Prioritizing disease-related rare variants by integrating gene expression data

Hanmin Guo
Stanford University
Alexander Eckehart Urban
Stanford University School of Medicine
Wing Hung Wong
whwong@stanford.edu

Stanford University

Method Article

Keywords:

Posted Date: May 10th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4355589/v1

Additional Declarations: No competing interests reported.

Abstract

Rare variants, comprising a vast majority of human genetic variations, are likely to have more deleterious impact on human diseases compared to common variants. Here we present carrier statistic, a statistical framework to prioritize disease-related rare variants by integrating gene expression data. By quantifying the impact of rare variants on gene expression, carrier statistic can prioritize those rare variants that have large functional consequence in the diseased patients. Through simulation studies and analyzing real multi-omics dataset, we demonstrated that carrier statistic is applicable in studies with limited sample size (a few hundreds) and achieves substantially higher sensitivity than existing rare variants association methods. Application to Alzheimer's disease reveals 16 rare variants within 15 genes with extreme carrier statistics. We also found strong excess of rare variants among the top prioritized genes in diseased patients compared to that in healthy individuals. The carrier statistic method can be applied to various rare variant types and is adaptable to other omics data modalities, offering a powerful tool for investigating the molecular mechanisms underlying complex diseases.

Introduction

Rare variants (minor allele frequency (MAF) < 1%) constitute a vast majority of human genetic variations[1,2]. They are on average more deleterious compared with common variants, and thus undergoes stronger selection and remains at low frequency in the general population. By analyzing large cohorts of whole genome sequencing/whole exome sequencing (WGS/WES) data, researchers have identified a handful of rare variants-trait association[3–6] and shown that rare variants contribute to a large proportion of missing heritability that cannot be explained by common variants[7].

Rare variants on average confer larger effects on gene expression and complex diseases and are easier to map to causal genes than common variants[8]. However, statistical power to identify disease-associated rare variants, especially for ultra-rare variants or even singletons, is limited, given that the sample size is not too high or the effect size is not too large. Variants collapsing methods (burden test, variance component test, omnibus test) are proposed to circumvent this obstacle[6,9–11], which evaluate association for multiple variants in a biologically relevant region, such as a gene, instead of testing the effect of single variant. These methods work well under the assumption that multiple variants in a gene cumulatively contribute to the disease risk with each individual allele explaining only a small fraction of the cases, but resolution to pinpoint the risk variants may be diluted when the assumption is violated. Recent studies based on large scale WES have revealed rare protein truncating variants associated with a wide range of phenotypes[4,5,12,13]. These variants exert extreme effects on the function of genes and their encoded proteins, underscoring the importance of considering how rare variants are related to gene expression.

Complementary to genome sequencing assay, RNA-seq can quantitatively measure gene expression level and provide molecular cause of complex diseases, especially rare diseases. Previous studies have shown that rare variants are enriched near genes with aberrant gene expression[14–16]. We posit that those rare variants will be more prone to disease pathology. In this work, we propose carrier statistic, a statistical framework for prioritizing those rare variants with large functional consequence in diseased patients by integrating gene expression data. We demonstrate superior performance of our method through extensive simulations and application study to the Alzheimer's disease, where given a limited sample size, existing rare variants association methods without functional gene expression data cannot provide positive findings.

Results

Method Overview

Our method stems from the expectation that diseased population shows enrichment in rare variants that have large impact on expression for disease-related genes. Suppose we have genotypes (e.g. variants call from WGS data) and gene expression measurements (e.g. reads count from RNA-seq data) on a disease relevant cell type for both diseased patients and healthy controls. For each rare variant-gene pair, we calculate the expression association z-score for rare variant carriers by using the gene expression from individuals without the variant as the null distribution (Fig. 1a). The expression association z-score, which we term as carrier statistic, is calculated separately within the case and the control group. We only consider rare variant-gene pair wherein the variant is located within the exon of the gene throughout this study. The carrier statistic quantifies the degree to which the rare variant impacts gene expression level. We assume that most rare variants do not have a large impact on the gene function, so the distribution of carrier statistic will be centered around 0. Now, if a gene is relevant for the disease, then conditioning on having the disease will bias the sampling towards people carrying rare variants with large functional impact on the gene expression. Thus, the carrier statistics for disease-related rare variant-gene pairs will tend to be more extreme compared to those for non-related pairs in the case group (Fig. 1b). We prioritize those rare variants and genes with outlier carrier statistic in the case group. False discovery rate (FDR) can be computed as the ratio of tail probability for carrier statistic between two groups (**Methods**).

Simulation Results

We first carried out simulations to assess whether the carrier statistic-based method would produce false positive findings. We simulated genotypes based on whole exome sequencing (WES) data from Genome Aggregation Database (gnomAD)[2] and simulated gene expression profiles based on RNA-seq data in whole blood tissue from the Genotype-Tissue Expression (GTEx) project (**Methods**). We perturbed the expression level of the causal genes for causal variants carriers by assuming that causal rare variants have large functional impact on disease-related genes. FDR for carrier statistic was well-calibrated in all simulation settings with varying penetrance of causal variant, prevalence in causal variant noncarriers, and number of causal variants per causal gene (**Supplementary Fig. 1**). We also checked if using gene expression from noncarriers in two groups together rather than separately as null distribution will produce false positive findings (**Methods**). In this case, we observed substantial inflation in FDR, particularly when diseased patients have systematic change in gene expression profile from healthy controls. In contrast, using gene expression from noncarriers in the same group as null distribution consistently gave well-calibrated error rates (**Supplementary Fig. 2**).

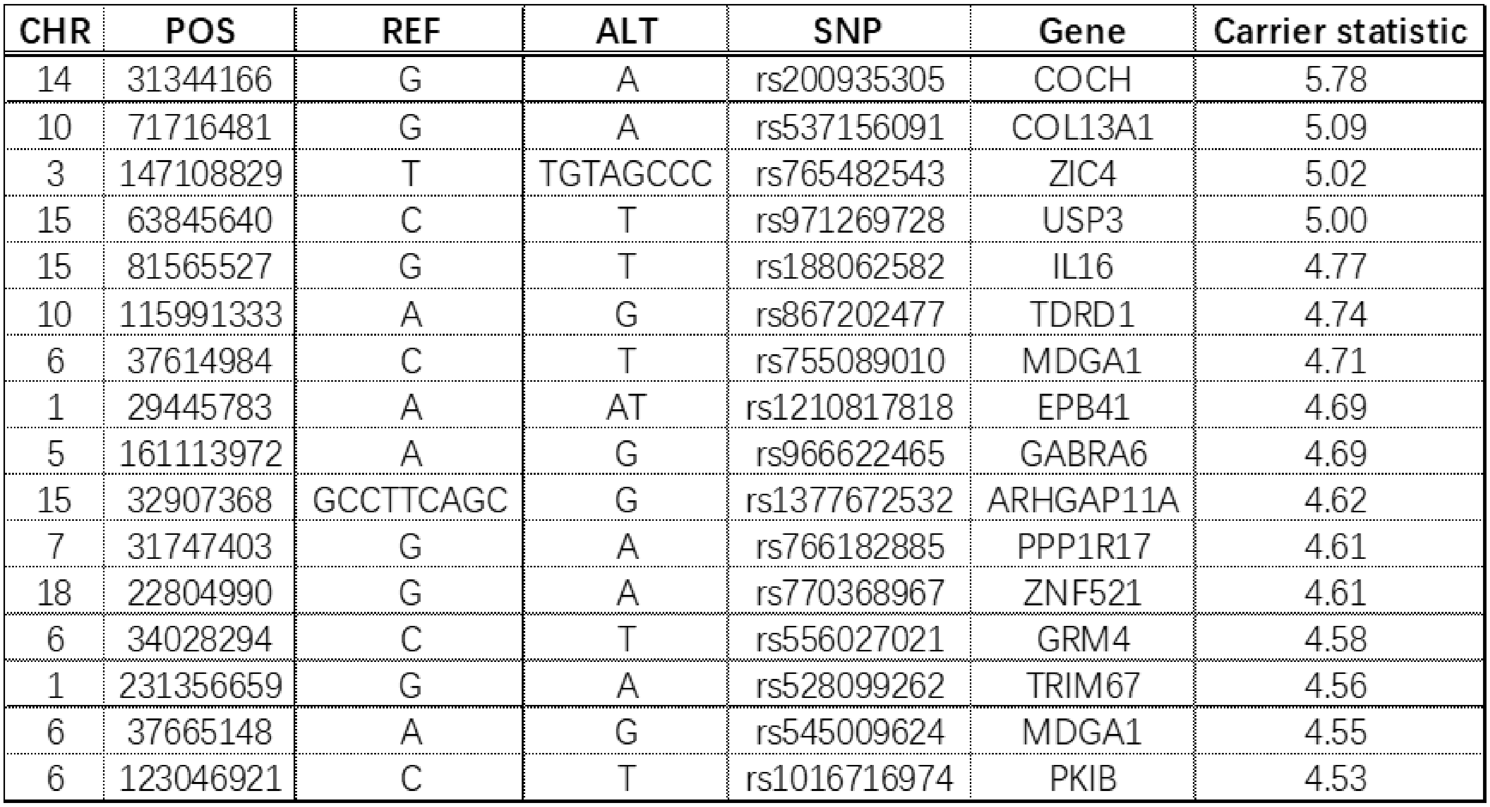
Next, based on the simulated data, we benchmarked the performance of carrier statistic with three existing rare variants association methods: burden test[9], Sequence Kernel Association Test (SKAT)[10], and SKAT-O[11]. Burden test counts the number of rare variants within a gene followed by the association test with the disease. SKAT computes a gene-level variance component score statistic which allows bidirectional effect of different variants. The unified test SKAT-O implements a linear combination of burden test statistic and variance component test statistic, which is preferred when the underlying genetic architecture of the disease is not known. While all four methods successfully controlled FDR at the nominal level (**Supplementary Fig. 1**), carrier statistic achieved higher sensitivity than the three variants collapsing methods under all simulation settings (Fig. 2). We believe the low sensitivity of burden-like statistics is due to the small number of case samples that can be attributed to the causal variants in any gene region, which makes it difficult to attain statistical significance of enrichment of rare variants burden for any causal gene region (**Supplementary Tables 1–3**).

We also performed empirical power analysis, which provides guidance for designing new disease association study with genome sequencing and RNA-seq data. Simulations were repeated 50 times to determine the sample size required for achieving 80% sensitivity, given that the penetrance of causal variant is 70%, the prevalence of causal variant noncarrier is 1%, and there are 5 causal variants per causal gene. When the effect size of causal variant on gene expression is large (e.g. Z = 5), 80% causal genes can be identified based on a cohort of 500 cases and 500 controls (Fig. 3). The necessary sample size gradually increases as causal variants become less deleterious and have smaller functional consequence on gene expression, e.g. 2,000 samples will be needed to achieve 80% sensitivity for Z = 4.5. In contrast, given the same sample sizes standard GWAS will not have sufficient power to detect rare causal variants regardless of their effect sizes.

Application to Alzheimer’s disease

Alzheimer’s disease is highly heritable, with heritability estimated to be as high as 60%-80% based on twin studies[17]. Large scale GWASs have identified multiple loci contributing to Alzheimer’s disease, but the genetic variance explained by these loci is far below the level suggested by the disease heritability[18]. Additionally, there is limited understanding regarding the molecular mechanism through which these GWAS variants affect the disease, with the exception of the well-known APOE locus. To investigate whether rare variants (single nucleotide variants [SNVs] and short indels) confer functional consequences in Alzheimer’s disease, we applied carrier statistic to a harmonized multi-omics dataset ($n_{case} = 444, n_{control} = 234$) consisting of WGS and RNA-seq from prefrontal cortex in four aging cohort studies: the Religious Orders Study (ROS) and Memory and Aging Project (MAP), the Mount Sinai Brain Bank (MSBB), and the Mayo Clinic (**Methods**). We found significant excess of large carrier statistic in the diseased patients (Fig. 4). Controlling FDR with a cutoff of 0.2, we prioritized 16 rare variants within 15 genes with large carrier statistic in the case group (Table 1), implicating them as candidate variants that may contribute to Alzheimer’s disease through up-regulating gene expression in the brain.

Table 1 16 rare variants within 15 genes with large carrier statistic in the Alzheimer’s disease patients.



CHR	POS	REF	ALT	SNP	Gene	Carrier statistic
14	31344166	G	A	rs200935305	COCH	5.78
10	71716481	G	A	rs537156091	COL13A1	5.09
3	147108829	T	TGTAGCCC	rs765482543	ZIC4	5.02
15	63845640	C	T	rs971269728	USP3	5.00
15	81565527	G	T	rs188062582	IL16	4.77
10	115991333	A	G	rs867202477	TDRD1	4.74
6	37614984	C	T	rs755089010	MDGA1	4.71
1	29445783	A	AT	rs1210817818	EPB41	4.69
5	161113972	A	G	rs966622465	GABRA6	4.69
15	32907368	GCCTTCAGC	G	rs1377672532	ARHGAP11A	4.62
7	31747403	G	A	rs766182885	PPP1R17	4.61
18	22804990	G	A	rs770368967	ZNF521	4.61
6	34028294	C	T	rs556027021	GRM4	4.58
1	231356659	G	A	rs528099262	TRIM67	4.56
6	37665148	A	G	rs545009624	MDGA1	4.55
6	123046921	C	T	rs1016716974	PKIB	4.53

To see if existing methods can also detect these variants, we applied burden test, SKAT, SKAT-O to the same Alzheimer’s disease dataset. These three variants collapsing methods did not identify any significant genes (FDR < 0.2), possibly due to insufficient sample size. To further evaluate the performance of carrier statistic, we assessed the enrichment of rare variants burden within the top prioritized genes in case group compared to that in controls. Among the top 100 genes with largest carrier statistic, 67 genes have fold enrichment larger than 1, 32 genes have fold enrichment larger than 4/3, while only 6 genes have fold enrichment smaller than 3/4 (**Supplementary Fig. 3**). Consistent with results from the simulations, enrichment of rare variants burden within each of those genes was moderate and did not pass significance threshold by the variants collapsing methods.

The significant genes prioritized by the carrier statistics may shed light on the genetic etiology of Alzheimer’s disease (Fig. 5). *COCH* has the largest carrier statistic of 5.78. Missense mutations within this gene were found to cause the late-onset DFNA9 deafness disorder[19,20]. Furthermore, deposits of Cochlin encoded by *COCH* is associated with age-related glaucomatous trabecular meshwork but absent in healthy controls. Additionally, SNPs inside *COCH* are associated with cortical thickness[21], changes in which through neuroimaging techniques is commonly used in early detection and monitoring of Alzheimer's disease progression[22,23]. Gene *ARHGAP11A* has a carrier statistic of 4.62. Transcribed mRNAs of the gene subcellularly localize and are locally translated in radial glia cells of human cerebral cortex and further regulate cortical development[24]. More importantly, *ARHGAP11A* may contribute to Alzheimer's disease pathology by mediating Amyloid-β generation and Amyloid-β oligomer neurotoxicity[25]. *PPP1R17* (carrier statistic = 4.61) functions as a suppressor of phosphatase complexes 1 (PP1) and 2A (PP2A). A recent study suggests that a subpopulation of neurons in the dorsomedial hypothalamus regulate aging and lifespan in mice through hypothalamic-adipose inter-tissue communication and this regulation depends on Ppp1r17 expression[26]. Interestingly, *PPP1R17* is also involved in human-specific cortical neurodevelopment regulated by enhancers in human accelerated regions[27]. *ZIC4* (carrier statistic = 5.02) plays an important role in the embryonal development of the cerebellum. Heterozygous deletions encompassing the *ZIC4* locus are associated with a rare congenital cerebellar malformation known as the Dandy–Walker malformation[28]. Notably, mutations in proximity to the *ZIC4* loci are implicated in multiple system atrophy, a rare neurodegenerative disease[29]. Large scale GWAS study of brain morphology has also identified associations with *ZIC4*, underscoring its significance in diverse neurological processes[30]. *MDGA1* (carrier statistic = 4.71) encodes a glycosylphosphatidylinositol (GPI)-anchored cell surface glycoprotein. It has been reported that *MDGA1* can contribute to cognitive deficits through altering inhibitory synapse development and transmission in the hippocampus[31]. Of note, *MDGA1* is one of the 96 genes from the Olink neurology panel with established links to neurobiological processes and neurological diseases.

Discussion

We presented carrier statistic, a statistical framework to perform multi-omics data analysis, for prioritization of disease-related rare variants and their regulated genes. Through simulations and analyses of real multi-omics dataset, we demonstrated that carrier statistic overcomes sample size limitation and achieves substantial gain in statistical power compared to existing variants collapsing methods. The superior performance of carrier statistic can be attributed to incorporation of functional gene expression data, which allows quantitatively measuring the impact of rare variants that cannot be determined by looking at the variants alone. We applied carrier statistic to Alzheimer's disease and highlighted several novel risk genes, providing insights into the molecular etiology of the complex disease.

Carrier statistic serves as a general approach to study how rare variants affect complex disease through mediating gene expression. There exist several methods such as transcriptome-wide association study (TWAS)[32,33] or colocalisation[34,35] that can also perform integrative analysis across multiple data modalities (genotype, gene expression, and phenotype). However, those methods focus exclusively on effects of common SNPs and will have limited power for rare variants. In addition to SNVs and short indels that we included in this study, the statistical framework can be also applied to other types of rare variants (simple structural variants [SVs], complex SVs, mobile element insertions, tandem repeat expansions), which in general have larger effect size than SNVs[36]. Finally, carrier statistic can be adapted to other omics data, such as epigenomics and proteomics.

Over the last fifteen years, abundant disease-associated loci have been identified based on genome sequences of biobank-scale sample size (e.g. hundreds of thousand)[37,38]. However, even with such large sample sizes it is still difficult for GWAS analysis to detect rare causal variants. Another important objective is to understand the biological roles of the detected loci, which remains challenging. We showed here that by integrating RNA-seq data, the carrier statistic approach offers a study design that may overcome the sample size limitation and may help to associate the functional rare variants and their target genes. This approach will be especially useful for the study of diseases for which biospecimen of disease relevant tissue are easy to obtain and RNA-seq can be performed, such as autoimmune disease (relevant to blood) or skin-related disease. Furthermore, when the tissue sample is available, adding RNA-seq to a WGS-based GWAS will not increase the cost of the study significantly. As multi-omics data accumulates alongside genome sequencing data, we anticipate that carrier statistics will become an effective approach to dissect the molecular mechanism of complex diseases.

Methods

Carrier statistic

For each rare variant-gene pair (the variant is located within the exon of the gene), we used the expression of that gene in the rare variant noncarriers as the null distribution and computed a z-score for each rare variant carrier, then average over carriers of that variant. The carrier statistic was computed separately within the case group and the control group. Rare variants were defined as SNVs and short indels whose allele count was no larger than 5 within the case group or within the control group. Therefore, the rare variants and thus the number of carrier statistics are not the same between two groups. The carrier statistic quantifies the degree to which the rare variant impacts gene expression level. We assume that diseased population shows enrichment in rare variants that have large impact on expression for disease-related genes, thus there will also be enrichment of extreme carrier statistic for disease-related rare variant-gene pairs in the case group. We prioritize those rare variants and genes with outlier carrier statistic in the case group. For rare variant-gene pairs with positive carrier statistic, false discovery rate at a given threshold of carrier statistic, denoted by z_0, can be computed as $\frac{\Pr[z_{ctrl} \geq z_0]}{\Pr[z_{case} \geq z_0]}$, where z_{case} and z_{ctrl} denote the carrier statistic in the case group and control group, respectively. Duplicative carrier statistics were removed (i.e. multiple rare variants occurring in the same individuals are counted as the same rare variant). Similarly, for rare variant-gene pairs with negative carrier statistic, false discovery rate at threshold of z_0 can be computed as $\frac{\Pr[z_{ctrl} \leq z_0]}{\Pr[z_{case} \leq z_0]}$.

Simulations

We simulated genotypes for a large population consisting of 125,748 individuals based on the alternative allele count from 125,748 exomes in the gnomAD v2.1.1 dataset[2]. Only exonic variants that passed all variant filters in the gnomAD dataset were retained. Then we simulated gene expression data for the large population as follows. We first simulated background gene expression profile for these 125,748 individuals while matching the mean and standard deviation of normalized gene expression in the reference expression dataset. In this study, we used $\log_2(\mathrm{readscount} + 1)$ as normalized gene expression and RNA-seq in the whole blood tissue from GTEx project v8 as the reference expression dataset[39]. Genes whose median number of reads count in the large population < 10 were removed. We randomly selected m causal genes and l causal variants for each causal gene, where m was set as 50 and l was set to vary from 1 to 10. For each causal gene, we perturbed the gene expression for causal variant carriers by $z * sd$ fold, where z was set as 4 and sd was the standard deviation of normalized gene expression in the reference dataset. Next, we simulated the disease status for the large population by assuming penetrance of causal variant as $p_{carrier}$ and prevalence in causal variant noncarrier as $p_{noncarrier}$. Here $p_{carrier}$ varied from 0.5 to 0.9 and $p_{noncarrier}$ varied from 0.005 to 0.02. Finally, we randomly sampled 500 cases and 500 controls from the affected and nonaffected population respectively to mimic the sample recruitment procedure in the disease study. Each simulation setting was repeated for 100 times.

We evaluated the performance of different methods using two metrics: FDR and sensitivity. FDR was defined as the proportion of falsely identified genes among all identified ones. If no gene was identified then FDR was set as 0. Sensitivity was defined as the proportion of truly identified genes among all underlying causal genes.

Note that carrier statistic was computed by using gene expression from rare variant noncarriers in the same group (i.e. case or control) as the carriers as null distribution. We also checked if using gene expression from noncarriers in both case group and control group as the null distribution will produce false positive findings. We perturbed expression level for all genes in the case group. In this case, FDR showed substantial inflation for using gene expression from all individuals in two groups as null distribution, especially when there is large systematic difference in the transcriptome between two groups (**Supplementary Fig. 2b**). On the contrary, using gene expression from rare variant noncarriers in the same group of carriers as null distribution consistently controlled FDR at the nominal level (**Supplementary Fig. 2a**).

Implementation of different methods

Variants collapsing methods were performed using the R package SKAT v.2.2.5. All the parameters were set as default value. Both common and rare variants were included in the analysis.

Multi-omics data analysis for Alzheimer's disease

WGS data for the four aging cohorts (ROS/MAP, MSBB, and the Mayo Clinic) were obtained from the Whole Genome Sequence Harmonization Study (Synapse ID: syn22264775). RNA-seq data for the same four cohorts were downloaded from the RNAseq Harmonization Study (syn21241740). Only white people with both WGS and RNA-seq data were included in the analysis.

We determined disease status following description in previous publications[40,41]. For the ROS/MAP cohorts, individuals with a Braak neurofibrillary tangle score $\geq$ 4, a CERAD neuritic and cortical plaque score $\leq$ 2, and a cognitive diagnosis of probable Alzheimer's disease with no other causes (cogdx = 4) were classified as cases, while individuals with a Braak score $\leq$ 3, a CERAD score $\geq$ 3, and a cognitive diagnosis of no cognitive impairment (cogdx = 1) were classified as controls. For MSBB, individuals with a Braak score $\geq$ 4, a CERAD score $\geq$ 2, and a Clinical Dementia Rating (CDR) score $\geq$ 1 were classified as cases, while individuals with a Braak score $\leq$ 3, a CERAD score $\leq$ 1, and a CDR score $\leq$ 0.5 were classified as controls. For the Mayo Clinic cohort, individuals with a Braak score $\geq$ 4 and a CERAD score $\geq$ 2 were classified as cases, while individuals with a Braak score $\leq$ 3 and a CERAD score $\leq$ 1 were classified as controls. Of note, definition of CERAD score in the ROS/MAP cohort is different from that in the MSBB and the Mayo Clinic cohorts. After harmonization across cohorts, 444 cases and 234 controls in total were identified and used for downstream analysis.

Next, we performed quality control on the WGS and RNA-seq data. For WGS data, only exonic variants with missing genotypes $<$ 10% were retained. For RNA-seq data, we selected prefrontal cortex as the target brain tissue. If a donor does not have RNA-seq in the prefrontal cortex, then RNA-seq in other tissues will be used based on the following order: dorsolateral prefrontal cortex > posterior cingulate cortex > head of caudate nucleus in the ROS/MAP cohort, prefrontal cortex > frontal pole > superior temporal gyrus > inferior frontal gyrus > parahippocampal gyrus in the MSBB cohort, and temporal cortex > cerebellum in the Mayo Clinic cohort. Genes with zero reads count in more than 10% of samples or with median reads count $<$ 10 across samples were excluded. $\log_2(\mathrm{readscount}+1)$ was used as normalized gene expression. Then we applied carrier statistic to perform downstream analysis.

Declarations

Declaration of interests

The authors declare no competing interests.

Author Contribution

W.H.W. and A.E.U. supervised the study. H.G. and W.H.W. designed the idea. H.G. performed data analysis. H.G., A.E.U., and W.H.W. interpreted the results. H.G. and W.H.W. wrote the manuscript.

Acknowledgement

We thank Dr. Bo Zhou and Dr. Hua Tang for discussion and providing feedbacks. This work was partially supported by NIH grants R01 HG010359, R01 MH116529, and P50 HG007735.

Data Availability

The WGS data in the Whole Genome Sequence Harmonization Study (https://www.synapse.org/#!Synapse:syn22264775) and the RNA-seq data in the RNAseq Harmonization Study (https://www.synapse.org/#!Synapse:syn21241740) are publicly available on the AD Knowledge Portal platform through completion of a data use certificate. The gnomAD v2.1.1 data consisting of 125,748 exomes were downloaded from https://gnomad.broadinstitute.org/. The gene expression data (gene reads count) from GTEx project version 8 were downloaded from the GTEx Portal, https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression.

References

1. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
2. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
3. Jurgens SJ, et al. Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. Nat Genet. 2022;54:240–50.
4. Wang Q, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature. 2021;597:527–32.
5. Flannick J, et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature. 2019;570:71–6.
6. Li X, et al. Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. Nat Genet. 2023;55:154–64.
7. Wainschtein P, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat Genet. 2022;54:263–73.
8. Claussnitzer M, et al. A brief history of human disease genetics. Nature. 2020;577:179–89.
9. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83:311–21.
10. Wu MC, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93.
11. Lee S, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91:224–37.
12. Liu D, et al. Schizophrenia risk conferred by rare protein-truncating variants is conserved across diverse human populations. Nat Genet. 2023;55:369–76.
13. DeBoever C, et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. Nat Commun. 2018;9:1612.

14. Li X, et al. The impact of rare variation on gene expression across tissues. Nature. 2017;550:239–43.
15. Zeng Y, et al. Aberrant gene expression in humans. PLoS Genet. 2015;11:e1004942.
16. Brechtmann F, et al. OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. Am J Hum Genet. 2018;103:907–17.
17. Gatz M, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006;63:168–74.
18. Wightman DP, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet. 2021;53:1276–82.
19. Bhattacharya SK. Focus on molecules: cochlin. Exp Eye Res. 2006;82:355.
20. Robertson NG, et al. Cochlin immunostaining of inner ear pathologic deposits and proteomic analysis in DFNA9 deafness and vestibular dysfunction. Hum Mol Genet. 2006;15:1071–85.
21. Van Der Meer D, et al. The genetic architecture of human cortical folding. Sci Adv. 2021;7:eabj9446.
22. Querbes O, et al. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain. 2009;132:2036–47.
23. Schwarz CG, et al. A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity. NeuroImage: Clin. 2016;11:802–12.
24. Pilaz L-J et al. Subcellular mRNA localization and local translation of Arhgap11a in radial glial cells regulates cortical development. *bioRxiv*, 2020.07. 30.229724 (2020).
25. Huang Y-r et al. ArhGAP11A mediates amyloid-β generation and neuropathology in an Alzheimer's disease-like mouse model. Cell Rep 42(2023).
26. Tokizane K, Brace CS, Imai S. -i. DMHPpp1r17 neurons regulate aging and lifespan in mice through hypothalamic-adipose inter-tissue communication. Cell Metabol. 2024;36:377–92. e11.
27. Girskis KM, et al. Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. Neuron. 2021;109:3239–51.
28. Grinberg I, et al. Heterozygous deletion of the linked genes ZIC1 and ZIC4 is involved in Dandy-Walker malformation. Nat Genet. 2004;36:1053–5.
29. Hopfner F, et al. Common Variants Near ZIC1 and ZIC4 in Autopsy-Confirmed Multiple System Atrophy. Mov Disord. 2022;37:2110–21.
30. Zhao B, et al. Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. Nat Genet. 2019;51:1637–44.
31. Connor SA, et al. Loss of synapse repressor MDGA1 enhances perisomatic inhibition, confers resistance to network excitation, and impairs cognitive function. Cell Rep. 2017;21:3637–45.
32. Gamazon ER, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47:1091–8.

33. Wainberg M, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51:592–9.
34. Giambartolomei C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10:e1004383.
35. Hormozdiari F, et al. Colocalization of GWAS and eQTL signals detects target genes. Am J Hum Genet. 2016;99:1245–60.
36. Chiang C, et al. The impact of structural variation on human gene expression. Nat Genet. 2017;49:692–9.
37. Tam V, et al. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20:467–84.
38. Abdellaoui A, Yengo L, Verweij KJ, Visscher PM. 15 years of GWAS discovery: realizing the promise. Am J Hum Genet. 2023;110:179–94.
39. Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
40. Wan Y-W et al. Meta-analysis of the Alzheimer’s disease human brain transcriptome and functional dissection in mouse models. Cell Rep 32(2020).
41. Vialle RA, de Paiva Lopes K, Bennett DA, Crary JF, Raj T. Integrating whole-genome sequencing with multi-omic data reveals the impact of structural variants on gene regulation in the human brain. Nat Neurosci. 2022;25:504–14.

Figures

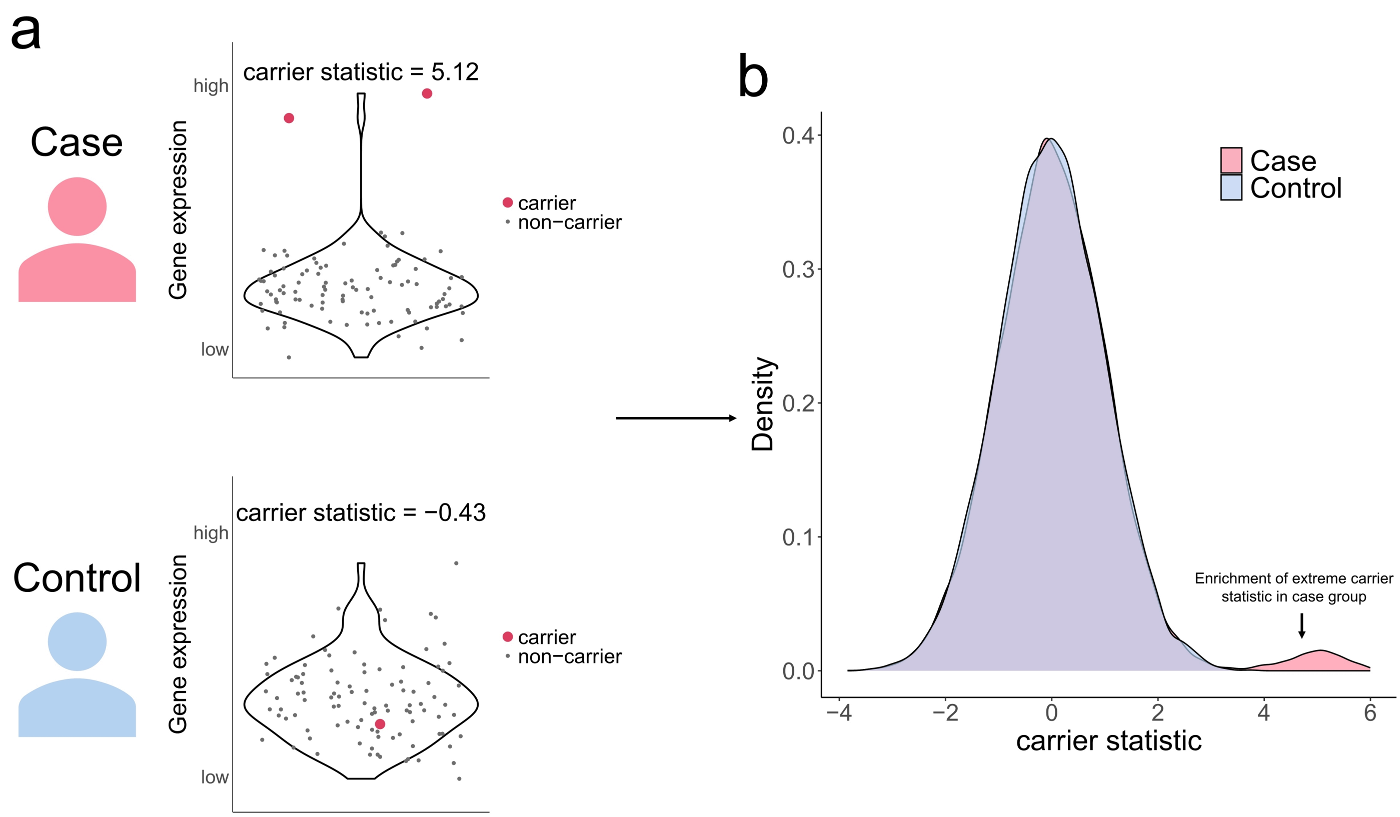


Figure 1

Carrier statistics for disease-related rare variant-gene pairs will tend to be more extreme compared to those for non-related ones in the case group

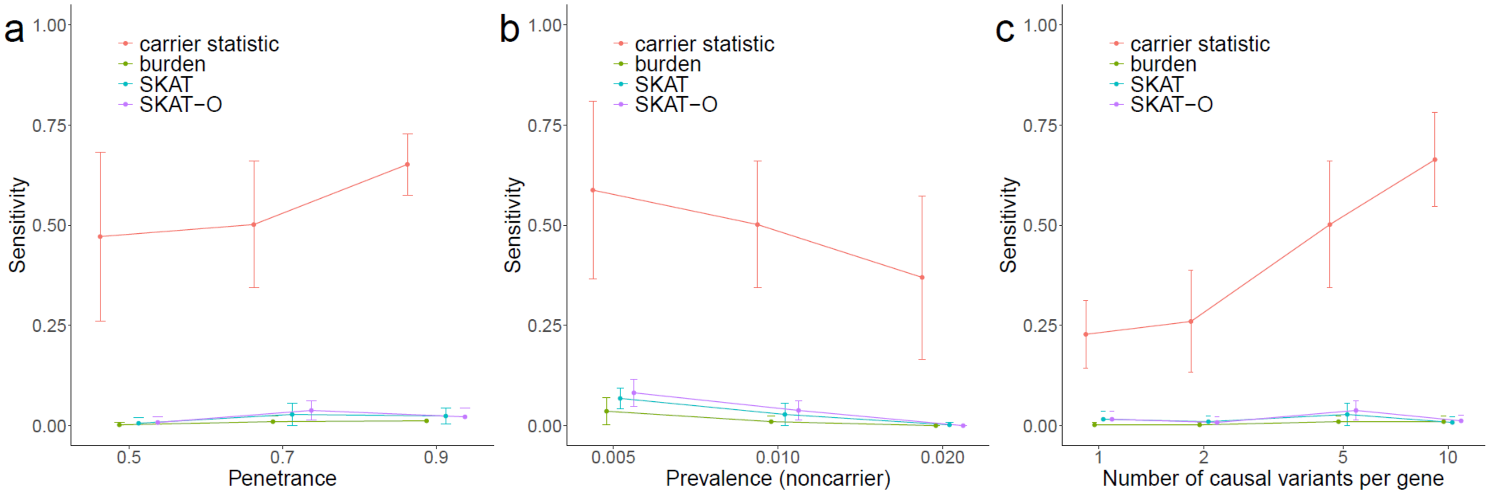


Figure 2

Carrier statistic achieves higher sensitivity than variants collapsing methods in simulations with varying (a) penetrance of causal variant, (b) prevalence in causal variant noncarriers, and (c) number of causal variants per causal gene

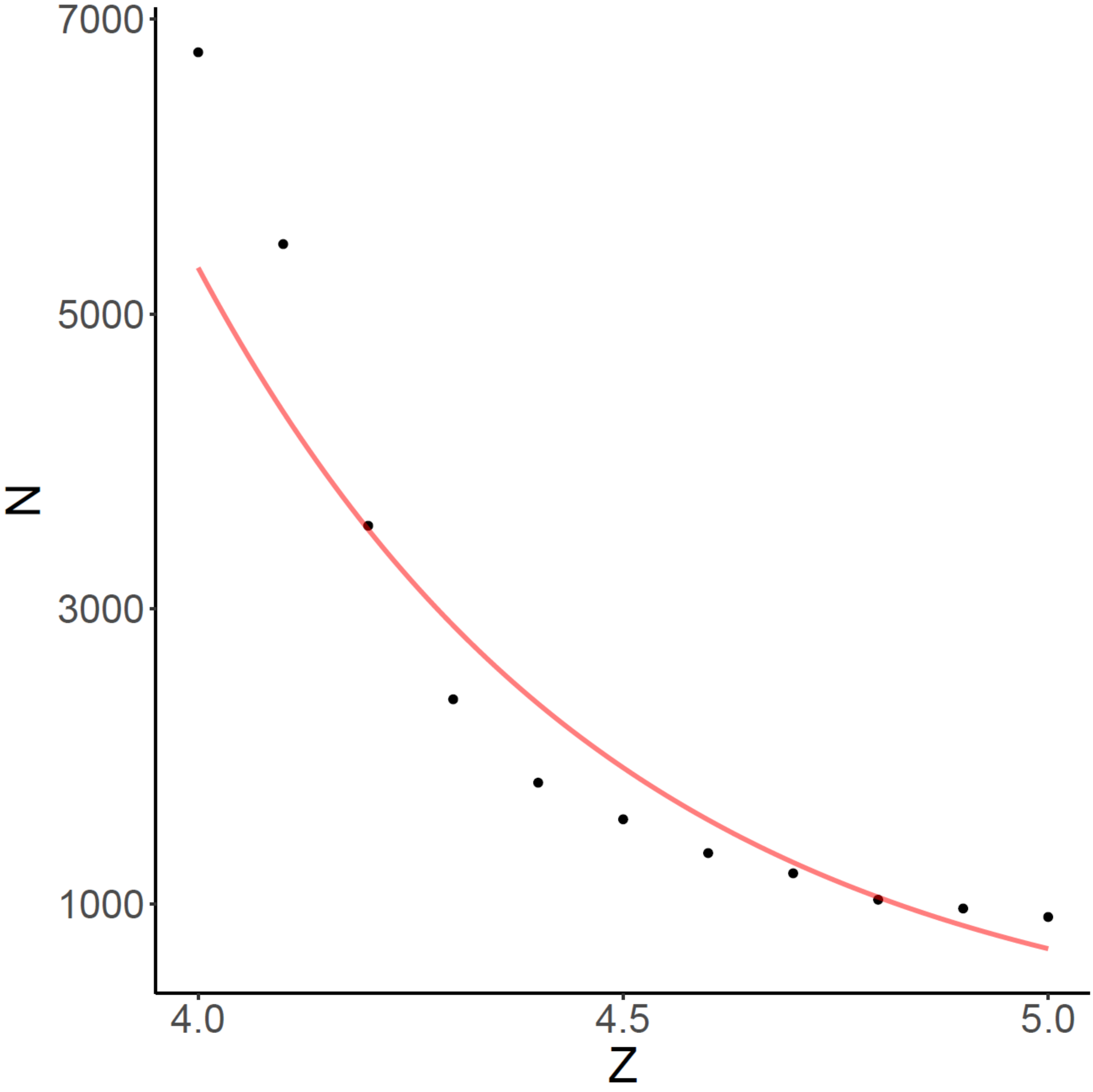


Figure 3

Required sample size for carrier statistic to attain 80% sensitivity based on simulations. Y-axis denotes total sample size with 50% case-control ratio and x-axis denotes effect size of causal variant on gene expression. Log-transformed sample size is regressed on Z and the fitted curve is shown in red.

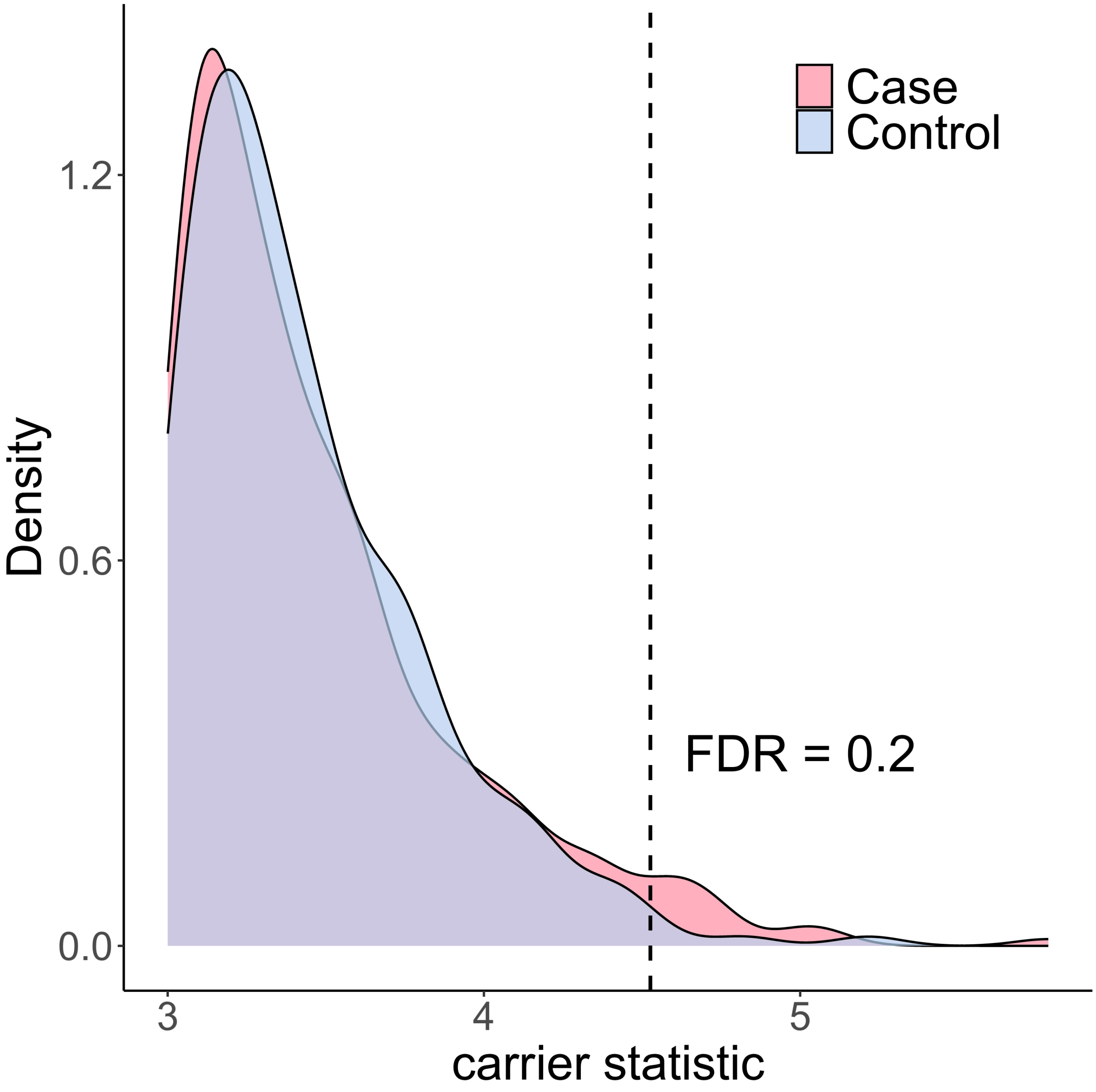


Figure 4

Alzheimer's disease patients show significant excess of large carrier statistic

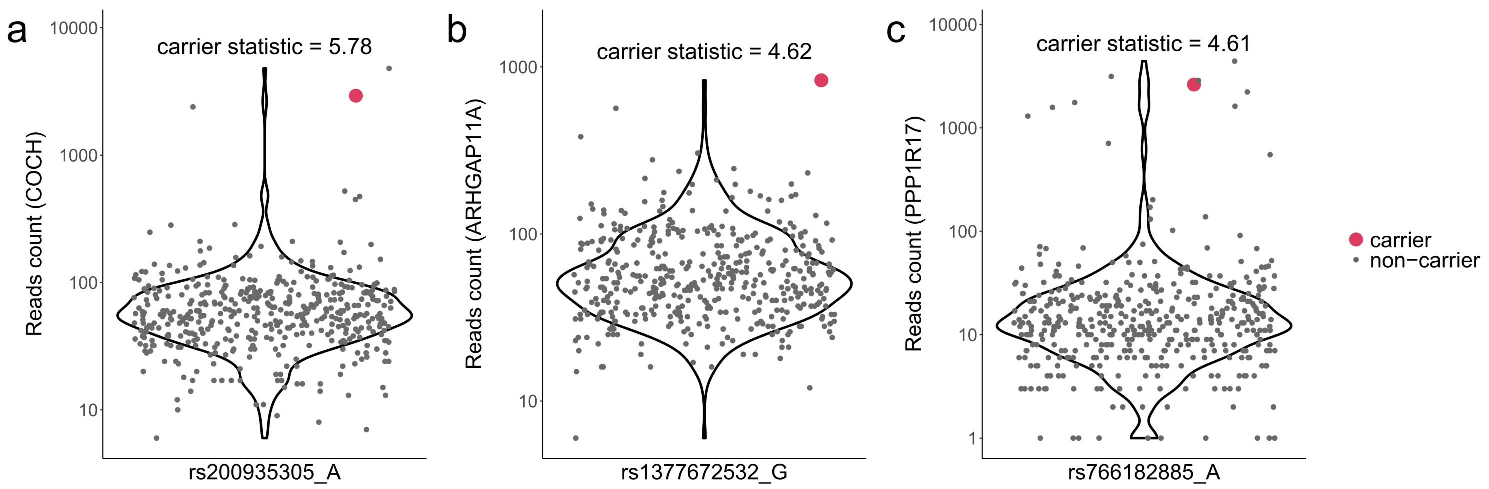


Figure 5

Rare variant carriers show outlier expression level for (a) *COCH*, (b) *ARHGAP11A*, and (c) *PPP1R17*. Pseudo count 1 was added to the RNA reads count for visualization purpose. Y-axis is on log scale

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures0430.pdf
- SupplementaryTables0430.xlsx