Supplementary Information

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) was 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 (HWE) p-value < $1x10^{-6}$. Finally, genotyped SNPs passing QC were used to impute SNPs in the HapMap III reference panel. Imputed SNPs underwent a separate QC regarding minor allele frequency (MAF > 0.01) and imputation quality (imputation R-squared > 0.3) in order 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 in order 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 group-wise mean and standard deviation computed in the discovery set.

On the genetic input, missing individual SNPs were replaced by the group-wise median of the discovery set. In concordance with the phenotype input, the resulting allele counts were standardized by group-wise mean and standard deviation 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 1,000,000 repetitions (Figure 1). Briefly, the 639 participants in the discovery set were randomly partitioned into two non-overlapping subgroups of equal size (here denoted 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:

$$f_1(j) = \text{argmax}_i \left(|\mathbf{v}_i^p \cdot \mathbf{u}_j^{p1}| \right),$$

$$f_2(j) = \text{argmax}_i \left(|\mathbf{v}_i^p \cdot \mathbf{u}_j^{p2}| \right)$$

This quantity takes values in [0,1], and is equal to 1 in case the components $\mathbf{v}_i^p$ and $\mathbf{u}_j^p$ are parallel (maximally similar); it equals 0 in case the components are orthogonal (maximally dissimilar). No matching was declared if no index *j* could fulfill the condition $|\mathbf{v}_i^p \cdot \mathbf{u}_j^p| > 0.6$, 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 the entire data, important and stable genetic loci were identified. First, the chromosomes were partitioned into 10kb sized bins. Among the resulting 277,889 bins, 90% of them contained at least one SNP, with on average 4.7 SNPs per bin.

Second, if a bin contained a SNP that received a large PLS weight (top 10% of absolute values) for both components $\mathbf{u}_l^{g1}$ and $\mathbf{u}_k^{g2}$ with $f_1(l)=f_2(k)$, then the bin was labeled 1; otherwise 0 (Figure 1A). 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 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 (Figure 2, left), 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 population, 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 CT and 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 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 cohort, 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 eigen-modes. 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 six to ten) 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 1,000,000 folds. These thresholds were not optimized to maximize specific statistical outcome (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 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 while 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 on the MCI cohort proposed in this study.

ADNI Acknowledgement List

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study

is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

Dataset S1. GENCODE gene annotation results.

Dataset S2. Functional prioritization through expression quantitative trait loci (eQTL) analysis based on the Genotype-Tissue-Expression project (GTEx) data GTEx-based eQTL

Table S1. Statistical testing (p-values) of prioritized genes with respect to the models estimated on the amyloid positive sub-cohort. When using the model estimated on the amyloid positive individuals only, TRIB3 still leads to significant differences between progressing and stable MCI, although non significant after Bonferroni correction for multiple comparison.

Figure S1: PLS framework.

Partial Least Squares (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**: $\mathbf{C} = \mathbf{X}^T\mathbf{Y}$. **C** has the dimension "number of SNPs" x "number of brain features" (~10^6 x ~10^5). SVD can be used decompose $\mathbf{C} = \mathbf{X}^T\mathbf{Y} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{Q}^T$. The diagonal matrix $\mathbf{\Lambda}$ contains the eigenvalues, the columns p_i of **P** (resp. columns q_i of **Q**) are the principal eigen-components that will be subsequently analyzed as detailed in manuscript section "Statistical Analysis". The projection of **X** (or **Y**) is achieved through multiplication with **P** (**Q**): $\mathbf{P_x} = \mathbf{XP}$ ($\mathbf{P_y} = \mathbf{YQ}$).

Figure S2: First PLS Component. The outer circular plots show the probability of a given genetic locus to be associated with the phenotype component 1. The inner circular plots show the PLS weights associated to 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 eQTLs are highlighted by red font. The red radial lines are located in correspondence of known AD genes.

Figure S3: Second PLS Component. The outer circular plots show the probability of a given genetic locus to be associated with the phenotype component 2. The inner circular plots show the PLS weights associated to 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 eQTLs are highlighted by red font. The red radial lines are located in correspondence of known AD genes.

Figure S4: Third PLS Component. The outer circular plots show the probability of a given genetic locus to be associated with the phenotype component 3. The inner circular plots show

the PLS weights associated to 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 eQTLs are highlighted by red font. The red radial lines are located in correspondence of known AD genes.

Figure S5: Gene expression by SNP for 14 genes from GTEx.

The Y-axis depicts rank normalized gene expression, while on the X-axis the status and sample size for each allele is 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).

Figure S6: Association strength with AD status for the genetic neighborhood of rs4813620.

Regional association plot generated with LocusZoom [6] for the neighborhood (+/- 30kb) of rs4813620. P-values were obtained from the stage 1 results from the International Genomics of Alzheimer's Project (IGAP) study comprising 17,008 cases and 37,154 controls [7]. Y-axis shows the –log10 p-value of the case-control association test, and X-axis the genomic location. The target SNP (rs4813620) is colored purple and other SNPs are colored corresponding to the LD with the target SNP.

Figure S7: Detailed eQTL overview of *TRIB3* provided by GTEx.

The right part of the figure 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 font. The SNP identified in the PLS model (rs4813620) is marked with a blue triangle. Transcription start site (TSS) and transcription end site (TES) of *TRIB3* are highlighted in the lower part. The left part of the figure depicts the LD structure of the upstream region of *TRIB3*.

Figure S8: Regional association with T2D.
Regional association plot generated with LocusZoom [6] for the neighborhood (+/- 30kb) of rs1555318. P-values were obtained from stage I of a large GWAS for type 2 diabetes comprising 12,171 cases and 56,862 controls [8]. Y-axis shows the –log10 p-value of the case-control association test, and X-axis the genomic location. The target SNP (rs1555318) is colored purple and other SNPs are colored corresponding to the LD with the target SNP.

Figure S9: Gene expression of *Il10ra* in transgenic mouse models from MOUSEAC.
Il10ra gene expression (left column) and AD pathology (right column) in transgenic and wild-type mice obtained from MOUSEAC [9] . Data is 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). X-axis depicts age in months and y-axis gene expression and plague/tangle density, respectively.

Bibliography

1 Fischl, B. FreeSurfer. In *NeuroImage* (2), 774-781.

2 Gutman BA, Madsen SK, Toga AW, Thompson PM. A family of fast Spherical registration algorithms for cortical shapes. In *International Workshop on Multimodal Brain Image Analysis* (2013), Springer International Publishing, 246-257.

3 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. *Am J Hum Genet*, 81, 3 (Sep 2007), 559-575.

4 Stage, E, Duran, T, Risacher, SL, Goukasian, N, Do, TM, West, JD, Wilhalme, H, Nho, K, Phillips, M, Elashoff, D, Saykin, AJ & Apostolova, LG. The effect of the top 20 Alzheimer disease risk genes on gray-matter density and FDG PET brain metabolism. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 5 (2016), 53-66.

5 Mendelson, A. F., Zuluaga, M. A., Lorenzi, M., Hutton, B. F., Ourselin, S.. Selection bias in the reported performances of AD classification pipelines. *NeuroImage: Clinical*, 14 (2016), 400-416.

6 Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* , 26, 18 (2010), 2336-2337.

7 EUROPEAN ALZHEIMER'S DISEASE INITIATIVE (EADI). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*, 45, 12 (2013), 1452-1458.

8 Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44, 9 (2012), 981.

9 Matarin, M., Salih, D. A., Yasvoina, M., Cummings, D. M., Guelfi, S., Liu, W., et al. A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. *Cell reports*, 10, 4 (2015), 633-644.