An association test of the spatial distribution of rare missense variants within protein structures identifies Alzheimer's disease–related patterns

Bowen Jin,[1] John A. Capra,[2] Penelope Benchek,[3] Nicholas Wheeler,[3] Adam C. Naj,[4] Kara L. Hamilton-Nelson,[5] John J. Farrell,[6] Yuk Yee Leung,[4] Brian Kunkle,[5,7] Badri Vadarajan,[8] Gerard D. Schellenberg,[4] Richard Mayeux,[8] Li-San Wang,[4] Lindsay A. Farrer,[6] Margaret A. Pericak-Vance,[5,7] Eden R. Martin,[5,7] Jonathan L. Haines,[3] Dana C. Crawford,[3] and William S. Bush[3]

[1]Graduate Program in Systems Biology and Bioinformatics, Department of Nutrition, School of Medicine, Case Western Reserve University, Cleveland, Ohio 44106, USA; [2]The Bakar Computational Health Sciences Institute, Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California 94143, USA; [3]Cleveland Institute for Computational Biology, Department for Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio 44106, USA; [4]Department of Pathology and Laboratory Medicine, Penn Neurodegeneration Genomics Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [5]The John P. Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, Florida 33136, USA; [6]Department of Medicine (Biomedical Genetics), Boston University School of Medicine, Boston, Massachusetts 02118, USA; [7]Dr. John T. Macdonald Foundation, Department of Human Genetics, Miller School of Medicine, University of Miami, Miami, Florida 33136, USA; [8]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Department of Neurology, Gertrude H. Sergievsky Center, Department of Neurology, Columbia University, New York, New York 10032, USA

More than 90% of genetic variants are rare in most modern sequencing studies, such as the Alzheimer's Disease Sequencing Project (ADSP) whole-exome sequencing (WES) data. Furthermore, 54% of the rare variants in ADSP WES are singletons. However, both single variant and unit-based tests are limited in their statistical power to detect an association between rare variants and phenotypes. To best use missense rare variants and investigate their biological effect, we examine their association with phenotypes in the context of protein structures. We developed a protein structure–based approach, protein optimized kernel evaluation of missense nucleotides (POKEMON), which evaluates rare missense variants based on their spatial distribution within a protein rather than their allele frequency. The hypothesis behind this test is that the three-dimensional spatial distribution of variants within a protein structure provides functional context to power an association test. POKEMON identified three candidate genes (*TREM2*, *SORL1*, and *EXOC3L4*) and another suggestive gene from the ADSP WES data. For *TREM2* and *SORL1*, two known Alzheimer's disease (AD) genes, the signal from the spatial cluster is stable even if we exclude known AD risk variants, indicating the presence of additional low-frequency risk variants within these genes. *EXOC3L4* is a novel AD risk gene that has a cluster of variants primarily shared by case subjects around the Sec6 domain. This cluster is also validated in an independent replication data set and a validation data set with a larger sample size.

[Supplemental material is available for this article.]

High-throughput DNA sequencing of diverse humans has identified millions of genetic variants, the vast majority of which are exceptionally rare. A survey of ~60,000 individuals from the Exome Aggregation Consortium (ExAC) found that out of ~7 million variants, 99% have a frequency <1% and 54% are singletons (Taliun et al. 2021). Similarly, in the Alzheimer's Disease Sequencing Project (ADSP) whole-exome sequencing (WES) of ~10,000 individuals, 97% of identified variants have a minor allele frequency <1%, and 23% are singletons (Butkiewicz et al. 2018). However, the effect of most rare variants on diseases of interest remains unknown because of insufficient statistical power to detect the associations between these variants and phenotypes.

We hypothesized that rare missense variants contribute to common diseases by disrupting the protein function and are likely to form clustered or dispersed patterns within protein structures when examined in population-based studies. Therefore, incorporating spatial context will improve rare variant association tests. Prior studies have shown that missense variants show nonrandom

Corresponding author: wsb36@case.edu

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276069.121. Freely available online through the *Genome Research* Open Access option.

patterns in protein structures, such as cancer-associated hotspot regions with a high density of missense somatic mutations (Tokheim et al. 2016). Our group (Sivley et al. 2018) also found that germline causal missense variants for Mendelian diseases show nonrandom patterns in three-dimensional (3D) space. These patterns include clusters that likely reflect disruption of a key functional region and dispersions that likely reflect depletion of variants within a sensitive protein core.

To test this hypothesis within sequencing studies of disease traits, we developed a kernel function to quantify genetic similarity among individuals by using protein structure information. When two individuals have different missense variants distal in genomic coordinates but close in 3D protein structure, these individuals will be assigned a high genetic similarity through our kernel function. When applied over an entire data set, our kernel function captures differences in the spatial patterns of rare missense variants among cases and controls or over continuous traits. Using a statistical framework similar to SKAT (Wu et al. 2011), we test the association of rare variants with quantitative and dichotomous phenotypes using this structure-based kernel. We call this approach protein optimized kernel evaluation of missense nucleotides (POKEMON). We validated that POKEMON can identify trait associations with spatial patterns formed by missense variants both in simulation studies and real-world data.

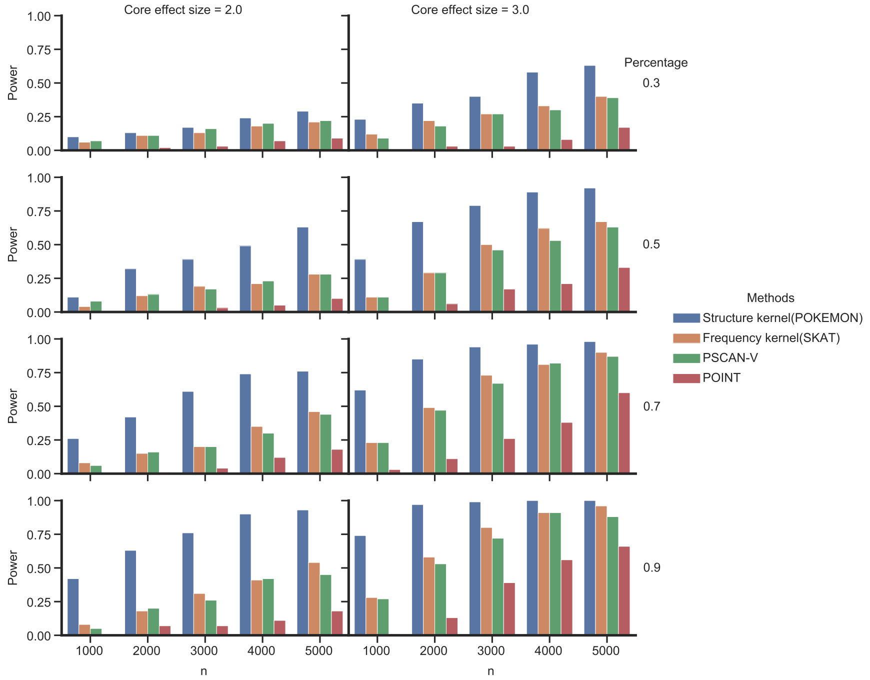

Figure 1. The empirical power for detecting the association between the phenotype and a core pattern on the protein among structure kernel (POKEMON), frequency kernel (SKAT), PSCAN with variance (PSCAN-V), and POINT. The core variant odds ratio is 2.0 or 3.0 (*left* to *right*). The percentage of pathological variants within the selected 50 variants ranges from 0.3 to 0.9 (*top* to *bottom*). The simulated phenotype is calculated based on the core variant odds ratio and the percentage of pathological variants. The empirical power is calculated by the percentage of tests with a *P*-value below the significance level out of 100 replicates.

Results

POKEMON can detect associations with spatially clustered or dispersed rare variants

As a proof of concept, we evaluated the performance of POKEMON using simulations that mimic real-world case/control studies (Supplemental Fig. S1A–D). The simulation data varied in sample sizes, the odds ratio of the core variants, and the proportion of influential to neutral variants. We simulated a cluster pattern of influential variants by establishing a maximum odds ratio decaying over a fixed distance of 7 Å. We limit the number of variants within the genotype profile to 50, the mean number of variants mapped per protein in the ADSP WES discovery data set.

We test the structure kernel approach implemented in POKEMON and compare it to two other structure-informed association methods, PSCAN and POINT (Marceau West et al. 2019; Tang et al. 2020) and a frequency-based kernel analogous to a SKAT analysis of missense variants (Fig. 1). We evaluated the false positive rate (FPR) of POKEMON with the structure kernel and found that the averaged FPR is 0.0455 for all simulated configurations (Supplemental Fig. S2).

Additionally, to evaluate POKEMON's ability to identify a dispersed pattern, we simulated a scenario in which influential variants are distributed on the protein's surface. None of the methods performed well when all influential variants on the surface had small odds ratios. When increasing the odds ratio to 1.5, POKEMON outperformed other methods in all scenarios (Supplemental Fig. S3).

We assessed POKEMON's power at a higher resolution for different core odds ratios and the proportion of influential to neutral variants. Figure 2 illustrates the dynamics of statistical power for the POKEMON test under the assumption of a spatial effect. POKEMON achieved a power of 0.8 with study designs commonly found in sequencing studies of complex disease: a population of 3000 cases/3000 controls, the core odds ratio of 3.0 (decaying to 1 within 7 Å), and 50% of the rare variants influential on the simulated phenotype with moderate effect. However, when the percentage of influential variants is low (<35%) and the core variant odds ratio is small (<1.8), POKEMON cannot reach 80% power. A small core odds ratio and a low percentage of influential variants are more challenging for POKEMON to assess because more control subjects will carry variants within the cluster region, making POKEMON less likely to identify associated patterns.

We further assessed if POKEMON can mitigate the confounding effect from population stratification typically seen in frequency-based tests. We simulated the scenarios from being highly correlated (with 95% subjects with ancestry-matched phenotype) to completely uncorrelated (with 50% subjects with ancestry-matched phenotype). When no covariates are included to adjust for population stratification, we found that tests with Protein Data Bank (PDB) or AlphaFold2 (Senior et al. 2020) structures have lower genomic inflation factors than the corresponding

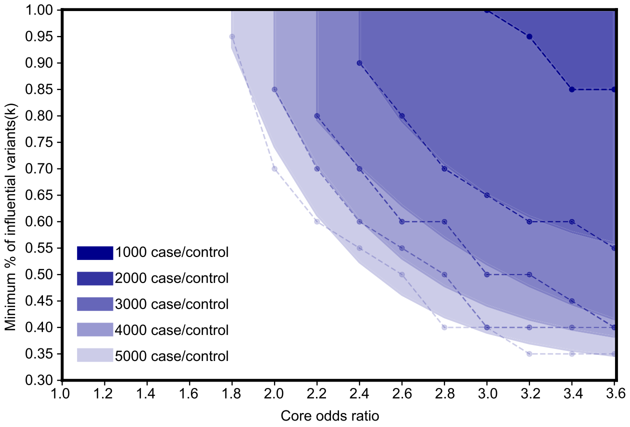

Figure 2. The power assessment for POKEMON with different configurations and the structure kernel. Each dashed line represents the minimum percentages of influential variants and minimum core variant odds ratios required to reach the power of 0.8 when the number of cases/controls is fixed. The empirical power is calculated by the percentage of tests with a *P*-value below the significance level out of 500 replicates. The edge of each shade is the inferred power boundary fit with an exponential function.

frequency tests (Fig. 3). Therefore, we conclude that although the POKEMON test is confounded by ancestry differences, it is less prone to population stratification than a frequency-based test.

POKEMON replicates the cancer-related spatial clusters from the TCGA data set

To show POKEMON's ability to identify spatial patterns from real-world data, we analyzed germline variants from The Cancer Genome Atlas (TCGA), which has previously been used to identify spatial clusters associated with cancer risk and metastasis (Huang et al. 2018). We constructed a case/control data set by combining 8647 subjects from TCGA across 33 cancer types with 4919 presumably cancer-free controls from the ADSP WES discovery data set. We restricted our POKEMON analysis to rare somatic and germline variants only and 31 proteins with functional assessment in the literature. Although the use of population-based controls is not ideal, this proof-of-concept analysis directly tested the hypothesis that cancer-related variants tend to cluster in a protein hotspot, whereas rare variants from cancer-free subjects are randomly distributed. We observed several highly significant associations within the 31 proteins evaluated and enriched significant results (20 with FDR corrected $P < 0.05$) (see Supplemental Table S1A).

From these results, we focus specifically on two genes highlighted in the literature—namely *RET* and *MET* (Table 1; Fig. 4A–F). We found similar clusters of variants for *RET* and *MET*, formed by somatic variants and pathological/likely pathological germline variants (Huang et al. 2018). For *MET*, POKEMON identified a cluster formed by V1088E, P1091L, C1009Y, V1110I, H1112Y, T1114S, F1142L, N1156K (case cluster 0 in Fig. 4B), which is around the pathological variant H1112R and overlap with the hotspot identified in Huang et al. (2018). For *RET*, POKEMON identified two clusters surrounding the pathological variants V804M and I852M (Fig. 4E,F).

POKEMON identified the clusters in *MET* via case/control analysis of rare germline and somatic variants while excluding known pathological variants. Apart from *MET*, we found seven genes (*BLM*, *MSH2*, *PMS2*, *POT1*, *PTPN11*, *TP53*, and *VHL*) with pathological variants identified in Huang et al. (2018) showing significant association even after the pathological variants are excluded (Supplemental Table S1B). Thus, our significant association statistic is driven by additional rare variants within *MET* surrounding those with known pathological effects.

POKEMON identifies known AD risk genes (*TREM2* and *SORL1*) and a novel candidate gene (*EXOC3L4*)

Next, to seek any spatial rare variant patterns associated with AD, we applied POKEMON with structure kernel to the ADSP WES discovery data set with 5522 AD cases and 4919 controls. We performed the POKEMON test on 5969 genes with structures from the PDB and 17,450 with AlphaFold2 predicted structures. All the structures are with five or more rare missense variants mapped ($\text{MAF} < 0.05$). *APOE* ε2 dosages, *APOE* ε4 dosages, PC1, PC2, and sex are included as covariates (model 0). The overall results of our discovery analysis did not show large genomic inflation (GC) in terms of the POKEMON analysis ($\text{GC} = 1.205$ with 5969 PDB structures and $\text{GC} = 1.169$ with 17,450 structures), which is comparable to 1.11 with SKAT-O model in Bis et al. (2020).

We used two significance thresholds to identify candidate genes: a Bonferroni correction threshold and an FDR threshold < 0.2. Overall, four genes meet our significance criteria. *TREM2* was identified with the Bonferroni correction, whereas *SORL1*, *EXOC3L4*, and *TAS2R39* were identified with the FDR threshold (Table 2). Full results with both model 0 and model 1 are in Supplemental Tables S2 and S3. We also note that *CSF1R*, a known dementia-associated gene, falls just below our FDR threshold (Supplemental Fig. S4A,B; Supplemental Table S8).

To determine if the cluster pattern we detected is stable even in the absence of the known associated variants within *SORL1* and *TREM2*, we excluded AD-related variants previously identified in GWAS studies and left only rare genetic variants with unknown effects on AD. A significant result from a POKEMON analysis of these

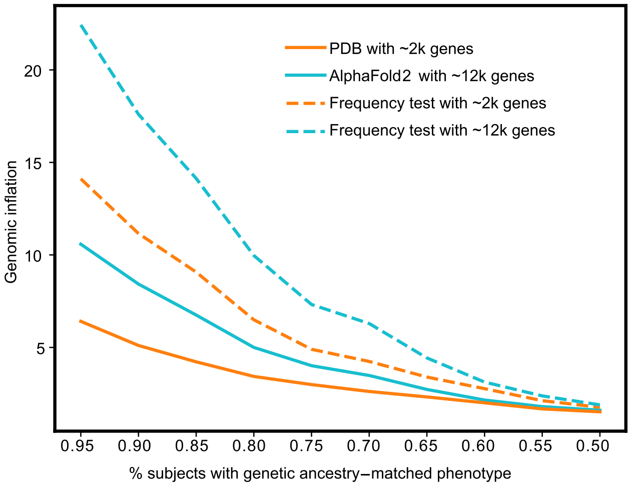

Figure 3. The genomic inflation assessment for POKEMON shows that POKEMON is less prone to population stratification than the frequency test. The genomic inflation is calculated for both approximately 2000 genes with available Protein Data Bank (PDB) structures and about 12,000 with available AlphaFold2 structures. The phenotype is simulated with varying percentages of subjects with genetic ancestry–matched phenotype. The results for frequency tests in the dashed lines are calculated with the same genes with available PDB structures or available AlphaFold2 structures.

Table 1. Results for genes from TCGA data set

Gene	Entry	Phenotype	Number of SNPs mapped	*P*-value	*P*-value (pathological variant excluded)
MET	PDB:1R0P	cancer	17	0.00488	0.00807
RET	PDB:2IVT	cancer	57	0.0174	0.0764

remaining variants indicates that additional rare variants within these genes contribute to AD risk.

Indeed, for *SORL1* (Ensembl: ENSG00000137642; PDB: 3WSY), although AD-related variants A528T (Overall MAC:439; MAF:0.0210), and E270K (Overall MAC:990; MAF:0.0474), respectively, were excluded (Vardarajan et al. 2015), the signals persist. The result indicates that the spatial pattern of variants within the 3WSY structure of *SORL1* is associated with AD (Table 3A). Similarly, for *TREM2* (Ensembl: ENSG00000095970; PDB: 6XDS), the signal persists after variant R47H (Guerreiro et al. 2013; Korvatska et al. 2015) is excluded (Table 3B).

We next tested the four significant genes (*TREM2*, *SORL1*, *EXOC3L4*, and *TAS2R39*) in two additional data sets, the ADSP WGS replication data set and the ADSP validation data set. The results for these four genes with model 0 are shown in Table 4. Additional results with Model0-10PCs, Model1-10PCs can be found in Supplemental Table S4. The ADSP WGS replication is independent of the ADSP WES discovery data set and contains non-Hispanic White, African American, and Hispanic individuals. The ADSP validation data set contains European descent subjects only, of which 9702 subjects are from the ADSP WES discovery data set and 5376 subjects are from the ADSP WGS replication data set. Furthermore, the joint genotype calling approach for the ADSP validation was updated; thus, the ADSP validation data set represents the largest consistently processed and ancestrally homogenous sequencing data set available for AD (Supplemental Fig. S5).

For *TREM2*, the signal regions are shown across three data sets (Fig. 5A–C; Supplemental Table S5). The replicated signal across three data sets contains a region from 16 to 66 amino acids (AA) with multiple variants, including Y38C, T66M, R47H, and R62H. These variants were found correlated with the loss of apo/lipoprotein binding (Yeh et al. 2016). For *SORL1*, we found that the signal regions are only identified in the ADSP WES discovery data set and replicated in the ADSP validation data set (Fig. 6A–C; Supplemental Table S6). One of the signal regions is case cluster 1 in Figure 6A and case cluster 6 in Figure 6C, which overlap with the 10CC-b subunit. The 10CC-b subunit has been found as a dynamic domain with large conformational change when propeptide binds (Kitago et al. 2015). Because the ADSP WES discovery data set and ADSP validation data set have European ancestry subjects only and the ADSP WGS replication data set includes multiancestry subjects, we infer that the signal region identified in *SORL1* is potentially population specific.

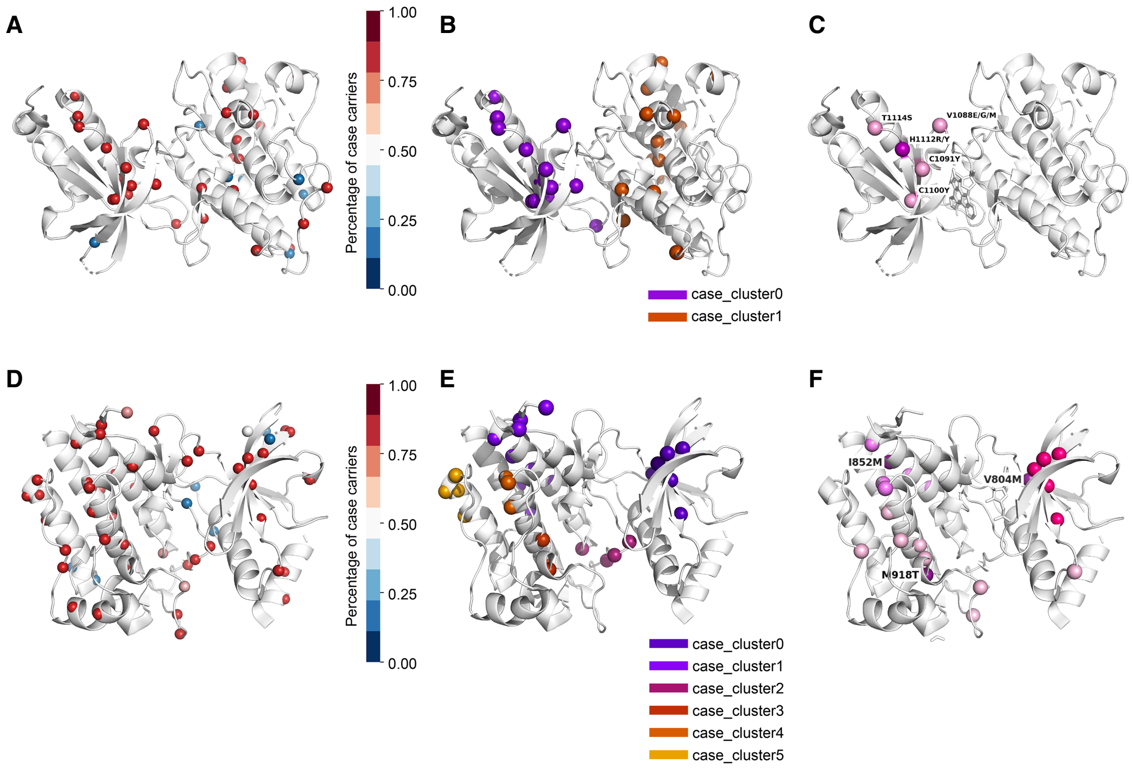

Figure 4. Spatial distribution of variants from TCGA data set within *MET* (PDB:1R0P) and *RET* (PDB:2IVT). (*A*) Rare missense variants mapped to the *MET*. The color scale indicates the percentage of case subjects that carry the variant of all subjects that have this variant. (*B*) Signal regions identified by POKEMON in MET. (*C*) A hotspot formed by germline and somatic variants is identified in Huang et al. (2018). Pathological variant H1112R/Y within the hotspot is highlighted with purple sphere models. (*D*) Rare missense variants mapped to the *RET*. (*E*) *RET* has signal regions identified by POKEMON. (*F*) Three hotspots formed by germline and somatic variants are identified in Huang et al. (2018). Three hotspots surrounding M918T, I852M, and V804M are colored pink, violet, and hot pink, respectively. M918T, I852M, and V804M are highlighted with purple sphere models.

Table 2. Genes associated with AD based on structure kernel

Gene	Entry	Number of SNPs mapped	*P*-value	FDR
PDB structures				
TREM2	PDB:6XDS	33	3.592×10^{-7}	3.351×10^{-5}
SORL1	PDB:3WSY	56	7.022×10^{-5}	6.701×10^{-5}
CSF1R[a]	PDB:4LIQ	38	5.186×10^{-4}	1.005×10^{-4}
AlphaFold2 structures				
TREM2	AlphaFold2: Q9NZC2	33	3.245×10^{-7}	1.146×10^{-5}
EXOC3L4	AlphaFold2: Q17RC7	68	2.504×10^{-5}	2.292×10^{-5}
TAS2R39	AlphaFold2: P59534	31	3.012×10^{-5}	3.438×10^{-5}

[a] *CSF1R* falls just below the FDR threshold.

EXOC3L4 has a case cluster 0 range within 581-670AA in the ADSP WES discovery data set, also shown in the ADSP WGS replication data set as case cluster 3 (577-714AA) and in the ADSP validation data set as case cluster 2 (555-714 AA), as shown in Figure 7A–C and Supplemental Table S7. Although the study of *EXOC3L4* function is limited, *EXOC3L4* belongs to the Sec6 protein family, and its C-terminal region is structurally and topologically similar to the Sec6 domain. The case cluster identified above overlaps with Sec6 domain, specifically the D and E regions (Fig. 8A–C), forming the exocyst complex. The exocyst complex involves multiple cellular processes, including exocytosis and cell growth cytokinesis, cell migration, and tumorigenesis (Miller et al. 2018). Miller et al. (2018) found rare variants in the splicing regulatory elements of *EXOC3L4* are associated with brain glucose metabolism measured by FDG PET-scans. Although the splice variant found by Miller et al. (2018) helps skip the second exon of *EXOC3L4*, which is the N terminal of the Sec6 domain, our finding is a cluster of case variants located in the C terminal of the Sec6 domain. Our results suggest alterations to the Sec6 domain of *EXOC3L4* may increase AD risk.

Discussion

We have shown that POKEMON improves the power to detect rare variant gene association in the context of protein structure. We found POKEMON outperforms other structure-based methods through simulation studies except in a small number of cases in which all existing methods have insufficient power. We applied POKEMON to the ADSP Discovery WES data set and identified spatial patterns of rare variants related to AD risk in two known AD genes: *SORL1* and *TREM2*. We also identified a potentially novel AD-associated cluster of variants within *EXOC3L4*, located around the C-terminal end of the Sec6 domain. Specifically, the cluster within *EXOC3L4* is validated both in the ADSP WGS replication and ADSP validation data sets.

An advantage for POKEMON over other rare variant analysis methods is that statistical power increases with the observation of any new variant, including singletons, assuming the existence of spatial patterns. In most rare variant association tests, increasing sample size only increases the power for nonsingleton variants in the resulting data. Even for those nonsingleton variants, the improvement in power is not necessarily proportional to the increased sample size. Moreover, additional neutral variants will be introduced, negatively impacting the statistical power when the sample size increases. In contrast, POKEMON can use rare variants and even singletons with its structure kernel, regardless of their low allele frequency. The increasing number of rare variants helps form the spatial pattern, which can be identified by POKEMON with a higher power (Supplemental Fig. S1D). We also showed that the spatial patterns are not driven by a single variant but rather a collection of variants with modest effects by excluding variants with known effects for *TREM2* and *SORL1* in the ADSP WES discovery data set.

POKEMON is designed to leverage preexisting biological information for sequencing data sets in which only variant counts or frequencies are typically considered. Although protein structure information of variants has been incorporated into association tests like POINT and PSCAN (Marceau West et al. 2019; Tang et al. 2020), they serve as guiding information for more traditional association tests ultimately based on allele frequency. Therefore, these approaches are still potentially subject to the limitations in unit-based or single variant tests. With the structure kernel, POKEMON uses the spatial information of a missense variant, which is independent of allele frequency. Assuming the rare variants form spatial patterns, POKEMON mitigates the power issue induced by increasing numbers of singleton variants as the sample size of sequencing studies increases.

We anticipate POKEMON will be helpful as a large-scale screening method to detect potentially disease-associated proteins in a proteome-wide fashion under the hypothesis that influential rare variants have a spatial pattern within protein structures. Currently, available protein structures deposited in the PDB only cover a small portion of the identified molecular functions in the human genome (Somody et al. 2017). We expect that the improvement in cryo-EM and advances in protein prediction methods like AlphaFold2 (Senior et al. 2020) will massively increase the availability and quality of structural information for proteins

Table 3A. Results for *SORL1* with and without known loci in ADSP WES discovery

Gene	PDB Entry	*P*-value
SORL1	PDB:3WSY	7.022×10^{-5}
SORL1 (exclude A528T)		1.481×10^{-2}
SORL1		7.022×10^{-5}
SORL1 (exclude E270K)		6.34×10^{-5}

Table 3B. Results for *TREM2* with and without known locus in ADSP WES discovery

Gene	PDB Entry	*P*-value
TREM2	PDB: 6XDS	3.592×10^{-7}
TREM2 (exclude R47H)		4.535×10^{-4}

Table 4. Results for candidate genes from the replication data sets

		ADSP WES discovery Overall: 10,441 Case/control: 5522/4919		ADSP WGS replication Overall: 7762 Case/control: 3757/4005		ADSP validation Overall: 15,078 Case/control: 8294/6784	
Gene	Entry	Number of SNPs mapped	*P*-value	Number of SNPs mapped	*P*-value	Number of SNPs mapped	*P*-value
TREM2	PDB:6XDS	33	3.592×10^{-7}	21	3.027×10^{-2}	33	1.484×10^{-8}
SORL1	PDB:3WSY	56	7.022×10^{-5}	53	0.16	82	0.00361
CSF1R[a]	PDB:4LIQ	38	5.186×10^{-4}	38	0.108	46	5.594×10^{-3}
TREM2	AlphaFold2: Q9NZC2	33	3.245×10^{-7}	21	1.194×10^{-2}	33	2.594×10^{-7}
EXOC3L4	AlphaFold2: Q17RC7	68	2.504×10^{-5}	83	5.104×10^{-3}	90	9.221×10^{-5}
TAS2R39	AlphaFold2: P59534	31	3.012×10^{-5}	27	0.522	37	0.0247

[a] *CSF1R* falls just below the FDR threshold.

and complexes. A key feature of POKEMON is to test if the structure kernel explains part of the variance of the phenotype; therefore, POKEMON only provides a single association statistic for the influence of missense variants within the protein on the phenotype. Follow-up analyses to assess specific variants or refine variant subsets may provide more detailed quantitative assessments of specific variant spatial patterns.

Methods

Derivation of the POKEMON method

We briefly review the linear mixed model used in association tests and then introduce the construction of a structure kernel for POKEMON. Assume we have n individuals for whom we have p nongenetic covariates, genotypes for m SNPs, and the

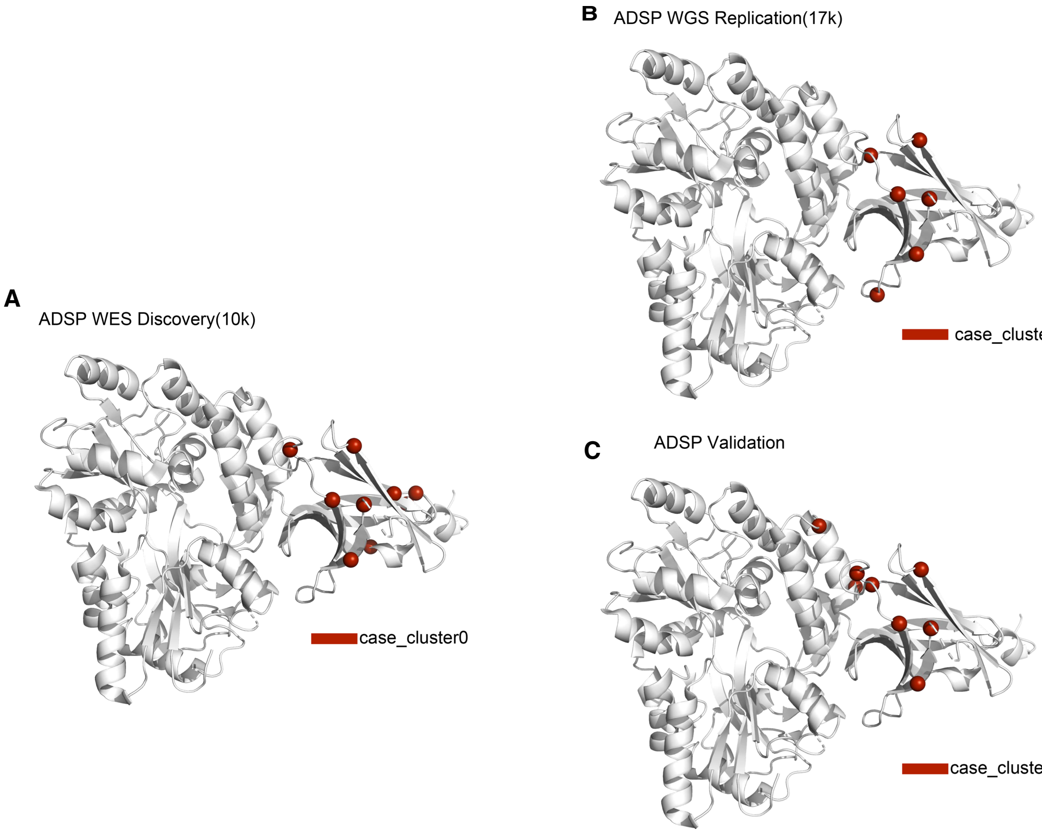

Figure 5. *TREM2* has the signal region identified in the ADSP WES discovery data set (*A*) and replicated both in the ADSP WGS replication (*B*) and the ADSP validation (*C*) data sets. The signal cluster is identified in the POKEMON test with the DBSCAN algorithm. All variants within the clusters are rare variants with MAF < 0.05. Clusters classified as case clusters are formed by variants carried primarily by AD subjects, and clusters classified as control clusters are formed by variants carried primarily by cognitively normal subjects. Variants assigned with a cluster label are shown, but all the other variants are not shown in the figure.

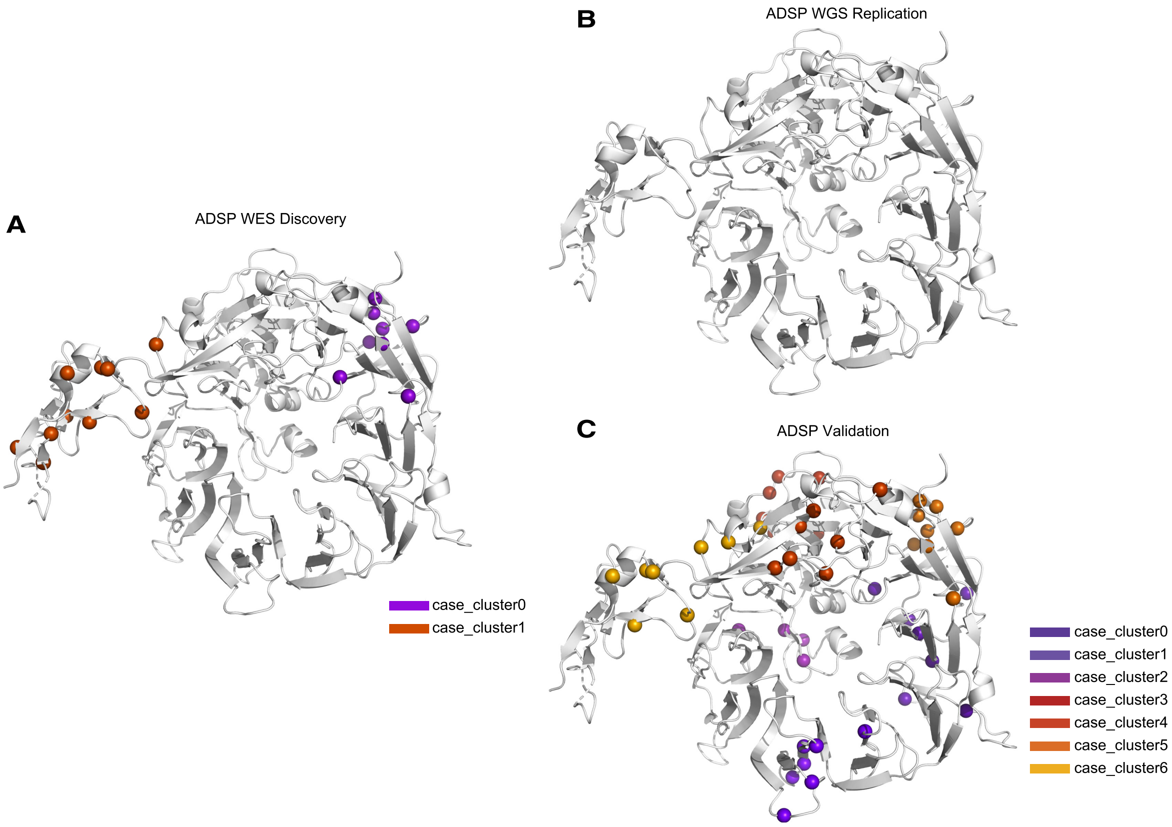

Figure 6. *SORL1* has a signal region identified in the ADSP WES discovery data set (*A*) and replicated in the ADSP validation data set (*C*) but not in the ADSP WGS replication data set (*B*). The signal cluster is identified in the POKEMON test with the DBSCAN algorithm. All variants within the clusters are rare variants with MAF < 0.05. Clusters classified as case clusters are formed by variants carried primarily by AD subjects and clusters classified as control clusters are formed by variants carried primarily by cognitively normal subjects. Variants assigned with a cluster label are shown, but all the other variants are not shown in the figure.

phenotype. Phenotype $\boldsymbol{y}$ is a $n \times 1$ vector. Genotype $\boldsymbol{G}$ is a $n \times m$ matrix. Covariate $\boldsymbol{X}$ is a $n \times p$ matrix.

A linear mixed model contains a fixed effect from covariates $\boldsymbol{X\beta}$, a random effect annotated by $\boldsymbol{Gu}$ with $\boldsymbol{u}$ being the unknown vector of random effects, and an unknown vector of random errors ε(Equation 1a). The $\boldsymbol{y}$ is fit with a high-dimension normal distribution (Equation 1b). The random effect contains two parts—namely, an environmental effect $\sigma_e^2 I$ and a genetic effect $\sigma_1^2\boldsymbol{K_g}$. $\boldsymbol{K_g}$ is the kernel containing the genetic similarity between individuals, and σ_1^2 is the amount of variance of y explained by $\boldsymbol{K_g}$. The null hypothesis $\sigma_1 = 0$ indicates that the $\boldsymbol{K_g}$ does not explain any variance of y.

$$y_{\boldsymbol{i}} = \boldsymbol{X_i\beta} + \boldsymbol{G_i u} + \epsilon_i \tag{1a}$$

$$\boldsymbol{y}\,\tilde{}\,N(\boldsymbol{X\beta},\ \sigma_1^2\boldsymbol{K_g} + \sigma_e^2\boldsymbol{I}) \tag{1b}$$

For continuous traits, the null model is a linear regression with covariates only:

$$\widehat{y_i} = \boldsymbol{X_i\beta} + \epsilon_i \tag{2}$$

$\hat{\boldsymbol{y}}$ is the vector with the ith value equivalent to $\widehat{y_i}$, so the score statistic Q is defined as:

$$\frac{\mathrm{Q}}{\sigma_e^2} = \boldsymbol{y^T S K_g S y} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T \boldsymbol{K_g}(\boldsymbol{y} - \hat{\boldsymbol{y}}) \tag{3}$$

Similarly, for dichotomous traits, the null model is a logistic regression with covariates only. $\widehat{\pi_i}$ is the estimated probability for $\boldsymbol{y_i} = 1$ under the null model.

$$\widehat{\pi_i} = \mathrm{logit}^{-1}(\boldsymbol{X_i\beta}) \tag{4}$$

$\hat{\boldsymbol{y}}$ is the vector with the ith value equivalent to $\widehat{y_i}$, so the score statistic Q is defined as:

$$\frac{\mathrm{Q}}{\sigma_e^2} = \boldsymbol{y^T S K_g S y} = (\boldsymbol{y} - \hat{\boldsymbol{\pi}})^T \boldsymbol{K_g}(\boldsymbol{y} - \hat{\boldsymbol{\pi}}) \tag{5}$$

Under the null hypothesis, Q follows a mixed χ^2 distribution (Equation 6), where $\boldsymbol{S}$ projects y into a space orthogonal to covariates, and λ_i are the eigenvalues of $\boldsymbol{SK_gS}$.

$$\frac{\mathrm{Q}}{\sigma_e^2}\,\tilde{}\,\sum_{i=1}^{n}\lambda_i\chi_1^2 \tag{6}$$

For POKEMON, we construct the $n \times n$ kernel $\boldsymbol{K_g}$ in the context of protein as follows: For $\boldsymbol{K_g}$, each entry is the genetic similarity between individuals based on the variants they carry, which is weighted by the variant's distance in the protein structure (Equation 7), where d_{kl} is the distance of pairwise single-nucleotide variants (SNVs) in angstroms (Å) within the protein, and k and l represent the kth variant from individual i and the lth variants from individual j.

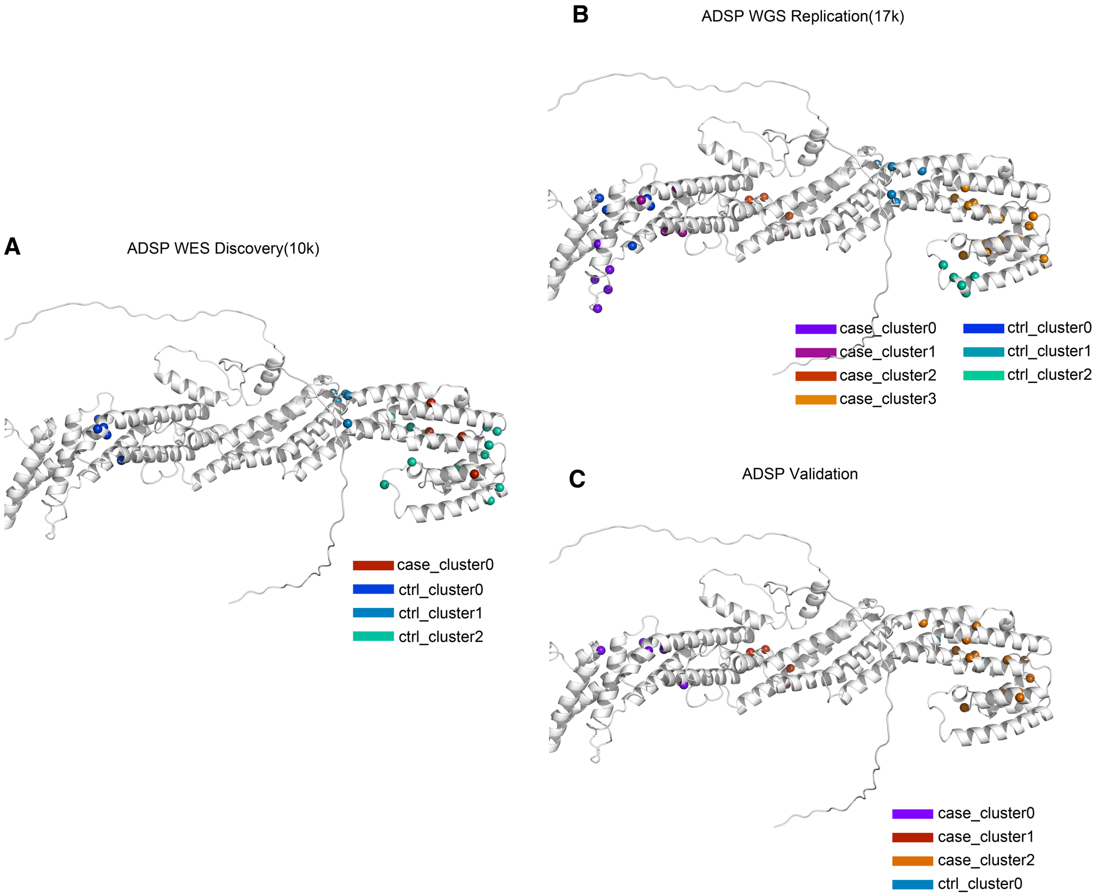

Figure 7. Signal regions on *EXOC3L4* (AlphaFold2: Q17RC7.A) are identified by POKEMON from the ADSP WES discovery data set (*A*) and validated in both the ADSP WGS replication (*B*) and the ADSP validation (*C*) data sets. The signal regions are identified in the POKEMON test with the DBSCAN algorithm. All variants within the clusters are rare variants with MAF < 0.05. Clusters classified as case clusters are formed by variants carried primarily by AD subjects and clusters classified as control clusters are formed by variants carried primarily by cognitively normal subjects.

$$K_{ij} = \sum_{i_k j_l} A_k A_l \min\{f(d_{kl})\}. \qquad (7)$$

Some protein structures are formed by identical subunits (homo-multimer), which introduces redundancy in the variant-to-amino acid projection (i.e., one variant can map to multiple amino acids located in different subunits). To eliminate the spatial similarity induced by multiple mapping locations of a single variant in a homo-multimer, we took d_{kl} to be the minimum distance among all pairwise distances. Function *f(d)* converts a Euclidean distance to the similarity score for a pair of variants.

$$f(d_{kl}) = e^{-\frac{d_{kl}^2}{2t^2}}. \qquad (8)$$

As a default, the exponential function for f is in Equation 8 with t set to a value of 14 Å; and 14 Å is a commonly adopted short-range nonbonded cutoff in molecular dynamic simulation (Monticelli et al. 2008).

Apart from spatial patterns, we also account for the magnitude of the protein change resulting from the different amino acid substitutions. We scaled the pairwise variants by their amino acid substitution, which is defined as A_k and A_l. A_k and A_l are the weights for amino acid substitution for variant k and variant l according to the BLOSUM62 matrix (Henikoff and Henikoff 1992), respectively. For a less conservative amino acid substitution, the score s_k in BLOSUM62 matrix will be negative; consequently, A_k will be greater than 1. In contrast, for a neutral or conservative amino acid substitution, s_k will be positive and A_k will be less than 1.

$$A_k = -\sqrt{e^{s_k}}. \qquad (9)$$

The structure kernel is nonlinear in contrast to the SKAT tests (Wu et al. 2011), which uses a linear kernel (e.g., $\boldsymbol{K} = \boldsymbol{GWW'G'}$) to calculate the genetic similarity between individuals. The genetic similarity in a linear kernel between individuals is the sum of weighted SNVs being shared. However, singletons are carried by only a single individual and thus fail to be included in calculating genetic similarity. With the structure kernel, a pair of singleton variants will be assigned non-zero weights if they are spatially proximate in the protein structure. The interpretation of the structure kernel is that case individuals are genetically similar because they share more spatially clustered or dispersed rare variants than the control individuals.

We also allow for incorporating allele frequency in the POKEMON test by a combined kernel function. One can consider that variants clustered in protein structure already contribute to a high genetic similarity based on structure kernel. With the combined kernel, those variants will be further up-weighted if they are rare in allele frequency and vice versa. The combined kernel function is based on the $\boldsymbol{K_g}$ and extended by further scaling variants by weights derived from the allele frequency. w_k = Beta

(MAF$_k$;a, b) is the weight for the kth variant characterized by beta density with $a = 1$ and $b = 25$ as default.

$$K_{ij} = \sum_{i_k j_l,\, k \neq l} A_k A_l \min\{f(d_{kl})\} + \sum_{i_k j_k} w_k A_k^2. \qquad (10)$$

The power of the frequency-based SKAT test is sensitive to the choice of beta weights. Therefore, although the default beta weights are generally acceptable, we suggest evaluating the beta weights based on the frequency distribution in the data of interest and selecting the optimal beta weights for a combined kernel (Chen et al. 2018).

POKEMON workflow

An overview of the POKEMON workflow is shown in Supplemental Figure S6. POKEMON requires a genotype matrix and consequence profile containing variant-to-amino acid mapping information as inputs. Additional covariate files are optional to adjust for covariates. POKEMON first maps the variants by the coordinates into the protein, which is accomplished with the consequence profile generated by Ensembl Variant Effect Predictor (VEP v95) and the reference from SIFTS mapping PDB entry to UniProt residue level (Dana et al. 2019). A single variant may be mapped to multiple amino acids for multimers with identical subunits. The protein structures are fetched from PDB during the analysis. If multiple protein structures are available for a single gene, the structure with the most variants mapped will be selected. However, if a PDB entry is given, POKEMON also allows the analysis of a specified protein structure. After mapping, the score between a pair of variants is calculated based on the minimum distance between them, which is further scaled by the amino acid substitution weight from the BLOSUM62 matrix by default. The pairwise genetic similarity between individuals is the summation of all pairwise scores of variants. The genetic similarity kernel $\boldsymbol{K_g}$ will be evaluated in the variance component test.

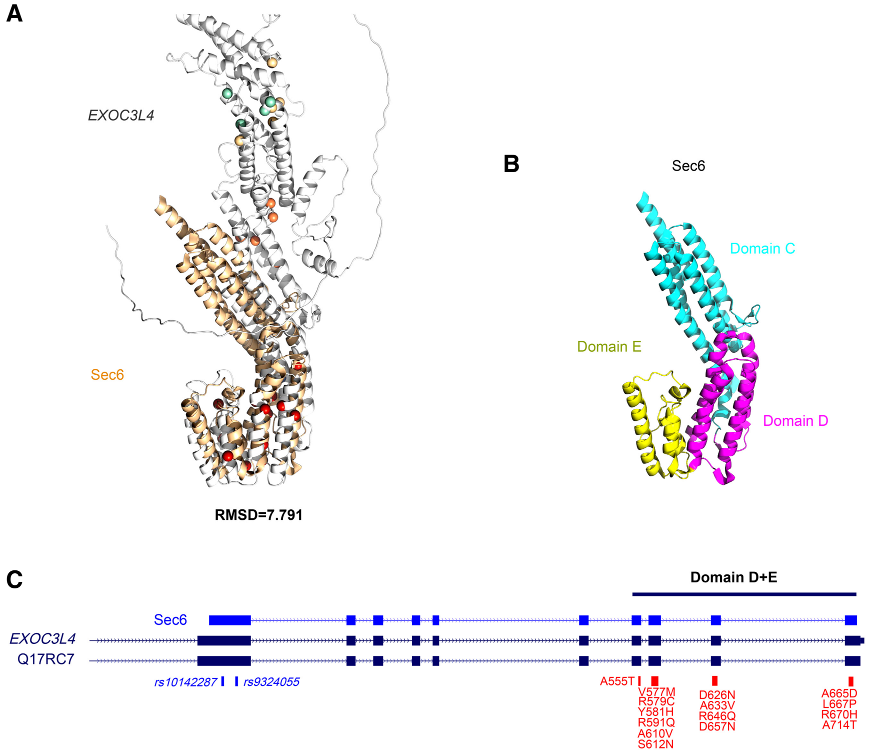

Figure 8. Sec6 domain in *EXOC3L4* contains a cluster of variants primarily carried by AD case individuals. (*A*) Alignment of the *EXOC3L4* and Sec6. The structure for *EXOC3L4* is from AlphaFold2 with entry Q17RC7, and the structure for Sec6 is PDB:2FJI. The alignment is performed with PyMOL. (*B*) The structure for the C-terminal domain of Sec6 is formed by three domains C, D, and E. (*C*) The genomic coordinate of Sec6, *EXOC3L4*, the splicing variants from Miller et al. (2018), and variants from case cluster 2 in Figure 7C. The splicing variants are colored blue and labeled with dbSNP Reference SNP number. The variants from case cluster 2 in Figure 7C are colored in red and labeled with the amino acid change.

Data simulation

Simulation strategy for power assessment

We conducted simulation studies to assess POKEMON's power in detecting disease-associated protein variant patterns. We hypothesized that variants with moderate effects on a phenotype form spatial patterns within a protein structure and alter the protein's function. To test the hypothesis, we established two patterns. The first pattern entails an embedded core within the protein disrupted by rare variants (i.e., variant clustering), whereas the other represents the localization of influential variants to the protein's surface (i.e., variant dispersion). Both patterns are shown in Supplemental Figure S1A,B. We randomly selected a protein PDB:2OGV to carry out simulations because the structural information for PDB:2OGV is available for both PSCAN and SKAT.

We simulated a clustering pattern by distributing influential variants within the core of the protein structure and scaling the variant odds ratios proportionally to their distance from the core. We then randomly sampled 50 variants from the protein. The minor allele frequencies for all the variants were randomly sampled from a log-transformed uniform distribution within an interval (−4, −2.3). This variant sampling strategy restricted the selected minor allele frequencies within the range (0.0001, 0.005) and generated singletons, which is consistent with ADSP WES studies (Supplemental Fig. S7). To investigate how neutral variants influence the power, we varied the percentage of influential variants out of all sampled variants (Supplemental Fig. S1D). For each set of parameters (e.g., sample size, core variant odds ratio, etc.), the empirical power was estimated by the percentage of successful tests out of 100 independent tests with a significance level of 0.05. We compared the empirical power of POKEMON with three other methods: SKAT, PSCAN-V, and POINT. The number of case and control subjects sampled is from 1000 to 5000. Additional details for the simulation can be found in Supplemental Figure S1A–D and Supplemental Methods.

We also simulated a dispersion pattern by distributing influential variants on the protein's surface. Considering the selected protein PDB:2OGV is about 40 Å in diameter, we defined the surface variants as those >21 Å away from the core, which yielded 33 variants. All the surface variants were assigned with the same odds ratio (e.g., 1.1), whereas the rest were considered neutral with an odds ratio of 1. The simulation settings were similar to the clustering pattern, with the only difference that we sampled 30 variants from the protein, which allowed us to tune the percentage of influential variants to as large as 90%.

Simulation strategy for genomic inflation assessment

We selected 671 subjects identified as African ancestry and 522 as European ancestry from the 1000 Genomes Project. When we chose different percentages of subjects with genetic ancestry–matched phenotype *r*, *r*% of the European ancestry subjects will be assigned a phenotype equivalent to 1. In contrast, the rest within European ancestry subjects will be assigned a phenotype equivalent to 0. Similarly, *r*% of the African ancestry subjects will be assigned a phenotype equivalent to 0, and the rest within African ancestry subjects will be assigned a phenotype equivalent to 1. Then we will test this phenotype with 2719 available protein structures from PDB and 13,691 structures from AlphaFold2. The genomic inflation is calculated for PDB structures and AlphaFold2 structures, respectively.

Applying POKEMON to ADSP data

The ADSP WES discovery data set, ADSP WGS replication data set used in this study is available at ADSP (https://www.niagads.org/adsp/content/home). An application to the NIAGADS Data Sharing Service is needed to access the data.

The model we used for all the results in Tables 1–4 is model 0, which adjusted for *APOE* ε2 and ε4 dosages, PC1, PC2, and sex. In model 0, *APOE* ε2 and ε4 dosages are to exclude signals induced by the well-known *APOE* association. PC1 and PC2 are included to avoid false positive signals owing to population structure.

We also evaluated other models that included additional covariates and all the results are in the Supplemental Tables. Model 1 adjusted for *APOE* ε2 and ε4 dosages, PC1, PC2, sex, and age at diagnosis or last follow-up. Model 0-10PCs adjusted for *APOE* ε2 and ε4 dosages, PC1-10, and sex. Model 1-10PCs adjusted for APOE ε2 and ε4 dosages, PC1-10, sex, and age.

ADSP WES discovery data set

We used the whole-exome sequencing (WES) data from the discovery phase case-control study under the Alzheimer's Disease Sequencing Project (ADSP). ADSP WES data contains 5740 late-onset AD cases and 5096 cognitively normal controls primarily of European ancestry, with 218 cases and 177 controls of Caribbean Hispanic ancestry. Cases were determined based on diagnosis using cognitive testing data and medical records, and controls were determined on their low risk of developing AD by age 85 yr (Beecham et al. 2017; Bis et al. 2020).

We selected 10,441 subjects of European ancestry from the ADSP as the study group (5522 late-onset AD cases and 4919 cognitively normal controls) and shown in Supplemental Figure S5. We retained the missense variants with minor allele frequency < 0.05 for our assessment. Overall, we selected 5969 genes with experimentally determined protein structures and 17,450 with AlphaFold2 predicted structures, all of which have five or more rare missense variants mapped to the structure. The mean number of rare missense variants mapped to the PDB structure per gene was approximately 50.

ADSP WGS replication data set

We used the whole-genome sequencing (WGS) data from the Alzheimer's Disease Sequencing Project (ADSP) as the replication data set. ADSP WGS contains 3757 AD cases and 4005 cognitively normal controls. Within these 7762 samples, 5375 are non-Hispanic White, 1571 are African American, and 803 are of Hispanic, Asian, or Native American ancestry (Supplemental Fig. S5). All the subjects in the ADSP WGS replication data set are independent of those in the ADSP WES discovery data set.

ADSP validation data set

The validation data set contains the 9702 subjects from the discovery phase case-control study plus an additional 5375 subjects from the replication data set for a total of 15,078 non-Hispanic White subjects (Supplemental Fig. S5). The WES data for the 9702 subjects were reprocessed using joint genotype calling approaches implemented in the VCPA pipeline (Leung et al. 2019), which were updated from the ATLAS genotype calling process implemented for the ADSP WES discovery data set. Therefore, we consider that this validation data set is valuable by expanding the sample size for a genetically homogenous population group and accounting for the variability in the variant calling process.

Applying POKEMON to TCGA data

The TCGA data is a real-world, true-positive example of spatial patterns of missense variants associated with phenotypes (Kamburov et al. 2015). To create a data set in the form of a case-control study, we combined 4919 control subjects from the ADSP WES discovery data set and 8647 subjects from TCGA data diagnosed with 33 cancer types (Huang et al. 2018). We assumed that 4919 cognitive normal control subjects from the ADSP WES discovery data set are cancer-free controls. Although this is not an ideal study design, any violation of this assumption would reduce statistical power rather than identifying spurious associations. The combined case/control data set provided a real-world assessment of our hypothesis that rare variants from cancer tissues would form spatial patterns. In contrast, those from control subjects would be randomly distributed within the protein.

Both germline and somatic variants from the TCGA are included. Moreover, we set a stringent MAF threshold as <0.01 to retain rare variants. In summary, we performed POKEMON tests on 31 genes with potential hotspots (Huang et al. 2018) and available protein structures with no covariate included.

Software availability

The code for this study is available as Supplemental Code and at GitHub (https://github.com/bushlab-genomics/POKEMON).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported in part by the National Institute of General Medical Sciences (NIGMS), grant 5R01GM126249-03 (W.S.B., D.C.C.), and National Institute on Aging (NIA) grants 5U01AG058654-03 (W.S.B., L.A.F., E.R.M., M.A.P.-V., J.L.H.) and 1RFAG061351-01 (W.S.B., A.C.N., J.E. Below).

The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI)-funded Large-Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC), funded by National Institute on Aging (NIA) (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National

Institutes of Health (NIH) institutes, and other foreign governmental and nongovernmental organizations. The discovery phase analysis of sequence data is supported through UF1AG047133 (to L.A.F., J.L.H., E.R.M., M.A.P.-V., and G.D.S.); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the discovery extension phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to M.A.P.-V.; and U01 AG052409 to Drs. Seshadri and Fornage.

Sequencing for the Follow Up Study (FUS) is supported through U01AG057659 (to M.A.P.-V., R.M., and B.V.) and U01A G062943 (to M.A.P.-V. and R.M.). Data generation and harmonization in the Follow-up Phase is supported by U54AG052427 (to G.D.S. and L.-S.W.). The FUS Phase analysis of sequence data is supported through U01AG058589 (to Drs. Destefano, Boerwinkle, De Jager, Fornage, Seshadri, and Wijsman), U01AG058654 (to Drs. J.L.H., W.S.B., L.A.F., E.R.M., and M.A.P.-V.), U01AG058635 (to Dr. Goate), RF1AG058066 (to J.L.H., M.A.P.-V., and Dr. Scott), RF1AG057519 (to L.A.F. and Dr. Jun), R01AG048927 (to L.A.F.), and RF1AG054074 (to M.A.P.-V. and Dr. Beecham).

The ADGC cohorts include: Adult Changes in Thought (ACT) (UO1 AG006781, UO1 HG004610, UO1 HG006375, U01 HG008657), the Alzheimer's Disease Centers (ADC) (P30 AG019610, P30 AG013846, P50 AG008702, P50 AG025688, P50 AG047266, P30 AG010133, P50 AG005146, P50 AG005134, P50 AG016574, P50 AG005138, P30 AG008051, P30 AG013854, P30 AG008017, P30 AG010161, P50 AG047366, P30 AG010129, P50 AG016573, P50 AG016570, P50 AG005131, P50 AG023501, P30 AG035982, P30 AG028383, P30 AG010124, P50 AG005133, P50 AG005142, P30 AG012300, P50 AG005136, P50 AG033514, P50 AG005681, and P50 AG047270), the Chicago Health and Aging Project (CHAP) (R01 AG11101, RC4 AG039085, K23 AG030944), Indianapolis Ibadan (R01 AG009956, P30 AG010133), the Memory and Aging Project (MAP) (R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, P50 AG016574), Mayo Parkinson's Disease controls (NS039764, NS071674, 5RC2HG005605), University of Miami (R01 AG027944, R01 AG028786, R01 AG019085, IIRG09133827, A2011048), the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE) (R01 AG09029, R01 AG025259), the National Cell Repository for Alzheimer's Disease (NCRAD) (U24 AG21886), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA-LOAD) (R01 AG041797), the Religious Orders Study (ROS) (P30 AG10161, R01 AG15819), the Texas Alzheimer's Research and Care Consortium (TARCC) (funded by the Darrell K Royal Texas Alzheimer's Initiative), Vanderbilt University/Case Western Reserve University (VAN/CWRU) (R01 AG019757, R01 AG021547, R01 AG027944, R01 AG028786, P01 NS026630, and Alzheimer's Association), the Washington Heights–Inwood Columbia Aging Project (WHICAP) (RF1 AG054023), the University of Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS: R01NS069719), the Columbia University HispanicEstudio Familiar de Influencia Genetica de Alzheimer (EFIGA) (RF1 AG015473), the University of Toronto (UT) (funded by Wellcome Trust, Medical Research Council, Canadian Institutes of Health Research), and Genetic Differences (GD) (R01 AG007584). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Dr. Psaty), RC2HL102419 (Dr. Boerwinkle), and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193.

The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904), the EU Joint Programme—Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435). ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA, and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607, and AG033040) and NINDS (R01NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam; the Netherlands Organization for Health Research and Development (ZonMw); the Research Institute for Diseases in the Elderly (RIDE); the Ministry of Education, Culture and Science; the Ministry for Health, Welfare and Sports; the European Commission (DG XII); and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies thank their participants, faculty, and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services.

The FUS cohorts include: the Alzheimer's Disease Centers (ADC) (P30 AG019610, P30 AG013846, P50 AG008702, P50 AG025688, P50 AG047266, P30 AG010133, P50 AG005146, P50

AG005134, P50 AG016574, P50 AG005138, P30 AG008051, P30 AG013854, P30 AG008017, P30 AG010161, P50 AG047366, P30 AG010129, P50 AG016573, P50 AG016570, P50 AG005131, P50 AG023501, P30 AG035982, P30 AG028383, P30 AG010124, P50 AG005133, P50 AG005142, P30 AG012300, P50 AG005136, P50 AG033514, P50 AG005681, and P50 AG047270), Alzheimer's Disease Neuroimaging Initiative (ADNI) (U19AG024904), Amish Protective Variant Study (RF1AG058066), Cache County Study (R01AG11380, R01AG031272, R01AG21136, RF1AG054052), Case Western Reserve University Brain Bank (CWRUBB) (P50AG008012), Case Western Reserve University Rapid Decline (CWRURD) (RF1AG058267, NU38CK000480), CubanAmerican Alzheimer's Disease Initiative (CuAADI) (3U01AG052410), Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA) (5R37AG015473, RF1AG015473, R56AG051876), Genetic and Environmental Risk Factors for Alzheimer Disease Among African Americans Study (GenerAAtions) (2R01AG09029, R01AG025259, 2R01AG048927), Gwangju Alzheimer and Related Dementias Study (GARD) (U01AG062602), Hussman Institute for Human Genomics Brain Bank (HIHGBB) (R01AG027944, Alzheimer's Association "Identification of Rare Variants in Alzheimer Disease"), Ibadan Study of Aging (IBADAN) (5R01AG009956), Mexican Health and Aging Study (MHAS) (R01AG018016), Multi-Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE) (2R01AG09029, R01AG025259, 2R01AG048927), Northern Manhattan Study (NOMAS) (R01NS29993), Peru Alzheimer's Disease Initiative (PeADI) (RF1AG054074), Puerto Rican 1066 (PR1066) (Wellcome Trust [GR066133/GR080002], European Research Council [340755]), Puerto Rican Alzheimer Disease Initiative (PRADI) (RF1AG054074), Reasons for Geographic and Racial Differences in Stroke (REGARDS) (U01NS041588), Research in African American Alzheimer Disease Initiative (REAAADI) (U01AG052410), Rush Alzheimer's Disease Center (ROSMAP) (P30AG10161, R01AG15819, R01AG17919), University of Miami Brain Endowment Bank (MBB), and University of Miami/Case Western/North Carolina A&T African American (UM/CASE/NCAT) (U01AG052410, R01AG028786).

The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079).

Biological samples and associated phenotypic data used in primary data analyses were stored at study investigator institutions and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated phenotypic data used in primary and secondary data analyses were provided by the study investigators, the NIA funded Alzheimer's Disease Centers (ADCs), the National Alzheimer's Coordinating Center (NACC, U01AG016976); and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. Contributors to the Genetic Analysis Data included the study investigators on projects that were individually funded by NIA and other NIH institutes, and by private US organizations, foreign governmental organizations, or nongovernmental organizations. We also acknowledge the following investigators who assembled and characterized participants of cohorts included in this study.

Adult Changes in Thought: James D. Bowen, Paul K. Crane, Gail P. Jarvik, C. Dirk Keene, Eric B. Larson, W. William Lee, Wayne C. McCormick, Susan M. McCurry, Shubhabrata Mukherjee.

Katie Rose Richmire Atherosclerosis Risk in Communities Study: Rebecca Gottesman, David Knopman, Thomas H. Mosley, B. Gwen Windham.

Austrian Stroke Prevention Study: Thomas Benke, Peter Dal-Bianco, Edith Hofer, Gerhard Ransmayr, Yasaman Saba.

Cardiovascular Health Study: James T. Becker, Joshua C. Bis, Annette L. Fitzpatrick, M. Ilyas Kamboh, Lewis H. Kuller, W.T. Longstreth, Jr, Oscar L. Lopez, Bruce M. Psaty, Jerome I. Rotter.

Chicago Health and Aging Project: Philip L. De Jager, Denis A. Evans.

Erasmus Rucphen Family Study: Hieab H. Adams, Hata Comic, Albert Hofman, Peter J. Koudstaal, Fernando Rivadeneira, Andre G. Uitterlinden, Dina Voijnovic.

Estudio Familiar de la Influencia Genetica en Alzheimer: Sandra Barral, Rafael Lantigua, Richard Mayeux, Martin Medrano, Dolly Reyes-Dumeyer, Badri Vardarajan.

Framingham Heart Study: Alexa S. Beiser, Vincent Chouraki, Jayanadra J. Himali, Charles C. White.

Genetic Differences: Duane Beekly, James Bowen, Walter A. Kukull, Eric B. Larson, Wayne McCormick, Gerard D. Schellenberg, Linda Teri.

Mayo Clinic: Minerva M. Carrasquillo, Dennis W. Dickson, Nilufer Ertekin-Taner, Neill R. Graff-Radford, Joseph E. Parisi, Ronald C. Petersen, Steven G. Younkin.

Mayo PD: Gary W. Beecham, Dennis W. Dickson, Ranjan Duara, Nilufer Ertekin-Taner, Tatiana M. Foroud, Neill R. Graff-Radford, Richard B. Lipton, Joseph E. Parisi, Ronald C. Petersen, Bill Scott, Jeffery M. Vance.

Memory and Aging Project: David A. Bennett, Philip L. De Jager.

Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study: Sanford Auerbach, Helan Chui, Jaeyoon Chung, L. Adrienne Cupples, Charles DeCarli, Ranjan Duara, Martin Farlow, Lindsay A. Farrer, Robert Friedland, Rodney C.P. Go, Robert C. Green, Patrick Griffith, John Growdon, Gyungah R. Jun, Walter Kukull, Alexander Kurz, Mark Logue, Kathryn L. Lunetta, Thomas Obisesan, Helen Petrovitch, Marwan Sabbagh, A. Dessa Sadovnick, Magda Tsolaki.

National Cell Repository for Alzheimer's Disease: Kelley M. Faber, Tatiana M. Foroud.

National Institute on Aging (NIA) Late Onset Alzheimer's Disease Family Study: David A. Bennett, Sarah Bertelsen, Thomas D. Bird, Bradley F. Boeve, Carlos Cruchaga, Kelley Faber, Martin Farlow, Tatiana M. Foroud, Alison M. Goate, Neill R. Graff-Radford, Richard Mayeux, Ruth Ottman, Dolly Reyes-Dumeyer, Roger Rosenberg, Daniel Schaid, Robert A. Sweet, Giuseppe Tosto, Debby Tsuang, Badri Vardarajan.

NIA Alzheimer Disease Centers: Erin Abner, Marilyn S. Albert, Roger L. Albin, Liana G. Apostolova, Sanjay Asthana, Craig S. Atwood, Lisa L. Barnes, Thomas G. Beach, David A. Bennett, Eileen H. Bigio, Thomas D. Bird, Deborah Blacker, Adam Boxer, James B. Brewer, James R. Burke, Jeffrey M. Burns, Joseph D. Buxbaum, Nigel J. Cairns, Chuanhai Cao, Cynthia M. Carlsson, Richard J. Caselli, Helena C. Chui, Carlos Cruchaga, Mony de Leon, Charles DeCarli, Malcolm Dick, Dennis W. Dickson, Nilufer Ertekin-Taner, David W. Fardo, Martin R. Farlow, Lindsay A. Farrer, Steven Ferris, Tatiana M. Foroud, Matthew P. Frosch, Douglas R. Galasko, Marla Gearing, David S. Geldmacher, Daniel H. Geschwind, Bernardino Ghetti, Carey Gleason, Alison M. Goate, Teresa Gomez-Isla, Thomas Grabowski, Neill R. Graff-Radford, John H. Growdon, Lawrence S. Honig, Ryan M. Huebinger, Matthew J. Huentelman, Christine M. Hulette, Bradley T. Hyman, Suman Jayadev, Lee-Way Jin, Sterling Johnson, M. Ilyas Kamboh, Anna Karydas, Jeffrey A. Kaye, C. Dirk Keene, Ronald Kim, Neil W. Kowall, Joel H. Kramer, Frank M. LaFerla, James J. Lah, Allan I. Levey, Ge Li, Andrew

P. Lieberman, Oscar L. Lopez, Constantine G. Lyketsos, Daniel C. Marson, Ann C. McKee, Marsel Mesulam, Jesse Mez, Bruce L. Miller, Carol A. Miller, Abhay Moghekar, John C. Morris, John M. Olichney, Joseph E. Parisi, Henry L. Paulson, Elaine Peskind, Ronald C. Petersen, Aimee Pierce, Wayne W. Poon, Luigi Puglielli, Joseph F. Quinn, Ashok Raj, Murray Raskind, Eric M. Reiman, Barry Reisberg, Robert A. Rissman, Erik D. Roberson, Howard J. Rosen, Roger N. Rosenberg, Martin Sadowski, Mark A. Sager, David P. Salmon, Mary Sano, Andrew J. Saykin, Julie A. Schneider, Lon S. Schneider, William W. Seeley, Scott Small, Amanda G. Smith, Robert A. Stern, Russell H. Swerdlow, Rudolph E. Tanzi, Sarah E. Tomaszewski Farias, John Q. Trojanowski, Juan C. Troncoso, Debby W. Tsuang, Vivianna M. Van Deerlin, Linda J. Van Eldik, Harry V. Vinters, Jean Paul Vonsattel, Jen Chyong Wang, Sandra Weintraub, Kathleen A. Welsh-Bohmer, Shawn Westaway, Thomas S. Wingo, Thomas Wisniewski, David A. Wolk, Randall L. Woltjer, Steven G. Younkin, Lei Yu, Chang-En Yu.

Religious Orders Study: David A. Bennett, Philip L. De Jager.

Rotterdam Study: Kamran Ikram, Frank J. Wolters.

Texas Alzheimer's Research and Care Consortium: Perrie Adams, Alyssa Aguirre, Lisa Alvarez, Gayle Ayres, Robert C. Barber, John Bertelson, Sarah Brisebois, Scott Chasse, Munro Culum, Eveleen Darby, John C. DeToledo, Thomas J. Fairchild, James R. Hall, John Hart, Michelle Hernandez, Ryan Huebinger, Leigh Johnson, Kim Johnson, Aisha Khaleeq, Janice Knebl, Laura J. Lacritz, Douglas Mains, Paul Massman, Trung Nguyen, Sid O'Bryant, Marcia Ory, Raymond Palmer, Valory Pavlik, David Paydarfar, Victoria Perez, Marsha Polk, Mary Quiceno, Joan S. Reisch, Monica Rodriguear, Roger Rosenberg, Donald R. Royall, Janet Smith, Alan Stevens, Jeffrey L. Tilson, April Wiechmann, Kirk C. Wilhelmsen, Benjamin Williams, Henrick Wilms, Martin Woon.

University of Miami: Larry D. Adams, Gary W. Beecham, Regina M. Carney, Katrina Celis, Michael L. Cuccaro, Kara L. Hamilton-Nelson, James Jaworski, Brian W. Kunkle, Eden R. Martin, Margaret A. Pericak-Vance, Farid Rajabli, Michael Schmidt, Jeffery M Vance.

University of Toronto: Ekaterina Rogaeva, Peter St. George-Hyslop.

University of Washington Families: Thomas D. Bird, Olena Korvatska, Wendy Raskind, Chang-En Yu.

Vanderbilt University: John H. Dougherty, Harry E. Gwirtsman, Jonathan L. Haines.

Washington Heights-Inwood Columbia Aging Project: Adam Brickman, Rafael Lantigua, Jennifer Manly, Richard Mayeux, Christiane Reitz, Nicole Schupf, Yaakov Stern, Giuseppe Tosto, Badri Vardarajan.

References

Beecham GW, Bis JC, Martin ER, Choi SH, DeStefano AL, van Duijn CM, Fornage M, Gabriel SB, Koboldt DC, Larson DE, et al. 2017. The Alzheimer's Disease Sequencing Project: study design and sample selection. *Neurol Genet* **3:** e194. doi:10.1212/NXG.0000000000000194

Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, Salerno WJ, Lancour D, Ma Y, Renton AE, et al. 2020. Whole exome sequencing study identifies novel rare and common Alzheimer's-associated variants involved in immune response and transcriptional regulation. *Mol Psychiatry* **25:** 1859–1875. doi:10.1038/s41380-018-0112-7

Butkiewicz M, Blue EE, Leung YY, Jian X, Marcora E, Renton AE, Kuzma A, Wang LS, Koboldt DC, Haines JL, et al. 2018. Functional annotation of genomic variants in studies of late-onset Alzheimer's disease. *Bioinformatics* **34:** 2724–2731. doi:10.1093/bioinformatics/bty177

Chen Z, Lu Y, Lin T, Liu Q, Wang K. 2018. Gene-based genetic association test with adaptive optimal weights. *Genet Epidemiol* **42:** 95–103. doi:10.1002/gepi.22098

Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, Velankar S. 2019. SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res* **47:** D482–D489. doi:10.1093/nar/gky1114

Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, Cruchaga C, Sassi C, Kauwe JSK, Younkin S, et al. 2013. *TREM2* variants in Alzheimer's disease. *N Engl J Med* **368:** 117–127. doi:10.1056/NEJMoa1211851

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci* **89:** 10915–10919. doi:10.1073/pnas.89.22.10915

Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al. 2018. Pathogenic germline variants in 10,389 adult cancers. *Cell* **173:** 355–370.e14. doi:10.1016/j.cell.2018.03.039

Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, Lander ES, Getz G. 2015. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci* **112:** E5486–E5495. doi:10.1073/pnas.1516373112

Kitago Y, Nagae M, Nakata Z, Yagi-Utsumi M, Takagi-Niidome S, Mihara E, Nogi T, Kato K, Takagi J. 2015. Structural basis for amyloidogenic peptide recognition by sorLA. *Nat Struct Mol Biol* **22:** 199–206. doi:10.1038/nsmb.2954

Korvatska O, Leverenz JB, Jayadev S, McMillan P, Kurtz I, Guo X, Rumbaugh M, Matsushita M, Girirajan S, Dorschner MO, et al. 2015. R47h variant of *TREM2* associated with Alzheimer disease in a large late-onset family: clinical, genetic, and neuropathological study. *JAMA Neurol* **72:** 920–927. doi:10.1001/jamaneurol.2015.0979

Leung YY, Valladares O, Chou YF, Lin HJ, Kuzma AB, Cantwell L, Qu L, Gangadharan P, Salerno WJ, Schellenberg GD, et al. 2019. VCPA: genomic variant calling pipeline and data management tool for Alzheimer's Disease Sequencing Project. *Bioinformatics* **35:** 1768–1770. doi:10.1093/bioinformatics/bty894

Marceau West R, Lu W, Rotroff DM, Kuenemann MA, Chang SM, Wu MC, Wagner MJ, Buse JB, Motsinger-Reif AA, Fourches D, et al. 2019. Identifying individual risk rare variants using protein structure guided local tests (POINT). *PLoS Comput Biol* **15:** e1006722. doi:10.1371/journal.pcbi.1006722

Miller JE, Shivakumar MK, Lee Y, Han S, Horgousluoglu E, Risacher SL, Saykin AJ, Nho K, Kim D, for the Alzheimer's Disease Neuroimaging Initiative. 2018. Rare variants in the splicing regulatory elements of EXOC3L4 are associated with brain glucose metabolism in Alzheimer's disease. *BMC Med Genomics* **11:** 76. doi:10.1186/s12920-018-0390-6

Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ. 2008. The MARTINI coarse-grained force field: extension to proteins. *J Chem Theory Comput* **4:** 819–834. doi:10.1021/ct700324x

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* **577:** 706–710. doi:10.1038/s41586-019-1923-7

Sivley RM, Sheehan JH, Kropski JA, Cogan J, Blackwell TS, Phillips JA, Bush WS, Meiler J, Capra JA. 2018. Three-dimensional spatial analysis of missense variants in *RTEL1* identifies pathogenic variants in patients with familial interstitial pneumonia. *BMC Bioinformatics* **19:** 18. doi:10.1186/s12859-018-2010-z

Somody JC, MacKinnon SS, Windemuth A. 2017. Structural coverage of the proteome for pharmaceutical applications. *Drug Discov Today* **22:** 1792–1799. doi:10.1016/j.drudis.2017.08.004

Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al. 2021. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* **590:** 290–299. doi:10.1038/s41586-021-03205-y

Tang ZZ, Sliwoski GR, Chen G, Jin B, Bush WS, Li B, Capra JA. 2020. PSCAN: spatial scan tests guided by protein structures improve complex disease gene discovery and signal variant detection. *Genome Biol* **21:** 217. doi:10.1186/s13059-020-02121-0

Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, Masica DL, Karchin R. 2016. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res* **76:** 3719–3731. doi:10.1158/0008-5472.CAN-15-3190

Vardarajan BN, Zhang Y, Lee JH, Cheng R, Bohm C, Ghani M, Reitz C, Reyes-Dumeyer D, Shen Y, Rogaeva E, et al. 2015. Coding mutations in *SORL1* and Alzheimer disease. *Ann Neurol* **77:** 215–227. doi:10.1002/ana.24305

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89:** 82–93. doi:10.1016/j.ajhg.2011.05.029

Yeh FL, Wang Y, Tom I, Gonzalez LC, Sheng M. 2016. TREM2 binds to apolipoproteins, including APOE and CLU/APOJ, and thereby facilitates uptake of amyloid-β by microglia. *Neuron* **91:** 328–340. doi:10.1016/j.neuron.2016.06.015

Received August 9, 2021; accepted in revised form February 17, 2022.

CSH PRESS

GENOME RESEARCH

An association test of the spatial distribution of rare missense variants within protein structures identifies Alzheimer's disease–related patterns

Bowen Jin, John A. Capra, Penelope Benchek, et al.

Genome Res. 2022 32: 778-790 originally published online February 24, 2022
Access the most recent version at doi:10.1101/gr.276069.121