized using the *densityPlot* method.

Cell-type enrichment analysis. Processed single-cell RNA sequencing as used in the FUMA web app[23] were downloaded from: https://github.com/Kyoko-wtnb/FUMA_scRNA_data. For each phenotype and single-cell RNA sequencing dataset, we performed cell-type enrichment analyses (one-sided tests) using the *geneSetAssoc* method, adjusting for the average expression across cell-types in addition to CDS length and the number of variants in the gene. Expression values were limited to 10 standard deviations from the mean.

Results

gdb format: efficient storage and retrieval of genotypes and annotations

RVAT efficiently stores large sequencing datasets, along with related sample and variant annotations, in a novel file format named *gdb*. The *gdb* format is built upon the widely-used SQLite database engine (Figure 1). A *gdb* can be generated directly from a VCF file using the *buildGdb* method, allowing parallelization by processing the VCF file in chunks. Each variant is assigned a unique indexed identifier ('VAR_id') to facilitate rapid queries, while sample genotypes are compressed per variant for efficient storage. As shown in more detail in the 'Performance' section, the resulting gdb-file is approximately 80 times smaller than a gz-compressed VCF file. Being a fully relational database, the gdb seamlessly integrates the genotype data with a wide array of annotations (e.g. variant effect predictions and sample phenotypes), whilst efficiently performing complex queries, including combinations of table joins, filtering, and grouping operations. These operations are facilitated through easy-to-use R methods such as *uploadAnno*, *getAnno*, *getCohort* and *getGT*, while advanced users can also directly execute SQL queries either via the command-line or through R SQL interfaces like RSQLite and dbplyr. Using the *subsetGdb* method, users can export subsets of a gdb, while retaining all links between genotype data and imported annotations. This simplifies the extraction of prioritized genes or

genomic regions for data sharing or subsequent in-depth exploratory analyses on a local machine. Together, the *gdb* format conveniently facilitates both large-scale analyses (i.e. on a high-performance cluster) as well as targeted, interactive queries.

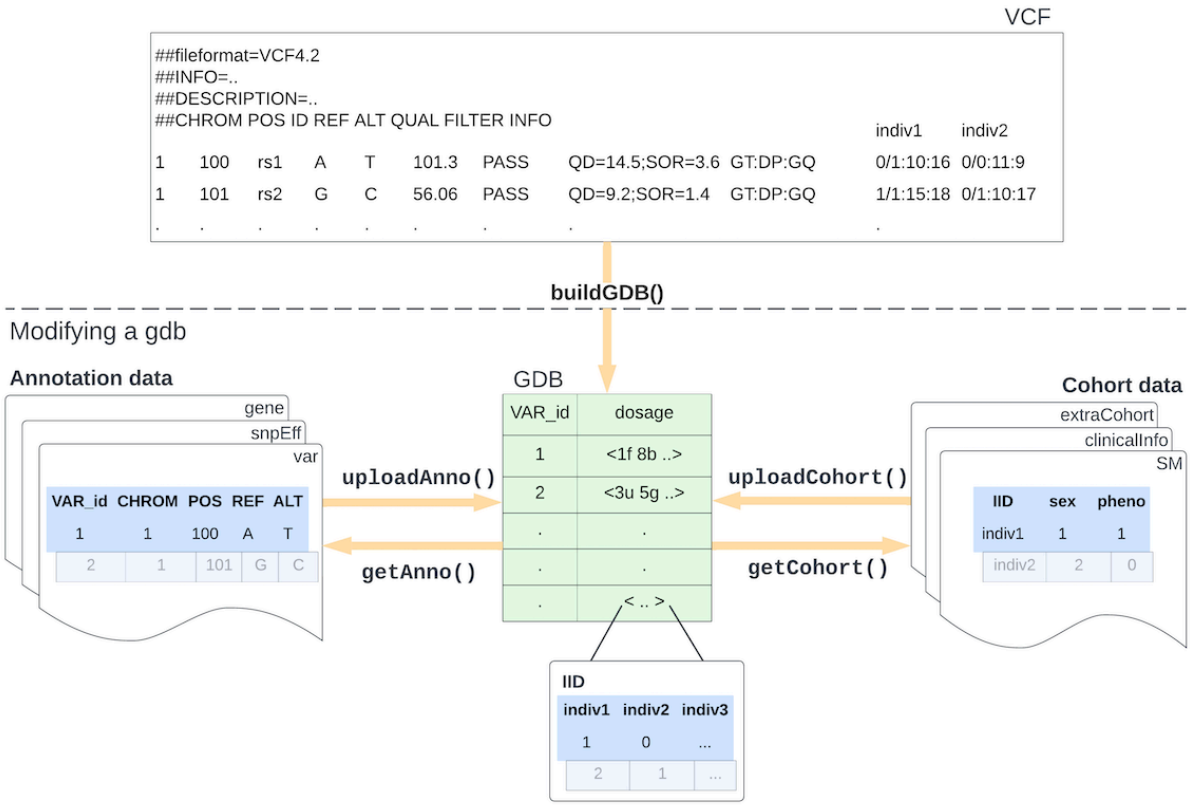

Figure 1. Schematic representation of the *gdb* file format. The gdb file can be created directly from a vcf file using the *buildGdb* method. Genotypes are compressed per-variant, and each variant is assigned a unique indexed identifier ('VAR_id') to enable rapid queries. Integration of sample and variant annotations (such as variant effect predictions and sample phenotypes) are easily integrated in the gdb using the uploadAnno/uploadCohort methods, while the getCohort/getAnno/getGT methods allow for complex queries including table joins, filtering, and grouping. The *gdb*-file format is complemented by an extensive range of rare single variant, region, gene and gene set analyses and visualizations available through the RVAT R and command-line interfaces.

genoMatrix class: intuitive handling of genotypes and annotations in R

Sequencing datasets often greatly exceed available computer memory, yet researchers frequently need to extract specific variants, regions, or genes of interest for further exploration, analysis, or visualization. To this end, RVAT allows rapid and easy access to genotypes and annotations of interest within the widely used R programming language, in which users can leverage both specialized RVAT methods described in this paper as well as the wider visualization and statistical capabilities of R. The RVAT *genoMatrix* class represent relevant subsets of the *gdb* in R (Figure 2), and is built upon the BioConductor *SummarizedExperiment* class[24] that is specifically designed to store and manage rectangular biological data along with sample- and feature-level metadata in a synchronized single instance. By using this and other standard BioConductor classes and methods, RVAT ensures an intuitive user experience and interoperability with other BioConductor packages. Genotypes can be retrieved from a *gdb* using the *getGT* method which returns variant, sample and genotype data as a single *genoMatrix* object. Users can specify the genome build in order to correctly code the genotypes in the non-pseudoautosomal regions of the sex chromosomes. A wide array of operations can be performed on a *genoMatrix* object, targeting specific cells, variants, samples, or entire tables. These include operations like variant and sample filtering, genotype masking, calculating minor allele frequencies, conducting single variant and aggregate association tests, generating burden scores, retrieving variant carriers and summarizing metrics such as call rate or carrier frequency across different groups.

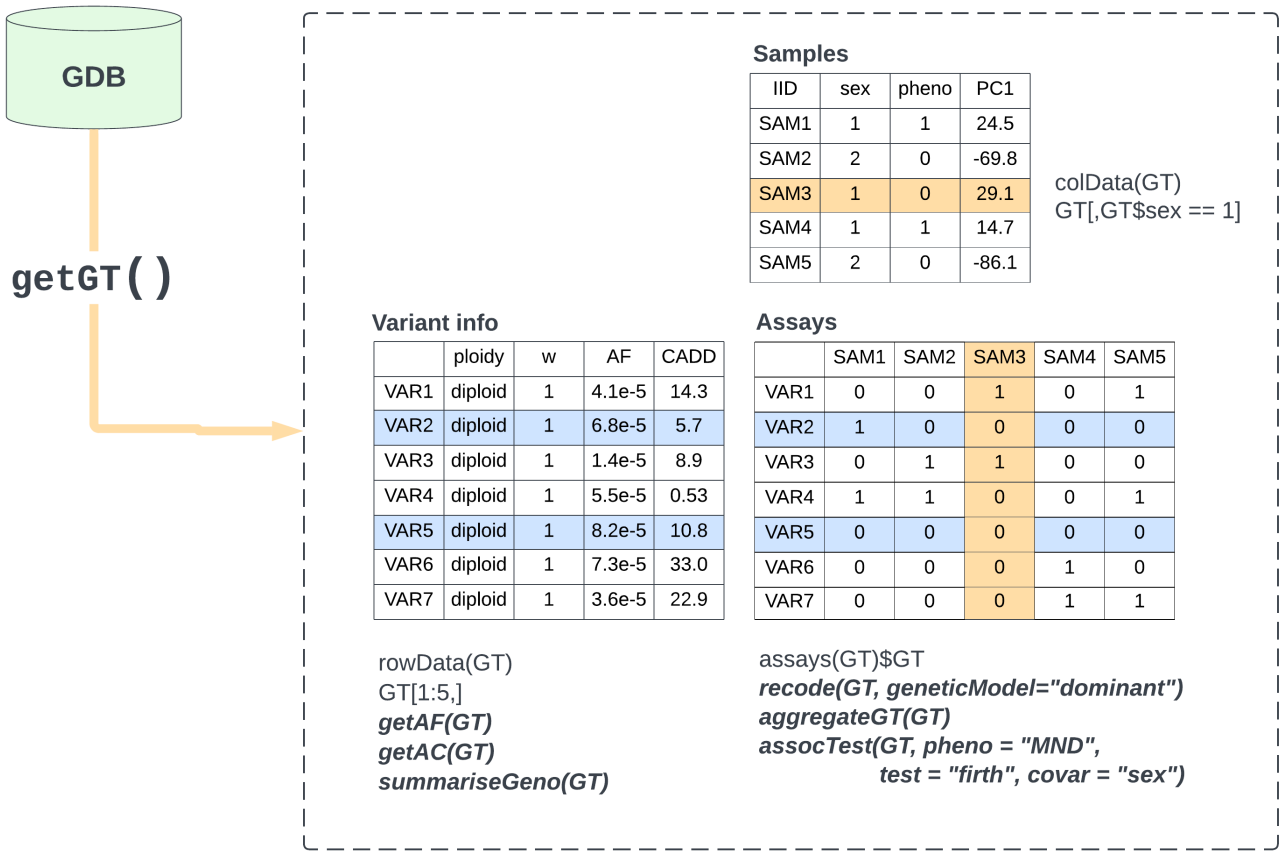

Figure 2. Schematic representation of the *genoMatrix* class. The *genoMatrix* class extends the BioConductor *SummarizedExperiment* class. Code snippets in bold represent methods specific to the *genoMatrix* class, code snippets in roman represent standard *SummarizedExperiment* methods that are inherited by the *genoMatrix* class. The rows of the *genoMatrix* class represent variants and the columns represent samples. Variant info is accessible through the *rowData* method and by default includes variant ploidy, variant weights and allele frequencies (automatically updated when samples are subsetted), along with user-supplied annotations such as variant pathogenicity scores (CADD in this example). Sample info is accessible through the *colData* method, which includes the mandatory sex information along with user-supplied sample phenotypes. A wide array of operations can be performed on a *genoMatrix* object, including operations like variant and sample filtering, genotype masking and imputation, conducting single variant and aggregate association tests, generating burden scores, and summarizing metrics such as call rate or carrier frequency across different groups.

varSets: define variant sets based on genomic annotations or unsupervised clustering

A key feature of RVAT is that it provides a unified interface for various types of aggregate variant tests. The first step in such analyses consists of defining sets of variants passing user-defined filters (e.g. based on functional consequence or minor allele frequency) across genomic regions of interest (e.g. genes, protein domains or clusters of variants). Consequently, this step typically involves the tedious process of processing, parsing, combining multiple sources of genomic annotations and variant effect predictions (VEP) RVAT reduces all this into an intuitive and fast workflow. First, the *mapVariants* method efficiently maps variants in the *gdb* onto genomic features of interest provided in standard formats such as bed, gff and gtf (used by e.g. Ensembl and RefSeq). Similarly, variant effect predictions can be incorporated in the *gdb* and the RVAT website includes tutorials on generating and integrating established tools such as snpEff, dbNSFP and AlphaMissense[16,25,26]. Second, the *buildVarSet* method then generates variant sets based on these feature and annotation tables, which can be weighted by features such as MAF and predicted pathogenicity. In addition to accommodating grouping variants by genomic features, RVAT also allows for partitioning the genome using the *spatialClust* method that implements an unsupervised spatial clustering algorithm reported by Loehlein-Fier *et al.*[27]. Variant sets are stored in the RVAT *varSet* format and collections of varSets (e.g. all genes, various filters) can be stored on-disk in the *varSetFile* format. This avoids the need to hold a large number of variant sets in memory, enables their reuse for subsequent analyses or queries, and enhances reproducibility. Genes and/or annotations of interest can be easily retrieved from a *varSetFile* using the *getVarSet* method. Variant sets can then be passed on to the *getGT* method to load the genotypes for the variant set(s) of interest or can be used as input for downstream analyses such as association tests.

assocTest: a unified interface for single variant and aggregate association testing

The *assocTest* method provides a unified and flexible interface to perform both single and aggregate variant tests. It implements a comprehensive set of statistical tests including widely used methods such as firth-based logistic regression (burden), the SKAT variance component test, the SKAT-O omnibus test and the Cauchy distribution based ACAT-v method[19,20,28,29]. Moreover, it supports different genetic models (i.e. additive, recessive and dominant) and allows for variant weighting based on annotations or minor allele frequency. In addition to existing statistical tests, a re-implementation of ACAT-V is included that is robust to case-control imbalance through either saddle-point-approximation (SPA) or firth correction[30,31]. Furthermore, we implemented a fast resampling procedure to calculate permutation-based empirical *P*-values for ACAT-v as well as other tests. Association tests can be performed either interactively on a *genoMatrix* in an interactive R session, or from the command-line allowing for (parallelized) iteration through multiple variant sets. Results from these analyses are stored in the *rvatResult* format, which is an extension of the BioConductor *DataFrame* class. Test-statistics in an *rvatResult* object can be visualized using (labeled) manhattan, qq- and forest plots. Finally, *P*-values in an *rvatResult* can be combined through a flexible implementation of the ACAT method[20], which allows users to combine *P*-values across, for example, complementary statistical tests, MAF bins, annotations or genomic features (e.g. combine transcript test-statistics into one *P*-value per gene).

geneSetAssoc: rare variant gene set and cell-type enrichment analyses

Gene set and cell-type enrichment analyses can aid in uncovering relevant biological mechanisms beyond what can be identified among single variants or genes.
RVAT implements a comprehensive set of rare variant gene set analysis methods (GSA), supported by an infrastructure to import and manage collections of gene sets.

Importing and managing gene sets. Gene sets can be imported into R from the GMT format that is used in The Molecular Signatures Database (MSigDB)[22], one of the most

widely used repositories that includes tens of thousands of gene sets including, among others, ontology (e.g. GO, HPO), oncogenic and cell-type signature gene sets. Alternatively, users can import custom datasets including user-curated datasets such as gene co-expression modules defined through WGCNA analyses of custom RNAseq datasets, lists of known disease genes or even complete single cell gene expression matrices for use in cell type enrichment analyses. After importing the gene sets, they can be easily managed, stored and retrieved using the *geneSet* format, analogous to the *varSet* format discussed earlier.

Broadly, the implemented GSA methods can be divided into competitive and self-contained tests, where the former tests whether genes *in* the gene set are more associated with the phenotype than genes *outside* the gene set, while the latter jointly tests whether genes *in* the gene set are associated with the phenotype without considering genes outside the set[32]. Consequently, through testing for an enrichment relative to the genes outside the gene set, competitive GSA controls for polygenicity as well as biases such as confounding and technical variability. Self-contained tests, on the other hand, are generally more powerful but may result in inflated test-statistics in case of polygenicity or residual biases in the data.

Competitive GSA. The *geneSetAssoc* method implements two types of competitive GSA methods which can be performed directly on results stored in an *rvatResult*. First, significant genes can be tested for overrepresentation among gene sets using Fisher's exact test, where non-significant genes are included as the background. The drawback of this method is that it relies on a *P*-value threshold, and therefore reduces the quantitative information contained in the test-statistics to a binary variable (significant or not). As a result, enrichment analyses of this kind fail to detect subtle effects across a larger number of genes where each individual gene fails to reach genome-wide significance by itself. Therefore, we also implemented a competitive method that doesn't rely on a *P*-value threshold, but instead directly tests for an association between the strength of association (inverse normal-transformed *P*-values) and pathway membership using linear regression. Such an approach is commonly called functional class scoring (FCS) in literature, and our approach is similar to that implemented in MAGMA for

GWAS[33,34]. Covariates can be included to account for potential confounding factors such as gene size and the number of variants. Potential gene-gene correlations can be accounted for using mixed linear models[35], in which the gene correlations can be estimated through correlation among burden scores or using a permutation approach. Additionally, cell-type enrichment analyses are supported in the same framework, testing for an association between test-statistics and gene expression values in cell-types of interest.

Self-contained GSA. Self-contained gene set analyses have been performed in several recent large whole-exome studies in the form of gene set burden analyses[1,36,37]. Gene set burden analyses extend the rationale behind gene burden tests to sets of genes, aggregating variants across gene sets rather than single genes. These types of analyses can be computationally demanding since the number of variants to aggregate, especially for large gene sets, is often many times larger than in single gene analyses. To address this, we implemented a two-step approach in RVAT. In the first step, burden scores are generated for each gene and stored in a compressed format. In the second step, gene set analyses are performed by aggregating the gene burden scores for each gene set, followed by testing for an association between the gene set burden score and the phenotype of interest.

Together, RVAT facilitates convenient import and managing of gene sets, and provides a wide range of GSA methods tailored to rare variant analyses.

Case studies

Case study 1: Recovering mutation hotspots in ALS

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease with a substantial genetic component[38,39]. Rare variants in over 30 genes have been linked to ALS, highlighting the genetic heterogeneity of the disease. In some ALS-linked genes, such as *SOD1,* pathogenic variants have been observed across the entire length of the gene, while in others, such as *FUS* and *TARDBP*, pathogenic variants appear to be concentrated in specific portions of the gene[40]. Identifying such mutation hotspots is highly relevant for unraveling pathophysiological mechanisms, improving genetic diagnostics and therapeutic development. Here we showcase how RVAT facilitates the identification of mutation hotspots using both a functional domain-based approach and an unsupervised spatial clustering approach.

For this case study, we combined WGS data of 9,600 individuals included in Project MinE with WXS data of 50,000 individuals included in the UK Biobank. Processing and quality control of the data was performed using an RVAT workflow that includes strict sample- and variant-level variant control and a tailored approach to eliminate batch effects in aggregated sequencing data (details of the workflow are described in a recent publication from our group[13]). This resulted in a total of 6,436 ALS cases and 48,436 controls after quality control (see Methods)[41–43]. We focused our inquiry on two known ALS genes: *SOD1* and *FUS*. We employed two distinct approaches to identify mutation hotspots in these genes: a supervised approach in which we grouped variants according to overlap with functional protein domains, and an unsupervised approach that clustered variants using a spatial clustering algorithm[27].

For both approaches, we first remapped the variants to coding sequence coordinates (CDS) using the *mapToCDS* RVAT method. Intronic variants within 12 base pairs of an exon border were mapped to the corresponding border. For the domain-based approach we obtained coordinates for Interpro domains, coiled coils, transmembrane helices, low complexity regions, and cleavage sites from Ensembl (v. 105), and mapped variants to the domains based on the CDS coordinates obtained previously. In the spatial clustering approach, consecutive groups of variants that colocalize along the coding sequence using

the *spatialClust* method. Both approaches result in variant sets stored in the *varSet* format, either for each cluster, or each domain. Subsequently, these variant sets were used as input for the *assocTest* method to test for an excess of rare (MAF<0.001) non-synonymous variants using firth logistic regression.

In *FUS*, we observed a strong enrichment towards the C-terminus of the coding sequence (Figure 3). Specifically, the 3'-end spatial cluster showed a substantially stronger association (OR = 7.05, P = 2.93 x 10^{-7}) than seen in the whole-gene analysis (OR = 1.78, P = 7.08 x 10^{-4}). This finding persists after accounting for the number of tested clusters using the ACAT method ($P_{\text{ACAT-clusters}}$ = 1.47 x 10^{-6} vs. $P_{\text{wholegene}}$ = 7.08 x 10^{-4}), confirming that the signal is localized. Moreover, in contrast to the whole-gene signal, the spatial cluster even reaches the typical multiple testing threshold used in exome-wide analyses (0.05/18000 = 2.8 x 10^{-6}). These findings are in accordance with previous literature, which showed that pathogenic *FUS* mutations tend to cluster in the C-terminal end of the gene, where they are thought to disrupt nuclear import of FUS[44].

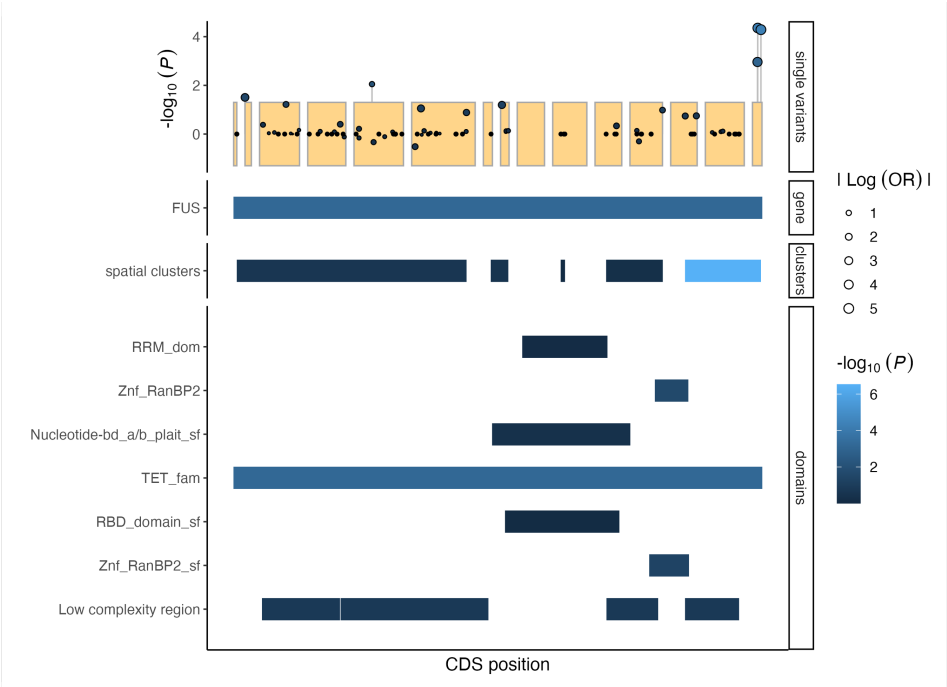

Figure 3. *FUS* mutation plot. The upper panel shows the coding sequence of FUS, with the y-axis showing the -log$_{10}$(*P*-value) for single variants. The panels below show the whole-gene, spatial clusters and domains respectively, colored by the -log$_{10}$(*P*-value) of the Firth burden test.

In contrast, for *SOD1* the whole-gene analysis was highly significant (OR = 8.01, 1.1 x 10^{-16}, Figure S1), while no meaningful spatial clusters were identified (one cluster, containing all variants and thus identical to the whole-gene analysis), and no meaningful domain signals were identified (all but one spanned the (near-)entirety of the gene and thus also resembled the whole-gene analysis). This aligns with previous literature, as pathogenic mutations have been reported across the entire length of the gene[45].

Together, this case study shows how the regional approaches implemented in RVAT successfully recover a proven mutation hotspot in ALS. Moreover, it is worth noting that the clustering approach may also prove useful for discovery of novel genes in a genome-wide setting. In *FUS*, the clustering approach showed a higher sensitivity (accounting for the number of tested clusters) than the whole-gene burden test. We additionally evaluated the SKAT and ACAT-v tests, which are both more powerful than burden tests when the proportion of causal variants is low{Citation}. Although in this case SKAT indeed had a higher sensitivity than a whole gene burden analysis (P_{SKAT} = 5.81 x 10^{-5}, P_{ACAT} = 1.49 x 10^{-3}), burden testing at the level of spatial clusters was more sensitive than either, substantiating that it can provide a powerful alternative to identify genes driven by regional hotspots.

Case study 2: Identification of biologically relevant gene sets and cell-types in the UK Biobank

In this case study we perform competitive gene set analyses on published rare variant gene statistics, illustrating how biologically relevant sets of genes can be discovered beyond exome-wide significant single genes. We obtained pLoF gene test-statistics from the Genebass browser, which includes exome-wide rare variant gene statistics of 4,529 phenotypes based WXS data of 394,841 individuals included in the UK Biobank[21]. Specifically, we focused on several traits highlighted in the corresponding article by Karczewski *et al.*: LDL-cholesterol, bone mineral density, red blood cell count and white

matter integrity. Ontology gene sets were downloaded from MSigDB[22], and were imported into the RVAT *geneSetFile* format using the *buildGeneSet* method. Competitive GSA was performed using *geneSetAssoc* method, adjusting for the number of variants included in the gene and the total coding sequence length.

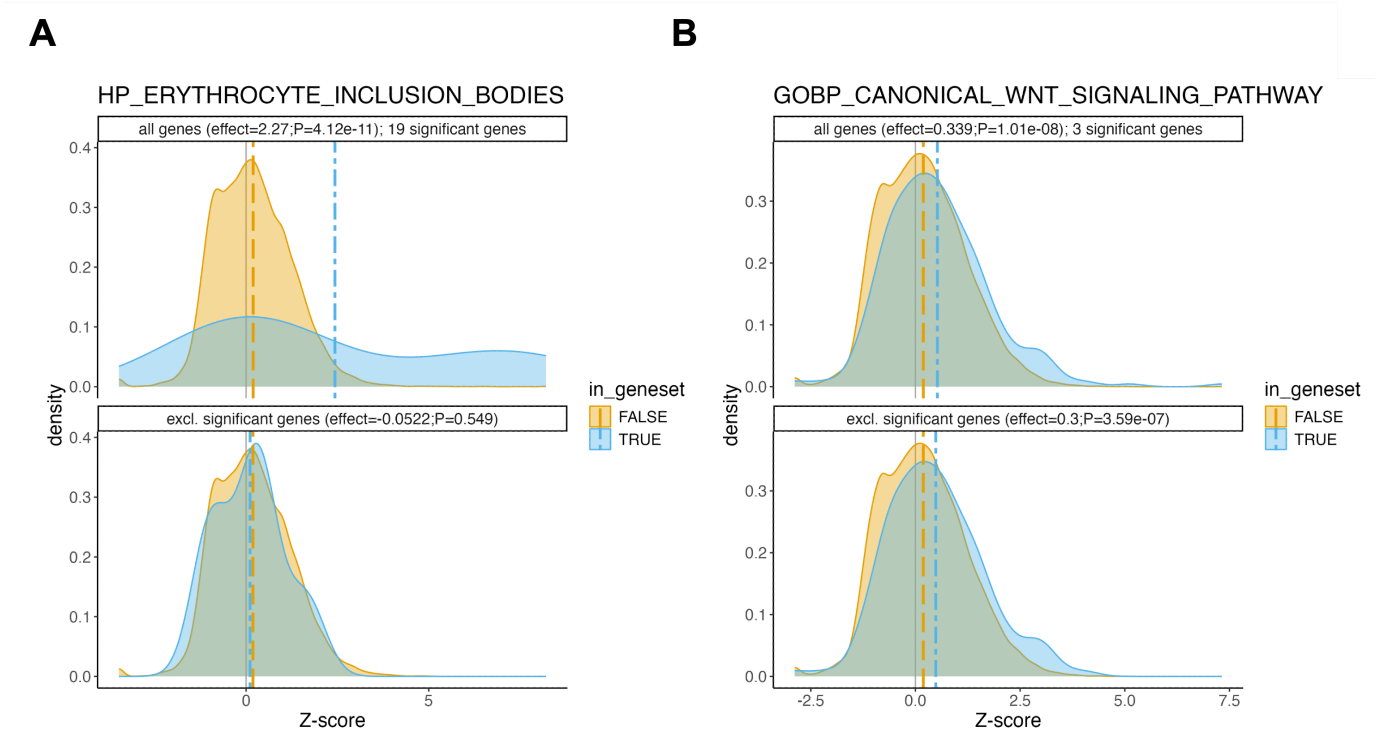

Figure 4. RVAT gene set summary plots. Density plots showing the Z-score distributions (inverse normal transformed *P*-values) of **(A)** red blood cell counts (RBC) for the gene set 'Erythrocyte Inclusion Bodies' and **(B)** bone mineral density for the gene set 'Canonical WNT Signaling'. The distribution of the background genes and genes within the gene set are shown in orange and blue respectively. The upper panels show the Z-score distribution including all genes, whereas the lower panels show the Z-score distribution excluding exome-wide significant genes ($P < 2.5 \times 10^{-7}$). The vertical lines represent the mean Z-scores. In panel **(A)**, the difference between Z-scores is large when including exome-wide significant genes but greatly diminishes upon rerunning the gene set analyses excluding them, thus indicating that the significant genes primarily drive the observed gene set signal. In panel **(B)** the Z-score difference largely remains upon excluding exome-wide significant genes, thus suggesting that a multitude of genes collectively drive the gene set association.

Numerous significant gene sets were identified ($P < 3.3 \times 10^{-6}$, Table S1), including many gene sets relevant to the respective traits. For example, the gene sets bone morphogenesis, WNT signaling, and ossification were associated with bone mineral density; genes related to polycythemia, increased hematocrit and oxygen binding were associated with red blood cell counts; genes related to sterol homeostasis, premature arteriosclerosis, and hypercholesterolemia were associated with LDL levels; and extrinsic component of (post-)synaptic membrane was associated with white matter integrity. We next sought to establish to what extent the identified gene sets were driven by exome-wide significant genes, as gene set associations that are primarily driven by one or two genes, may not be sufficient to draw broad conclusions about the involvement of the entire gene set.To that end, we reran the gene set analyses, excluding genes passing the significance threshold employed by Karczewski *et al.* ($P < 2.5 \times 10^{-7}$)[21]. This showed that a substantial proportion of findings were primarily driven by a small number of significant genes (Figures S2-3). Examples include the RBC-associated gene sets 'Erythrocyte Inclusion Bodies' (Figure 4A, *densityPlot* method) and 'Decreased mean corpuscular volume' (Figure S4), as can be clearly observed from the Z-score distributions. On the other hand, several relevant gene sets remained significant, indicating that the observed association was not solely driven by individual genes, thus suggesting a multitude of genes collectively drive the gene set association. Examples include the bone density-associated gene sets 'Canonical WNT signaling' ($P = 3.59 \times 10^{-7}$, Figure 4B) and 'Ossification' ($P = 4.26 \times 10^{-6}$, Figure S4), both showing a clear mean shift in Z-scores that is not solely driven by a small number of genes. Among the top sub-significant genes in these gene sets were several relevant genes including *LRP4* ($P = 7.12 \times 10^{-4}$; mediates bone formation inhibition and is related to the highly significant gene *LRP5*), *SOST* ($P = 1.55 \times 10^{-5}$; encodes the sclerostin protein), *COL1A2* ($P = 4.52 \times 10^{-5}$; encodes the pro-alpha2 chain of type I collagen) and *MRC2* ($P = 1.56 \times 10^{-5}$; encodes a protein involved in extracellular matrix remodeling).

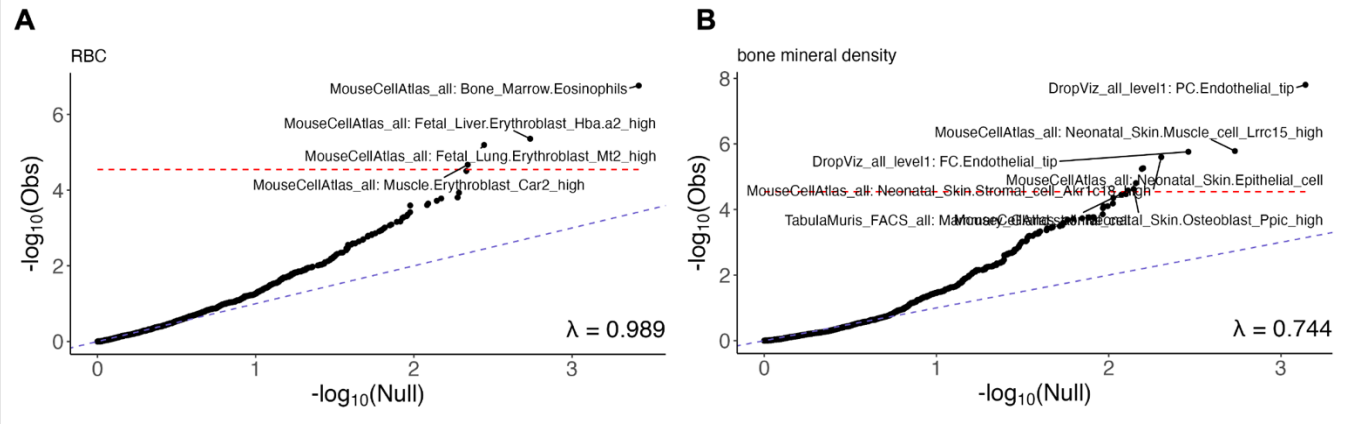

Figure 5. Single cell-type enrichment analyses. Quantile-quantile (qq) plots showing observed single cell enrichment test statistics (-$\log_{10}$(*P*-value)) versus expected −$\log_{10}$ (*P*-values) under the null model for **(A)** red blood cell counts (RBC) **(B)** bone mineral density. Labels indicate the experiment name and cell-type respectively, separated by a colon. The red line indicates the significance threshold ($P < 2.9 \times 10^{-5}$).

In a similar vein, we performed single-cell RNAseq (scRNA-seq) enrichment analyses using 51 publicly available scRNA-seq datasets[23]. Enrichment analyses were performed using the *geneSetAssoc* method, adjusting for average expression across cell-types in addition to the covariates employed in GSA. This revealed several significantly enriched cell-types relevant to the traits studied ($P < 2.9 \times 10^{-5}$; Figure 5, Figure S5 and Table S2). These include, among others, erythroblast cell subtypes (enriched for gene level association signals with red blood cell counts, $P = 4.4 \times 10^{-6}$), endothelial tip (enriched for gene level association signals with bone mineral density, $P = 1.8 \times 10^{-8}$) and liver hepatocytes (enriched for gene level association signals with LDL levels, $P = 5.4 \times 10^{-6}$). While the liver hepatocyte enrichment in genes exhibiting rare variant associations with LDL was primarily driven by exome-wide significant genes, several other enriched cell-types were not (e.g. the endothelial tip and erythroblast enrichments), suggesting a concentration of rare variant signal in relevant genes that have yet to reach exome-wide significance (Figure S6).

Together, these analyses highlight how RVAT's rare variant competitive gene set and cell-type enrichment analyses can aid in uncovering relevant biology beyond single genes.

Performance

We tested the performance of RVAT by benchmarking several key aspects of RVAT using an exome dataset (Project MinE[42] and UK Biobank[43]), consisting of ~60,000 samples and ~15 million variants (59,546 samples and 14,889,039 variants respectively). We first converted the gz-compressed vcf (3.7TB) to the RVAT *gdb* format, which completed in 734 CPU hours and required less than 8GB RAM on a computer cluster equipped with AMD EPYC 7702P 64-Core Processors. The resulting gdb file was significantly smaller, approximately 80x, with a size of 47GB compared to the input gz-compressed VCF-file of 3.7TB. We should note that the gdb format is not lossless as it retains variant info and genotype calls, while omitting other per-call metrics such as genotype quality and depth. None of the RVAT workflows, however, require these metrics.

To demonstrate the efficiency of RVAT, we show that the equivalent of a modern laptop is sufficient to perform a range of analyses on the ~60K samples and ~15M variants included in the *gdb*. All analyses below were performed on a single 2.60GHz processor (6 cores) with 16GB of RAM. First, we used the *mapVariants* method to map all 15M variants to Ensembl gene models (gff format), which completed within 2 minutes (114 seconds). Subsequently, the *uploadAnno* method was used to map variant effect predictions generated by snpEff[16] and dbscSNV[17] onto the gdb, which took approximately 7 minutes (431 seconds)

With the gene models and annotations in place, we generated non-synonymous variant sets for all protein-coding genes using the *buildVarSet* method, which took less than a half a minute (22 seconds). These variant sets were then used as input for exome-wide burden tests, employing firth logistic regression to test for an association between aggregated rare variant count (MAF < 0.001) and MND status (motor neuron disease). Burden tests were performed for a total of 17,149 protein-coding genes, and took approximately 11.3 hours, averaging around 2.4 seconds per gene. The resulting test-

statistics were used as input for competitive GSA across 15,473 ontology gene sets (MSigDB[22]), which took less than a minute to complete (47s). Finally, we benchmarked loading genotypes into a *genoMatrix* in R, for interactive downstream analyses or lookups, which took about half a second for 100 variants and ~60K samples.

Combined, the above benchmarks show that various steps involved in an exome-wide burden analysis, such as variant annotation, variant set generation, burden analyses and gene set analysis, can be completed overnight on a single CPU. To speed up analyses further and/or accommodate larger datasets, RVAT enables parallelization via its command-line interface (CLI). The RVAT website provides tutorials and pipelines demonstrating how to use RVAT in parallel on a computer cluster.

Discussion

Through a series of example use cases we have demonstrated that RVAT addresses unmet needs for functions that enable flexible and interactive mining of large genetic datasets for analyses of rare variant contributions to health and disease.

The RVAT *gdb* file format is built upon SQLite, a highly efficient and reliable database engine that is widely used across scientific disciplines and industries. Consequently, the *gdb* file format shares the strengths of a relational database, providing consistent storage of multiple linked data tables and facilitating complex queries that involve various combinations of these data types. We demonstrate that the *gdb* format is well-suited for storing and analyzing genetic data, which typically involves a variety of variant annotations, gene annotations, variant pathogenicity scores and phenotypic / clinical data for individual subjects. The indexed per-variant genotype compression ensures that the gdb is both storage-efficient (~80x smaller than a compressed vcf-file) while facilitating rapid genotype querying (loading 100 variants for 60K samples into R takes about half a second). *gdb* files are also highly portable and support easy resharing of linked data fields while subsetting operations make it easy to export reduced gdb for regions of interest while still maintaining all desired links between genotype, variant and sample level data.

RVAT avoids specific hardware dependencies, and we demonstrate that the framework is sufficiently lightweight and efficient to enable genome-wide rare variant association

screens of large sample numbers on the equivalent of a modern laptop. We also provide extensive documentation and multiple tutorials with example data to support ease of use. Moreover, RVAT is written in the open-source R programming language and both its functions and object types are fully integrated with the Bioconductor ecosystem. This decreases the learning curve and time investments needed by users to start new analyses and to further extend the functionalities of RVAT for niche applications with new custom scripts or other BioConductor packages. As example, in previous work we demonstrated that the core functionalities of unpublished RVAT prototypes enabled successful large scale rare variant association analyses of tens of thousands of subjects[10–13]. These studies respectively yielded the discovery of genes with high effect size associations to amyotrophic lateral sclerosis and Parkinson's disease. A key strength in the use of RVAT within these studies was the ease of deploying new subcohort analyses, adaptive variant quality control and target gene sensitivity analyses that enabled us to overcome confounding technical artefacts.

In this study we also further extended RVAT with gene partitioning and gene aggregating analyses to tackle key challenges in not only the discovery of rare variant associations, but also the correct interpretation of these associations. The first case study showcases how relevant mutation hotspots can be identified by utilizing several integrated components of RVAT, including spatial variant clustering, domain mapping, rare variant testing and RVAT visualizations. This framework enables researchers to fine-map genetic risk and estimate variant effect sizes within subgenic regions. Our results demonstrate how this can influence the interpretation of variants of unknown significance, and thus models to be considered for genetic testing and experimental analyses of disease gene biology. The second case study illustrates how our rare variant adaptations of competitive gene set and cell-type enrichment analyses can use publicly available summary statistics to reveal relevant biology beyond what is found in individual genes. While gene and cell-type analyses are widely applied in GWAS[32], their adoption in the context of rare variant analyses is limited. RVAT addresses this gap by providing a comprehensive framework for managing and importing gene sets, performing gene-set and cell-type enrichment analyses, and visualizing the results.

The RVAT package has some limitations. First is that simultaneous writing to a gdb by multiple users is not currently possible. This ensures that conflicting user operations do not cause data corruption but does introduce a requirement for coordination amongst users wishing to simultaneously upload new annotations to a shared gdb. RVAT is also specifically optimized for analyses relating to rare variant association testing. As such, while it is possible to use RVAT within custom code to analyze variant segregation in families, we have not optimized RVAT for this purpose. Finally, tools such as Regenie efficiently calculate regression offsets that are reported to aid rare variant analyses of large-scale biobank data. If desired, RVAT provides functionality to import these offsets generated using a dedicated tool such as Regenie. In this setting users can adjust their analyses for these offsets while still leveraging unique advantages of RVAT including its data structure, convenient R interface, optimized data querying functions and unique functionalities such as gene partitioning and geneset / cell type analyses.

As part of on-going work additional functionalities are being incorporated in future versions of RVAT. These will include an interactive result browser, analyses to support survival analyses, additional analyses for unsupervised variant clustering and additional analyses tailored to addressing challenges of studying rare variation in the non-coding genome. RVAT also already incorporates functions to facilitate use cases beyond the scope of this study, including analyses of recessive disease models and sex-linked chromosomes, a dedicated CLI interface and features to support effective deployment on high-performance computing environments. Full details of these aspects as well as function documentation and tutorials with accompanying example datasets are available on the RVAT website (https://kennalab.github.io/rvat/).

Acknowledgements

K.K. is supported by grants from the Dutch Research Council (grant no. ZonMW-VIDI 91719350) and the ALS Foundation Netherlands. This research has been conducted using the UK Biobank Resource under application number 48361.

References

1. Palmer, D. S. *et al.* Exome sequencing in bipolar disorder identifies AKAP11 as a risk gene shared with schizophrenia. *Nat. Genet.* **54**, 5 (2022).
2. Broad Genomics Platform *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
3. Wilcox, N. *et al.* Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk. *Nat. Genet.* **55**, 1435–1439 (2023).
4. Holstege, H. *et al.* Exome sequencing identifies rare damaging variants in ATP8B4 and ABCA1 as risk factors for Alzheimer's disease. *Nat. Genet.* **54**, 12 (2022).
5. Sazonovs, A. *et al.* Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility. *Nat. Genet.* **54**, 9 (2022).
6. Kaisinger, L. R. *et al.* Large-scale exome sequence analysis identifies sex- and age-specific determinants of obesity. *Cell Genomics* **3**, 100362 (2023).
7. Rajagopal, V. M. *et al.* Rare coding variants in CHRNB2 reduce the likelihood of smoking. *Nat. Genet.* **55**, 1138–1148 (2023).
8. Hindy, G. *et al.* Rare coding variants in 35 genes associate with circulating lipid levels—A multi-ancestry analysis of 170,000 exomes. *Am. J. Hum. Genet.* **109**, 81–96 (2022).
9. Dhindsa, R. S. *et al.* Rare variant associations with plasma protein levels in the UK Biobank. *Nature* (2023).
10. Smith, B. N. *et al.* Exome-wide Rare Variant Analysis Identifies TUBA4A Mutations Associated with Familial ALS. *Neuron* **84**, 324–331 (2014).
11. Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 9 (2016).
12. Nicolas, A. *et al.* Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* **97**, 1268-1283.e6 (2018).

13. Hop, P. J. *et al.* Systematic rare variant analyses identify RAB32 as a susceptibility gene for familial Parkinson's disease. *Nat. Genet.* **56**, 1371–1376 (2024).
14. The All of Us Research Program Genomics Investigators *et al.* Genomic data in the All of Us Research Program. *Nature* (2024).
15. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
16. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118} ; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
17. Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* **42**, 13534–13544 (2014).
18. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
19. Zhao, Z. *et al.* UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. *Am. J. Hum. Genet.* **106**, 3–12 (2020).
20. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am. J. Hum. Genet.* **104**, 410–421 (2019).
21. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022).
22. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
23. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
24. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).

25. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).
26. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
27. Loehlein Fier, H. *et al.* On the association analysis of genome-sequencing data: A spatial clustering approach for partitioning the entire genome into nonoverlapping windows. *Genet. Epidemiol.* **41**, 332–340 (2017).
28. Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
29. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
30. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).
31. Firth, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80**, 27–38 (1993).
32. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364 (2016).
33. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
34. Maleki, F., Ovens, K., Hogan, D. J. & Kusalik, A. J. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front. Genet.* **11**, 654 (2020).
35. Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346–5348 (2019).

36. ALSGENS Consortium *et al.* Exome sequencing in amyotrophic lateral sclerosis implicates a novel gene, DNAJC7, encoding a heat-shock protein. *Nat. Neurosci.* **22**, 1966–1974 (2019).
37. Chen, C.-Y. *et al.* The impact of rare protein coding genetic variation on adult cognitive function. *Nat. Genet.* **55**, 927–938 (2023).
38. van Es, M. A. *et al.* Amyotrophic lateral sclerosis. *The Lancet* **390**, 2084–2098 (2017).
39. Ryan, M., Heverin, M., McLaughlin, R. L. & Hardiman, O. Lifetime risk and heritability of amyotrophic lateral sclerosis. *JAMA Neurol.* **76**, 1367–1374 (2019).
40. Akçimen, F. *et al.* Amyotrophic lateral sclerosis: translating genetic discoveries into therapies. *Nat. Rev. Genet.* (2023).
41. Project MinE ALS Sequencing Consortium. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *Eur. J. Hum. Genet.* **26**, 1537–1546 (2018).
42. van Rheenen, W. *et al.* Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat. Genet.* **53**, 1636–1648 (2021).
43. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).
44. Shang, Y. & Huang, E. J. Mechanisms of FUS mutations in familial amyotrophic lateral sclerosis. *Brain Res.* **1647**, 65–78 (2016).
45. Dilliott, A. A. *et al.* Clinical testing panels for ALS: global distribution, consistency, and challenges. *Amyotroph. Lateral Scler. Front. Degener.* **24**, 420–435 (2023).

Supplementary figures

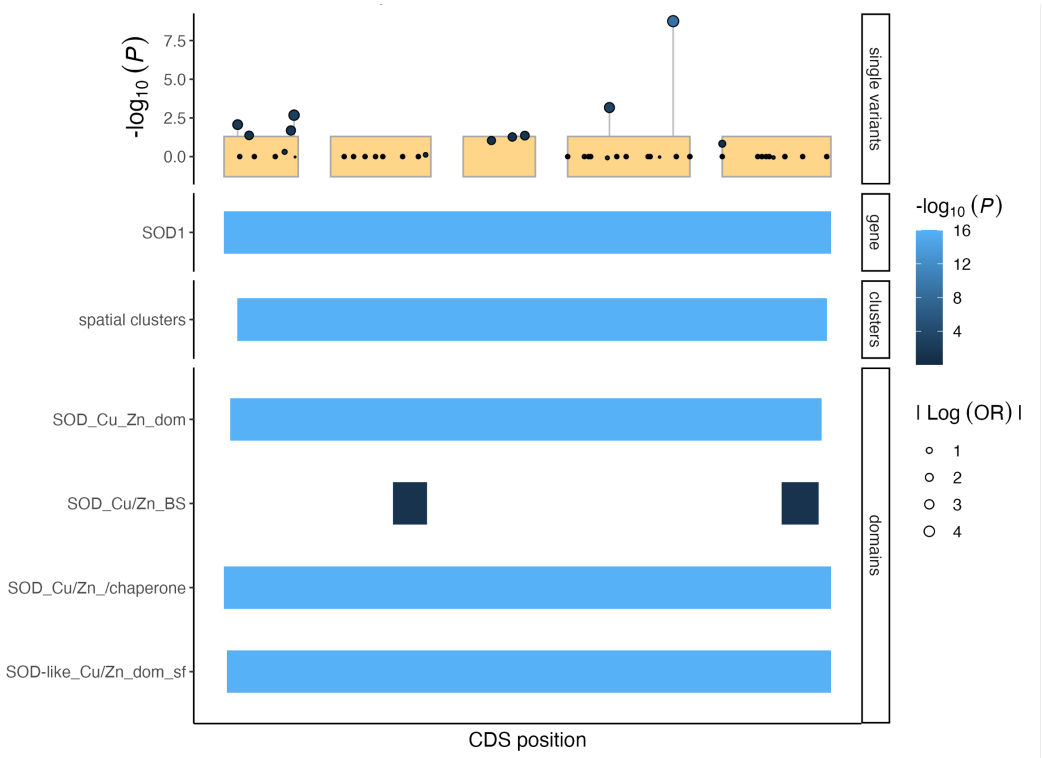

Figure S1. *SOD1* mutation plot. The upper panel shows the coding sequence of *SOD1*, with the y-axis showing the -log$_{10}$(*P*-value) for single variants. The panels below show the whole-gene, spatial clusters and domains respectively, colored by the -log$_{10}$(*P*-value) of the firth burden test.

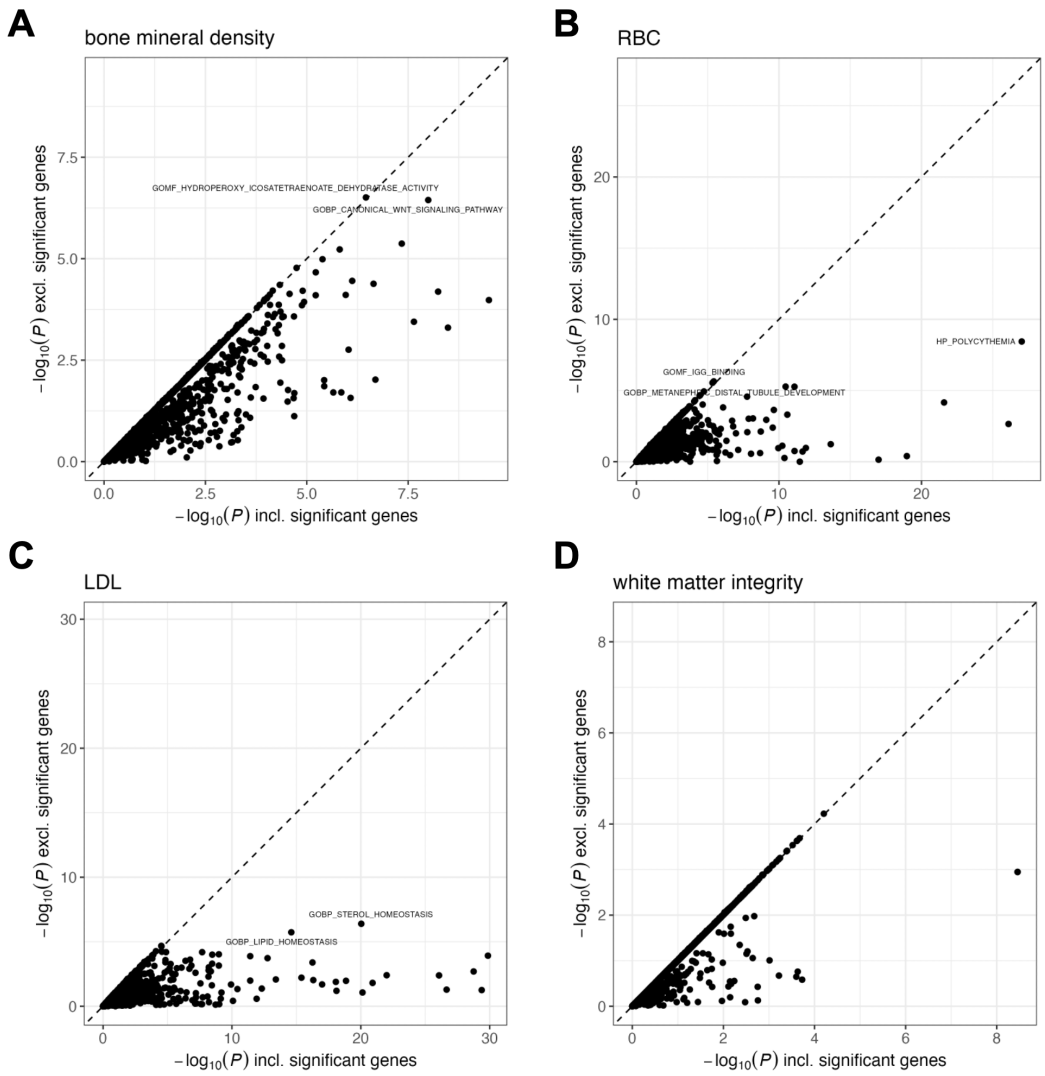

Figure S2. Competitive GSA test statistics (-$\log_{10}$(*P*-value)) excluding (y-axis) and including (x-axis) significant ($P < 2.5 \times 10^{-7}$) genes for **(A)** bone mineral density **(B)** red blood cell counts **(C)** LDL levels **(D)** white matter integrity.

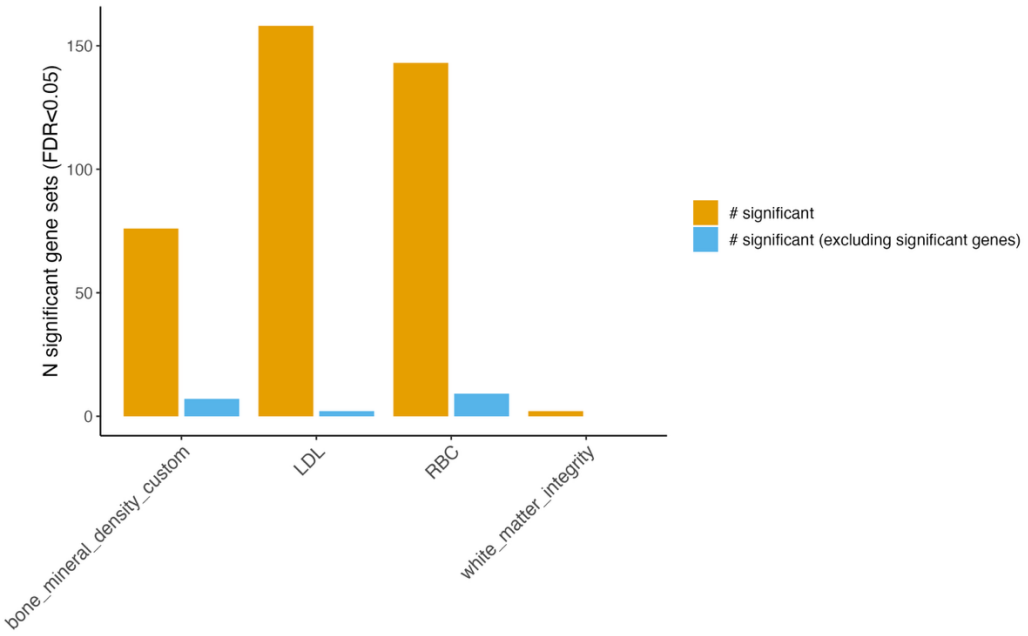

Figure S3. Number of significant gene sets (P < 3.3 x 10^{-6}) identified per phenotype.

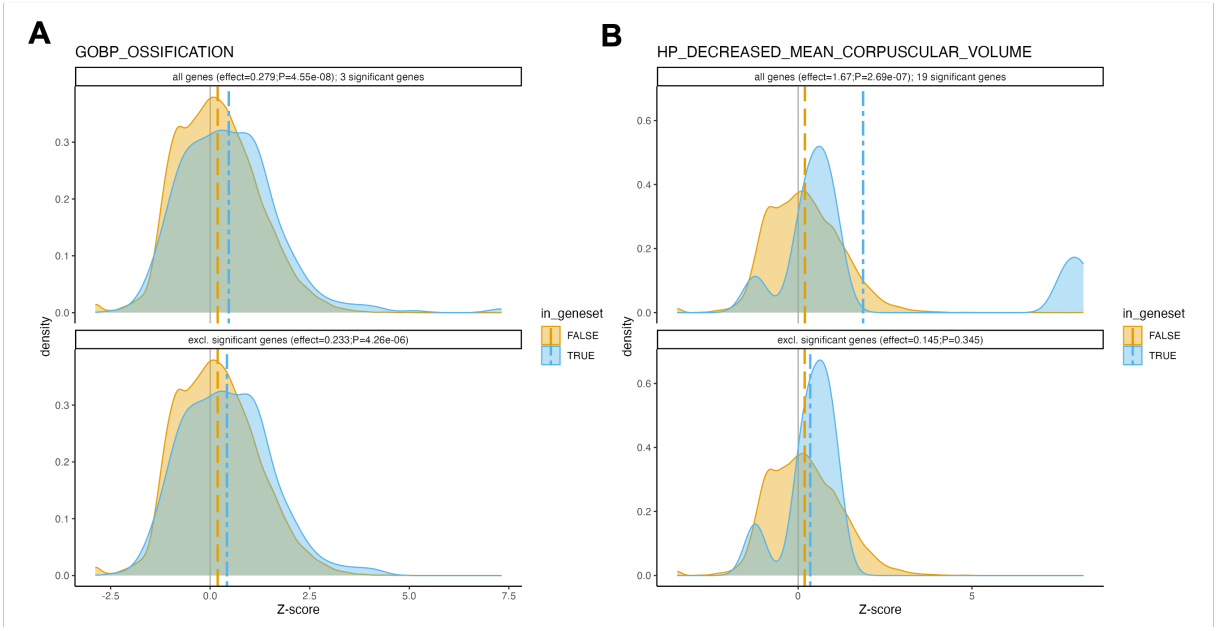

Figure S4. Density plots showing the Z-score distributions of **(A)** bone mineral density for the GO gene set 'Ossification' and **(B)** red blood cell counts (RBC) for the HP gene set 'Decreased Mean Corpuscular Volume'. The distribution of the background genes and genes within the geneset are shown in orange and blue respectively. The upper panels show the Z-score distribution including all genes, whereas the lower panels show the Z-score distribution excluding exome-wide significant genes (P < 2.5 x 10^{-7}). The vertical lines represent the mean Z-scores.

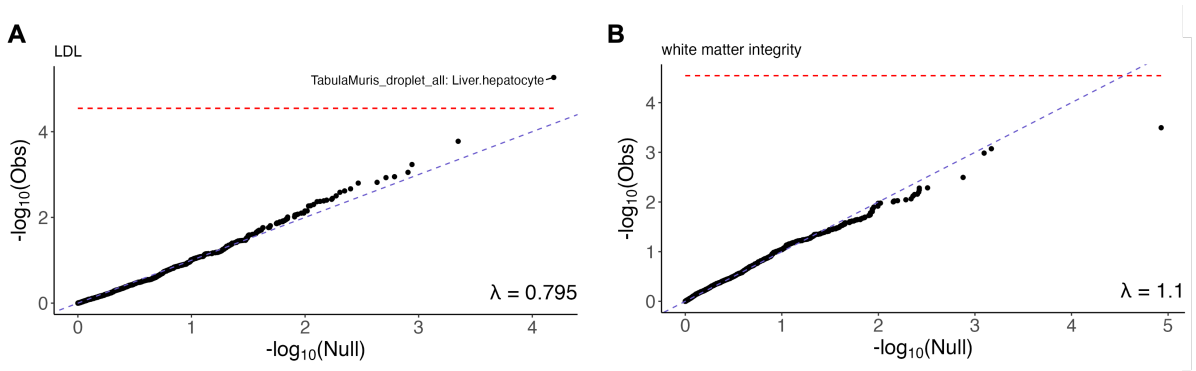

Figure S5. Single cell-type enrichment analyses. Quantile-quantile (qq) plots showing observed single cell enrichment test statistics (-$\log_{10}$(*P*-value)) versus expected −$\log_{10}$ (*P*-values) under the null model for **(A)** LDL-cholesterol **(B)** white matter integrity. Labels indicate the experiment name and cell-type respectively, separated by a colon. The red line indicates the significance threshold ($P < 2.9 \times 10^{-5}$).

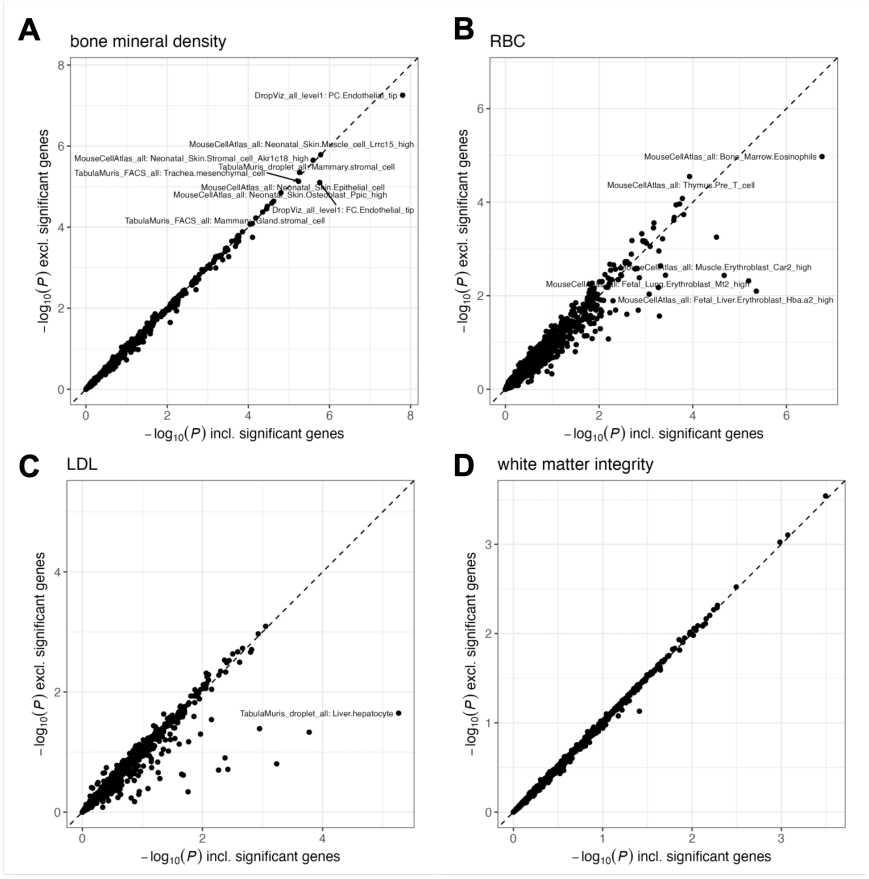

Figure S6. Single cell enrichment test statistics (-$\log_{10}$(*P*-value)) excluding (y-axis) and including (x-axis) significant ($P_{gene} < 2.5 \times 10^{-7}$) genes for **(A)** bone mineral density **(B)** red blood cell counts **(C)** LDL levels **(D)** white matter integrity.

Supplementary tables

Table S1: Gene set analysis results per trait.
Table S2: Single cell enrichment analysis results per trait.