Supporting Information

Lorenzi et al. 10.1073/pnas.1706100115

SI Materials and Methods

Data Processing. The imaging phenotype comprised the baseline brain cortical thickness maps estimated with FreeSurfer 5.3 (1) and the bilateral radial thickness maps for hippocampi, amygdalae, thalami, caudate, putamen, globus pallidus, and nucleus accumbens. In detail, radial thickness of each subcortical surface model was based on the distance to a medial curve. We fit the medial curve using curve evolution individually for each shape (2). Surfaces are then registered parametrically to achieve point-to-point correspondence by matching curvature and medial curve-based features. The procedure resembles the cortical surface registration on the sphere performed in FreeSurfer. Finally, the full imaging component comprises 327,684 cortical and 27,120 subcortical features per subject.

SNP genotype data (Illumina Human610-Quad BeadChip for ADNI-1 and Illumina Human Omni Express for ADNI-GO/2) were downloaded from the ADNI website and preprocessed with PLINK (3). Standard quality control (QC) parameters were used to filter SNPs: minor allele frequency (MAF) < 0.01, genotype call rate <95%, and Hardy–Weinberg equilibrium P value $< 1 \times 10^{-6}$. Finally, genotyped SNPs passing QC were used to impute SNPs in the HapMap III reference panel. Imputed SNPs underwent a separate QC regarding MAF (>0.01) and imputation quality (imputation $r^2 > 0.3$) to exclude poorly imputed SNPs. For the analysis, the individuals' minor allele counts for each of the resulting 1,167,126 SNPs in the 22 autosomes were used.

Data matrices were preprocessed to remove effects from confounding variables (such as age) and make them eligible for PLS analysis. The influence of age, total intracranial volume, and sex was regressed from the raw thickness values. Next, they were standardized by groupwise mean and SD computed in the discovery set.

On the genetic input, missing individual SNPs were replaced by the groupwise median of the discovery set. In concordance with the phenotype input, the resulting allele counts were standardized by groupwise mean and SD in the discovery set.

PLS Modeling and Relevance Assessment. PLS was applied for modeling the joint variation between phenotype and genotype observed in the discovery set. The first five PLS components $\boldsymbol{v} = \{\boldsymbol{v}_i^g, \boldsymbol{v}_i^p\}$, $i \in \{1,2,3,4,5\}$, of joint genotype ($\boldsymbol{v}_i^g$) and phenotype ($\boldsymbol{v}_i^p$) variation were initially estimated, and their reproducibility and robustness were assessed through a stability selection scheme with split-half cross-validation based on 1 million repetitions (Fig. 1). Briefly, the 639 participants in the discovery set were randomly partitioned into two nonoverlapping subgroups of equal size (here denoted as G1 and G2). On each subgroup, PLS was independently estimated to compute the first five components of joint phenotype and genotype variation: $\boldsymbol{u_{G1}} = \{\boldsymbol{u}_j^{p1}, \boldsymbol{u}_j^{g1}\}$ and $\boldsymbol{u_{G2}} = \{\boldsymbol{u}_j^{p2}, \boldsymbol{u}_j^{g2}\}$, $j \in \{1,2,3,4,5\}$. Although the main patterns are often preserved, changes to the dataset may alter the order of the latent components with respect to the ones estimated in the whole cohort ($\{\boldsymbol{v}_i^g, \boldsymbol{v}_i^p\}$). Thus, component mappings $f_1(j)$ and $f_2(j)$ between $\boldsymbol{v}$ and the sets $\boldsymbol{u_{G1}}$ and $\boldsymbol{u_{G2}}$, respectively, were assessed by evaluating the similarity in the phenotype components (i.e., by measuring the absolute value of the dot product):

$$f_1(j) = \mathrm{argmax}_i\left(\left|\boldsymbol{v}_i^p \cdot \boldsymbol{u}_j^{p1}\right|\right),$$
$$f_2(j) = \mathrm{argmax}_i\left(\left|\boldsymbol{v}_i^p \cdot \boldsymbol{u}_j^{p2}\right|\right).$$

This quantity takes values in [0,1]. It is equal to one in the case that the components $\boldsymbol{v}_i^p$ and $\boldsymbol{u}_j^p$ are parallel (maximally similar); it equals zero in the case that the components are orthogonal (maximally dissimilar). No matching was declared if no index j could fulfill the condition $|\boldsymbol{v}_i^p \cdot \boldsymbol{u}_j^p| > 0.6$ (i.e., no component estimated on a data split could be mapped with sufficient similarity to one of the original components).

Assessing the Importance of a Genetic Locus. After mapping the components of the splits G1 and G2 to the components identified on all of the data, important and stable genetic loci were identified. First, the chromosomes were partitioned into 10-kb-sized bins. Among the resulting 277,889 bins, 90% of them contained at least 1 SNP, with, on average, 4.7 SNPs per bin.

Second, if a bin contained an SNP that received a large PLS weight (top 10% of absolute values) for both components $\boldsymbol{u}_l^{g1}$ and $\boldsymbol{u}_k^{g2}$ with $f_1(l) = f_2(k)$, then the bin was labeled one; otherwise, it was labeled zero (Fig. 1*A*). In particular, under the null hypothesis of independence between loci, the 10% threshold translates, by definition, to 0.1 probability for selecting a locus. Consequently, the chance to identify a locus in both split-half is $0.1^2 = 0.01$.

The resulting array averaged across all repetitions takes values in [0,1] and provides us with the null sampling distribution via permutation testing. This value thus indicates for each bin the selection probability in the PLS model in both independent random splits (Fig. 2) and serves as a measure of importance of the genomic location.

Methodological Considerations. The experimental setting proposed in this study is based on the investigation of potential genetic candidates in the AD and healthy training populations and on their testing in the MCI cohort. This experimental choice was motivated by clinical and practical considerations.

From the clinical point of view, although we cannot exclude that the imaging–genetics association patterns could be modulated by state-specific factors throughout the development of the disease (4), the heterogeneity of the MCI label is likely to lead to the inclusion in the discovery dataset of individuals with non-AD pathologies. Thus, including MCIs in the discovery cohort bears the risk of diluting the gene finding (especially considering the relatively low sample size of the study cohort). Likewise, GWAS in AD carried out to date focus on comparing healthy controls (CT) with AD. Moreover, the paradigm proposed in this study is rather conservative, since it explores associations present throughout the progression of the pathology (i.e., associations were discovered by comparing CT and AD subjects and validated on disease progression in the intermediate MCI cohort). This consideration, while being more conservative, may play in favor of the robustness of the reported results. From a practical point of view, the proposed scheme allowed for the validation of the model on a clinically relevant testing cohort by taking advantage of the full sample available in the ADNI dataset. Splitting the available AD and CT subjects into discovery and validation cohorts would have dramatically reduced the sample size, thus increasing the uncertainty of the PLS findings.

Concerning the number of components analyzed in the PLS model, we limited the study to the exploration of the first five eigenmodes. As shown in the experimental results, the stability of PLS parameters of the high-order components was generally quite low and did not lead to any significant results after permutation testing. For this reason, we believe that extending the analysis to higher-order components (e.g., components 6–10) would not change the proposed analysis and subsequent results.

The relevance assessment procedure proposed in this study relies on the choice of statistical significance thresholds, such as the 10% cutoff on the magnitude of the PLS weights and $P < 0.05$ for the selection frequency over the 1 million folds. These thresholds were not optimized to maximize specific statistical outcomes (e.g., the ratio between true and false positives). Indeed, the optimization of these parameters may lead to important methodological issues, such as overfitting and selection bias (5), and ultimately, they may lead to poor generalization of the statistical findings. This is particularly true in the challenging setting proposed in this work characterized by large dimensions and low sample size. For this reason, we chose to use standard cutoffs for significance assessment as a compromise between minimizing this important source of bias and still identifying meaningful genotype and phenotype features. Furthermore, we believe that the ultimate approach to assess the validity of the findings is through testing on genuinely independent data, such as the MCI cohort proposed in this study.

1. Fischl B (2012) FreeSurfer. *Neuroimage* 62:774–781.
2. Gutman BA, Madsen SK, Toga AW, Thompson PM (2013) A family of fast Spherical registration algorithms for cortical shapes. *Proceedings of the International Workshop on Multimodal Brain Image Analysis*, eds Shen L, Liu T, Yap PT, Huang H, Shen D, Westin CF (Springer, Berlin), pp 246–257.
3. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
4. Stage E, et al. (2016) The effect of the top 20 Alzheimer disease risk genes on gray-matter density and FDG PET brain metabolism. *Alzheimers Dement (Amst)* 5:53–66.
5. Mendelson AF, Zuluaga MA, Lorenzi M, Hutton BF, Ourselin S; Alzheimer's Disease Neuroimaging Initiative (2016) Selection bias in the reported performances of AD classification pipelines. *Neuroimage Clin* 14:400–416.

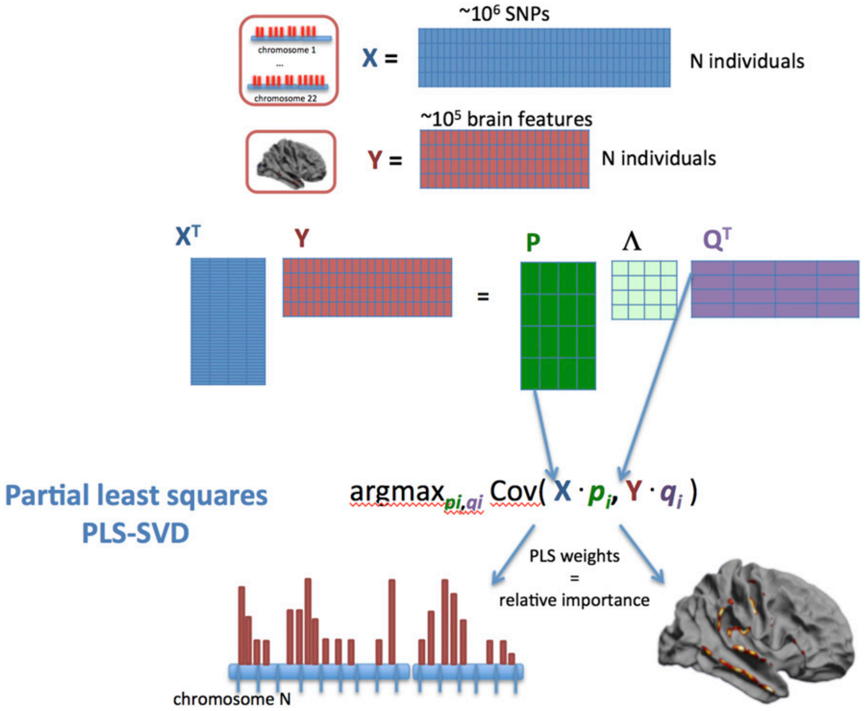

Fig. S1. PLS framework. PLS modeling rationale. The latent PLS components are obtained through the singular value decomposition (SVD) of the covariance matrix (**C**) between genetic features **X** and phenotype features **Y**: $\mathbf{C} = \mathbf{X}^\mathbf{T}\mathbf{Y}$. **C** has the dimension "number of SNPs" × "number of brain features" ($\sim 10^6 \times \sim 10^5$). SVD can be used decompose $\mathbf{C} = \mathbf{X}^\mathbf{T}\mathbf{Y} = \mathbf{P}\ \boldsymbol{\Lambda}\ \mathbf{Q}^\mathbf{T}$. The diagonal matrix $\boldsymbol{\Lambda}$ contains the eigenvalues, and the columns p_i of **P** (columns q_i of **Q**) are the principal eigencomponents that will be subsequently analyzed as detailed in the text. The projection of **X** (or **Y**) is achieved through multiplication with **P** (**Q**): $\mathbf{P_x} = \mathbf{XP}$ ($\mathbf{P_y} = \mathbf{YQ}$).

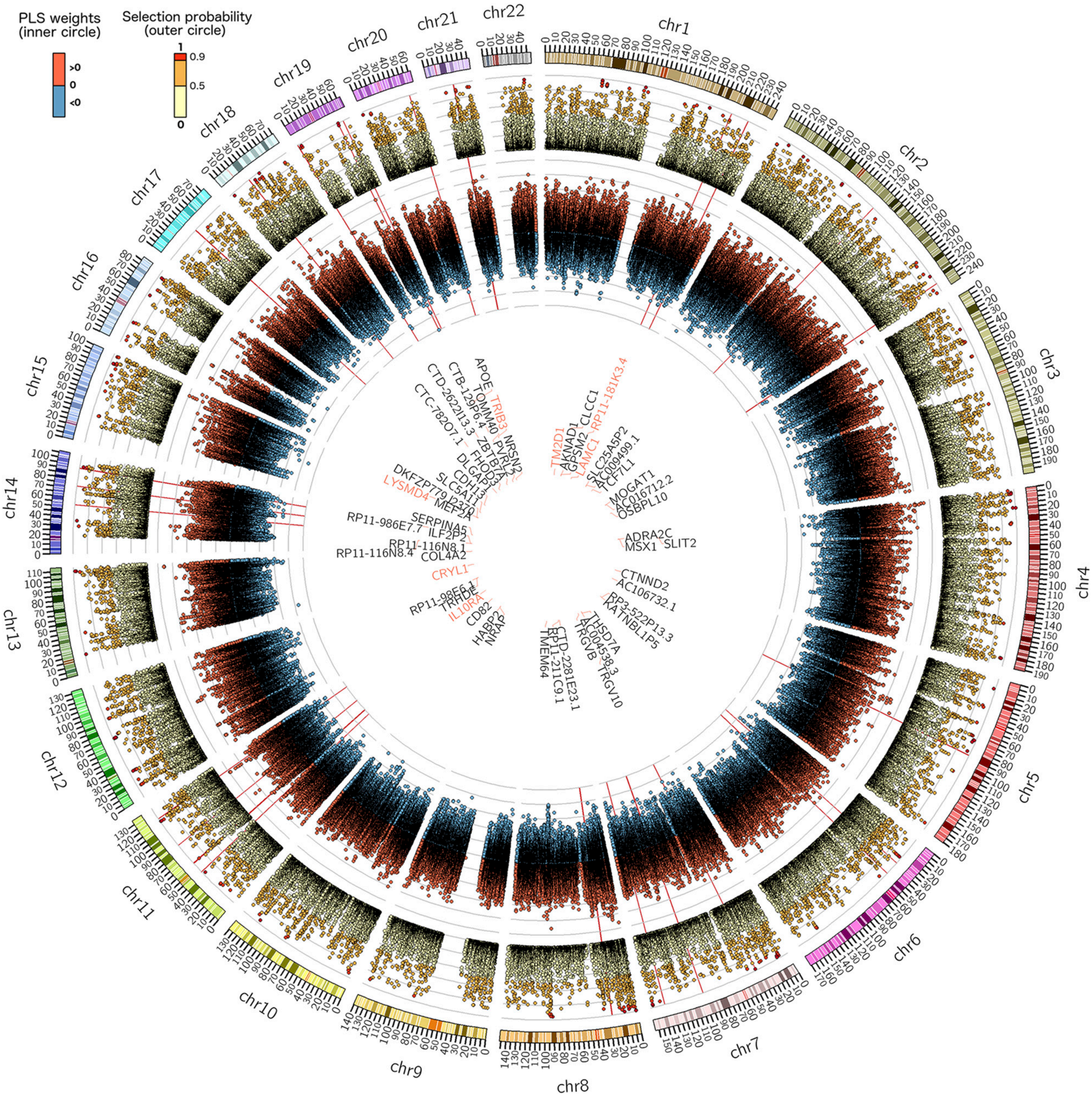

Fig. S2. First PLS component. The outer circular plots show the probability of a given genetic locus being associated with the phenotype component 1. The inner circular plots show the PLS weights associated with each genetic locus (red, positive; blue, negative). The genes close to the important loci ($P > 0.95$) are listed in the innermost circle depending on their genomic position; genes with eQTL are highlighted in red. The red radial lines are located in correspondence of known AD genes.

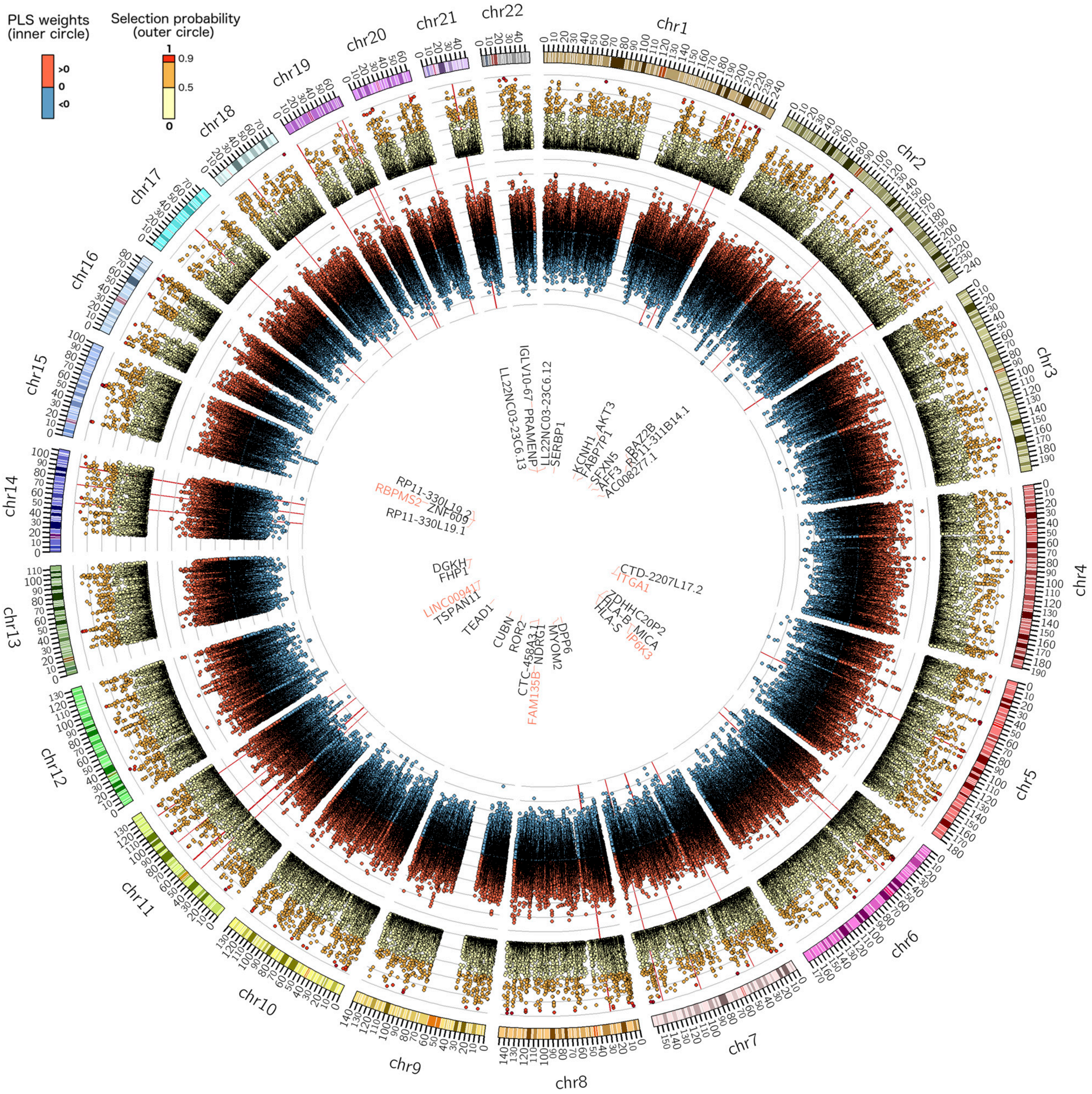

Fig. S3. Second PLS component. The outer circular plots show the probability of a given genetic locus being associated with the phenotype component 2. The inner circular plots show the PLS weights associated with each genetic locus (red, positive; blue, negative). The genes close to the important loci ($P > 0.95$) are listed in the innermost circle depending on their genomic position; genes with eQTL are highlighted in red. The red radial lines are located in correspondence of known AD genes.

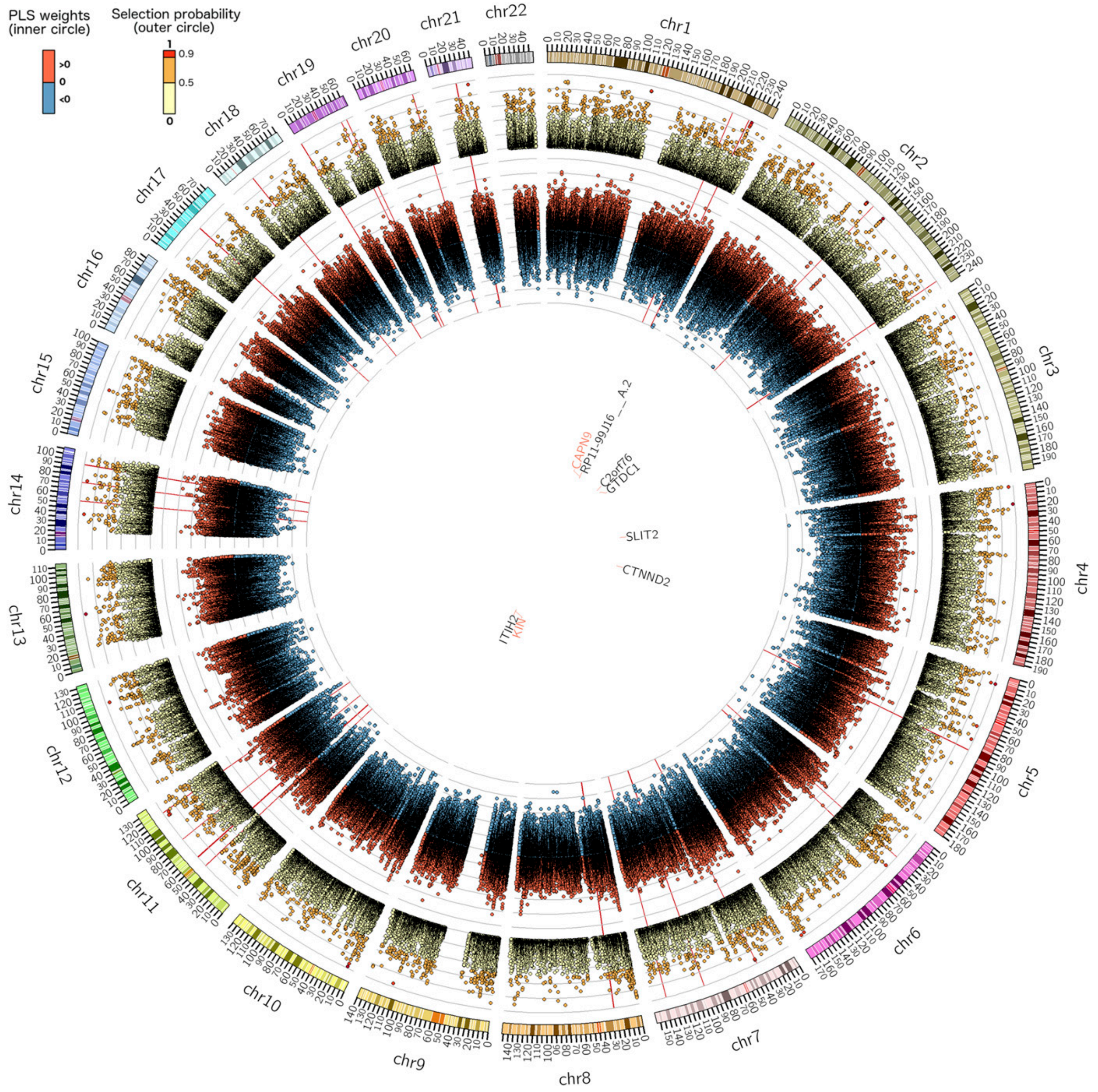

Fig. S4. Third PLS component. The outer circular plots show the probability of a given genetic locus being associated with the phenotype component 3. The inner circular plots show the PLS weights associated with each genetic locus (red, positive; blue, negative). The genes close to the important loci ($P > 0.95$) are listed in the innermost circle depending on their genomic position; genes with eQTL are highlighted in red. The red radial lines are located in correspondence of known AD genes.

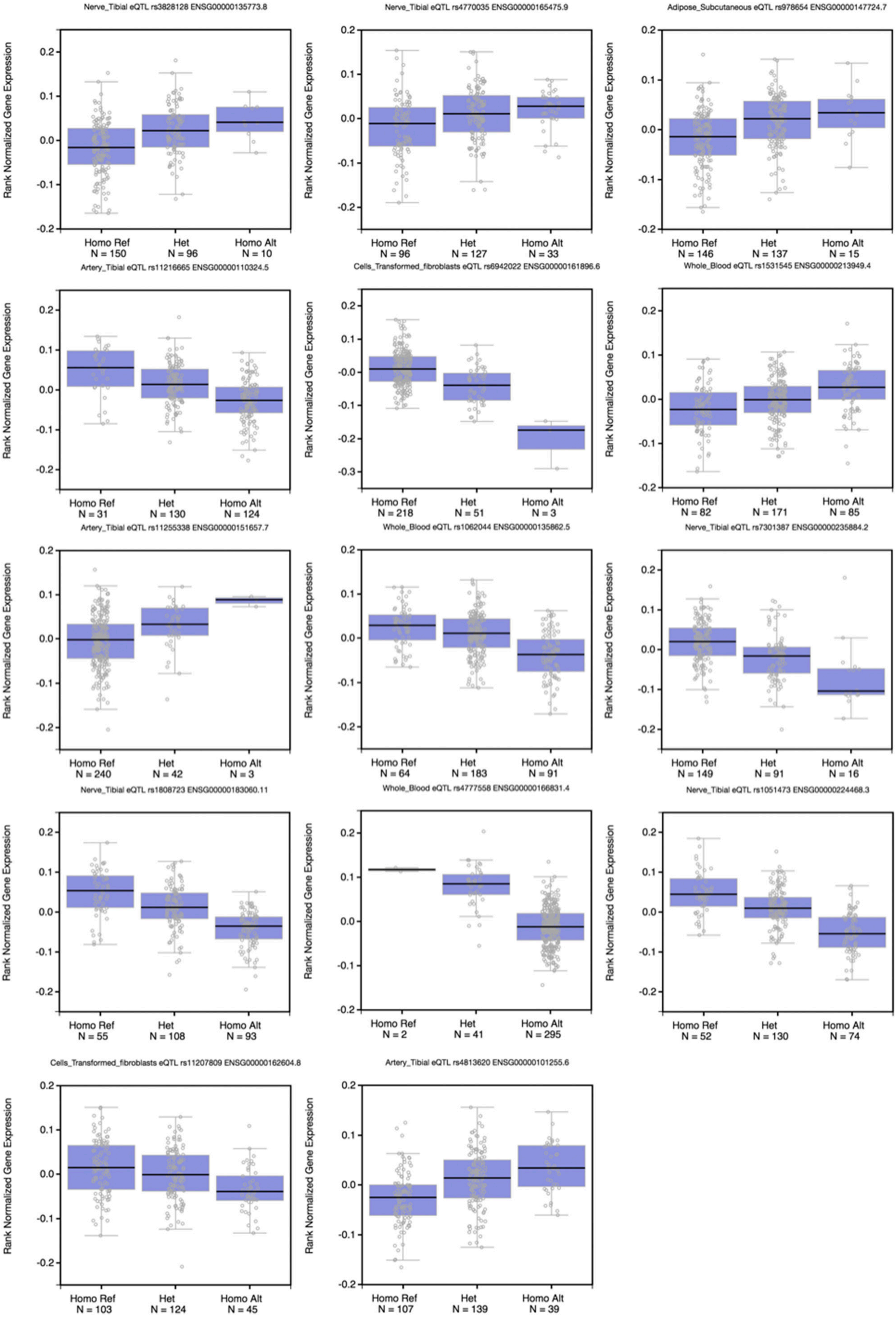

Fig. S5. Gene expression by SNP for 14 genes from GTEx. The *y* axes depict rank normalized gene expression, while on the *x* axes, the status and sample size for each allele are provided. The caption of each subfigure states the tissue, rs number of the SNP, and gene (Ensembl identifier). The genes in left to right and top to bottom order with corresponding *P* values in parentheses are *CAPN9* ($P = 5.3e{-}9$), *CRYL1* ($P = 1.5e{-}5$), *FAM135B* ($P = 2.1e{-}8$), *IL10RA* ($P = 1.5e{-}14$), *IP6K3* ($P = 5.7e{-}16$), *ITGA1* ($P = 4.8e{-}10$), *KIN* ($P = 1.6e{-}5$), *LAMC1* ($P = 1.5e{-}15$), *LINC00941* ($P = 7.1e{-}13$), *LYSMD4* ($P = 2.9e{-}24$), *RBPMS2* ($P = 2.0e{-}38$), *RP11-181K3.4* ($P = 1.5e{-}25$), *TM2D1* ($P = 1.2e{-}6$), and *TRIB3* ($P = 6.3e{-}12$). Het, heterozygous; Homo Alt, homozygote for the alternative allele; Homo Ref, homozygote for the reference allele used in GTEx.

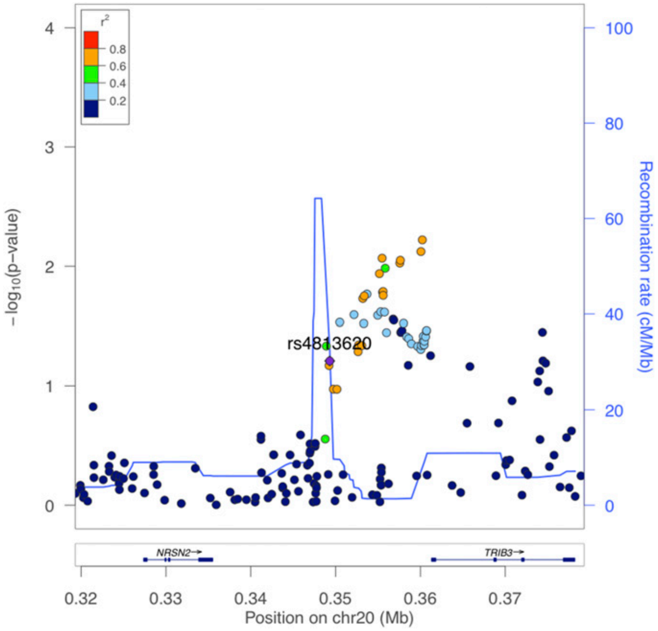

Fig. S6. Association strength with AD status for the genetic neighborhood of rs4813620. Regional association plot generated with LocusZoom (1) for the neighborhood (±30 kb) of rs4813620. *P* values were obtained from the stage 1 results from the International Genomics of Alzheimer's Project Study comprising 17,008 cases and 37,154 controls (2). The *y* axis shows the −log10 *P* value of the case–control association test, and the *x* axis shows the genomic location. The target SNP (rs4813620) is colored purple, and other SNPs are colored corresponding to the LD with the target SNP.

1. Pruim RJ, et al. (2010) LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* 26:2336–2337.
2. Lambert JC, et al.; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 45:1452–1458.

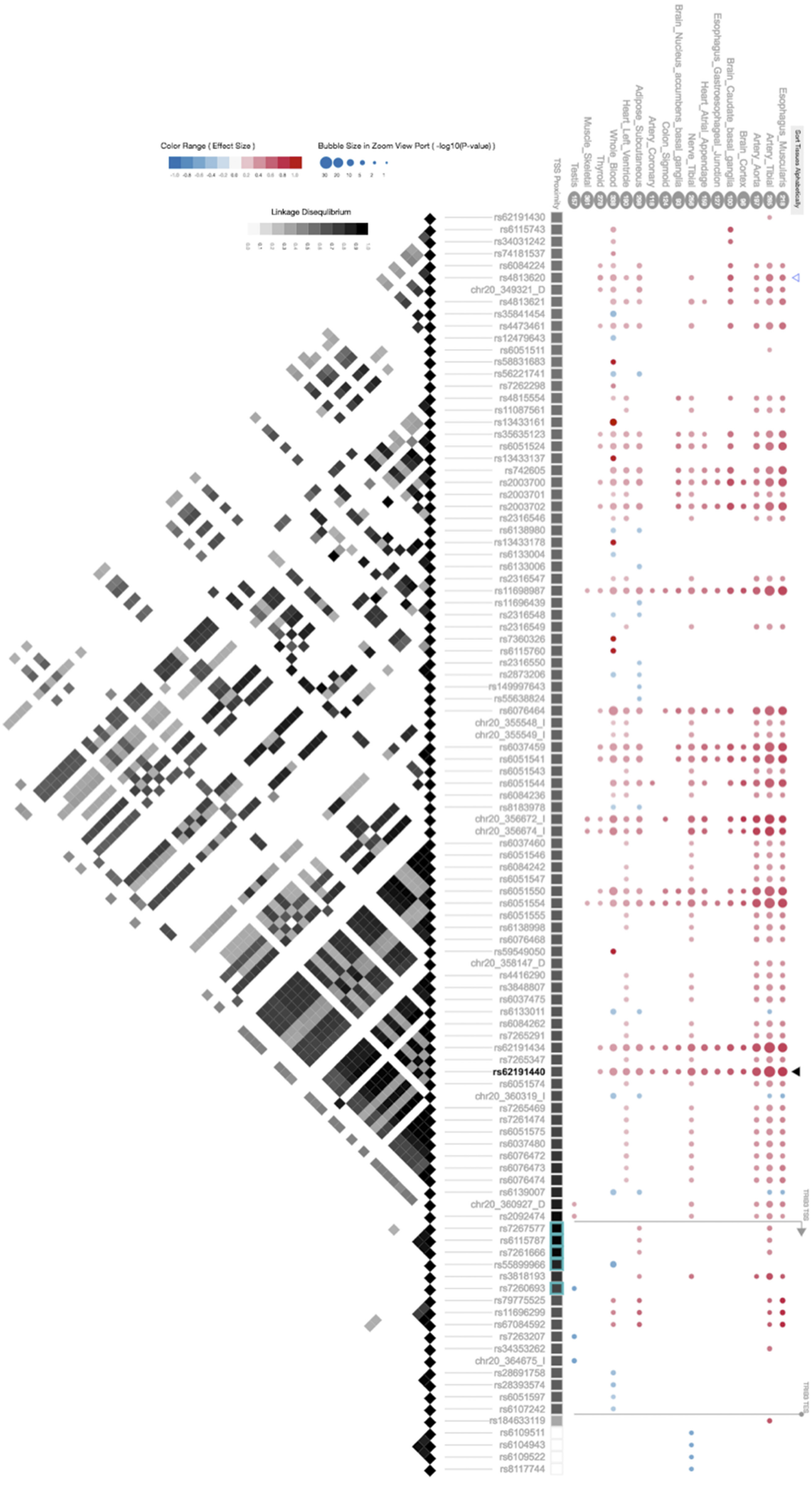

Fig. S7. Detailed eQTL overview of *TRIB3* provided by GTEx. *Right* shows the association strength between SNPs upstream of *TRIB3* and *TRIB3* expression in 17 tissues. Association direction is color-coded, and association strength is expressed in bubble sizes. Tissues are ordered with respect to the effect size of rs62191440, which is also highlighted in bold. The SNP identified in the PLS model (rs4813620) is marked with a blue triangle. Transcription start site and transcription end site of *TRIB3* are highlighted in the lower part. *Left* depicts the LD structure of the upstream region of *TRIB3*.

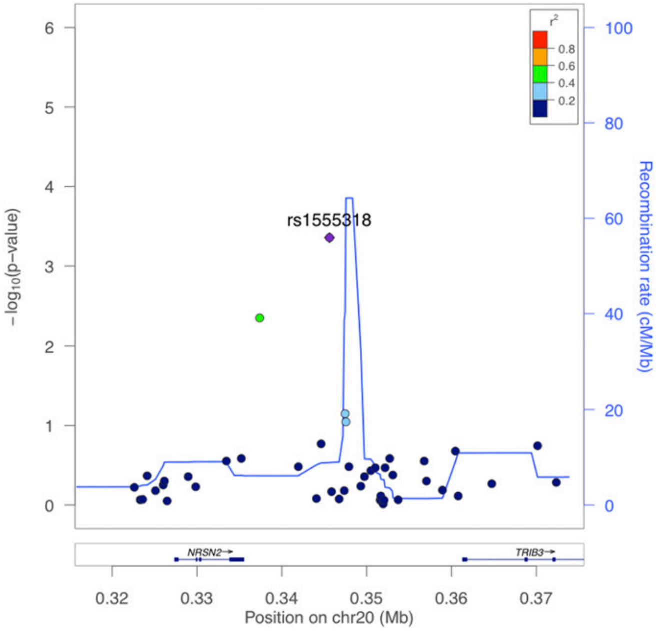

Fig. S8. Regional association with type 2 diabetes. Regional association plot generated with LocusZoom (1) for the neighborhood (±30 kb) of rs1555318. *P* values were obtained from stage 1 of a large GWAS for type 2 diabetes comprising 12,171 cases and 56,862 controls (2). The *y* axis shows the −log10 *P* value of the case–control association test, and the *x* axis shows the genomic location. The target SNP (rs1555318) is colored purple, and other SNPs are colored corresponding to the LD with the target SNP.

1. Pruim RJ, et al. (2010) LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* 26:2336–2337.

2. Morris AP, et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 44:981–990.

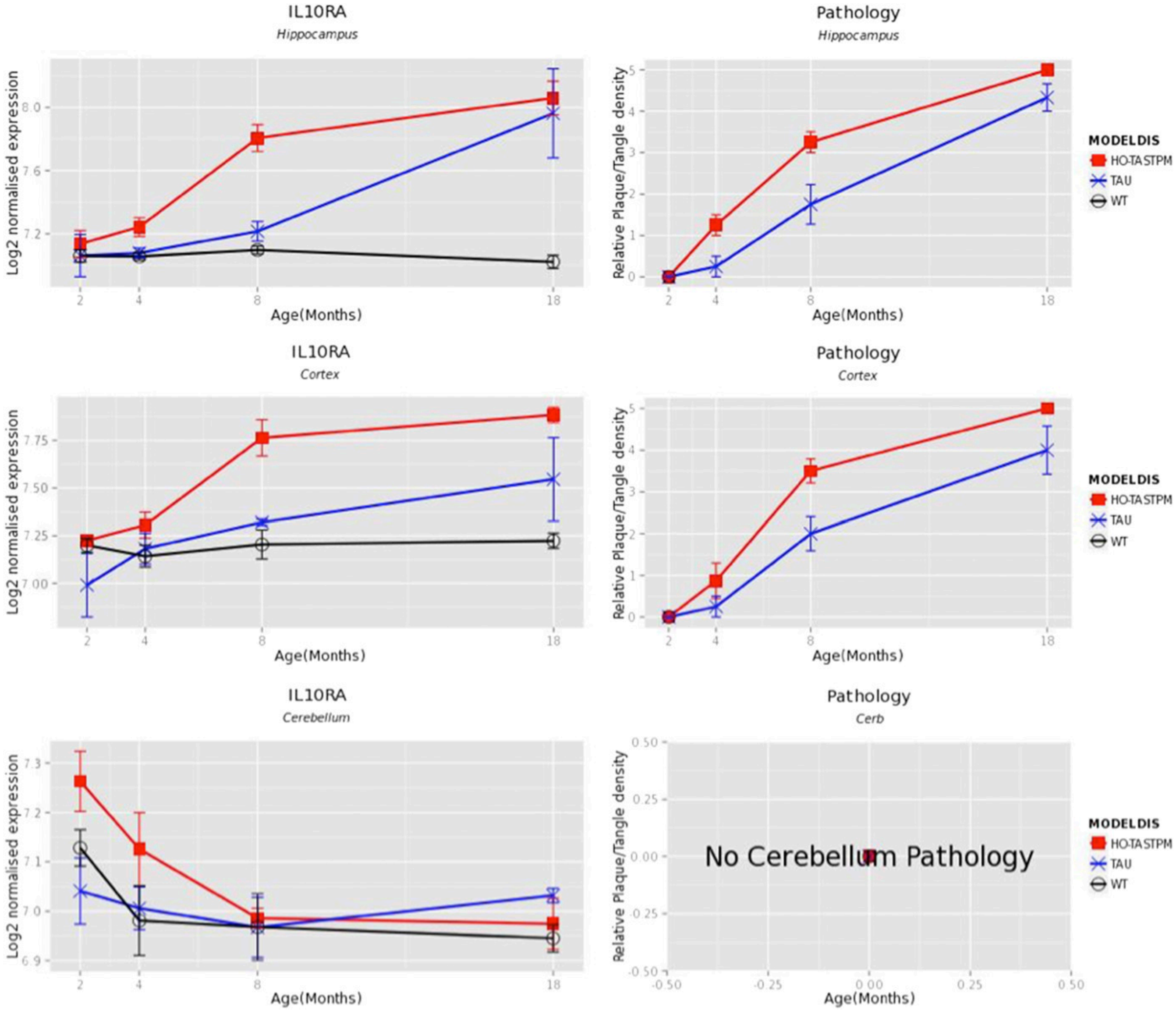

Fig. S9. Gene expression of *Il10ra* in transgenic mouse models from MOUSEAC. Il10ra gene expression (*Left*) and AD pathology (*Right*) in transgenic and wild-type mice were obtained from MOUSEAC (1). Data are shown for wild-type mice (black), transgenic mice with MAPT mutation P301L (blue), and transgenic mice with homozygous mutations in APP (K670N and M671L) and PSEN1 (M146V; red). The *x* axes depict age in months, and the *y* axes show gene expression (*Left*) and plague/tangle density (*Right*).

1. Matarin M, et al. (2015) A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. *Cell Rep* 10:633–644.

Table S1. Statistical testing (*P* values) of prioritized genes with respect to the models estimated on the amyloid-positive subcohort

Gene	Full	Amyloid positive
TM2D1	0.0528	0.0717
IL10RA	0.6198	0.0777
TRIB3	0.0034	0.0134
ZBTB7A	0.9135	0.557
LYSMD4	0.2057	0.7208
CRYL1	0.1176	0.1176
FAM135B	0.5588	0.0506
IP6K3	0.4646	0.2907
ITGA1	0.731	0.9677
KIN	0.2061	0.1736
LAMC1	0.0618	0.0665
LINC00941	0.6896	0.17
RBPMS2	0.2149	0.3547
RP11-181K3.4	0.0527	0.0756

When using the model estimated on the amyloid-positive individuals only, *TRIB3* still leads to significant differences between progressing and stable MCIs, although it was not significant after Bonferroni correction for multiple comparison.

Dataset S1. GENCODE gene annotation results

Dataset S1

Dataset S2. Functional prioritization through eQTL analysis based on the GTEx data GTEx-based eQTL

Dataset S2