



31 have not had sufficient power to examine rare non-coding effects that incorpo-  
32 rate such cell-type specific information. To address this gap, we 1) use predicted  
33 enhancer and promoter regions in microglia and other potentially relevant cell  
34 types (oligodendrocytes, astrocytes, and neurons) to define per-gene non-coding  
35 RV test sets and 2) include cell-type specific variant effect predictions (VEPs)  
36 as functional annotations. `gruyere` identifies 15 significant genetic associations  
37 not detected by other RV methods and finds deep learning-based VEPs for splic-  
38 ing, transcription factor binding, and chromatin state are highly predictive of  
39 functional non-coding RVs. Our study establishes a novel and robust framework  
40 incorporating functional annotations, coding RVs, and cell-type associated non-  
41 coding RVs, to perform genome-wide association tests, uncovering AD-relevant  
42 genes and annotations.

43 **Keywords:** Rare variants, Alzheimer’s Disease, Bayesian probabilistic model,  
44 whole-genome sequencing

## 45 1 Main

46 The recent increase in available whole-genome sequencing (WGS) data has facilitated  
47 the study of rare variants (RVs), particularly in understanding their effects on complex  
48 diseases like late-onset Alzheimer’s disease (AD). AD is a neurodegenerative disorder  
49 with an estimated heritability between 59% and 74% [1]. While genome-wide associ-  
50 ation studies (GWAS) have identified over 100 loci linked to AD, with the *APOE-ε4*  
51 allele as the strongest genetic risk factor, they are restricted to common variant asso-  
52 ciations [2, 3]. Despite considerable efforts to quantify the polygenic nature of AD, a  
53 significant portion of genetic heritability remains unaccounted for. Some of this miss-  
54 ing heritability may be recovered with RVs [4]. RVs generally exhibit larger effect sizes  
55 than common variants, but their role is not yet well understood [5]. Studies have shown  
56 that integrating RVs into cumulative polygenic risk scores (PRS) can enhance predic-  
57 tive performance [6], but existing methods have identified fewer gene associations and  
58 have lower predictive power compared to common variant approaches. While a num-  
59 ber of genes, including *TREM2*, *ABCA7* and *SORL1* [7], have known RV associations  
60 in AD, the majority of these findings are restricted to coding variants. As most GWAS  
61 signals lie in the non-coding genome, expanding RV association studies beyond coding  
62 variants is critical. However, the study of non-coding RVs poses challenges due to the  
63 vast number of these variants, most of which likely have no functional impact [8]. It is  
64 therefore of substantial interest to use functional annotations for variant filtering and  
65 prioritization. To develop a more robust understanding of the contributions of both  
66 coding and non-coding RVs to AD, we propose a novel method that not only weights  
67 variants according to annotations but also prioritizes functional annotations that are  
68 most trait-relevant.

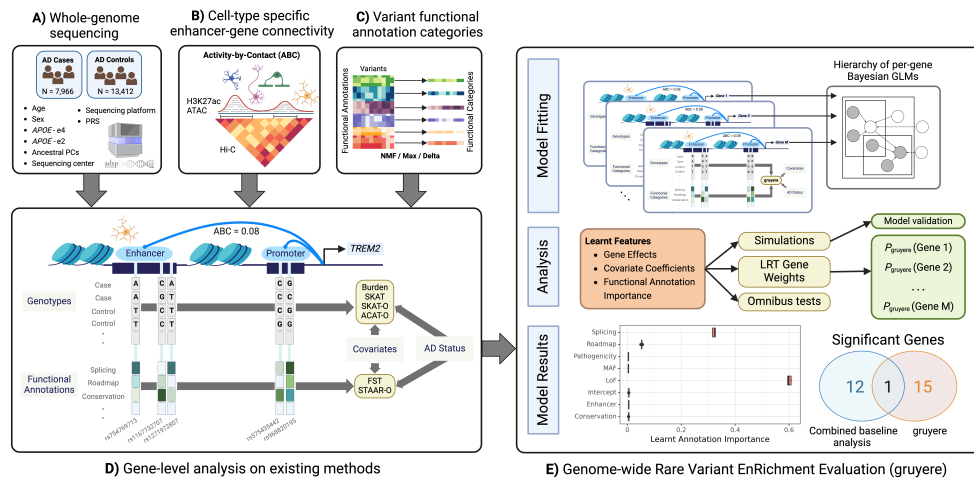
69 Applying traditional variant-level approaches like GWAS to RVs has low statistical  
70 power due to sparsity and a high multiple testing burden due to the large number of  
71 RVs compared to common variants. To address these limitations, RV methods aggre-  
72 gate variants in biologically related regions, typically by gene, to increase power [9].

73 More recent RV methods additionally account for functional annotations to prioritize  
74 relevant variants and filter out those predicted to have no function, which otherwise  
75 reduce power [10, 11]. Despite growing efforts to accurately predict which variants will  
76 affect particular molecular phenotypes (e.g., enhancer activation, RNA splicing) [12–  
77 16], there is a limited understanding of *which functions* are the most disease-relevant.  
78 Using functional annotations that have no phenotypic associations to weight RVs can  
79 add noise to models and decrease their power. This motivated us to develop a method  
80 that learns a genome-wide mapping from functional annotations to variant importance.

81 Growing evidence suggests that disrupted gene regulation in central nervous system  
82 (CNS) cell types, particularly microglia, is associated with the development and  
83 progression of AD [1, 17]. The majority of RV tests are developed for coding variant  
84 associations because 1) predicting functional coding variants is comparatively straight-  
85 forward (at least for loss-of-function), 2) population-scale whole-exome sequencing  
86 predates WGS, and 3) defining non-coding regions for testing is challenging in itself.  
87 Some methods use sliding windows, but testing overlapping windows of varying sizes  
88 can result in loss of power due to multiple testing [18]. Other methods use predicted  
89 *cis*-regulatory elements (CREs), in particular enhancers and promoters, to construct  
90 testing regions [19–21]. Given their modest size (typically less than 2kb), testing indi-  
91 vidual CREs still has limited statistical power. Combining multiple CREs that regulate  
92 a gene could help address this limitation but relies on accurate predictions of enhancer-  
93 gene links. We leverage the Activity-by-Contact (ABC) model, which predicts cell-type  
94 specific enhancer-gene connectivity using chromatin state and conformation data [22].  
95 We aggregate ABC-predicted enhancer-gene pairs to determine non-coding, cell-type  
96 and gene specific RV testing regions.

97 Due to the large number of genes and several million RVs found in population-scale  
98 WGS, existing methods are primarily restricted to per-gene models. This limits our  
99 understanding of disease-associated functional annotations. Most existing RV meth-  
100 ods are explicitly, or can be viewed as, generalized linear mixed models (GLMM). We  
101 instead develop a Bayesian generalized linear model (GLM) Genome-wide Rare Vari-  
102 ant EnRichment Evaluation, or **gruyere**, to model cell-type specific, non-coding RV  
103 associations on a genome-wide scale. In **gruyere**, a variant’s effect is a deterministic  
104 function of its annotations and the estimated AD-relevance of the gene it is linked  
105 to (if any). Our Bayesian model iteratively learns AD-relevant gene effects, covariate  
106 weights, and functional annotation importance while quantifying uncertainty, provid-  
107 ing increased flexibility to capture the complex, hierarchical structure of genetic data.  
108 We test our model using simulation analyses and compare results to several existing RV  
109 methods. We apply **gruyere** to WGS data from the Alzheimer’s Disease Sequencing  
110 Project (ADSP), consisting of 7,966 cases and 13,412 controls. Our model determines  
111 splicing, transcription factor (TF) binding, and chromatin state annotations most  
112 enriched for AD-associated non-coding RVs and identifies 16 significant genes, 15 of  
113 which are uniquely identified by **gruyere**. Of these, four – *C9orf78*, *MAF1*, *NUP93*,  
114 and *GALNT9* – remain significant in omnibus tests.

## 115 2 Results



**Fig. 1 Overview of the application of gruyere to AD.** Input data includes **A)** WGS and clinical information for AD cases and controls, **B)** Enhancer-gene interactions predicted by the ABC model for microglia, oligodendrocytes, astrocytes, and neurons, and **C)** variant functional annotations. **D)** Example analysis for the *TREM2* gene and microglia cell-type on existing methods. Columns represent RVs; light grey rectangles represent individual-level genotypes from WGS data for cases and controls; functional annotations for each RV are shown below genotypes; Burden, SKAT, SKAT-O, and ACAT-O are existing tests that use genotype, covariate, and AD status; FST and STAAR-O additionally use functional annotations. **E)** Workflow for gruyere. Per-gene RVs are aggregated and used for fitting the hierarchical Bayesian GLM. gruyere learns weights for covariates, genes, and functional annotations. We use simulations to assess gruyere at different heritabilities. Likelihood ratio tests are used to calculate gene-level *p*-values. Optionally, the gruyere *p*-values can be integrated with existing methods through omnibus testing.

### 116 2.1 Genome-wide Rare Variant EnRichment Evaluation 117 (gruyere) overview

**Table 1** Summary of gruyere Variables

Variable	Shape	Description
$Y$	$n \times 1$	Phenotypes for $n$ samples
$X$	$n \times c$	$c$ Covariates for $n$ samples
$G_g$	$n \times p_g$	Genotypes for $p$ variants in gene $g$ for $n$ samples
$Z_g$	$p_g \times q$	$q$ Functional annotations for $p$ variants in gene $g$
$\alpha_g$	$c \times 1$	$c$ Covariate coefficients
$\beta_{gj}$	$p_g \times 1$	Variant effect sizes for $p$ variants in gene $g$
$w_j$	$p_g \times 1$	Variant weight $\text{Beta}(MAF_j 1, 25)$
$w_g$	$1 \times 1$	Gene importance weights
$\tau$	$q \times 1$	Functional annotation weight



Current RV methods rely on independent per-gene models and, therefore, cannot capture genome-wide functional annotation importance. **gruyere** serves as a complementary method to existing RV tests by learning trait-specific functional annotation weights, covariate coefficients, and gene effects under a Bayesian framework (Supplemental Figure 1, Table 1). Rather than modeling each gene separately, we jointly fit **gruyere** as a hierarchy of per-gene GLMs using stochastic variational inference (SVI) [23]. We model AD risk for each gene  $g$ ,

$$\text{logit}(\mu_{ig}) = X_i\alpha_g + G_{ig}\beta_{gj}$$

where  $\mu_{ig}$  is the probability of AD for individual  $i$  given the genotypes for RVs associated with gene  $g$ ,  $X_i$  is a vector of covariates (e.g. sex, age, *APOE-ε4* genotype), and  $G_{ig}$  is a genotype vector. We learn covariate weights  $\alpha_g$  and variant effects  $\beta_{gj}$ . We set  $\beta_{gj}$  to be a deterministic function of a learned gene effect  $w_g$ , transformed minor allele frequencies (MAFs)  $w_j$ , functional annotations  $Z$  (detailed in Methods 3.4 and Figure 1C), and learned annotation importance weights  $\tau$ ,

$$\beta_{gj} = w_g w_j (\tau_0 + \sum_{k=1}^q Z_{gjk} \tau_k).$$

118 In our analyses, **gruyere** learns annotation weights  $\tau$  for a range of annotations  $Z$ ,  
119 including in silico mutagenesis deep learning model predictions of splicing disruption  
120 (derived as the maximum of four individual SpliceAI scores [24]) and cell-type specific  
121 TF binding and chromatin state (derived from the Enformer model [14]). A larger  
122 magnitude of  $w_g$  indicates that disruption of gene  $g$  is associated with a higher pre-  
123 dicted risk of AD. Similar to a burden test, **gruyere** assumes all variants within a gene  
124 have the same direction of effect [25, 26]. However, because our functional annotations  
125 include both loss- and gain-of-function predictions, we are able to capture additional  
126 dispersion-based signal. To ensure robust generalization of learned parameters, we  
127 split data into training (80%) and test (20%) sets, where model weights are optimized  
128 using the training set and assessed on the unseen test set. We apply **gruyere** to both  
129 coding and non-coding RVs for AD, defining four cell-type specific non-coding groups  
130 for AD-relevant cell types (microglia, oligodendrocytes, astrocytes, and neurons [17])  
131 and testing each group individually.

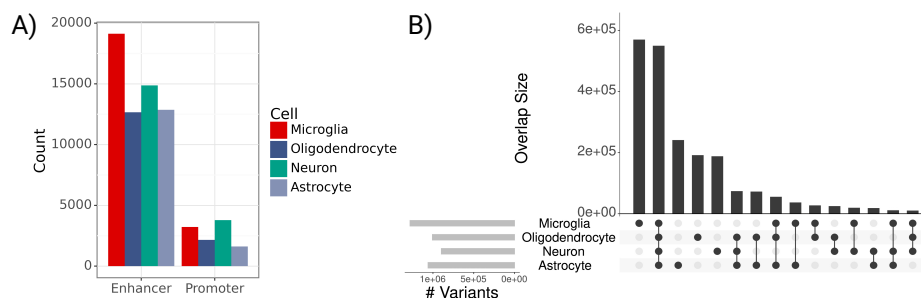
### 132 **Step 1. Estimating global annotation weights $\tau$ .**

133 Fitting  $\tau$  jointly across the entire genome would be 1) computationally challenging  
134 due to the large number of RVs and 2) statistically inefficient, as only AD-associated  
135 genes will contribute relevant signal. We therefore estimate  $\tau$  under the **gruyere** model  
136 from a subset of potentially AD-relevant genes identified using a lenient significance  
137 threshold (nominal  $p < 0.01$ ) for the Functional Score Test (FST)[27]. We assess  
138 the robustness of **gruyere** estimates when selecting genes with varied significance  
139 thresholds and for a number of existing RV tests and find annotation weights  $\tau$  are  
140 broadly consistent ( $+/- 0.02$ ).

141 **Step 2. Per-gene analysis.**

142 Once genome-wide estimates for  $\tau$  are obtained, **gruyere** simplifies to a logistic  
 143 regression that learns covariate  $\alpha_g$  and gene  $w_g$  weights. We efficiently fit **gruyere**  
 144 separately and in parallel for all genes, holding  $\tau$  fixed. We perform likelihood ratio  
 145 tests (LRT) to compare a covariate-only regression against combined covariate and  
 146 genotype regression models to determine gene-level significance for  $w_g$ .

147 **2.2 Constructing cell-type & gene specific variant sets using**  
 148 **predicted CREs**



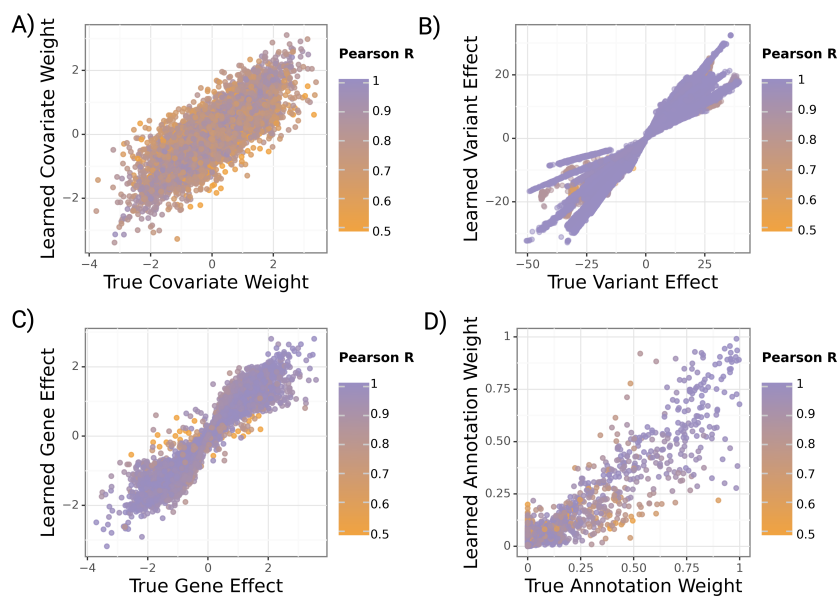
**Fig. 2 Predicted regulatory element and variant counts across cell types. A)** Bar plot of predicted CRE counts by cell type ( $ABC > 0.02$ ). **B)** Upset plot of variant overlap across 4 cell types in ADSP data; Light grey bars on left indicate total RV counts for each cell type; Vertically connected dots represents groups and corresponding bars indicate variant overlap for that group.

149 We grouped non-coding RVs by gene and CNS cell type using the Activity-by-  
 150 Contact (ABC) model (Figure 1B) [22]. The ABC model uses epigenomic profiles and  
 151 chromatin conformation to determine cell-type specific enhancer-gene interactions,  
 152 filtering out genes that are not expressed. We use publicly available ATAC-seq and  
 153 H3K27ac ChIP-seq signals for microglia, oligodendrocytes, astrocytes, and neurons  
 154 [28], as well as Hi-C averaged across ten CNS cell types to account for 3-dimensional  
 155 chromatin interactions. For each gene, we analyze RVs aggregated across all CREs  
 156 interacting with that gene ( $ABC > 0.02$ ). We test each cell type separately and also  
 157 analyze rare coding variants for comparison. In total, ABC defines 70,300 CREs across  
 158 all four cell types and 17,929 genes, with higher relative counts of microglia-predicted  
 159 enhancers (Figure 2A). Predicted CREs frequently co-occur across cell types, with  
 160 39.4% of CREs found in more than one cell type. Promoter regions tend to have higher  
 161 ABC scores than enhancers (mean  $ABC = 0.07$  vs. mean  $ABC = 0.04$ ), but their  
 162 genomic lengths are similar, with an average length of 632bp and standard deviation  
 163 of 132bp. ABC accounts for interactions of a single enhancer with multiple genes so  
 164 one RV can be linked to multiple genes. In our analysis, an ABC enhancer maps to  
 165 an average of 3.8 genes in microglia and between 5.4 and 5.9 genes in the other three  
 166 cell types. Our non-coding variant sets contain an average of 376 RVs per gene. There  
 167 are a total of 2,092,931 RVs in CREs across the four cell types, 901,570 of which are

168 included for more than one cell type, and 550,001 that are in all four cell types (Figure  
169 2B).

### 170 2.3 Simulation studies confirm accurate estimation of model 171 parameters

172 We generate synthetic phenotypes (see Methods 3.5) and fit `gruyere` on 100 sets of  
173 simulated data with estimated heritability between 5% and 30% (detailed in Supple-  
174 mental Methods [29]) using 500 randomly selected genes. We find that all variables  
175 are well recovered, with average Pearson correlations  $R = 0.81, 0.95, 0.98, 0.97$  for  
176  $\alpha_g, \beta_{gj}, w_g$  and  $\tau$  respectively (Figure 3). Covariates  $\alpha_g$  have the lowest  $R$ , possibly  
177 due to correlated covariates. Average recovery across all variables remains high when  
178 varying the prior distributions (Pearson  $R > 0.78$ ) as well as when simulated distri-  
179 butions differ from the priors used during inference (Pearson  $R > 0.66$ ). Results are  
180 robust to the number of covariates, genes, and annotations modeled.

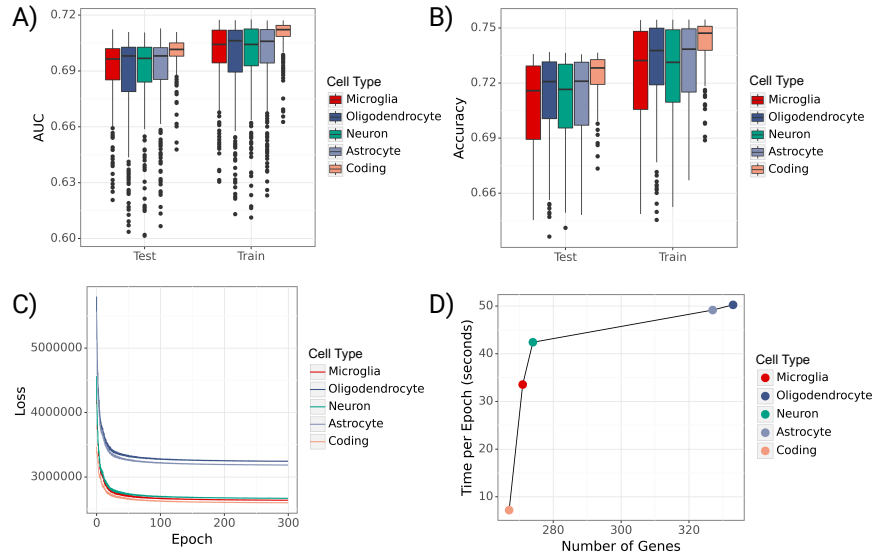


**Fig. 3** Learned versus true `gruyere` parameters across 100 simulations. Points are colored by the Pearson correlation coefficient of a parameter for a given simulation. **A)** Covariate regression coefficients  $\alpha$  ( $c = 30$  covariates). **B)** Variant effect  $\beta$ . **C)** Gene effect  $w_g$  ( $M = 500$  randomly selected genes). **D)** Annotation weight  $\tau$  ( $q = 13$  annotations).

181 We analyze how simulation performance correlates with overall and genetic her-  
182 itability for each simulation. This allows us to more meaningfully evaluate model  
183 performance for complex diseases like AD where estimated heritability is low. As  
184 expected, we find that `gruyere` is better able to recover  $\beta_{gj}$  and  $w_g$  with increased  
185 genetic heritability (Supplemental Figure 2). However, even when total heritability

186 is as low as 5%, the minimum correlation between true and estimated parameters  
 187 remains quite high (Pearson  $R = 0.68$ ).

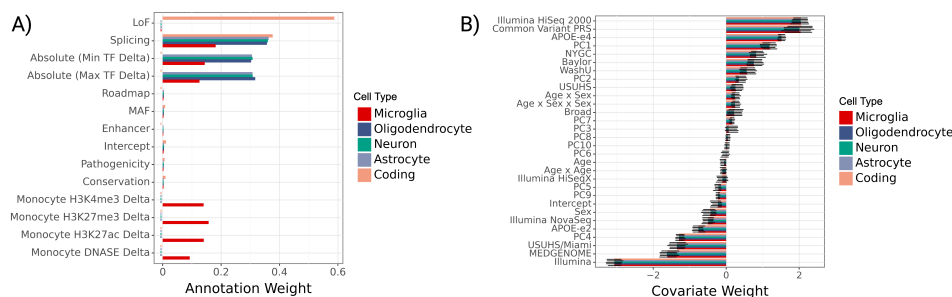
188 **2.4 Applying gruyere to AD WGS data reveals novel disease**  
 189 **associations**



**Fig. 4 Performance of gruyere.** **A)** Boxplots of per-gene AUROCs for train and test sets across cell types. **B)** Boxplots of per-gene accuracies for train and test sets across cell types. **C)** Trace ELBO loss over 300 epochs across cell types. **D)** Average training time per epoch (seconds) versus number of genes used in joint model for each cell type.

190 **Performance on ADSP WGS data.** After validating model performance through  
 191 simulations, we fit *gruyere* to the ADSP WGS data. We analyze coding and non-  
 192 coding (microglia, oligodendrocyte, neuron, astrocyte) groups separately, and refer to  
 193 each set as a cell type. For each cell type, we use a subset of genes for joint fitting  
 194 (FST  $p$ -value  $< 0.01$ ), leading to between 267 and 333 genes per cell-type. AD predic-  
 195 tion performance is fairly consistent across non-coding variants (average test set Area  
 196 Under the Receiver Operating Characteristic, or AUROC, of 0.69) and slightly higher  
 197 for coding variants (average  $AUROC_{test} = 0.70$ ) (Figure 4A-B). When averaging predic-  
 198 ted probabilities across genes, performance further improves ( $AUROC_{test} = 0.72$ )  
 199 for all cell types. These metrics are in line with current AD literature and outperform  
 200 a covariate-only regression model ( $AUROC_{test} = 0.65$ ) [30]. There is a substantial  
 201 range in prediction performance for each cell type (e.g. minimum  $AUROC_{test} = 0.62$ ,  
 202 maximum  $AUROC_{test} = 0.71$  for microglia), highlighting the varying degrees of asso-  
 203 ciation with AD across genes. We find that gene-level performance is consistent across  
 204 model refitting and that the loss converges reliably (4C) [31]. Fitting time increases

205 approximately logarithmically with the total number of genes (Figure 4D). On average,  
 206 it takes 37 seconds per epoch and three hours total to jointly fit `gruyere` across  
 207 300 epochs. Per-gene estimation is much faster, taking an average of 4.3 seconds per  
 208 gene to complete.



**Fig. 5** `gruyere` parameters learned from ADSP WGS. **A)** Bar plot of genome-wide annotation weights  $\tau$  learned in jointly fit model across cell types. We denote crosses (X) to the left of bars where an annotation is not included for a cell type. **B)** Bar plot of per-gene covariate weights ( $\alpha_g$ ) learned in jointly fit model across cell types. Error bars illustrate the minimum and maximum values learned across genes.

209 **Learned annotation and covariate weights.** We find that the top `gruyere`  
 210 functional annotation weights come from splicing across all non-coding RV groups and  
 211 loss-of-function (LoF) for coding variants (Figure 5A). LoF variants can be highly  
 212 disruptive to gene function and are often used as a variant filtering method in gene-  
 213 based tests. Therefore, it is predictable that we find `gruyere` places a large weight  
 214 on LoF coding RVs. It is perhaps not surprising that `gruyere` also prioritizes RVs  
 215 predicted to disrupt normal splicing, as they can substantially change the protein  
 216 product or have large effects on gene dosage via nonsense mediated decay. For all  
 217 non-coding regions, cell-type specific TF binding predictions from Enformer contain  
 218 the next largest annotation weights. This suggests RVs associated with an increase  
 219 (Max TF Delta) or decrease (Min TF Delta) in binding are predicted to have larger  
 220 effects on AD, at least in AD-relevant genes. For microglia RV sets, we additionally  
 221 find increased AD association for variants related to histone modification (H3K4me3,  
 222 H3K27me3, H3K27ac) and DNASE in monocytes (often used as a proxy for microglia  
 223 [32]). We restrict Enformer annotations to non-coding variants as they are specific  
 224 to cell-types. The enhancer category has very small weights across cell-types. Since  
 225 all variants included in the non-coding analyses are in putative CREs, it is perhaps  
 226 not too surprising that cell-type specific enhancer annotations are lowly prioritized by  
 227 `gruyere`.

228 Covariate effects are learned consistently across genes and cell types, with sequenc-  
 229 ing center, common variant polygenic risk score, and Illumina HiSeq 2000 sequencing  
 230 platform as the top three covariates (Figure 5B). As expected, *APOE-e4* is learned to  
 231 have a large positive risk effect while the *APOE-e2* allele has a negative (protective)  
 232 effect [2]. These effects agree well with those of a simple logistic regression predicting  
 233 AD status from covariates.



**Fig. 6 Top gruyere genes. A)** Manhattan plot across cell types. The Y-axis shows  $-\log_{10}(p\text{-value})$  for each gene and X-axis shows gene position. Each color is a cell type. **gruyere**-significant genes are labeled. **B)** Stacked upset plot of significant gene overlap across all tests after pruning for coregulation. Dark grey bars on left indicate total number of significant genes for each test. Vertically connected dots represent groups and corresponding bars indicate the number of overlapping significant genes identified for that group. Each bar is colored by cell type.

234 **Learned gene effects and associations:** Estimated gene effects  $w_g$  are fit from  
 235 a per-gene logistic regression, where we use LRTs to determine gene-level **gruyere**  $p$ -  
 236 values (Methods 3.2). Significant genes after Bonferroni correction for each cell type  
 237 are shown in Figure 6A, where a total of 16 genes reach genome-wide significance.  
 238 The well-established *TREM2* RV association [33], as well as *MAF1*, *C9orf78*, and  
 239 *GRIK3* are found significant for coding variants. Although not as widely recognized as  
 240 *TREM2*, *MAF1* has been previously reported in association with AD [34], and *C9orf78*  
 241 has been identified in an AD dementia meta-analysis [35]. *GRIK3* has emerged as a  
 242 gene of interest due to the role of kainate receptors in neuroinflammation, a key feature  
 243 of AD. Inflammatory responses can amplify glutamate release and disrupt receptor  
 244 functioning, which may further accelerate neurodegeneration. This makes *GRIK3*,  
 245 and glutamate signaling more broadly, potential targets for therapies [36, 37]. The  
 246 identification of these genes by **gruyere** highlights their potential as candidate genes  
 247 for further study in AD.

248 **gruyere** identified 12 non-coding RV associations across cell types with 2 in  
 249 microglia, 6 in astrocytes, 1 in neurons, and 3 in oligodendrocytes. The most significant  
 250 of these is *NUP93* (microglia), which, although not specifically linked to AD, is part of  
 251 a group of nucleoporin (Nup) mutations associated with neurodegenerative disorders  
 252 like AD [38]. Three significant genes, *GALNT9*, *FBRSL1*, and *LOC101928416*, are  
 253 closely located on chromosome 12q24.33 and share over 80% of their ABC-predicted  
 254 CREs, indicating that their associations are driven by the same set of RVs. Although



255 variants can map to multiple genes in our model framework, making our analysis sus-  
256 ceptible to coregulation, we are able to investigate and identify the specific CREs  
257 driving these associations. Of the overlapping promoters for these three genes, regions  
258 have higher ABC scores for *FBRSL1* (ABC = 0.20) compared to *GALNT9* (ABC =  
259 0.04) and *LOC101928416* (ABC = 0.06), suggesting a stronger regulatory impact on  
260 *FBRSL1* and further isolating overlapping signal. *FBRSL1* (neuron), has not been  
261 linked to AD, but it presents a strong candidate gene for its distinctive neuronal expres-  
262 sion profile and involvement in neurogenesis and transcriptional regulatory networks  
263 [39]. Multiple associations (*GALNT18*, *CTR9*, *EIF4G2*, *ZBED5-AS1*, *LOC101928053*,  
264 and *MRV11*) specific to astrocyte and oligodendrocyte cell-types are coregulated in  
265 chromosome 11p15.4, and the strongest signal, *GALNT18*, has been connected to AD  
266 in more than one study [40, 41]. After pruning coregulated signals, **gruyere** identifies  
267 8 significant genes.

268 We identify both known and novel AD-associated risk genes with **gruyere**. Sig-  
269 nificant **gruyere** genes are associated with increased gene expression across thirteen  
270 brain tissues found in GTEx (2-sample t-test  $p = 9.2 \times 10^{-31}$ , Supplemental Figure  
271 3A) [42]. 5 of our 16 significant genes have expression QTLs in our microglia genomic  
272 atlas (isoMiGA) that colocalize with AD or Parkinson’s disease (PD) GWAS (Sup-  
273 plemental Figure 3B) [43–46]. Specifically, *TREM2* and *MAF1* have significant SNPs  
274 in a recent AD GWAS [44] while *FBRSL1*, *EIF4G2*, and *ZBED5-AS1* are signifi-  
275 cant in a large PD GWAS [45]. AD and PD have known genetic overlap, motivating  
276 QTL colocalization of both traits [47]. Finally, we compare **gruyere**  $p$ -values with the  
277 Alzheimer’s Disease Variant Portal (ADVP) catalog of 956 reported AD genes, find-  
278 ing that **gruyere** yields more significant  $p$ -values for ADVP versus non-ADVP genes  
279 (2-sample t-test  $p = 7.1 \times 10^{-9}$ , Supplemental Figure 3C) [48].

#### 280 *Comparing gruyere to existing methods.*

281 We compare pruned **gruyere** results with AD associations identified by a number of  
282 existing RV methods: burden test, sequence kernel association test (SKAT), optimal  
283 unified test (SKAT-O), functional score test (FST), aggregated Cauchy association test  
284 (ACAT), and variant-set test for association using annotation information (STAAR)  
285 [10, 25, 27, 49–51]. Burden, SKAT, SKAT-O, and ACAT-O tests do not include  
286 functional annotations, while FST and STAAR incorporate them (description in Sup-  
287 plemental Table 1 and detailed in Figure 1D for *TREM2* and microglia). We use the  
288 same set of functional annotations for FST and STAAR as for **gruyere**. We find that  
289 **gruyere**  $-\log_{10}(p\text{-values})$  have the highest correlation with Burden tests (Pearson  
290  $R = 0.86$ ) and show moderate to high correlation with combination methods STAAR-  
291 O, SKAT-O, FST, and ACAT-O (Pearson  $R = 0.45 - 0.58$ ) (Supplemental Figure  
292 4A). The higher observed correlation with burden tests is expected, as **gruyere** also  
293 assumes unidirectional variant effects within a gene. We examine overlap of significant  
294 genes across all tests and find that there is minimal overlapping signal across meth-  
295 ods (Figure 6B). Of the 16 (8 pruned) AD associations identified by **gruyere**, 15 (7  
296 pruned) are unique to **gruyere**, while *TREM2* (coding) is detected across all tests but  
297 SKAT where it narrowly misses significance. In total, burden, SKAT, SKAT-O, and



298 ACAT-O tests identify only two significant associations, highlighting the importance  
299 of including functional annotations, particularly for non-coding RV associations.

300 **Integrating gruyere into omnibus tests.** We combine *gruyere*  $p$ -values with  
301 existing methods using ACAT (Supplemental Figure 4B). Comparing  $ACAT(\text{burden,}$   
302  $SKAT, SKAT-O, FST, ACAT, STAAR)$  to  $ACAT(\text{gruyere, burden, SKAT, SKAT-}$   
303  $O, FST, ACAT, STAAR)$ , we find that the inclusion of *gruyere* in omnibus tests  
304 boosts the number of significant associations from 12 to 16, adding *C9orf78, MAF1,*  
305 *NUP93* and *GALNT9*. There is no loss of power with this method, as all existing  
306 signals remain after including *gruyere*; we simply increase the total number of AD  
307 associations identified.

## 308 3 Methods

### 309 3.1 Data Overview

310 **Whole-Genome Sequencing Data:** We analyze the latest release of WGS data  
311 from the Alzheimer’s disease sequencing project (ADSP) consisting of 21,378 unre-  
312 lated individuals over the age of 65 (7,966 cases, 13,412 controls) after QC [52, 53]. We  
313 follow a standard pipeline to QC WGS data. First, we combine phenotype information  
314 across multiple cohorts and remove genetically identical duplicates ( $IBD \hat{\pi} > 0.95$ ) and  
315 technical replicate samples, selecting samples with the highest call rates. We priori-  
316 tize phenotype information for individuals in family studies over case-control studies.  
317 Related individuals are removed using Kinship-based Inference for Gwas (KING) [54],  
318 keeping AD cases where possible. In PLINK [55], we remove individuals with more  
319 than 10% genotype missingness, variants with less than 90% genotyping rate, and  
320 keep only biallelic variants with an observed  $MAF \leq 0.05$ . Missing genotypes are  
321 imputed as the average observed MAF. For analysis, we randomly split samples into  
322 80% train and 20% unseen test sets, stratifying by ancestry. ADSP samples are pri-  
323 marily of European ( $N = 9,133$ ), African ( $N = 5,173$ ) and Hispanic (5,059) ancestry,  
324 with smaller South Asian ( $N = 1,951$ ) and East Asian ( $N = 62$ ) groups.

325 **Clinical Information:** We use 30 available covariates in our model: sex, age,  
326  $\text{age}^2$ ,  $\text{age} \times \text{sex}$ ,  $\text{age} \times \text{sex}^2$ , *APOE-e4* genotype, *APOE-e2* genotype, 10 ancestry  
327 principal components calculated from the 1000 Genomes Project [56], a common vari-  
328 ant PRS [57], one-hot encoded sequencing platform (Illumina HiSeq 2000, HiSeqX,  
329 Nova Seq), one-hot encoded sequencing center (Illumina, USUHS, USUHS/Miami,  
330 NYGC, MEDGENOME, Baylor, Broad, WashU), and an intercept term (Figure 1A).  
331 Covariates are min-max scaled to a range of 0 to 1.

### 332 3.2 Proposed Bayesian rare variant model: gruyere

We develop *gruyere*, a hierarchy of per-gene GLMs (Supplemental Figure 1). We  
define our model jointly as

$$\text{logit}(\mu_{ig}) = X_i \alpha_g + G_{ig} \beta_{gj}$$

where  $\mu_{ig}$  is the probability of AD for individual  $i$  associated with gene  $g$ ,  $X_i$  is a vector of covariates, and  $G_{ig}$  is a vector of genotype dosages for each RV. Covariate coefficients  $\alpha_g$  are modeled from prior,

$$\alpha_g \sim \text{Normal}(0, 1).$$

The key innovation in the **gruyere** model is the construction of per-variant genetic effects for gene  $g$ , or  $\beta_{gj} = (\beta_{gj1}, \dots, \beta_{gjp})^T$ , which is defined as the product of gene effects, transformed MAF, and weighted functional annotations. Of note, if a variant  $j$  is included in the RV set for both genes  $g_1$  and  $g_2$ , the variant effect can differ for  $\beta_{g_1j}$  and  $\beta_{g_2j}$ . We define  $\beta_{gj}$  as

$$\beta_{gj} = w_g w_j (\tau_0 + \sum_{k=1}^q Z_{gjk} \tau_k)$$

where  $w_g$  are gene importance weights,  $w_j$  are variant weights based on observed MAF as suggested by Wu et al. [49],  $\tau$  are genome-wide annotation importance scores, and  $Z_{gjk}$  is a scaled functional annotation  $k$  for RV  $j$  and gene  $g$ . The variables are modeled as,

$$\begin{aligned} w_j &= \text{Beta}(MAF_j | 1, 25) \\ w_g &\sim \text{Normal}(0, 1) \\ \tau &\sim \text{Dirichlet}(\mathbf{1}_q) \end{aligned}$$

333 For each gene, we use a Bernoulli likelihood to sample  $\sigma(X_i \alpha_g + G_{ig} \beta_{gj})$ , and  
 334 aggregate loss across each  $g \in M$ . Learned parameters are  $\alpha_g, w_g$ , and  $\tau$ . We select  
 335 a Dirichlet prior for annotation weights  $\tau$  to ensure identifiability between  $\tau$  and  $w_g$ .  
 336 Without constraining  $\tau$  to a fixed sum,  $w_g$  can be swapped for  $w_g/c$  and  $\tau$  for  $c\tau$  for  
 337 any positive constant  $c$  without changing the likelihood, leading to non-identifiability  
 338 between gene and annotation weights.

339 We approximate the true posterior distribution for **gruyere** by minimizing the  
 340 Kullback-Leibler (KL) divergence, which is equivalent to maximizing the Evidence  
 341 Lower Bound (ELBO) [31]. To maximize the ELBO, we use SVI, implemented in the  
 342 `pyro` probabilistic programming language [58]. We approximate the posterior distribu-  
 343 tion of latent variables  $\alpha_g$  and  $w_g$  with mean field normal distributions (AutoNormal  
 344 guide), while optimizing annotation weights  $\tau$  as point estimates with a Delta distribu-  
 345 tion (AutoDelta guide). We apply the Adam optimizer, a learning rate of 0.1, train  
 346 for 300 epochs, and draw 50 samples from the posterior to estimate standard devi-  
 347 ations of the learned parameters. We explore different prior distributions for all key  
 348 parameters.

349 Once global  $\tau$  is learned, we streamline **gruyere** with a per-gene analysis. Holding  $\tau$   
 350 fixed, our model simplifies to a logistic regression where only  $\alpha_g$  and  $w_g$  are estimated.  
 351 **gruyere** efficiently computes gene-level  $p$ -values using LRTs comparing a covariate-  
 352 only regression to a combined covariate and genotype regression model:

$$LR_g = -2 \times (LL_{\text{combined}_g} - LL_{\text{covariate only}}), \quad df = 1$$

$$\text{gruyere } p\text{-value}(g) = P(\chi^2 > LR_g)$$

where  $LL$  are the log-likelihoods for each logistic regression. For each cell type, we use Bonferroni correction to define the  $p$ -value significance threshold:

$$p < \frac{0.05}{\# \text{ genes per cell type}} \Rightarrow 2.88 \times 10^{-6} < p < 3.64 \times 10^{-6}$$

### 3.3 Cell-type specific RV gene set prediction using the ABC model

We calculate enhancer-gene connectivity using publicly available ATAC-seq and H3K27ac ChIP-seq data for human microglia, oligodendrocytes, astrocytes, and neurons [28]. We apply ABC to this data following the guidelines and default parameters provided at <https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction>. This involves first calling candidate peak regions for the ATAC-seq DNase hypersensitive sites (DHS) using MACS2 (peakExtendFromSummit = 250, nStrongestPeaks = 150000). Then we quantify enhancer activity as the geometric mean of the read counts of DHS and H3K27ac ChIP-seq in candidate enhancer regions. Lastly, we compute the ABC score using averaged Hi-C data (hic\_resolution = 5000) fit to the power-law model. The omics data is aligned to hg19, so we converted the ABC-predicted start and end positions of enhancers to hg38 for analysis. For each gene and separately each cell type, we aggregate all elements  $E$  for gene  $G$  that have an  $ABC \geq 0.02$  and extract RVs from within these regions to determine our cell-type specific non-coding RV gene sets.

### 3.4 Calculating functional annotation groups

We use a range of variant-level functional annotations primarily from the Whole Genome Sequencing Annotation database [59]. Annotations with greater than 5% missingness in our RVs are removed, resulting in 50 coding variant and 52 non-coding variant functional annotations listed in Supplemental Table 2. To reduce dimensionality of related annotations while accounting for their diverse biological effects, we apply non-negative matrix factorization (NMF) to summarize groups of related annotations, inspired by STAAR [10]. We use NMF to retain interpretable directional scaling of annotations. Based on correlation structure and a priori knowledge, we define six major functional categories – splicing, conservation, integrative deleterious predictions, brain-related Roadmap epigenetics, population-specific MAF, and enhancer activity [15, 60, 61]. Because the splicing annotation group is derived from four SpliceAI predictions that are not highly correlated and sparsely distributed, we instead use the maximum score for this category. We also include a binary LoF prediction calculated with Loss-Of-Function Transcript Effect Estimator (LOFTEE) [62] for coding variants along with an intercept term. All annotations are scaled between 0 and 1, where a larger value represents increased predicted variant function.

**Deep Learning Delta Scores:** For all four cell types, we include additional cell-type specific functional annotations: absolute maximum and absolute minimum TF delta scores derived from Enformer [14], a deep learning genomics model. We calculate

389 variant delta scores for 5,318 functional genomics assays. The Enformer model predicts  
390 read counts (in 128 BP bins) of these assays as a function of 196,608 BP input DNA  
391 sequence. For a particular variant, we compare the model output of the reference  
392 sequence, centered around the variant position, with the output of the alternative  
393 sequence which replaces the reference allele with the alternative allele. For a particular  
394 genomics assay, the delta score is the difference between the sum of reference sequence  
395 predictions for the middle 32 bins and the sum of the alternate sequence predictions  
396 for the same bins. We normalize these scores by first calculating the delta scores for  
397 the approximately 18 million variants from the UK Biobank cohort used in PolyFun  
398 [63, 64], and then Z-score normalize each assay according based on this collection of  
399 variants. We apply this normalization to the delta scores used in our analysis. We  
400 aggregate delta scores to determine composite maximum and minimum predictions  
401 for each variant, highlighting the delta scores of only the enriched TFs within each of  
402 the four cell types ([28]). For microglia non-coding variant sets, we additionally use  
403 delta scores for 4 epigenetic marks (H3K4me3, H3K27ac, H3K27me3, and DNASE)  
404 for monocytes, a proxy for microglia.

### 405 3.5 Data Simulation

We use simulations to evaluate **gruyere** performance. We randomly sample values for each parameter and use these simulated variables in the GLM framework,  $\text{logit}(\mu_{ig}) = X_i\alpha_g + G_{ig}\beta_g$ . The real ADSP genotypes  $G_{ig}$ , covariates  $X_i$ , and functional annotations  $Z_{gjk}$  along with simulated parameters  $\alpha_g^S, w_g^S$ , and  $\tau^S$ , generate simulated phenotypes  $Y_{ig}^S$ . Simulations are restricted to a maximum estimated heritability of 30% to realistically evaluate complex diseases. For each simulation, we draw **gruyere** parameters from the following distributions:

$$\alpha_g^S \sim \text{Normal}(0, 1), \quad w_g^S \sim \text{Normal}(0, 1), \quad \tau^S \sim \text{Dirichlet}(\mathbf{1}_q)$$

We define  $\beta_{gj}$  in the same way, simply using simulated variables:

$$\beta_{gj}^S = w_g^S w_j \sum_{k=1}^q Z_{gjk} \tau_k^S$$

406 Using this combination of true data and simulated parameters, we sample synthetic  
407 phenotypes  $y_g^S$  from a Bernoulli distribution. We perform 100 simulations on  $M =$   
408 500 randomly selected genes. In general, we sample **gruyere** parameters from the  
409 same distribution that they are learned. We have also tested model performance when  
410 simulated data comes from a different distribution than its learned counterpart (e.g.  
411 sampling  $w_g$  from a Normal distribution in simulations but fitting from a Gamma  
412 prior).

## 413 4 Discussion

414 We develop **gruyere**, a functionally-informed RV association test that fits a hierar-  
415 chy of Bayesian GLMs to estimate genome-wide functional annotation importance,

416 gene effects, and covariate coefficients. **gruyere** builds upon existing RV methods with  
417 two key advancements: 1) a genome-wide approach that enables trait-specific weight-  
418 ing of functional annotations, and 2) a flexible, powerful and calibrated probabilistic  
419 framework that estimates uncertainty. We incorporate an innovative methodology for  
420 analyzing RVs in the non-coding genome. Using the Activity-by-Contact model, we  
421 predict cell-type-specific enhancer-gene connectivity from chromatin state and con-  
422 formation data, aggregating predictions by gene to define interpretable non-coding  
423 RV testing regions. We use in silico mutagenesis under state-of-the-art deep learning  
424 models of pre (SpliceAI) and post (Enformer) transcription gene regulation to pre-  
425 dict RV effects. Simulation analyses validate **gruyere** and show it is able to recover  
426 ground truth parameters across diverse model specifications and even for realistically  
427 low heritability.

428 We apply **gruyere**, along with a number of established RV association tools, to  
429 the most recent WGS release from ADSP. Our analysis identifies both known (e.g.,  
430 *TREM2*) and novel (e.g., *NUP93*) candidate AD genes. Specifically, **gruyere** uniquely  
431 identifies 15 genes, of which *C9orf78*, *MAF1*, *NUP93* and *GALNT9* remain signif-  
432 icant in aggregated Cauchy tests. Our analysis additionally provides an improved  
433 understanding of AD-relevant functional annotations. **gruyere** confirms the expecta-  
434 tion that LoF is the most informative annotation for coding variants, but additionally  
435 finds deep learning-based predictions for splicing, TF binding and chromatin state are  
436 highly predictive of functional non-coding RVs.

437 We use ancestry principal components as covariates to account for population  
438 diversity, but one area for future work would be integrating a random effect term to  
439 better account for relatedness and population structure [65]. Another possible exten-  
440 sion to **gruyere** would be incorporating gene-level features as priors [66]. While we  
441 focus our analysis on AD, **gruyere** can be applied to any complex disease with suf-  
442 ficient WGS data. As the quality of functional annotations continues to improve,  
443 **gruyere** will become an increasingly valuable tool for identifying disease-associated  
444 genes and annotations.

445 **Supplementary information.** Supplementary Figures 1-4; Supplementary Table  
446 1-2; Supplementary Methods; Supplementary Acknowledgements; Supplementary  
447 Data: Gene  $p$ -values for each cell-type.

448 **Acknowledgements.** Research reported in this paper was supported by  
449 Alzheimer’s Disease Sequencing Project of the National Institutes of Health under  
450 award number 5 U01 AG068880-02. The content is solely the responsibility of the  
451 authors and does not necessarily represent the official views of the National Institutes  
452 of Health. We are thankful to Tulsi Patel, Edoardo Marcora, Alison Goate, and Iuliana  
453 Ionita-Laza for helpful feedback and discussions.

## 454 **Declarations**

- 455 • Funding: Research reported in this paper was supported by Alzheimer’s Disease  
456 Sequencing Project of the National Institutes of Health under award number 5 U01  
457 AG068880-02.
- 458 • Competing interests: We have no competing interests to declare.
- 459 • Ethics approval and consent to participate: Not applicable
- 460 • Consent for publication: Not applicable
- 461 • Data availability: This paper uses the ADSP Release 4 WGS data and AD phenotype  
462 data.
- 463 • Materials availability: Not applicable
- 464 • Code availability: Code and details for running `gruyere` is available on GitHub:  
465 <https://github.com/daklab/gruyere>
- 466 • Author contribution: D.A.K. conceived and supervised the project. A.D. and D.A.K.  
467 developed the methods. A.D. and D.A.K. wrote the manuscript. A.D. wrote the  
468 software code and performed the analyses. All authors read, reviewed, and approved  
469 the final manuscript.

470 Editorial Policies for:

471 Springer journals and proceedings: <https://www.springer.com/gp/editorial-policies>

472 Nature Portfolio journals: <https://www.nature.com/nature-research/editorial-policies>

473 *Scientific Reports*: <https://www.nature.com/srep/journal-policies/editorial-policies>

474 BMC journals: <https://www.biomedcentral.com/getpublished/editorial-policies>

## 475 **References**

- 476 [1] Sims, R., Hill, M., Williams, J.: The multiplex model of the genetics of alzheimer’s  
477 disease. *Nat. Neurosci.* **23**(3), 311–322 (2020)
- 478 [2] Andrews, S.J., Renton, A.E., Fulton-Howard, B., Podlesny-Drabiniok, A., Mar-  
479 cora, E., Goate, A.M.: The complex genetic architecture of alzheimer’s disease:  
480 novel insights and future directions. *EBioMedicine* **90**, 104511 (2023)

- 481 [3] Auer, P.L., Lettre, G.: Rare variant association studies: considerations, challenges  
482 and opportunities. *Genome Med.* **7**(1), 16 (2015)
- 483 [4] Wainschtein, P., Jain, D., Zheng, Z., TOPMed Anthropometry Working Group,  
484 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Cupples,  
485 L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty,  
486 B.M., Kooperberg, C., Liu, C.-T., Albert, C.M., Roden, D., Chasman, D.I., Dar-  
487 bar, D., Lloyd-Jones, D.M., Arnett, D.K., Regan, E.A., Boerwinkle, E., Rotter,  
488 J.I., O’Connell, J.R., Yanek, L.R., Andrade, M., Allison, M.A., McDonald, M.-  
489 L.N., Chung, M.K., Fornage, M., Chami, N., Smith, N.L., Ellinor, P.T., Vasan,  
490 R.S., Mathias, R.A., Loos, R.J.F., Rich, S.S., Lubitz, S.A., Heckbert, S.R., Red-  
491 line, S., Guo, X., Chen, Y.-D.I., Laurie, C.A., Hernandez, R.D., McGarvey, S.T.,  
492 Goddard, M.E., Laurie, C.C., North, K.E., Lange, L.A., Weir, B.S., Yengo, L.,  
493 Yang, J., Visscher, P.M.: Assessing the contribution of rare variants to complex  
494 trait heritability from whole-genome sequence data. *Nat. Genet.* **54**(3), 263–273  
495 (2022)
- 496 [5] Kosmicki, J.A., Churchhouse, C.L., Rivas, M.A., Neale, B.M.: Discovery of rare  
497 variants for complex phenotypes. *Hum. Genet.* **135**(6), 625–634 (2016)
- 498 [6] Fiziev, P.P., McRae, J., Ulirsch, J.C., Dron, J.S., Hamp, T., Yang, Y., Wain-  
499 schtein, P., Ni, Z., Schraiber, J.G., Gao, H., Cable, D., Field, Y., Aguet,  
500 F., Fasnacht, M., Metwally, A., Rogers, J., Marques-Bonet, T., Rehm, H.L.,  
501 O’Donnell-Luria, A., Khera, A.V., Farh, K.K.-H.: Rare penetrant mutations  
502 confer severe risk of common diseases. *Science* **380**(6648), 1131 (2023)
- 503 [7] Hoogmartens, J., Cacace, R., Van Broeckhoven, C.: Insight into the genetic eti-  
504 ology of alzheimer’s disease: A comprehensive review of the role of rare variants.  
505 *Alzheimers. Dement.* **13**(1), 12155 (2021)
- 506 [8] Li, Z., Li, X., Zhou, H., Gaynor, S.M., Selvaraj, M.S., Arapoglou, T., Quick,  
507 C., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Auer, P.L., Bielak, L.F.,  
508 Bis, J.C., Blackwell, T.W., Blangero, J., Boerwinkle, E., Bowden, D.W., Brody,  
509 J.A., Cade, B.E., Conomos, M.P., Correa, A., Cupples, L.A., Curran, J.E., Vries,  
510 P.S., Duggirala, R., Franceschini, N., Freedman, B.I., Göring, H.H.H., Guo, X.,  
511 Kalyani, R.R., Kooperberg, C., Kral, B.G., Lange, L.A., Lin, B.M., Manichaikul,  
512 A., Manning, A.K., Martin, L.W., Mathias, R.A., Meigs, J.B., Mitchell, B.D.,  
513 Montasser, M.E., Morrison, A.C., Naseri, T., O’Connell, J.R., Palmer, N.D.,  
514 Peyser, P.A., Psaty, B.M., Raffield, L.M., Redline, S., Reiner, A.P., Reupena,  
515 M.S., Rice, K.M., Rich, S.S., Smith, J.A., Taylor, K.D., Taub, M.A., Vasan, R.S.,  
516 Weeks, D.E., Wilson, J.G., Yanek, L.R., Zhao, W., NHLBI Trans-Omics for Preci-  
517 sion Medicine (TOPMed) Consortium, TOPMed Lipids Working Group, Rotter,  
518 J.I., Willer, C.J., Natarajan, P., Peloso, G.M., Lin, X.: A framework for detect-  
519 ing noncoding rare-variant associations of large-scale whole-genome sequencing  
520 studies. *Nat. Methods* **19**(12), 1599–1611 (2022)



- 521 [9] Lee, S., Abecasis, G.R., Boehnke, M., Lin, X.: Rare-variant association analysis:  
522 study designs and statistical tests. *Am. J. Hum. Genet.* **95**(1), 5–23 (2014)
- 523 [10] Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R.,  
524 Arnett, D.K., Aslibekyan, S., Ballantyne, C.M., Bielak, L.F., Blangero, J., Boer-  
525 winkle, E., Bowden, D.W., Broome, J.G., Conomos, M.P., Correa, A., Cupples,  
526 L.A., Curran, J.E., Freedman, B.I., Guo, X., Hindy, G., Irvin, M.R., Kardia,  
527 S.L.R., Kathiresan, S., Khan, A.T., Kooperberg, C.L., Laurie, C.C., Liu, X.S.,  
528 Mahaney, M.C., Manichaikul, A.W., Martin, L.W., Mathias, R.A., McGarvey,  
529 S.T., Mitchell, B.D., Montasser, M.E., Moore, J.E., Morrison, A.C., O’Connell,  
530 J.R., Palmer, N.D., Pampana, A., Peralta, J.M., Peyser, P.A., Psaty, B.M., Red-  
531 line, S., Rice, K.M., Rich, S.S., Smith, J.A., Tiwari, H.K., Tsai, M.Y., Vasan,  
532 R.S., Wang, F.F., Weeks, D.E., Weng, Z., Wilson, J.G., Yanek, L.R., NHLBI  
533 Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Lipids  
534 Working Group, Neale, B.M., Sunyaev, S.R., Abecasis, G.R., Rotter, J.I., Willer,  
535 C.J., Peloso, G.M., Natarajan, P., Lin, X.: Dynamic incorporation of multiple in  
536 silico functional annotations empowers rare variant association analysis of large  
537 whole-genome sequencing studies at scale. *Nat. Genet.* **52**(9), 969–983 (2020)
- 538 [11] Clarke, B., Holtkamp, E., Öztürk, H., Mück, M., Wahlberg, M., Meyer, K.,  
539 Munzlinger, F., Brechtmann, F., Hölzlwimmer, F.R., Gagneur, J., Stegle, O.:  
540 Integration of variant annotations using deep set networks boosts rare variant  
541 association genetics (2023)
- 542 [12] Wagner, N., Çelik, M.H., Hölzlwimmer, F.R., Mertes, C., Prokisch, H., Yépez,  
543 V.A., Gagneur, J.: Aberrant splicing prediction across human tissues. *Nat. Genet.*  
544 **55**(5), 861–870 (2023)
- 545 [13] Zeng, T., Li, Y.I.: Predicting RNA splicing from DNA sequence using pangolin.  
546 *Genome Biol.* **23**(1), 103 (2022)
- 547 [14] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Tay-  
548 lor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R.: Effective gene expression  
549 prediction from sequence by integrating long-range interactions. *Nat. Methods*  
550 **18**(10), 1196–1203 (2021)
- 551 [15] Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., Siepel, A.: Detection of nonneu-  
552 tral substitution rates on mammalian phylogenies. *Genome Res.* **20**(1), 110–121  
553 (2010)
- 554 [16] Ionita-Laza, I., McCallum, K., Xu, B., Buxbaum, J.D.: A spectral approach inte-  
555 grating functional genomic annotations for coding and noncoding variants. *Nat.*  
556 *Genet.* **48**(2), 214–220 (2016)
- 557 [17] Skene, N.G., Grant, S.G.N.: Identification of vulnerable cell types in major  
558 brain disorders using single cell transcriptomes and expression weighted cell type  
559 enrichment. *Front. Neurosci.* **10**, 16 (2016)

- 560 [18] Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A.C., Boerwin-  
561 kle, E., Lin, X.: Dynamic scan procedure for detecting Rare-Variant association  
562 regions in Whole-Genome sequencing studies. *Am. J. Hum. Genet.* **104**(5),  
563 802–814 (2019)
- 564 [19] Sey, N.Y.A., Hu, B., Mah, W., Fauni, H., McAfee, J.C., Rajarajan, P., Bren-  
565 nand, K.J., Akbarian, S., Won, H.: A computational tool (H-MAGMA) for  
566 improved prediction of brain-disorder risk genes by incorporating brain chromatin  
567 interaction profiles. *Nat. Neurosci.* **23**(4), 583–593 (2020)
- 568 [20] Ma, S., Dagleish, J., Lee, J., Wang, C., Liu, L., Gill, R., Buxbaum, J.D., Chung,  
569 W.K., Aschard, H., Silverman, E.K., Cho, M.H., He, Z., Ionita-Laza, I.: Powerful  
570 gene-based testing by integrating long-range chromatin interactions and knockoff  
571 genotypes. *Proc. Natl. Acad. Sci. U. S. A.* **118**(47) (2021)
- 572 [21] Zhang, S., Moll, T., Rubin-Sigler, J., Tu, S., Li, S., Yuan, E., Liu, M., Butt, A.,  
573 Harvey, C., Gornall, S., Alhalthli, E., Shaw, A., Souza, C.D.S., Ferraiuolo, L.,  
574 Hornstein, E., Shelkownikova, T., Dijk, C.H., Timpanaro, I.S., Kenna, K.P., Zeng,  
575 J., Tsao, P.S., Shaw, P.J., Ichida, J.K., Cooper-Knock, J., Snyder, M.P.: Deep  
576 learning modeling of rare noncoding genetic variants in human motor neurons  
577 defines *CCDC146* as a therapeutic target for ALS (2024)
- 578 [22] Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian,  
579 V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., Nguyen,  
580 T.H., Kane, M., Perez, E.M., Durand, N.C., Lareau, C.A., Stamenova, E.K.,  
581 Aiden, E.L., Lander, E.S., Engreitz, J.M.: Activity-by-contact model of enhancer-  
582 promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*  
583 **51**(12), 1664–1669 (2019)
- 584 [23] Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference.  
585 *J. Mach. Learn. Res.* (2013)
- 586 [24] Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi,  
587 S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz,  
588 G.B., Chow, E.D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S.J.,  
589 Farh, K.K.-H.: Predicting splicing from primary sequence with deep learning. *Cell*  
590 **176**(3), 535–54824 (2019)
- 591 [25] Madsen, B.E., Browning, S.R.: A groupwise association test for rare mutations  
592 using a weighted sum statistic. *PLoS Genet.* **5**(2), 1000384 (2009)
- 593 [26] Morgenthaler, S., Thilly, W.G.: A strategy to discover genes that carry multi-  
594 allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST).  
595 *Mutat. Res.* **615**(1-2), 28–56 (2007)
- 596 [27] He, Z., Xu, B., Lee, S., Ionita-Laza, I.: Unified Sequence-Based association tests  
597 allowing for multiple functional annotations and meta-analysis of noncoding

- 598 variation in metabochip data. *Am. J. Hum. Genet.* **101**(3), 340–352 (2017)
- 599 [28] Nott, A., Holtman, I.R., Coufal, N.G., Schlachetzki, J.C.M., Yu, M., Hu, R.,  
600 Han, C.Z., Pena, M., Xiao, J., Wu, Y., Keulen, Z., Pasillas, M.P., O’Connor, C.,  
601 Nickl, C.K., Schafer, S.T., Shen, Z., Rissman, R.A., Brewer, J.B., Gosselin, D.,  
602 Gonda, D.D., Levy, M.L., Rosenfeld, M.G., McVicker, G., Gage, F.H., Ren, B.,  
603 Glass, C.K.: Brain cell type-specific enhancer-promoter interactome maps and  
604 disease-risk association. *Science* **366**(6469), 1134–1139 (2019)
- 605 [29] Dempster, E.R., Lerner, I.M.: Heritability of threshold characters. *Genetics* **35**(2),  
606 212–236 (1950)
- 607 [30] Zhou, X., Chen, Y., Ip, F.C.F., Jiang, Y., Cao, H., Lv, G., Zhong, H., Chen,  
608 J., Ye, T., Chen, Y., Zhang, Y., Ma, S., Lo, R.M.N., Tong, E.P.S., Alzheimer’s  
609 Disease Neuroimaging Initiative, Mok, V.C.T., Kwok, T.C.Y., Guo, Q., Mok,  
610 K.Y., Shoai, M., Hardy, J., Chen, L., Fu, A.K.Y., Ip, N.Y.: Deep learning-based  
611 polygenic risk analysis for alzheimer’s disease prediction. *Commun. Med. (Lond.)*  
612 **3**(1), 49 (2023)
- 613 [31] Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for  
614 statisticians. *J. Am. Stat. Assoc.* **112**(518), 859–877 (2017)
- 615 [32] Gopinath, A., Collins, A., Khoshbouei, H., Streit, W.J.: Microglia and other  
616 myeloid cells in central nervous system health and disease. *J. Pharmacol. Exp.*  
617 *Ther.* **375**(1), 154–160 (2020)
- 618 [33] Gratuze, M., Leyns, C.E.G., Holtzman, D.M.: New insights into the role of  
619 TREM2 in alzheimer’s disease. *Mol. Neurodegener.* **13**(1), 66 (2018)
- 620 [34] Han, Y., Chen, K., Yu, H., Cui, C., Li, H., Hu, Y., Zhang, B., Li, G.: Maf1 loss  
621 regulates spinogenesis and attenuates cognitive impairment in alzheimer’s disease.  
622 *Brain* **147**(6), 2128–2143 (2024)
- 623 [35] Bottero, V., Potashkin, J.A.: Meta-analysis of gene expression changes in  
624 the blood of patients with mild cognitive impairment and alzheimer’s disease  
625 dementia. *Int. J. Mol. Sci.* **20**(21), 5403 (2019)
- 626 [36] Izadi, F., Soheilifar, M.H.: Exploring potential biomarkers underlying pathogene-  
627 sis of alzheimer’s disease by differential co-expression analysis. *Avicenna J. Med.*  
628 *Biotechnol.* **10**(4), 233–241 (2018)
- 629 [37] Xiao, B., Kuang, Z., Zhang, W., Hang, J., Chen, L., Lei, T., He, Y., Deng, C.,  
630 Li, W., Lu, J., Qu, J., Zhou, Q., Hao, W., Sun, Z., Li, L.: Glutamate ionotropic  
631 receptor kainate type subunit 3 (GRIK3) promotes epithelial-mesenchymal tran-  
632 sition in breast cancer cells by regulating SPDEF/CDH1 signaling. *Mol. Carcinog.*  
633 **58**(7), 1314–1323 (2019)

- 634 [38] Spead, O., Zaepfel, B.L., Rothstein, J.D.: Nuclear pore dysfunction in neurode-  
635 generation. *Neurotherapeutics* **19**(4), 1050–1060 (2022)
- 636 [39] Bukvic, N., De Rinaldis, M., Chetta, M., Trabacca, A., Bassi, M.T., Marsano,  
637 R.M., Holoubkova, L., Rivieccio, M., Oro, M., Resta, N., Kerkhof, J., Sadikovic,  
638 B., Viggiano, L.: De novo pathogenic variant in FBRSL1, non OMIM gene  
639 paralogue AUTS2, causes a novel recognizable syndromic manifestation with  
640 intellectual disability; an additional patient and review of the literature. *Genes*  
641 (Basel) **15**(7), 826 (2024)
- 642 [40] Mishra, R., Li, B.: The application of artificial intelligence in the genetic study  
643 of alzheimer’s disease. *Aging Dis.* **11**(6), 1567–1584 (2020)
- 644 [41] Lutz, M.W., Sprague, D., Barrera, J., Chiba-Falek, O.: Shared genetic etiology  
645 underlying alzheimer’s disease and major depressive disorder. *Transl. Psychiatry*  
646 **10**(1), 88 (2020)
- 647 [42] GTEx Consortium: The GTEx consortium atlas of genetic regulatory effects  
648 across human tissues. *Science* **369**(6509), 1318–1330 (2020)
- 649 [43] Humphrey, J., Brophy, E., Kosoy, R., Zeng, B., Coccia, E., Mattei, D., Ravi, A.,  
650 Efthymiou, A.G., Navarro, E., Muller, B.Z., Snijders, G.J., Allan, A., Münch, A.,  
651 Kitata, R.B., Kleopoulos, S.P., Argyriou, S., Shao, Z., Francoeur, N., Tsai, C.-F.,  
652 Gritsenko, M.A., Monroe, M.E., Paurus, V.L., Weitz, K.K., Shi, T., Sebra, R., Liu,  
653 T., Witte, L.D., Goate, A.M., Bennett, D.A., Haroutunian, V., Hoffman, G.E.,  
654 Fullard, J.F., Roussos, P., Raj, T.: Long-read RNA-seq atlas of novel microglia  
655 isoforms elucidates disease-associated genetic regulation of splicing. medRxiv,  
656 2023–120123299073 (2023)
- 657 [44] Bellenguez, C., Küçükali, F., Jansen, I.E., Klei, L., Moreno-Grau, S.,  
658 Amin, N., Naj, A.C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., Hol-  
659 mans, P.A., Boland, A., Damotte, V., Lee, S.J., Costa, M.R., Kuulasmaa, T.,  
660 Yang, Q., Rojas, I., Bis, J.C., Yaqub, A., Prokic, I., Chapuis, J., Ahmad, S.,  
661 Giedraitis, V., Aarsland, D., Garcia-Gonzalez, P., Abdelnour, C., Alarcón-Martín,  
662 E., Alcolea, D., Alegret, M., Alvarez, I., Álvarez, V., Armstrong, N.J., Tzolaki,  
663 A., Antúnez, C., Appollonio, I., Arcaro, M., Archetti, S., Pastor, A.A., Arosio,  
664 B., Athanasiu, L., Bailly, H., Banaj, N., Baquero, M., Barral, S., Beiser, A.,  
665 Pastor, A.B., Below, J.E., Benček, P., Benussi, L., Berr, C., Besse, C., Bessi,  
666 V., Binetti, G., Bizarro, A., Blesa, R., Boada, M., Boerwinkle, E., Borroni, B.,  
667 Boschi, S., Bossù, P., Bråthen, G., Bressler, J., Bresner, C., Brodaty, H., Brookes,  
668 K.J., Brusco, L.I., Buiza-Rueda, D., Bürger, K., Burholt, V., Bush, W.S., Calero,  
669 M., Cantwell, L.B., Chene, G., Chung, J., Cuccaro, M.L., Carracedo, , Cecchetti,  
670 R., Cervera-Carles, L., Charbonnier, C., Chen, H.-H., Chillotti, C., Ciccone, S.,  
671 Claassen, J.A.H.R., Clark, C., Conti, E., Corma-Gómez, A., Costantini, E., Cus-  
672 todero, C., Daian, D., Dalmaso, M.C., Daniele, A., Dardiotis, E., Dartigues,  
673 J.-F., Deyn, P.P., Paiva Lopes, K., Witte, L.D., D’Ebneth, S., Deckert, J., Del Ser,

674 T., Denning, N., DeStefano, A., Dichgans, M., Diehl-Schmid, J., Diez-Fairen,  
675 M., Rossi, P.D., Djurovic, S., Duron, E., Düzel, E., Dufouil, C., Eiriksdottir,  
676 G., Engelborghs, S., Escott-Price, V., Espinosa, A., Ewers, M., Faber, K.M.,  
677 Fabrizio, T., Nielsen, S.F., Fardo, D.W., Farotti, L., Fenoglio, C., Fernández-  
678 Fuertes, M., Ferrari, R., Ferreira, C.B., Ferri, E., Fin, B., Fischer, P., Fladby,  
679 T., Fließbach, K., Fongang, B., Fornage, M., Fortea, J., Foroud, T.M., Fostinelli,  
680 S., Fox, N.C., Franco-Macías, E., Bullido, M.J., Frank-García, A., Froelich, L.,  
681 Fulton-Howard, B., Galimberti, D., García-Alberca, J.M., García-González, P.,  
682 Garcia-Madrona, S., Garcia-Ribas, G., Ghidoni, R., Giegling, I., Giorgio, G.,  
683 Goate, A.M., Goldhardt, O., Gomez-Fonseca, D., González-Pérez, A., Graff, C.,  
684 Grande, G., Green, E., Grimmer, T., Grünblatt, E., Grunin, M., Gudnason, V.,  
685 Guetta-Baranes, T., Haapasalo, A., Hadjigeorgiou, G., Haines, J.L., Hamilton-  
686 Nelson, K.L., Hampel, H., Hanon, O., Hardy, J., Hartmann, A.M., Hausner, L.,  
687 Harwood, J., Heilmann-Heimbach, S., Helisalmi, S., Heneka, M.T., Hernández, I.,  
688 Herrmann, M.J., Hoffmann, P., Holmes, C., Holstege, H., Vilas, R.H., Hulsman,  
689 M., Humphrey, J., Biessels, G.J., Jian, X., Johansson, C., Jun, G.R., Kastumata,  
690 Y., Kauwe, J., Kehoe, P.G., Kilander, L., Ståhlbom, A.K., Kivipelto, M., Koivisto,  
691 A., Kornhuber, J., Kosmidis, M.H., Kukull, W.A., Kuksa, P.P., Kunkle, B.W.,  
692 Kuzma, A.B., Lage, C., Laukka, E.J., Launer, L., Lauria, A., Lee, C.-Y., Lehtisalo,  
693 J., Lerch, O., Lleó, A., Longstreth, W. Jr, Lopez, O., Munain, A.L., Love,  
694 S., Löwemark, M., Luckcuck, L., Lunetta, K.L., Ma, Y., Macías, J., MacLeod,  
695 C.A., Maier, W., Mangialasche, F., Spallazzi, M., Marquié, M., Marshall, R.,  
696 Martin, E.R., Montes, A.M., Rodríguez, C.M., Masullo, C., Mayeux, R., Mead,  
697 S., Mecocci, P., Medina, M., Meggy, A., Mehrabian, S., Mendoza, S., Menéndez-  
698 González, M., Mir, P., Moebus, S., Mol, M., Molina-Porcel, L., Montreal, L.,  
699 Morelli, L., Moreno, F., Morgan, K., Mosley, T., Nöthen, M.M., Muchnik, C.,  
700 Mukherjee, S., Nacmias, B., Ngandu, T., Nicolas, G., Nordestgaard, B.G., Olasso,  
701 R., Orellana, A., Orsini, M., Ortega, G., Padovani, A., Paolo, C., Papenberg, G.,  
702 Parnetti, L., Pasquier, F., Pastor, P., Peloso, G., Pérez-Cordón, A., Pérez-Tur,  
703 J., Pericard, P., Peters, O., Pijnenburg, Y.A.L., Pineda, J.A., Piñol-Ripoll, G.,  
704 Pisanu, C., Polak, T., Popp, J., Posthuma, D., Priller, J., Puerta, R., Quenez,  
705 O., Quintela, I., Thomassen, J.Q., Rábano, A., Rainero, I., Rajabli, F., Ramak-  
706 ers, I., Real, L.M., Reinders, M.J.T., Reitz, C., Reyes-Dumeyer, D., Ridge, P.,  
707 Riedel-Heller, S., Riederer, P., Roberto, N., Rodriguez-Rodriguez, E., Rongve,  
708 A., Allende, I.R., Rosende-Roca, M., Royo, J.L., Rubino, E., Rujescu, D., Sáez,  
709 M.E., Sakka, P., Saltvedt, I., Sanabria, , Sánchez-Arjona, M.B., Sanchez-Garcia,  
710 F., Juan, P.S., Sánchez-Valle, R., Sando, S.B., Sarnowski, C., Satizabal, C.L.,  
711 Scamosci, M., Scarmeas, N., Scarpini, E., Scheltens, P., Scherbaum, N., Scherer,  
712 M., Schmid, M., Schneider, A., Schott, J.M., Selbæk, G., Seripa, D., Serrano,  
713 M., Sha, J., Shadrin, A.A., Skrobot, O., Slifer, S., Snijders, G.J.L., Soininen,  
714 H., Solfrizzi, V., Solomon, A., Song, Y., Sorbi, S., Sotolongo-Grau, O., Spalletta,  
715 G., Spottke, A., Squassina, A., Stordal, E., Tartan, J.P., Tárraga, L., Tesí, N.,  
716 Thalamuthu, A., Thomas, T., Tosto, G., Traykov, L., Tremolizzo, L., Tybjærg-  
717 Hansen, A., Uitterlinden, A., Ullgren, A., Ulstein, I., Valero, S., Valladares, O.,  
718 Van Broeckhoven, C., Vance, J., Vardarajan, B.N., Lugt, A., Van Dongen, J.,

719 Rooij, J., Swieten, J., Vandenberghe, R., Verhey, F., Vidal, J.-S., Vogelgsang, J.,  
720 Vyhnalek, M., Wagner, M., Wallon, D., Wang, L.-S., Wang, R., Weinhold, L.,  
721 Wiltfang, J., Windle, G., Woods, B., Yannakoulia, M., Zare, H., Zhao, Y., Zhang,  
722 X., Zhu, C., Zulaica, M., EADB, GR@ACE, DEGESCO, EADI, GERAD, Dem-  
723 gene, FinnGen, ADGC, CHARGE, Farrer, L.A., Psaty, B.M., Ghanbari, M., Raj,  
724 T., Sachdev, P., Mather, K., Jessen, F., Ikram, M.A., Mendonça, A., Hort, J.,  
725 Tsolaki, M., Pericak-Vance, M.A., Amouyel, P., Williams, J., Frikke-Schmidt, R.,  
726 Clarimon, J., Deleuze, J.-F., Rossi, G., Seshadri, S., Andreassen, O.A., Ingelsson,  
727 M., Hiltunen, M., Sleegers, K., Schellenberg, G.D., Duijn, C.M., Sims, R., Flier,  
728 W.M., Ruiz, A., Ramirez, A., Lambert, J.-C.: New insights into the genetic eti-  
729 ology of alzheimer's disease and related dementias. *Nat. Genet.* **54**(4), 412–436  
730 (2022)

731 [45] Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bel-  
732 lenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., Russo,  
733 G., Thorton-Wells, T.A., Jones, N., Smith, A.V., Chouraki, V., Thomas, C.,  
734 Ikram, M.A., Zelenika, D., Vardarajan, B.N., Kamatani, Y., Lin, C.F., Gerrish,  
735 A., Schmidt, H., Kunkle, B., Dunstan, M.L., Ruiz, A., Bihoreau, M.T., Choi,  
736 S.H., Reitz, C., Pasquier, F., Cruchaga, C., Craig, D., Amin, N., Berr, C., Lopez,  
737 O.L., De Jager, P.L., Deramecourt, V., Johnston, J.A., Evans, D., Lovestone,  
738 S., Letenneur, L., Morón, F.J., Rubinsztein, D.C., Eiriksdottir, G., Sleegers, K.,  
739 Goate, A.M., Fiévet, N., Huentelman, M.W., Gill, M., Brown, K., Kamboh, M.I.,  
740 Keller, L., Barberger-Gateau, P., McGuinness, B., Larson, E.B., Green, R., Myers,  
741 A.J., Dufouil, C., Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J.,  
742 St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D.W., Yu, L.,  
743 Tsolaki, M., Bossù, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-  
744 Garcia, F., Fox, N.C., Hardy, J., Deniz Naranjo, M.C., Bosco, P., Clarke, R.,  
745 Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's  
746 Disease Initiative (EADI), Genetic and Environmental Risk in Alzheimer's Dis-  
747 ease, Alzheimer's Disease Genetic Consortium, Cohorts for Heart and Aging  
748 Research in Genomic Epidemiology, Moebus, S., Mecocci, P., Del Zompo, M.,  
749 Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias,  
750 B., Gilbert, J.R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Car-  
751 rasquillo, M.M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O.,  
752 Younkin, S.G., Coto, E., Hamilton-Nelson, K.L., Gu, W., Razquin, C., Pastor, P.,  
753 Mateo, I., Owen, M.J., Faber, K.M., Jonsson, P.V., Combarros, O., O'Donovan,  
754 M.C., Cantwell, L.B., Soininen, H., Blacker, D., Mead, S., Mosley, T.H. Jr, Ben-  
755 nett, D.A., Harris, T.B., Fratiglioni, L., Holmes, C., Bruijn, R.F., Passmore, P.,  
756 Montine, T.J., Bettens, K., Rotter, J.I., Brice, A., Morgan, K., Foroud, T.M.,  
757 Kukull, W.A., Hannequin, D., Powell, J.F., Nalls, M.A., Ritchie, K., Lunetta,  
758 K.L., Kauwe, J.S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltunen,  
759 M., Martin, E.R., Schmidt, R., Rujescu, D., Wang, L.S., Dartigues, J.F., Mayeux,  
760 R., Tzourio, C., Hofman, A., Nöthen, M.M., Graff, C., Psaty, B.M., Jones, L.,  
761 Haines, J.L., Holmans, P.A., Lathrop, M., Pericak-Vance, M.A., Launer, L.J.,  
762 Farrer, L.A., Duijn, C.M., Van Broeckhoven, C., Moskva, V., Seshadri, S.,



- 763 Williams, J., Schellenberg, G.D., Amouyel, P.: Meta-analysis of 74,046 individu-  
764 als identifies 11 new susceptibility loci for alzheimer’s disease. *Nat. Genet.* **45**(12),  
765 1452–1458 (2013)
- 766 [46] Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S.,  
767 Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., Bras, J., Young, E., Coelln,  
768 R., Simón-Sánchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlström, L., Siiton-  
769 en, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J.R., Hernandez, D.G., Scholz,  
770 S.W., Botia, J.A., Martinez, M., Corvol, J.-C., Lesage, S., Jankovic, J., Shulman,  
771 L.M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O.A.,  
772 Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman,  
773 J.M., Wood, N.W., Hinds, D.A., Hardy, J.A., Morris, H.R., Gratten, J., Visscher,  
774 P.M., Graham, R.R., Singleton, A.B., 23andMe Research Team, System Genomics  
775 of Parkinson’s Disease Consortium, International Parkinson’s Disease Genomics  
776 Consortium: Identification of novel risk loci, causal insights, and heritable risk for  
777 parkinson’s disease: a meta-analysis of genome-wide association studies. *Lancet*  
778 *Neurol.* **18**(12), 1091–1102 (2019)
- 779 [47] Desikan, R.S., Schork, A.J., Wang, Y., Witoelar, A., Sharma, M., McEvoy, L.K.,  
780 Holland, D., Brewer, J.B., Chen, C.-H., Thompson, W.K., Harold, D., Williams,  
781 J., Owen, M.J., O’Donovan, M.C., Pericak-Vance, M.A., Mayeux, R., Haines,  
782 J.L., Farrer, L.A., Schellenberg, G.D., Heutink, P., Singleton, A.B., Brice, A.,  
783 Wood, N.W., Hardy, J., Martinez, M., Choi, S.H., DeStefano, A., Ikram, M.A.,  
784 Bis, J.C., Smith, A., Fitzpatrick, A.L., Launer, L., Duijn, C., Seshadri, S., Ulstein,  
785 I.D., Aarsland, D., Fladby, T., Djurovic, S., Hyman, B.T., Snaedal, J., Ste-  
786 fansson, H., Stefansson, K., Gasser, T., Andreassen, O.A., Dale, A.M., ADNI,  
787 ADGC, GERAD, CHARGE and IPDGC Investigators: Genetic overlap between  
788 alzheimer’s disease and parkinson’s disease at the MAPT locus. *Mol. Psychiatry*  
789 **20**(12), 1588–1595 (2015)
- 790 [48] Kuksa, P.P., Liu, C.-L., Fu, W., Qu, L., Zhao, Y., Katanic, Z., Clark, K., Kuzma,  
791 A.B., Ho, P.-C., Tzeng, K.-T., Valladares, O., Chou, S.-Y., Naj, A.C., Schel-  
792 lenberg, G.D., Wang, L.-S., Leung, Y.Y.: Alzheimer’s disease variant portal: A  
793 catalog of genetic findings for alzheimer’s disease. *J. Alzheimers. Dis.* **86**(1),  
794 461–477 (2022)
- 795 [49] Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X.: Rare-variant association  
796 testing for sequencing data with the sequence kernel association test. *Am. J. Hum.*  
797 *Genet.* **89**(1), 82–93 (2011)
- 798 [50] Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson,  
799 D.A., NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Chris-  
800 tiani, D.C., Wurfel, M.M., Lin, X.: Optimal unified approach for rare-variant  
801 association testing with application to small-sample case-control whole-exome  
802 sequencing studies. *Am. J. Hum. Genet.* **91**(2), 224–237 (2012)



- 803 [51] Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., Lin, X.: ACAT: A fast  
804 and powerful p value combination method for Rare-Variant analysis in sequencing  
805 studies. *Am. J. Hum. Genet.* **104**(3), 410–421 (2019)
- 806 [52] Beecham, G.W., Bis, J.C., Martin, E.R., Choi, S.-H., DeStefano, A.L., Duijn,  
807 C.M., Fornage, M., Gabriel, S.B., Koboldt, D.C., Larson, D.E., Naj, A.C., Psaty,  
808 B.M., Salerno, W., Bush, W.S., Foroud, T.M., Wijsman, E., Farrer, L.A., Goate,  
809 A., Haines, J.L., Pericak-Vance, M.A., Boerwinkle, E., Mayeux, R., Seshadri, S.,  
810 Schellenberg, G.: The alzheimer’s disease sequencing project: Study design and  
811 sample selection. *Neurol Genet* **3**(5), 194 (2017)
- 812 [53] Leung, Y.Y., Lee, W.-P., Kuzma, A.B., Gangadharan, P., Issen, H., Qu, L., Sreed-  
813 hara, M., Cantwell, L.B., Valladares, O., Zhao, Y., Vardarajan, B.N., Kunkle,  
814 B.W., Martin, E.R., Naj, A.C., Pericak-Vance, M.A., Wang, L.-S., Schellenberg,  
815 G.D., Alzheimer’s disease Sequencing Project: ADSP whole genome sequenc-  
816 ing (WGS) release 3 data update from genome center for alzheimer’s disease.  
817 *Alzheimers. Dement.* **19**(S1), 065290 (2023)
- 818 [54] Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., Chen, W.-M.:  
819 Robust relationship inference in genome-wide association studies. *Bioinformatics*  
820 **26**(22), 2867–2873 (2010)
- 821 [55] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender,  
822 D., Maller, J., Sklar, P., Bakker, P.I.W., Daly, M.J., Sham, P.C.: PLINK: a tool  
823 set for whole-genome association and population-based linkage analyses. *Am. J.*  
824 *Hum. Genet.* **81**(3), 559–575 (2007)
- 825 [56] Siva, N.: 1000 genomes project. *Nat. Biotechnol.* **26**(3), 256 (2008)
- 826 [57] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., Smoller, J.W.: Polygenic prediction  
827 via bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**(1),  
828 1776 (2019)
- 829 [58] Eli, B., Jonathan, P.C., Martin, J., Fritz, O., Neeraj, P., Theofanis, K., Rohit, S.,  
830 Paul, S., Paul, H., Noah, D.G.: Pyro: Deep universal probabilistic programming.  
831 *arXiv [cs.LG]* (2018) [cs.LG]
- 832 [59] Liu, X., White, S., Peng, B., Johnson, A.D., Brody, J.A., Li, A.H., Huang, Z.,  
833 Carroll, A., Wei, P., Gibbs, R., Klein, R.J., Boerwinkle, E.: WGSA: an annotation  
834 pipeline for human genome sequencing studies. *J. Med. Genet.* **53**(2), 111–112  
835 (2016)
- 836 [60] Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic,  
837 A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham,  
838 P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., Thomson, J.A.: The NIH roadmap  
839 epigenomics mapping consortium. *Nat. Biotechnol.* **28**(10), 1045–1048 (2010)

- 840 [61] Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q.,  
841 Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., Poterba, T., Wilson, M.W.,  
842 Tarasova, Y., Phu, W., Grant, R., Yohannes, M.T., Koenig, Z., Farjoun, Y.,  
843 Banks, E., Donnelly, S., Gabriel, S., Gupta, N., Ferriera, S., Tolonen, C., Novod,  
844 S., Bergelson, L., Roazen, D., Ruano-Rubio, V., Covarrubias, M., Llanwarne, C.,  
845 Petrillo, N., Wade, G., Jeandet, T., Munshi, R., Tibbetts, K., Genome Aggregation  
846 Database Consortium, O'Donnell-Luria, A., Solomonson, M., Seed, C.,  
847 Martin, A.R., Talkowski, M.E., Rehm, H.L., Daly, M.J., Tiao, G., Neale, B.M.,  
848 MacArthur, D.G., Karczewski, K.J.: A genomic mutational constraint map using  
849 variation in 76,156 human genomes. *Nature* **625**(7993), 92–100 (2024)
- 850 [62] Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang,  
851 Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D.,  
852 Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England,  
853 E.M., Seaby, E.G., Kosmicki, J.A., Walters, R.K., Tashman, K., Farjoun, Y.,  
854 Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha,  
855 K.E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A.H., Minikel, E.V.,  
856 Weisburd, B., Lek, M., Ware, J.S., Vittal, C., Armean, I.M., Bergelson, L., Cibul-  
857 skis, K., Connolly, K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S.,  
858 Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod,  
859 S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto,  
860 J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M.E., Genome Aggregation  
861 Database Consortium, Neale, B.M., Daly, M.J., MacArthur, D.G.: The muta-  
862 tional constraint spectrum quantified from variation in 141,456 humans. *Nature*  
863 **581**(7809), 434–443 (2020)
- 864 [63] Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S.,  
865 Schoech, A.P., Geijn, B., Reshef, Y., Márquez-Luna, C., O'Connor, L., Pirinen,  
866 M., Finucane, H.K., Price, A.L.: Functionally informed fine-mapping and poly-  
867 genic localization of complex trait heritability. *Nat Genet* **52**(12), 1355–1363  
868 (2020)
- 869 [64] Szustakowski, J.D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P.G.,  
870 Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., Katrina Loomis,  
871 A., Mikkilineni, R., Noh, H.J., Wadhawan, S., Bai, X., Hawes, A., Krasheninina,  
872 O., Ulloa, R., Lopez, A.E., Smith, E.N., Waring, J.F., Whelan, C.D., Tsai, E.A.,  
873 Overton, J.D., Salerno, W.J., Jacob, H., Szalma, S., Runz, H., Hinkle, G., Nioi,  
874 P., Petrovski, S., Miller, M.R., Baras, A., Mitnau, L.J., Reid, J.G., Moiseyenko,  
875 O., Rios, C., Saha, S., Abecasis, G., Banerjee, N., Beechert, C., Boutkov, B.,  
876 Cantor, M., Coppola, G., Economides, A., Eom, G., Forsythe, C., Fuller, E.D.,  
877 Gu, Z., Habegger, L., Jones, M.B., Lanche, R., Lattari, M., LeBlanc, M., Li, D.,  
878 Lotta, L.A., Manoochehri, K., Mansfield, A.J., Maxwell, E.K., Mighty, J., Nafde,  
879 M., O'Keeffe, S., Orelus, M., Padilla, M.S., Panea, R., Polanco, T., Pradhan, M.,  
880 Rasool, A., Schleicher, T.D., Sharma, D., Shuldiner, A., Staples, J.C., Van Hout,  
881 C.V., Widom, L., Wolf, S.E., John, S., Chen, C.-Y., Sexton, D., Kupelian, V.,  
882 Marshall, E., Swan, T., Eaton, S., Liu, J.Z., Loomis, S., Jensen, M., Duraisamy,

- 883 S., Tetrault, J., Merberg, D., Badola, S., Reppell, M., Grundstad, J., Zheng,  
884 X., Deaton, A.M., Parker, M.M., Ward, L.D., Flynn-Carroll, A.O., Austin, C.,  
885 March, R., Pangalos, M.N., Platt, A., Snowden, M., Matakidou, A., Wasilewski,  
886 S., Wang, Q., Deevi, S., Carss, K., Smith, K., Sogaard, M., Hu, X., Chen, X.,  
887 Ye, Z.: Advancing human genetics research and drug discovery through exome  
888 sequencing of the UK biobank. *Nat Genet* **53**(7), 942–948 (2021)
- 889 [65] Zhou, W., Zhao, Z., Nielsen, J.B., Fritsche, L.G., LeFaive, J., Gagliano Taliun,  
890 S.A., Bi, W., Gabrielsen, M.E., Daly, M.J., Neale, B.M., Hveem, K., Abecasis,  
891 G.R., Willer, C.J., Lee, S.: Scalable generalized linear mixed model for region-  
892 based association tests in large biobanks and cohorts. *Nat. Genet.* **52**(6), 634–639  
893 (2020)
- 894 [66] Londhe, S., Lindner, J., Chen, Z., Holtkamp, E., Hölzlwimmer, F.R., Casale, F.P.,  
895 Brechtmann, F., Gagneur, J.: Functional gene embeddings improve rare variant  
896 polygenic risk scores. *bioRxiv*, 2024–0722604535 (2024)