PLOS | GENETICS

RESEARCH ARTICLE

Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease

Qiongshi Lu[1☯¤], Ryan L. Powles[2☯], Sarah Abdallah[3], Derek Ou[3], Qian Wang[2], Yiming Hu[1], Yisi Lu[4], Wei Liu[5], Boyang Li[1], Shubhabrata Mukherjee[6], Paul K. Crane[6], Hongyu Zhao[1,2,7]*

1 Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America, **2** Program of Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **3** Yale University School of Medicine, New Haven, Connecticut, United States of America, **4** Department of Immunobiology, Yale University School of Medicine, New Haven, Connecticut, United States of America, **5** School of Life Sciences, Peking University, Beijing, China, **6** Division of General Internal Medicine, Department of Medicine, University of Washington, Seattle, Washington, United States of America, **7** VA Cooperative Studies Program Coordinating Center, West Haven, Connecticut, United States of America

☯ These authors contributed equally to this work.
¤ Current address: Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America
* hongyu.zhao@yale.edu

OPEN ACCESS

Citation: Lu Q, Powles RL, Abdallah S, Ou D, Wang Q, Hu Y, et al. (2017) Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. PLoS Genet 13(7): e1006933. https://doi.org/10.1371/journal.pgen.1006933

Editor: Xiaofeng Zhu, Case Western Reserve University, UNITED STATES

Received: March 20, 2017

Accepted: July 18, 2017

Published: July 24, 2017

Data Availability Statement: All annotation data and processing code are available from the GenoSkyline server at http://genocanyon.med.yale.edu/GenoSkyline.

Funding: This study was supported in part by the National Institutes of Health grants R01 GM59507, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, and the Yale World Scholars Program sponsored by the China Scholarship

Abstract

Continuing efforts from large international consortia have made genome-wide epigenomic and transcriptomic annotation data publicly available for a variety of cell and tissue types. However, synthesis of these datasets into effective summary metrics to characterize the functional non-coding genome remains a challenge. Here, we present GenoSkyline-Plus, an extension of our previous work through integration of an expanded set of epigenomic and transcriptomic annotations to produce high-resolution, single tissue annotations. After validating our annotations with a catalog of tissue-specific non-coding elements previously identified in the literature, we apply our method using data from 127 different cell and tissue types to present an atlas of heritability enrichment across 45 different GWAS traits. We show that broader organ system categories (e.g. immune system) increase statistical power in identifying biologically relevant tissue types for complex diseases while annotations of individual cell types (e.g. monocytes or B-cells) provide deeper insights into disease etiology. Additionally, we use our GenoSkyline-Plus annotations in an in-depth case study of late-onset Alzheimer's disease (LOAD). Our analyses suggest a strong connection between LOAD heritability and genetic variants contained in regions of the genome functional in monocytes. Furthermore, we show that LOAD shares a similar localization of SNPs to monocyte-functional regions with Parkinson's disease. Overall, we demonstrate that integrated genome annotations at the single tissue level provide a valuable tool for understanding the etiology of complex human diseases. Our

Council. PKC's and SM's efforts were supported by grant R01 AG042437 and U01 AG006781. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

GenoSkyline-Plus annotations are freely available at http://genocanyon.med.yale.edu/GenoSkyline.

Author summary

After years of community efforts, many experimental and computational approaches have been developed and applied for functional annotation of the human genome, yet proper annotation still remains challenging, especially in non-coding regions. As complex disease research rapidly advances, increasing evidence suggests that non-coding regulatory DNA elements may be the primary regions harboring risk variants in human complex diseases. In this paper, we introduce GenoSkyline-Plus, a principled annotation framework to identify tissue and cell type-specific functional regions in the human genome through integration of diverse high-throughput epigenomic and transcriptomic data. Through validation of known non-coding tissue-specific regulatory regions, enrichment analyses on 45 complex traits, and an in-depth case study of neurodegenerative diseases, we demonstrate the ability of GenoSkyline-Plus to accurately identify tissue-specific functionality in the human genome and provide unbiased, genome-wide insights into the genetic basis of human complex diseases.

Introduction

Large consortia such as ENCODE [1] and Epigenomics Roadmap Project [2] have generated a rich collection of high-throughput genomic and epigenomic data, providing unprecedented opportunities to delineate functional structures in the human genome. As complex disease research rapidly advances, evidence has emerged that disease-associated variants are enriched in regulatory DNA elements [3, 4]. Therefore, functional annotation of the non-coding genome is critical for understanding the genetic basis of human complex diseases. Unfortunately, categorizing the complex regulatory machinery of the genome requires integration of diverse types of annotation data as no single annotation captures all types of functional elements [5]. Recently, we have developed GenoSkyline [6], a principled framework to identify tissue-specific functional regions in the human genome through integrative analysis of various chromatin modifications. In this work, we introduce GenoSkyline-Plus, a comprehensive update of GenoSkyline that incorporates RNA sequencing and DNA methylation data into the framework and extends to 127 integrated annotation tracks covering a spectrum of human tissue and cell types.

To demonstrate the ability of GenoSkyline-Plus to systematically provide novel insights into complex disease etiology, we jointly analyzed summary statistics from 45 genome-wide association studies (GWAS; $N_{total} \approx 3.8M$) and identified biologically relevant tissues for a broad spectrum of complex traits. We next performed an in-depth, annotation-driven investigation of Alzheimer's disease (AD), a neurodegenerative disease characterized by deposition of amyloid-β (Aβ) plaques and neurofibrillary tangles in the brain. Late-onset AD (LOAD) includes patients with onset after 65 years of age and has a complex mode of inheritance [7]. Around 20 risk-associated genetic loci have been identified in LOAD GWAS [8]. However, our understanding of LOAD's genetic architecture and disease etiology is still far from complete. Through integrative analysis of GWAS summary data and GenoSkyline-Plus annotations, we identified strong enrichment for LOAD associations in immune cell-related DNA

elements, consistent with other data suggesting a crucial role for the immune system in AD etiology [9–11]. Jointly analyzing GWAS summary data for LOAD and Parkinson's disease (PD), we identified substantial enrichment for pleiotropic associations in the monocyte functional genome. Our findings provide support for the critical involvement of the immune system in the etiology of neurodegenerative diseases, and suggest a previously unsuspected role for an immune-mediated pleiotropic effect between LOAD and PD.

Results

Identify tissue and cell type-specific functionality in the human genome

We use our previously established statistical framework to calculate the posterior probability of functionality for each nucleotide in the human genome [12]. Integrating tissue and cell-specific genomic functional data available through Epigenomics Roadmap Project [2], we make available GenoSkyline-Plus scores for 127 individual tissue annotation tracks (Methods; S1 Table). H3K4me3 and H3K9ac, known markers of open chromatin and active transcription [13], are shown to have the largest odds ratios of predicting functionality across the genome (Fig 1A). Identifying H3K4me3 and H3K9ac as strong indicators of genomic functionality is a finding consistent with previous studies of gene regulation through chromatin marks [14]. In contrast, H3K9me3, a well established repressive mark [13], has a reversed effect on genome functionality. The bimodal pattern of GenoSkyline scores [6] allows us to impose a score cutoff to robustly define the functional genome. Using a cutoff of 0.5, 3% of the genome is considered functional on average across all annotation tracks (Fig 1B). This functionality percentage varies from 1% in pancreatic islet cells to 8% in PMA-I stimulated T-helper cells. Our findings on functionality across all tracks are consistent with previous findings [12]; 34% of the intergenic human genome is predicted to be functional in at least one annotation track (Fig 1C). Additionally, coding regions of the genome are predicted to have much greater proportions of functionality in multiple tissues than intronic and intergenic regions.

To assess the ability of GenoSkyline-Plus to capture tissue and cell-specific, non-coding functionality in the human genome, we consider a diverse set of known non-coding regulatory elements studied across the genome. To start, we examined microRNAs (miRNA), which are known to regulate a variety of cellular processes through the translational repression and

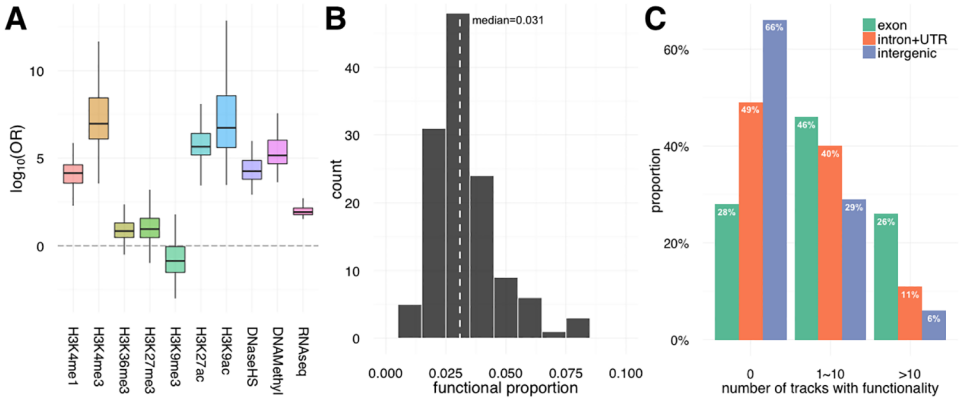

Fig 1. Basic characteristics of GenoSkyline-Plus annotation. (A) Odds ratio of predicting functionality. Each box represents the odds ratio for the same data type across 127 GenoSkyline-Plus tracks. **(B)** Histogram of predicted functional proportion across 127 annotation tracks. Dashed line marks the mean functional proportion. **(C)** Distribution of tracks with predicted functionality. For example, 26% of exon regions are predicted to be functional in more than 10 GenoSkyline-Plus tracks.

https://doi.org/10.1371/journal.pgen.1006933.g001

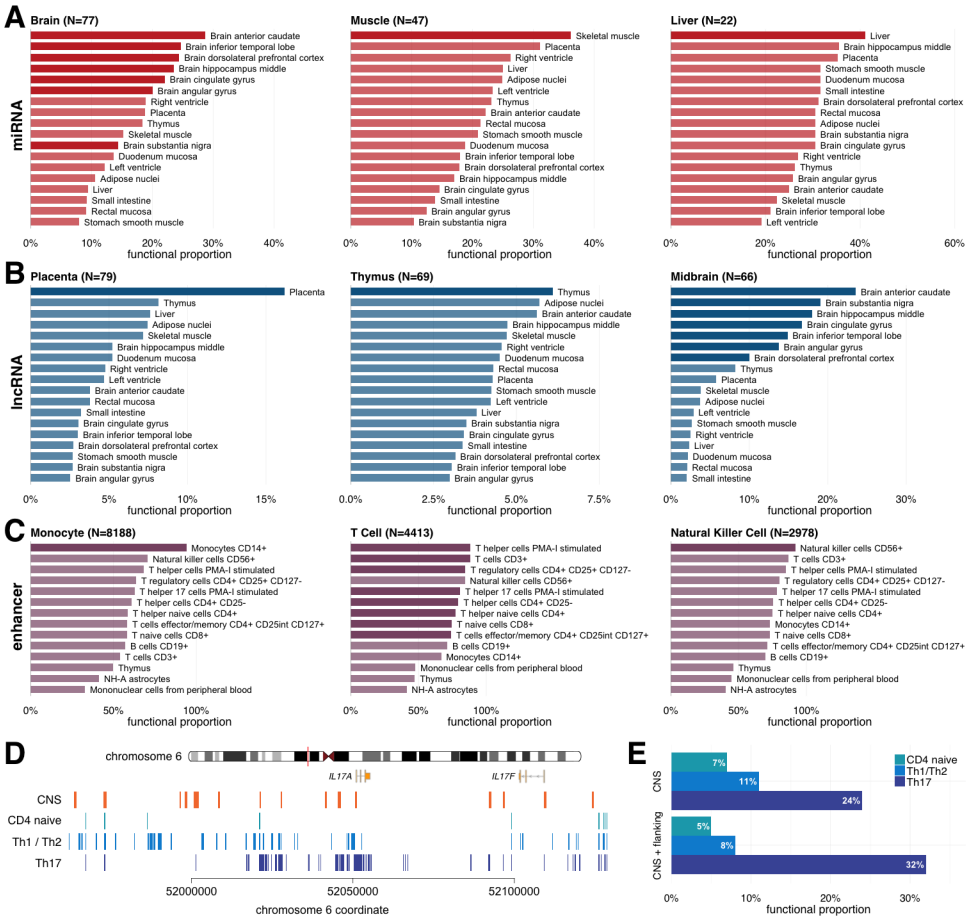

Fig 2. Identify tissue and cell type-specific functionality. Predicted functional proportions of different classes of previously identified tissue-specific non-coding elements. Sample sizes count the number of non-coding elements with specificity for the titled tissue or cell type (see Methods). Darker bars represent annotation tracks that physiologically match the tissue to which the corresponding set of non-coding elements are specific. **(A)** miRNAs with TSI > 0.75 identified in Ludwig et al. **(B)** lncRNAs with TSI > 0.75 identified in Derrien et al. **(C)** Enhancers with differential expression within a cell type facet identified by Andersson et al. **(D)** Predicted functional elements based on GenoSkyline-Plus annotations in the *IL-17A* LCR. Orange boxes mark identified CNS sites. **(E)** Predicted functional proportion in CNS sites and their 200-bp flanking regions across different T-cell subsets.

https://doi.org/10.1371/journal.pgen.1006933.g002

degradation signaling of transcripts [15]. Recent work by Ludwig et al. profiled miRNA expression in 61 different human tissues and identified miRNAs with functionality unique to single tissues through a tissue specific index [16, 17] (TSI; Methods). We applied GenoSkyline-Plus scores to miRNA with tissue-specific functionality by calculating the total proportion of nucleotides predicted to be functional in each tissue. We next looked for which annotation tracks are able to predict the highest proportion of functionality for these known functional regions. The best predictors of high functionality for the three tissues with the largest sample sizes (i.e. brain, liver, and muscle) are tracks for brain structures, the liver track, and the muscle track, respectively (Fig 2A).

We next examined long non-coding RNAs (lncRNA), another non-coding element known for its tissue-specific regulatory action [18]. Using a custom-designed microarray targeting GENCODE lncRNA, Derrien et al. profiled the activity of 9,747 lncRNA transcripts [19]. In order to reidentify and validate the set of lncRNA transcripts that are specific to their

respective tissues, we calculated the previously described TSI and selected lncRNAs with expression specific to only a few cell types. Physiologically matching tracks show a higher proportion of predicted functionality than unmatched tracks in complex, heterogeneous tissue structures like the midbrain. More functionally uniform tissues, such as the thymus or placenta, show the highest functional proportion in matching annotation tracks (Fig 2B).

We also assessed enhancers, non-coding elements that can remotely regulate transcription of an associated promoter elsewhere on the genome with important roles in cell-type specificity [20]. We extracted tissue and cell type-specific enhancer facets identified through the FANTOM5 cap analysis of gene expression (CAGE) atlas and positive differential expression when compared against other defined facets [21]. To determine the utility of the large library of immune cells available in the Epigenomics Roadmap Project for which we developed annotation tracks, we focused on enhancer facets with differential CAGE expression in immune cells. While the method by which enhancers are defined to be differential in a facet is liberal (Methods) and does not imply facet-specific expression, GenoSkyline-Plus still showed outstanding ability to identify matching cell types. Indeed, matched annotation tracks for T-cells, natural killer cells, and monocytes show consistently higher functional proportions than other, non-matched immune cell annotation tracks (Fig 2C).

Finally, we present a case study of the *IL17A-IL17F* locus control region (LCR) in humans, a ~200kb regulatory region surrounding the *IL17A* gene locus. *IL17A* encodes the primary secreted cytokine effector molecule IL-17 of T helper 17 (Th17) cells [22]. The LCR has been studied in mouse models and is found to contain many potential human-conserved intergenic regulatory elements that bind transcription factors that are essential for Th17 cell differentiation and effector function [23, 24]. Experimentally, these conserved noncoding sequences (CNS) acquire functionally permissive H3 acetylation marks at much greater magnitudes under Th17-inducing conditions than naïve or combined Th1 and Th2 populations [25]. Comparing annotation tracks for naïve $CD4^+$ T-cells, differentiated Th17 cells, and differentiated Th1/Th2 cell populations, we identified highly Th17-specific functionality in the conserved regions of the human genome corresponding to known murine CNS regions (Fig 2D and 2E). CNS sites and their flanking regions showed substantially higher functional proportion in Th17 cells than in naïve $CD4^+$ T-cells or Th1/Th2 cell subsets.

Stratify heritability by tissue and cell type for 45 human complex traits

We jointly analyzed three tiers of annotation tracks that respectively represent the overall functional genome, 7 broad tissue clusters, and 66 tissue and cell types (Methods; S2 Table), with summary statistics from 45 GWAS covering a variety of human complex traits (S3 Table). We applied LD score regression [26] to stratify trait heritability by tissue and cell type, and identified a total of 226 significantly enriched annotation tracks for 34 traits after correcting for multiple testing (S4–S7 Tables). In general, GWAS with a large number of significant SNP-level associations showed stronger heritability enrichment in the predicted functional genome (Fig 3A and 3B). Tissue and cell tracks refined the resolution of heritability stratification and provided additional insights into the genetic basis of complex traits (Fig 3C and 3D).

The immune annotation track was significantly enriched for 7 immune diseases, namely celiac disease (CEL), Crohn's disease (CD), ulcerative colitis (UC), primary biliary cirrhosis (PBC), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and multiple sclerosis (MS). Using tracks for cell types, we identified several significant enrichments, including monocytes for CD ($p = 2.9e\text{-}11$) and B cells for PBC ($p = 2.3e\text{-}6$), RA ($p = 1.2e\text{-}5$), and MS ($p = 2.2e\text{-}6$). Inflammatory bowel diseases showed significant enrichment in the gastrointestinal (GI) annotation track (CD: $p = 1.4e\text{-}4$; UC: $p = 5.6e\text{-}5$). Another autoimmune disease with

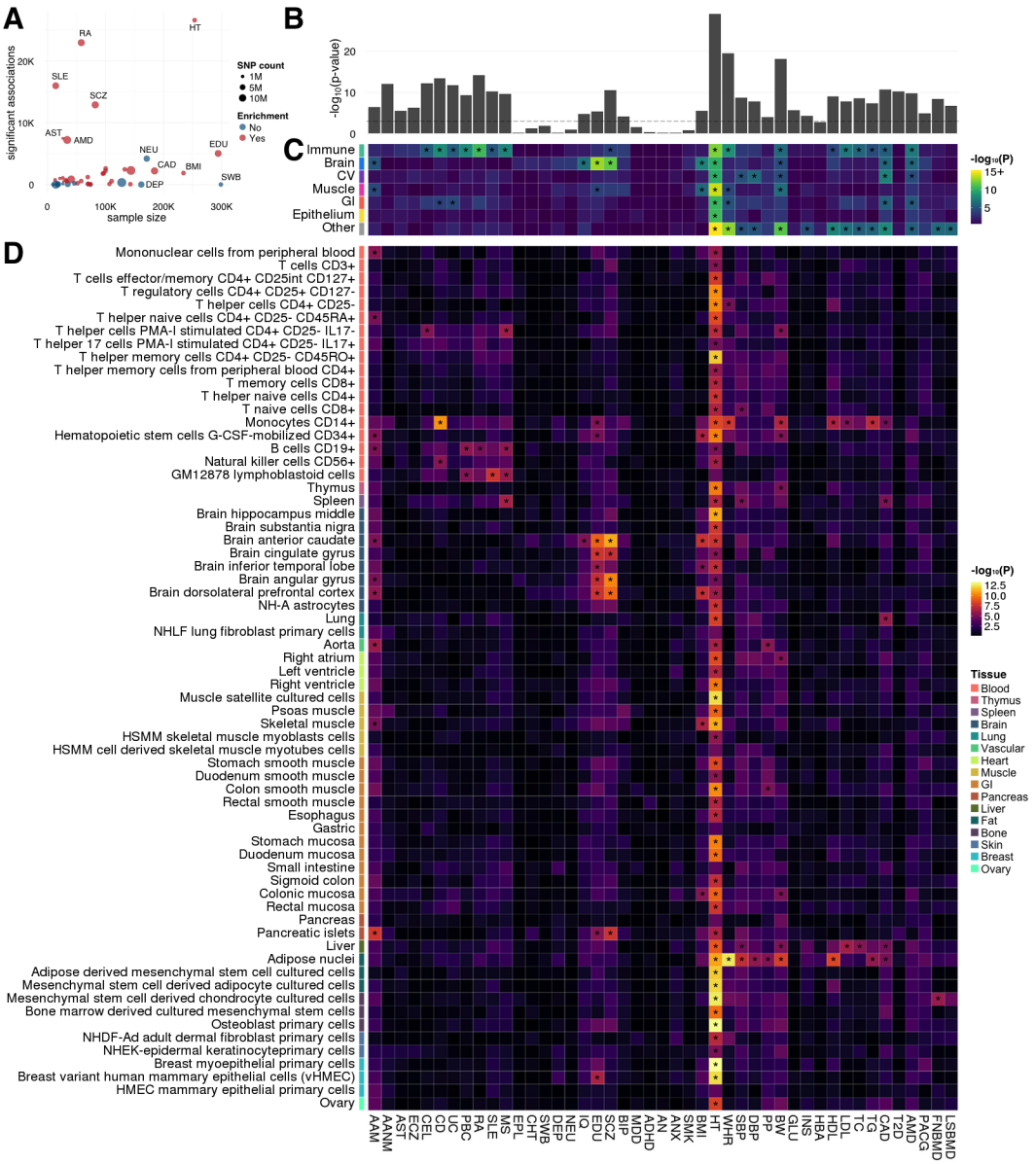

Fig 3. Enrichment analysis for 45 human complex traits. (A) Relationship between GWAS sample size, total count of significant associations, and signal enrichment in the functional genome. Traits significantly enriched in at least one annotation are highlighted in red. **(B)** Enrichment in the general functional genome predicted by GenoCanyon annotation. The dashed line marks the Bonferroni-corrected significance cutoff. **(C)** Enrichment across 7 broadly defined tissue tacks. Asterisks highlight significance after correcting for 45 traits and 7 tissues. **(D)** Enrichment in 66 tissue and cell tracks. Asterisks highlight significant enrichment after correcting for 45 traits and 66 annotations. Details for annotation tracks and different traits are summarized in S2 and S3 Tables.

https://doi.org/10.1371/journal.pgen.1006933.g003

a well-established GI component, CEL, also showed nominal enrichment in the GI annotation track (p = 3.7e-4).

Several brain annotation tracks were significantly enriched for associations of schizophrenia (SCZ), education years (EDU), and cognitive performance (IQ). Bipolar disorder (BIP), neuroticism (NEU), and chronotype (CHT) all showed nominally significant enrichment in the anterior caudate annotation track. Body mass index (BMI) and age at menarche (AAM)

were significantly enriched in multiple brain annotation tracks. Compared to other brain regions, the substantia nigra annotation track showed weaker enrichment for these brain-based traits, which is consistent with its primary function of controlling movement.

Hundreds of height-associated loci have been identified in GWAS [27]. Such a highly polygenic genetic architecture is also reflected in our analysis. 59 of 66 tier-3 tissue and cell annotation tracks were significantly enriched for height associations, with breast myoepithelial cell (p = 6.2e-14) and osteoblast (p = 8.5e-14) being the most significant. Waist-hip ratio (WHR), birth weight (BW), and three blood pressure traits showed significant enrichment in the adipose annotation track. Overall, cardiovascular (CV) annotation tracks showed strong enrichment for blood pressure and coronary artery disease (CAD). Interestingly, the aorta annotation track is significantly enriched for pulse pressure (PP) but not systolic or diastolic blood pressure (SBP and DBP). CAD and 4 lipid traits, i.e. high and low density lipoprotein (HDL and LDL), total cholesterol (TC), and triglycerides (TG), shared a similar enrichment pattern in liver, adipose, and monocyte annotation tracks, which is consistent with the causal relationship among these traits [28].

Our results demonstrated that annotations with refined specificity could provide insights into disease etiology while broader annotations have greater statistical power. Age-related macular degeneration (AMD) was significantly enriched in broadly defined annotation tracks including immune, brain, CV, and GI, despite the non-significant enrichment results using tier-3 annotation tracks. Analyses based on all three tiers of annotations could systematically provide the most interpretable results for most traits. Importantly, we note that greater GWAS sample sizes will effectively increase statistical power in the enrichment analysis while leaving the overall enrichment pattern stable (S1 Fig). Therefore, many more suggestive enrichment results are likely to become significant as GWAS sample sizes grow. Finally, some traits, e.g. type-II diabetes (T2D) and age at natural menopause (AANM), showed strong enrichment in the general functional genome but not in specific tissues, suggesting that we may be able to gain a better understanding of these traits when annotation data for tissues or cell types more relevant to these traits are made available.

Identify enrichment in immune-related DNA elements for neurodegenerative diseases

Next, we performed an integrative analysis of stage-I GWAS summary statistics from the International Genomics of Alzheimer's Project [8] (IGAP; n = 54,162) with GenoSkyline-Plus annotations (Methods). SNPs located in the broadly defined immune annotation track, which account for 24.4% of the variants in the IGAP data, could explain 98.7% of the LOAD heritability estimated using LD score regression (enrichment = 4.0; p = 1.5e-4). Somewhat surprisingly, the signal enrichment in DNA elements functional in immune cells was substantially stronger than the enrichment in brain and other tissue types (Fig 4A). To investigate if immune-related DNA elements are also enriched for associations of other neurodegenerative diseases, we analyzed a publicly accessible GWAS summary dataset for PD [29] (n = 5,691; Methods). Again, the immune annotation track was the most significantly enriched annotation (enrichment = 6.3; p = 7.5e-6), followed by epithelium and CV (Fig 4A).

Analysis based on 66 tissue and cell tracks further refined the resolution of our enrichment study. Monocyte (enrichment = 10.9; p = 2.0e-5) and liver (enrichment = 16.6; p = 4.1e-4) annotation tracks were significantly enriched for LOAD associations (Fig 4B). In fact, the combined functional regions in monocyte and liver covered 8.8% of the SNPs in the IGAP data, but could account for 99.6% of the LOAD heritability currently captured in the IGAP stage-I

Fig 4. Tissue and cell type-specific enrichment for AD and PD. (A) Enrichment in 7 broadly defined tissue tracks. **(B)** Enrichment analysis using 66 GenoSkyline-Plus tissue and cell tracks. Dashed lines indicate Bonferroni-corrected significance cutoff. **(C)** Percentage of variants covered by each annotated category and percentage of heritability explained by variants in that category.

https://doi.org/10.1371/journal.pgen.1006933.g004

GWAS (Fig 4C). In PD GWAS, signal enrichment in liver was absent, but monocyte-functional regions remained strongly enriched (enrichment = 16.3; p = 8.5e-7).

Our findings support the critical role of innate immunity in neurodegenerative diseases [10]. Significant enrichment for LOAD associations in liver-specific DNA elements also provides additional support for the possible involvement of cholesterol metabolism in LOAD

etiology [30, 31]. LOAD signal enrichment in liver remained significant after removing the *APOE* region (chr19: 45,147,340–45,594,595; hg19) from the analysis (S2 Fig), suggesting a polygenic architecture in this pathway. Finally, some adaptive immune cells also showed enrichment for AD and PD associations. LOAD signal enrichment in the B cell annotation track was nominally significant, while multiple T cell annotation tracks were significantly enriched for PD associations. These results not only suggest the involvement of adaptive immunity in neurodegenerative diseases, but also hint at distinct mechanisms of such involvement between AD and PD. Finally, for comparison, we applied several other annotations including CADD [32], GWAVA [33], and EIGEN [34] to the LOAD GWAS data. GenoCanyon and GenoSkyline annotations for seven tissues were also included in the comparison. Our annotations outperformed these methods, showing stronger fold enrichment and more significant p-values (S8 Table).

Identify shared genetic components between AD and PD

Our results showed strong enrichment for both AD and PD in the monocyte functional genome. Next, we investigate if the enrichment for both diseases is through shared or distinct genetic components. Recent studies have failed to identify statistically significant genome-wide pleiotropic effects between AD and PD [35]. We instead hypothesize that the same set of immune-related genetic components are involved in both diseases. Therefore, we aim to identify enrichment for pleiotropic effects in the genome localized to regions of monocyte functionality.

We first partitioned AD and PD heritability by chromosome. Chromosome-wide heritability showed moderate correlation between the two diseases (correlation = 0.65; Fig 5A). When focusing on monocyte functional elements, chromosome-wide heritability showed high concordance between AD and PD (correlation = 0.96; Fig 5B). Interestingly, such high concordance cannot be fully explained by chromosome size. In fact, the correlation between chromosome size and per-chromosome heritability estimates is 0.56 for AD and 0.59 for PD, both lower than the correlation between AD and PD's per-chromosome heritability estimates, especially in the monocyte functional genome. The percentage of explained LOAD heritability on chromosome 19 is lower than previous estimation [36] due to removal of SNPs with large effects in the *APOE* region (Methods). Next, to quantify the shared genetics between AD and PD, we identified significant enrichment for pleiotropic effects in monocyte functional regions (enrichment = 1.8; p = 9.4e-4) using a window-based approach (Methods). To account for potential bias due to the moderate sample overlap between the two GWAS as well as other confounding factors, we applied a permutation-based testing approach (Methods). Enrichment for pleiotropic effects in the monocyte functional genome remained significant (p = 4.6e-3). In addition, these results were robust with the choice of window size.

We identified 15 candidate loci for pleiotropic effects (Methods; S9 Table), among which signals at *SLC9A9* and *AIM1* are the clearest (Fig 5C and 5D). *SLC9A9*, whose encoded protein localizes to the late recycling endosomes and plays an important role in maintaining cation homeostasis (RefSeq, Mar 2012), is associated with multiple pharmacogenomic traits related to neurological diseases, including response to cholinesterase inhibitor in AD [37], response to interferon beta in MS [38], response to angiotensin II receptor blockade therapy [39], and multiple complex diseases including attention-deficit/hyperactivity disorder [40], autism [41], and non-alcoholic fatty liver [42]. Gene *AIM1* is associated with stroke [43], human longevity [44], and immune diseases including RA [45] and SLE [46].

A few candidate loci pointed to clear gene candidates but showed unclear or distinct peaks of association (S3 Fig). These include an inflammatory bowel disease risk gene *ANKRD33B*

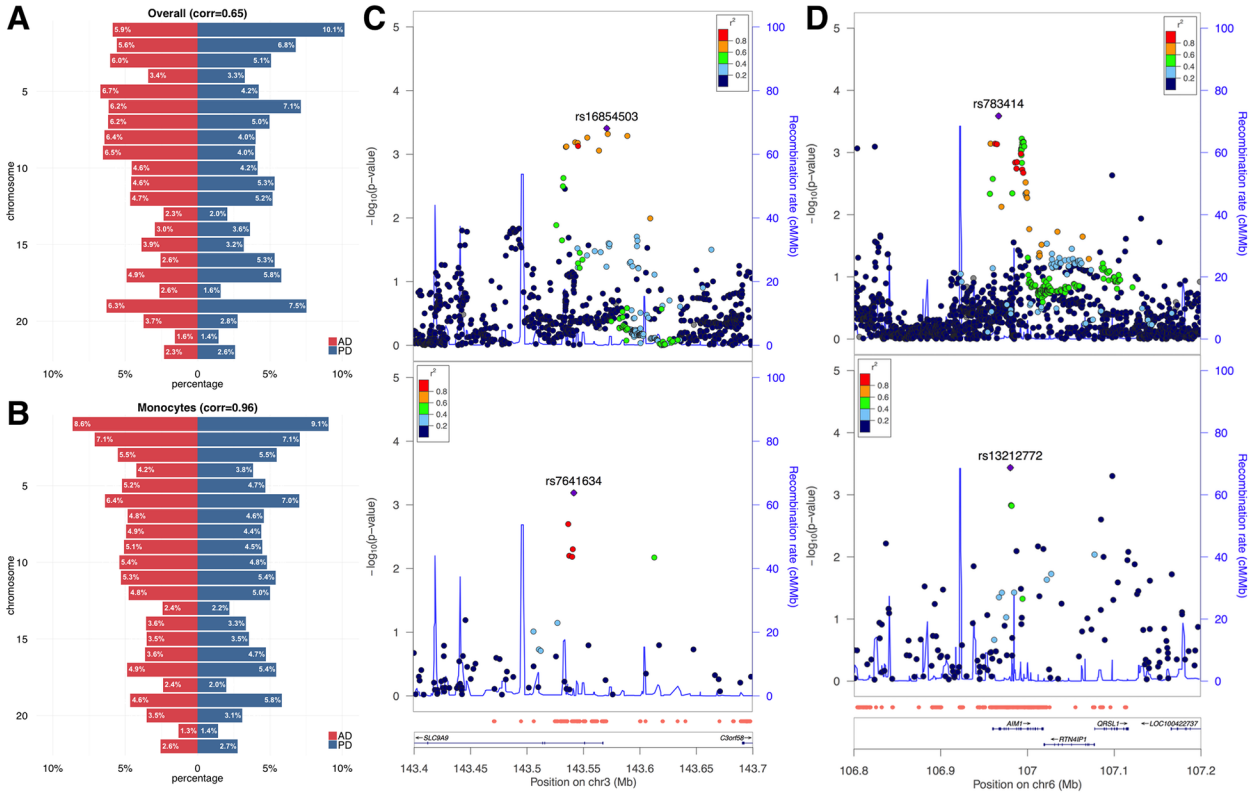

Fig 5. Identify genetic correlation between LOAD and PD. (A) estimated chromosome-by-chromosome heritability percentage for LOAD and PD. **(B)** chromosome-by-chromosome heritability in the monocyte functional genome. **(C-D)** Association peaks in pleiotropic loci *SLC9A9* and *AIM1*. The upper and the lower panels represent associations for LOAD and PD, respectively. Monocyte-specific functional regions are highlighted by red dots at the bottom of the figure above the gene annotations.

https://doi.org/10.1371/journal.pgen.1006933.g005

[47]. *PRUNE2* is a gene associated with response to amphetamine [48] and hippocampal atrophy which is a quantitative trait for AD [49]. *HBEGF* is associated with AD in *APOE* ε4- population [50] and involved in Aβ clearance [51]. *PROK2* is a gene involved in Aβ-induced neurotoxicity [52]. Additionally, the protein product of *AXIN1* negatively affects phosphorylation of tau protein [53]. Other gene candidates include *CCDC158*, *PRSS16*, and *ZNF615*, which are previously identified risk genes for PD, SCZ, and BIP, respectively [54–56]. Some other windows showed complex structures of linkage disequilibrium (LD) and contained large association peaks spanning a number of genes (S4 Fig), which include the region near PD risk gene *PRSS8* [54] and the *HLA* region. Interestingly, we also identified the surrounding region of *MAPT*, a gene that encodes the tau protein which is a critical component of both AD and PD pathologies [50, 54, 57, 58].

Pathway enrichment analysis for genes in 15 pleiotropic candidate loci identified significant enrichment in immune-related pathways staphylococcus aureus infection (KEGG:05150; p = 1.9e-5) and systemic lupus erythematosus (KEGG:05322; p = 3.7e-04; Methods). Both pathways remained significant after removing two *HLA* loci from our analysis.

Reprioritize AD risk loci through integrative analysis of functional annotation

Finally, we reprioritize AD risk loci using monocyte and liver annotation tracks. We integrated IGAP stage-I summary statistics with GenoSkyline-Plus using genome-wide association

prioritizer (GenoWAP [59]), and ranked all SNPs based on their GenoWAP posterior scores (Methods). Under a posterior cutoff of 0.95, we identified 8 loci that were not reported in the IGAP GWAS meta-analysis using monocyte annotation and 4 loci using the liver annotation track (S10 Table).

We then sought replication for SNPs with the highest posterior score at each of these loci using inferred IGAP stage-II z-scores (Methods). After removing shared SNPs between monocyte- and liver-based analyses, 10 SNPs remained in the analysis, 7 of which showed consistent effect directions between the discovery and the replication cohorts (Fig 6A). One SNP was successfully replicated in the inferred IGAP stage-II dataset, i.e. rs4456560 (p = 0.013). SNP rs4456560 is located in *SCIMP* (Fig 6B), a gene that encodes a lipid tetraspanin-associated transmembrane adaptor protein that is expressed in antigen-presenting cells and localized in the immunological synapse [60].

A moderate replication rate in the IGAP stage-II cohort was expected since we focused on loci that did not reach genome-wide significance in the IGAP meta-analysis and the IGAP stage-II cohort is relatively small (n = 19,884) compared to the data in the discovery stage. Furthermore, data from IGAP stage-II cohort are not publicly available and we were limited to the inverse inference approach shown here. It is possible additional loci will replicate when IGAP stage-II summary or individual-level data are made available. However, all identified loci have been linked to AD or relevant phenotypes in the literature. *RPN1* was linked to AD through a network-based technique [61]. Association between *ECHDC3* and AD risk was established through a joint analysis of AD and lipid traits [62]. Association between *DLST* and AD has also been previously reported [63]. *BZRAP1* and *MINK1* were shown to be associated with cognitive function and blood metabolites, respectively [64, 65]. A pleiotropic effect candidate gene *HBEGF* showed up again in the SNP reprioritization analysis. Multiple genes in the sorting nexin family have been found to participate in *APP* metabolism and Aβ generation [66]. Association between *SNX1* and AD has also been previously identified using gene-based tests [67]. Finally, during the peer review process of this paper, three new genome-wide significant loci (i.e. *PFDN1/HBEGF*, *USP6NL/ECHDC3*, and *BZRAP1-AS1*) were reported in a trans-ethnic

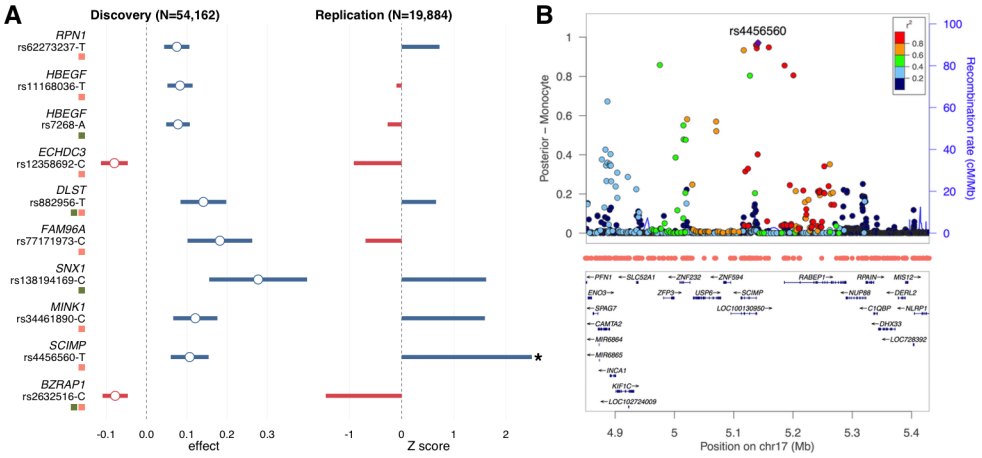

Fig 6. Reprioritize AD GWAS loci using functional annotations. (A) Effect size estimates for 10 SNPs of interest in the discovery and the replication cohort. Intervals in the discovery stage indicate 95% confidence. Asterisk indicates significant effects in the replication cohort. Red and green squares highlight loci identified using monocyte or liver annotation track, respectively. **(B)** The successfully replicated *SCIMP* locus. The vertical axis shows the GenoWAP posterior probability based on monocyte annotation track. Functional regions in monocyte are highlighted by red dots.

https://doi.org/10.1371/journal.pgen.1006933.g006

GWAS meta-analysis for AD [68], all of which were among our reprioritized list of risk loci. Further, the most significant SNPs at loci *PFDN1/HBEGF* (rs11168036, p = 7.1e-9) and *BZRAP1-AS1* (rs2632516, p = 4.4e-8) matched with our top reprioritized SNPs (Fig 6A).

Discussion

Increasing evidence suggests that non-coding regulatory DNA elements may be the primary regions harboring risk variants in human complex diseases. In this work, we have substantially expanded our previously established GenoSkyline annotation by incorporating RNA-seq and DNA methylation into its framework, imputing incomplete epigenomic and transcriptomic annotation tracks, and extending it to more than 100 human tissue and cell types. With the help of integrative functional annotations, we identified strong enrichment for LOAD heritability in functional DNA elements related to innate immunity and liver tissue using hypothesis-free tissue-specific enrichment analysis. This enrichment was also found in immune-related DNA elements using PD data. Our analysis also clearly indicated that monocyte functional elements in particular appear to be highly relevant in explaining AD and PD heritability. Of note, we analyzed 45 complex diseases and traits in addition to AD and PD. The substantial enrichment for multiple psychiatric and neurological traits in the brain functional genome shows that the lack of brain enrichment in neurodegeneration is not due to poor quality of brain annotations. Further, the monocytes annotation track was the most significantly enriched for Crohn's disease among the 45 GWAS, and was not ubiquitously enriched for a large number of traits. Consistent and biologically interpretable enrichment results on a large collection of complex traits demonstrate the effectiveness of our approach and increase the validity of novel findings.

It is worth noting that multiple studies have highlighted the role of myeloid cells in the genetic susceptibility of neurodegenerative diseases [11]. Several genes expressed in myeloid cells (e.g. ABCA7, CD33, and TREM2) have been identified in GWAS and sequencing-based association studies for AD [8, 69, 70]. Further, AD risk alleles identified in GWASs have been shown to enrich for cis-eQTLs in monocytes [9]. In addition, two recent papers identified enrichment for AD heritability in active genome regions in myeloid cells [71, 72], which suggested a polygenic genetic architecture for immune-related DNA elements in AD etiology and hinted at a large number of unidentified, immune-related genes for AD. Compared to the aforementioned work, our study utilizes a better set of tissue-specific genome annotations and explicitly accounts for the similarity between different cell types through a multiple regression model. One major limitation in our analysis is lack of data for other potentially AD-relevant cell types such as microglia. Whether our findings correctly reflected the direct involvement of peripheral immune cells in neurodegenerative diseases rather than the detection of epigenomic similarities between monocytes and microglia remains to be carefully investigated in the future.

Furthermore, we successfully identified enrichment for shared genetic components between AD and PD in the monocyte functional genome, which hints at a shared neuroinflammation pathway between these two neurodegenerative diseases. We note that several candidate loci with potential pleiotropic effects showed fairly marginal associations with AD and PD, which explains why they have been missed in traditional SNP-based association analysis. Importantly, SNPs in immune-related DNA elements explain a large proportion of AD and PD heritability in total. These results suggest that weak but pervasive associations related with immunity still remain unidentified. Further evaluations of these relationships using GWAS with larger sample sizes may provide insights into the shared biology of these neurodegenerative conditions.

Through multi-tier enrichment analyses on 45 GWAS, an in-depth case study of neurodegenerative diseases, and validation of known non-coding tissue-specific regulatory machinery, we have demonstrated the ability of GenoSkyline-Plus to provide unbiased, genome-wide insights into the genetic basis of human complex diseases. The analyzed GWAS represent a variety of human complex diseases and traits, highlighting the effectiveness of our method in different contexts and genetic architecture. However, while our non-coding validation study demonstrated that GenoSkyline-Plus annotations indeed captured tissue-specific activity in a variety of intergenic machinery, there is a need to develop a more statistically robust framework to identify new non-coding elements rather than validate existing ones. Our approach of identifying the functionally active proportion of all elements in aggregate is only able to identify tissue specificity while considering large groups of highly specific non-coding elements. The availability of over 100 different annotation tracks introduces many multiple-testing issues that should be addressed in the case of a statistically sound analysis for tissue-specificity. We have also demonstrated how GenoSkyline-Plus and its explanatory power improve with the addition of more data. Currently, functionality in 28% of exonic regions still remains to be identified. As the quantity and quality of high-throughput epigenomic data continue to grow, GenoSkyline-Plus has the potential to further evolve and provide even more comprehensive annotations of tissue-specific functionality in the human genome. We will update our annotations when data for new tissue and cell types from the Roadmap consortium become available. Finally, several recent papers have introduced novel models to integrate functional annotations in tissue-specific enrichment analysis [73, 74]. Many models that do not explicitly incorporate functional annotation information have also emerged in transcriptome-wide association studies and other closely-related applications in human genetics research [75–78]. Our annotations, in conjunction with rapidly advancing statistical techniques and steadily increasing sample sizes in genetics studies, may potentially benefit a variety of human genetics applications and promise a bright future for complex disease genetics research.

Methods

Annotation data preprocessing

Chromatin data were extracted from the Epigenomics Roadmap Project's consolidated reference epigenomes database (http://egg2.wustl.edu/roadmap/). Specifically, ChIP-seq peak calls were collected for each epigenetic mark (H3k4me1, H3k4me3, H3k36me3, H3k27me3, H3k9me3, H3k27ac, H3k9ac, and DNase I Hypersensitivity) in each Roadmap consolidated epigenome where available. Peak calls imputed using ChromImpute [79] were used in place of missing data. Next, peak files were reduced to a per-nucleotide binary encoding of presence or absence of contiguous regions of strong ChIP-seq signal enrichment compared to input (Poisson p-value threshold of 0.01).

DNA methylation data were also collected from the Roadmap's reference epigenomes database. CpG islands were identified in each sample using the CpG Islands Track of the UCSC Genome Browser (http://genome.ucsc.edu/), and unmethylated islands were those CpG islands with less than 0.5 fractionated methylation based on imputed methylation signal tracks in the Roadmap reference epigenomes database. Presence of an unmethylated CpG island was then encoded for each nucleotide as a binary variable. Finally, Roadmap's RNA-seq data were dichotomized using an rpkm cutoff of 0.5 at 25-bp resolution and included in our annotations.

GenoSkyline-Plus model

We adapt the existing framework established by Lu et al. to a broader set of genomic data [12]. Briefly, given a set of Annotations $\boldsymbol{A}$ and a binary indicator of genomic functionality $\boldsymbol{Z}$, the

joint distribution of $\boldsymbol{A}$ along the genome is assumed to be a mixture of annotations at functional nucleotides and non-functional nucleotides. Assuming that each of the annotations in $\boldsymbol{A}$ is conditionally independent given $\boldsymbol{Z}$, we factorize the conditional joint density of $\boldsymbol{A}$ given $\boldsymbol{Z}$ as:

$$f(\boldsymbol{A}|Z=c)=\prod_{i=1}^{10} f_i(A_i|Z=c),\ c=0,\ 1 \tag{1}$$

All annotations have been preprocessed into binary classifiers, and the marginal functional likelihood given each individual annotation can be modeled with a Bernoulli distribution

$$f_i(A_i|Z=c)=p_{ic}^{A_i}(1-p_{ic})^{1-A_i},\ \ i=1,\ldots,10;\ c=0,\ 1 \tag{2}$$

With an assumed prior probability π of functionality, the parameter p_{ic} of each individual annotation can be estimated with the Expectation-Maximization (EM) algorithm. The posterior probability of functionality at a nucleotide, known as the GenoSkyline-Plus score, is then:

$$P(Z=1|\boldsymbol{A})=\frac{\pi\prod_{i=1}^{10} f_i(A_i|Z=1)}{\pi\prod_{i=1}^{10} f_i(A_i|Z=1)+(1-\pi)\prod_{i=1}^{10} f_i(A_i|Z=0)} \tag{3}$$

Giving us with 21 parameters for each annotation track:

$$\Theta=(\pi,p_{1,0},p_{2,0},\ \ldots\ ,p_{10,0},p_{1,1},p_{2,1},\ \ldots\ ,p_{10,1}) \tag{4}$$

These parameters were estimated using the GWAS Catalog, downloaded from the NHGRI website (http://www.genome.gov/gwastudies/). 13,070 unique SNPs found to be significant in at least one published GWAS were expanded into 1kb bp intervals and formed a sampling covering 12,801,840 bp of the genome. This sampling method has been shown to be a robust representation of functional and non-functional regions along the genome [6]. Notably, other models have been recently developed to predict functional non-coding SNPs [34].

Data for validating annotation quality

Quantile-normalized expression values were downloaded for all mature miRNAs profiled in Ludwig et al [17]. Due to inconsistent levels of miRNA specificity in the two donors in this study and to avoid diluting miRNA specificity, we used miRNA data from only body 1, which had a higher fraction of tissue specific miRNAs. TSI values were calculated as described in the study:

$$TSI_j=\ \frac{\sum_{i=1}^{N}(1-x_{j,i})}{N-1} \tag{5}$$

Where N is the total number of tissues measured, $x_{j,i}$ is the expression intensity of tissue i divided by the maximum expression across all tissues for miRNA j. We extract any miRNAs with a TSI score greater than the median value of 0.75 to produce a sufficiently large collection of miRNAs with expression highly specific to only a few tissues that we can then attempt re-identify using GenoSkyline-Plus. We next download genomic positions and identify the highest expressed tissue for each TSI-filtered miRNA. miRNA coordinates were extracted from miRbase (http://mirbase.org/) and mapped to hg19 using the UCSC liftover tool (http://genome.ucsc.edu/). lncRNA data was prepared similarly to miRNA. Expression data of 9,747 lncRNA transcripts based on GENCODE v3c annotation across 31 human tissues [19] (GEO accession: GSE34894) was downloaded. As above, the TSI of each lncRNA transcript was

calculated, and transcripts with a TSI greater than 0.75 were labeled for genomic position and maximally expressed tissue.

Pre-defined enhancer differentially expressed cell facets [21] were downloaded from PrESSto database (http://enhancer.binf.ku.dk/presets/). Andersson et al. define their enhancer sets via bi-directional CAGE expression collected by the FANTOM consortium [80]. Cell facets were manually constructed using hierarchical FANTOM5 cell ontology term mappings to create mutually exclusive and broadly covered histological and functional annotations. Enhancers were considered differentially expressed in a facet using Kruskal-Wallis rank sum test and subsequent pair-wise post-hoc tests to identify enhancers with significantly differential expression between pairs of facets. Based on this method, an enhancer is considered differentially expressed in a facet if it is significantly differentially expressed compared to any other facet and has overall positive standard linear statistics.

For each of the three data validation sets, functional specificity is assessed by calculating the per-nucleotide functional proportion of all non-coding elements across a tissue. Functionality is defined by a Genoskyline-Plus score greater than 0.5 at that nucleotide. For Roadmap samples with multiple donors (e.g. skeletal muscle and rectal mucosa) we took the average GenoSkyline-Plus score at each nucleotide across the samples. For each set of non-coding elements we selected the top three tissues with the largest sample size that had matching annotations in Genoskyline-Plus. For example, we did not calculate scores for enhancers with maximal expressions in human testis because there is no corresponding Roadmap sample in which we would detect tissue-specific functionality.

To examine cell-specific functionality of the *IL17A* LCR in T-cell subsets, we extracted GenoSkyline-Plus scores for each nucleotide along the ~200 kilobase region between the genes *PKHD1* and *MCM3* [23]. While scores for Th17 and Th1/Th2 subsets (i.e. 'CD4+ CD25- IL17 + PMA-Ionomycin stimulated Th17 Primary Cells' and 'CD4+ CD25- IL17- PMA-Ionomycin stimulated MACS purified Th Primary Cells'; S1 Table) were extracted as-is, we took the average score of the two available CD4+ naïve T-cell subsets (i.e. 'CD4 Naïve Primary Cells' and 'CD4+ CD25- CD45RA+ Naïve Primary Cells'). We identified the analogous human regions of previously identified functional murine CNS regions [25] by taking the top 20 most conserved intergenic sites between mouse and human in the LCR region using the VISTA browser (http://pipeline.lbl.gov/cgi-bin/gateway2). GenoSkyline-Plus scores in the 20 CNS sites and their 200-bp flanking regions were compared across different cell types.

GWAS data details

Summary statistics for 45 GWAS are publicly accessible. Details for these studies are summarized in S3 Table. IGAP is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage-I, IGAP used genotyped and imputed data on 7,055,881 SNPs to meta-analyze four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative–EADI, the Alzheimer Disease Genetics Consortium–ADGC, The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium–CHARGE, and The Genetic and Environmental Risk in AD consortium–GERAD). In stage-II, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 AD cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages I and II. IGAP stage-I GWAS summary data is publicly accessible from IGAP consortium website (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php). GWAS summary statistics for PD was acquired from dbGap (accession: pha002868.1). Details for AD and PD studies have been previously reported [8, 29].

Stratify heritability by tissue and cell type

Heritability stratification and enrichment analyses were performed using LD score regression implemented in the LDSC software (https://github.com/bulik/ldsc/). Annotation-stratified LD scores were estimated using dichotomized annotations, 1000 Genomes (1KG) samples with European ancestry [81], and a default 1-centiMorgan window. Enrichment was defined as the ratio between the percentage of heritability explained by variants in each annotated category and the percentage of variants covered by that category.

$$Enrichment = \frac{\%\ Heritability\ explained}{\%\ Genome\ covered} \quad (6)$$

A resampling-based approach was used to assess standard error estimates [26]. Three tiers of annotations of different resolutions were used in enrichment analyses:

1. Generally functional genome predicted by GenoCanyon annotation smoothed along 10-kb windows.
2. Seven unique tissue and cell type clusters (i.e. immune, brain, CV, muscle, GI, epithelium, and other), representing common, physiologically related organ systems. Each category is defined as the union of functional regions in related tissue and cell types (S2 Table).
3. GenoSkyline-Plus annotations for 66 selected tissue and cell types (S2 Table).

The smoothing strategy for GenoCanyon improves its ability to identify general functionality in the human genome [59]. GenoSkyline-Plus and smoothed GenoCanyon annotations were dichotomized using a cutoff of 0.5. Such dichotomization is robust to the cutoff choice due to the bimodal nature of annotation scores [6]. We selected 66 annotation tracks in the tier-3 analysis by removing all the fetal and embryonic cells, and taking the union of different Roadmap epigenomes for the same cell type (S2 Table). The 53 baseline annotations of LD score regression were always included in the model across all analyses as suggested in the LDSC user manual. Smoothed GenoCanyon annotation track was also included in tier-2 and tier-3 analyses to account for unobserved tissue and cell types. Of note, the proposed multiple regression model explicitly takes the overlapped functional regions across biologically-related cell types into account. Further, the linear mixed-effects model in LDSC does not assume linkage equilibrium, and therefore LD will most likely not introduce bias into heritability estimation and enrichment calculation. We removed the MHC region from our analysis due to its unique LD patterns.

A slightly different strategy was adopted when comparing the performance of different computation annotation tools. To make fair comparison, we dichotomized all annotation tracks using each score's top 90% quantile calculated from SNPs with minor allele count greater than five in 1000 Genomes samples with European ancestry. We then followed the suggested protocol of LDSC and kept baseline annotations in the model while adding each annotation track one at a time.

Pleiotropy analysis

We calculated chromosome-by-chromosome heritability percentage through summing up and normalizing per-SNP heritability estimated using LD score regression and tier-3 annotation tracks. Of note, since only GWAS summary statistics were used as the input, popular heritability estimation tools such as GCTA [82] could not be applied. The sums over complete chromosomes are compared with the sums over monocyte functional regions only. Notably, LDSC is conceptually different from some other tools (e.g. GCTA [82]) in its estimation of

trait heritability. GCTA estimates the proportion of phenotypic variability that can be explained by SNPs in the GWAS dataset while LDSC aims to estimate the proportion of phenotypic variability explained by all the SNPs in samples from the 1KG Project. In practice, LDSC only uses HAPMAP SNPs to fit the LD score regression model and assumes that HAPMAP SNPs are sufficient for tagging all 1KG SNPs through LD [26]. Additionally, LDSC applies a few stringent SNP filtering steps for quality control reasons, e.g. removing SNPs with very large effect sizes (i.e. $\chi^2 > 80$), which leads to the removal of some SNPs in the *APOE* region in our analysis. Finally, we note that a recent method may potentially improve the heritability estimates based on LDSC [75].

To evaluate enrichment of pleiotropic sites in the monocyte functional genome, we partition the genome into windows with length of 1M bases. Sex chromosomes and windows without SNPs are removed in our datasets. For each disease (i.e. AD and PD), we label a window 1 if the following criteria are met.

1. There is at least one SNP with p-value $<$ 1e-3 in the window.
2. Among SNPs that meet condition 1, at least one is located in the monocyte-specific functional genome.

Otherwise, the window is labeled 0. This labeling results in two binary vectors, one for each disease. A window marked as 1 for both AD and PD is a window of interest that suggests a possible association in monocytes-related DNA for both diseases in that region. We use a hypergeometric test to assess if such a pattern of local association appears more often than by chance. Windows marked as 1 for both diseases are subsequently curated to identify the association peaks that potentially have pleiotropic effects for AD and PD.

There is a moderate overlap of control samples between IGAP AD GWAS and the PD GWAS (KORA controls, N~480). To account for the bias introduced by sample overlap and other confounding factors, we designed a permutation-based approach. In each permutation step, we shuffle the annotation status while keeping the total proportion of annotated regions, and then pick out windows that meet condition 2. We calculate the p-value through comparing the observed number of windows that meet conditions 1 and 2 for both diseases with the empirical distribution acquired in permutations.

Of note, we also applied this approach using a window size of 500K bases. Results in all related tests remained similar.

GWAS loci reprioritization

We briefly describe the SNP reprioritization approach implemented in the GenoWAP software available on our server (http://genocanyon.med.yale.edu/GenoSkyline). First, we identify three disjoint cases for SNPs in a given GWAS dataset.

1. The SNP is in a genomic region that is functional for the given phenotype and tissue ($Z_D = 1, Z_T = 1$).
2. The SNP is in a genomic region that is functional in the given tissue, but that region has no functionality for the phenotype ($Z_D = 0, Z_T = 1$).
3. The SNP is in a genomic region that is not functional in the given tissue ($Z_T = 0$).

A useful metric for prioritizing SNPs is the conditional probability that the SNP is classified under case-1 given its p-value in the GWAS study, i.e. $P(Z_D = 1, Z_T = 1 \mid p)$. We can denote this

probability using Bayes formula as follows:

$$P(Z_D = 1, Z_T = 1 \mid p) = P(Case\,1 \mid p) = \frac{f(p|Case\,1) \times P(Case\,1)}{\sum_{k=1}^{3} f(p|Case\,k) \times P(Case\,k)} \quad (7)$$

First, $P(Case\,3) = 1 - P(Z_T = 1)$ can be directly identified using GenoSkyline-Plus scores. We partition all the SNPs into two subgroups based on a mean GenoSkyline-Plus score threshold of 0.1. Notably, these probabilities are not sensitive to changing threshold [6]. In this way, we can directly estimate $f(p|Case\,3) = f(p|Z_T = 0)$ by applying a histogram approach on the SNP subgroup with low GenoSkyline-Plus scores.

Next, we assume that SNPs that are functional in a tissue but not relevant to the phenotype will have the same p-value distribution to all other SNPs that are not relevant to the phenotype, which in turn behave similarly to SNPs that are not functional at all. We have previously demonstrated that this assumption is backed by empirical evidence [6]. More formally, this relationship is denoted as follows:

$$f(p|Case\,2) = f(p|Z_D = 0,\ Z_T = 1) = f(p|Z_D = 0) = f(p|Z = 0) \quad (8)$$

We estimate the distribution $f(p|Z = 0)$ by using a similar approach to estimating $f(p|Z_T = 0)$, but partitioning SNPs using the general functionality GenoCanyon score instead of tissue-specific GenoSkyline-Plus score.

Finally, all remaining terms in Formula 6 can be estimated using the EM algorithm. The p-value distribution of the subset of SNPs located in tissue-specific functional regions (i.e. $Z_T = 1$) is the following mixture:

$$f(p|Z_T = 1) = P(Z_D = 1|Z_T = 1) \times f(p|Case\,1) + P(Z_D = 0|Z_T = 1) \times f(p|Case\,2) \quad (9)$$

Density function $f(p|Case\,2)$ has been estimated in Formula (8) and $f(p|Case\,2)$ is assumed to follow a beta distribution, which guarantees a closed-form expression in the EM algorithm.

$$(p|Z_D = 1, Z_T = 1) \sim Beta(\alpha, 1),\ \ 0 < \alpha < 1 \quad (10)$$

Notably, the *APOE* region was removed in the SNP reprioritization analysis for LOAD.

Inverse inference of IGAP stage-II z-scores

Summary statistics from both IGAP stage-I GWAS and stage-I+II meta-analysis are publicly available (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php). We inferred z-scores from IGAP stage-II replication cohort using the following formula.

$$\frac{Z_{1+2} \times \sqrt{N_1 + N_2} - Z_1 \times \sqrt{N_1}}{\sqrt{N_2}} \quad (11)$$

In this formula, Z_1 and Z_{1+2} indicate z-scores from the stage-I GWAS and the combined meta-analysis, respectively. N_i indicates the sample size from the i^{th} stage. This formula was derived from the sample size based meta-analysis model, an approach known to be asymptotically equivalent to inverse variance based meta-analysis [83].

Data accessibility and other bioinformatics tools

GenoSkyline-Plus annotation tracks, tiers 1–3 LD score files, and scripts for generating GenoSkyline-Plus scores are freely available on the GenoSkyline server (http://genocanyon.med.yale.edu/GenoSkyline). All annotation tracks can be visualized using UCSC genome browser. Web server g:Profiler was used to perform pathway enrichment analysis [84]. The g:SCS threshold

implemented in g:Profiler was applied to account for multiple testing. Locus plots were generated using LocusZoom [85]. Gene plots were generated using R package "Gviz".

Supporting information

S1 Fig. Enrichment analysis for three schizophrenia studies. Overall enrichment pattern for schizophrenia remains stable as sample size increases. Sample sizes for the three studies shown below are 21,856, 32,143, and 82,315, respectively.
(PNG)

S2 Fig. Enrichment analysis for LOAD after removing the *APOE* region.
(PNG)

S3 Fig. Nine pleiotropic loci for LOAD and PD. For each locus, the upper and lower panels show associations for LOAD and PD, respectively. Monocyte functional regions are marked by red dots above gene names.
(PNG)

S4 Fig. Four pleiotropic loci that span a large number of candidate genes. For each locus, the upper and lower panels show associations for LOAD and PD, respectively. Monocyte functional regions are marked by red dots above gene names.
(PNG)

S1 Table. List of 127 GenoSkyline-Plus annotation tracks. Epigenome ID and cell type list are acquired from Epigenomics Roadmap Project and ENCODE.
(XLSX)

S2 Table. Curated annotation tracks used in tissue-specific enrichment analysis.
(XLSX)

S3 Table. Details of 45 genome-wide association studies.
(PDF)

S4 Table. P-values for tissue-specific enrichment analysis (tiers 1 and 2).
(XLSX)

S5 Table. Fold change for tissue-specific enrichment analysis (tiers 1 and 2).
(XLSX)

S6 Table. P-values for tissue-specific enrichment analysis (tier 3).
(XLSX)

S7 Table. Fold change for tissue-specific enrichment analysis (tier 3).
(XLSX)

S8 Table. AD enrichment in other computational annotations.
(XLSX)

S9 Table. Candidate loci for pleiotropic effect between AD and PD.
(PDF)

S10 Table. Top loci based on GenoWAP posterior scores. Subscripts indicate stage-I discovery cohort, stage-II replication cohort, or stage-I+II meta-analysis. Each locus may contain multiple genes. SNP rs143560707 was not genotyped in the IGAP stage-II cohort.
(XLSX)

Acknowledgments

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. We are also grateful for all the consortia and investigators that provided publicly accessible GWAS summary statistics.

Author Contributions

Conceptualization: Qiongshi Lu, Hongyu Zhao.

Data curation: Ryan L. Powles, Qian Wang, Shubhabrata Mukherjee, Paul K. Crane.

Formal analysis: Qiongshi Lu, Ryan L. Powles, Sarah Abdallah, Derek Ou, Yiming Hu, Yisi Lu, Wei Liu, Boyang Li.

Funding acquisition: Hongyu Zhao.

Methodology: Qiongshi Lu.

Project administration: Hongyu Zhao.

Resources: Shubhabrata Mukherjee, Paul K. Crane, Hongyu Zhao.

Software: Ryan L. Powles.

Supervision: Hongyu Zhao.

Visualization: Qiongshi Lu.

Writing – original draft: Qiongshi Lu, Ryan L. Powles.

Writing – review & editing: Sarah Abdallah, Derek Ou, Qian Wang, Yiming Hu, Yisi Lu, Wei Liu, Boyang Li, Shubhabrata Mukherjee, Paul K. Crane, Hongyu Zhao.

References

1. Bernstein B.E., et al. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012. 489(7414): p. 57–74. https://doi.org/10.1038/nature11247 PMID: 22955616
2. Kundaje A., et al. Integrative analysis of 111 reference human epigenomes. Nature, 2015. 518(7539): p. 317–330. https://doi.org/10.1038/nature14248 PMID: 25693563
3. Schaub M.A., et al. Linking disease associations with regulatory information in the human genome. Genome research, 2012. 22(9): p. 1748–1759. https://doi.org/10.1101/gr.136127.111 PMID: 22955986
4. Maurano M.T., et al. Systematic localization of common disease-associated variation in regulatory DNA. Science, 2012. 337(6099): p. 1190–1195. https://doi.org/10.1126/science.1222794 PMID: 22955828
5. Kellis M., et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A, 2014.
6. Lu Q., et al. Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. PLoS Genet, 2016. 12(4): p. e1005947. https://doi.org/10.1371/journal.pgen.1005947 PMID: 27058395
7. Myers A.J. and Goate A.M., The genetics of late-onset Alzheimer's disease. Current opinion in neurology, 2001. 14(4): p. 433–440. PMID: 11470958
8. Lambert J.-C., et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature genetics, 2013. 45(12): p. 1452–1458. https://doi.org/10.1038/ng.2802 PMID: 24162737

9. Raj T., et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science, 2014. 344(6183): p. 519–523. https://doi.org/10.1126/science.1249547 PMID: 24786080
10. Heneka M.T., Golenbock D.T., and Latz E., Innate immunity in Alzheimer's disease. Nat Immunol, 2015. 16(3): p. 229–236. https://doi.org/10.1038/ni.3102 PMID: 25689443
11. Heppner F.L., Ransohoff R.M., and Becher B., Immune attack: the role of inflammation in Alzheimer disease. Nature Reviews Neuroscience, 2015. 16(6): p. 358–372. https://doi.org/10.1038/nrn3880 PMID: 25991443
12. Lu Q., et al. A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. Sci. Rep., 2015. 5.
13. Fischle W., Wang Y., and Allis C.D., Binary switches and modification cassettes in histone biology and beyond. Nature, 2003. 425(6957): p. 475–479. https://doi.org/10.1038/nature02017 PMID: 14523437
14. Trynka G., et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature genetics, 2013. 45(2): p. 124–130. https://doi.org/10.1038/ng.2504 PMID: 23263488
15. Wienholds E. and Plasterk R.H., MicroRNA function in animal development. FEBS letters, 2005. 579 (26): p. 5911–5922. https://doi.org/10.1016/j.febslet.2005.07.070 PMID: 16111679
16. Yanai I., et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics, 2005. 21(5): p. 650–659. https://doi.org/10.1093/bioinformatics/bti042 PMID: 15388519
17. Ludwig N., et al. Distribution of miRNA expression across human tissues. Nucleic acids research, 2016. 44(8): p. 3865–3877. https://doi.org/10.1093/nar/gkw116 PMID: 26921406
18. Mattick J.S., The genetic signatures of noncoding RNAs. PLoS Genet, 2009. 5(4): p. e1000459. https://doi.org/10.1371/journal.pgen.1000459 PMID: 19390609
19. Derrien T., et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research, 2012. 22(9): p. 1775–1789. https://doi.org/10.1101/gr.132159.111 PMID: 22955988
20. Bulger M. and Groudine M., Enhancers: the abundance and function of regulatory sequences beyond promoters. Developmental biology, 2010. 339(2): p. 250–257. https://doi.org/10.1016/j.ydbio.2009.11.035 PMID: 20025863
21. Andersson R., et al. An atlas of active enhancers across human cell types and tissues. Nature, 2014. 507(7493): p. 455–461. https://doi.org/10.1038/nature12787 PMID: 24670763
22. Korn T., et al. IL-17 and Th17 Cells. Annual review of immunology, 2009. 27: p. 485–517. https://doi.org/10.1146/annurev.immunol.021908.132710 PMID: 19132915
23. Wilson C.B., Rowell E., and Sekimata M., Epigenetic control of T-helper-cell differentiation. Nature Reviews Immunology, 2009. 9(2): p. 91–105. https://doi.org/10.1038/nri2487 PMID: 19151746
24. Dong C., Genetic controls of Th17 cell differentiation and plasticity. Experimental & molecular medicine, 2011. 43(1): p. 1–6.
25. Akimzhanov A.M., Yang X.O., and Dong C., Chromatin remodeling of interleukin-17 (IL-17)-IL-17F cytokine gene locus during inflammatory helper T cell differentiation. Journal of Biological Chemistry, 2007. 282(9): p. 5969–5972. https://doi.org/10.1074/jbc.C600322200 PMID: 17218320
26. Finucane H.K., et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics, 2015.
27. Wood A.R., et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet, 2014. 46(11): p. 1173–86. https://doi.org/10.1038/ng.3097 PMID: 25282103
28. Group, S.S.S.S., Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). The Lancet, 1994. 344(8934): p. 1383–1389.
29. Simon-Sanchez J., et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nature genetics, 2009. 41(12): p. 1308–1312. https://doi.org/10.1038/ng.487 PMID: 19915575
30. Hardy J., et al. Pathways to Alzheimer's disease. Journal of internal medicine, 2014. 275(3): p. 296–303. https://doi.org/10.1111/joim.12192 PMID: 24749173
31. Jones L., et al. Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. PloS one, 2010. 5(11): p. e13950. https://doi.org/10.1371/journal.pone.0013950 PMID: 21085570
32. Kircher M., et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics, 2014. 46(3): p. 310–315. https://doi.org/10.1038/ng.2892 PMID: 24487276
33. Ritchie G.R., et al. Functional annotation of noncoding sequence variants. Nature methods, 2014. 11 (3): p. 294–296. https://doi.org/10.1038/nmeth.2832 PMID: 24487584

34. Ionita-Laza I., et al. A spectral approach integrating functional genomic annotations for coding and non-coding variants. Nature genetics, 2016. 48(2): p. 214–220. https://doi.org/10.1038/ng.3477 PMID: 26727659

35. Anttila V., et al. Analysis of shared heritability in common disorders of the brain. bioRxiv, 2016: p. 048991.

36. Ridge P.G., et al. Alzheimer's disease: analyzing the missing heritability. PloS one, 2013. 8(11): p. e79771. https://doi.org/10.1371/journal.pone.0079771 PMID: 24244562

37. Martinelli-Boneschi F., et al. Pharmacogenomics in Alzheimer's disease: a genome-wide association study of response to cholinesterase inhibitors. Neurobiology of aging, 2013. 34(6): p. 1711. e7–1711. e13.

38. Esposito F., et al. A pharmacogenetic study implicates SLC9a9 in multiple sclerosis disease activity. Annals of neurology, 2015. 78(1): p. 115–127. https://doi.org/10.1002/ana.24429 PMID: 25914168

39. Turner S.T., et al. Genomic association analysis identifies multiple loci influencing antihypertensive response to an angiotensin II receptor blocker. Hypertension, 2012. 59(6): p. 1204–1211. https://doi.org/10.1161/HYP.0b013e31825b30f8 PMID: 22566498

40. Mick E., et al. Family-based genome-wide association scan of attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry, 2010. 49(9): p. 898–905. e3.

41. Kondapalli K.C., et al. Functional evaluation of autism-associated mutations in NHE9. Nature communications, 2013. 4.

42. Chalasani N., et al. Genome-wide association study identifies variants associated with histologic features of nonalcoholic fatty liver disease. Gastroenterology, 2010. 139(5): p. 1567–1576. e6. https://doi.org/10.1053/j.gastro.2010.07.057 PMID: 20708005

43. Matarín M., et al. A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. The Lancet Neurology, 2007. 6(5): p. 414–420. https://doi.org/10.1016/S1474-4422(07)70081-9 PMID: 17434096

44. Yashin A.I., et al. Joint influence of small-effect genetic variants on human longevity. Aging, 2010. 2(9): p. 612–620. https://doi.org/10.18632/aging.100191 PMID: 20834067

45. Okada Y., et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature, 2014. 506(7488): p. 376–381. https://doi.org/10.1038/nature12873 PMID: 24390342

46. Martin J.-E., et al. A systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals new shared susceptibility loci. Human molecular genetics, 2013. 22(19): p. 4021–4029. https://doi.org/10.1093/hmg/ddt248 PMID: 23740937

47. Day F.R., et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. Nat Genet, 2015. 47(11): p. 1294–303. https://doi.org/10.1038/ng.3412 PMID: 26414677

48. Hart A.B., et al. Genome-wide association study of d-amphetamine response in healthy volunteers identifies putative associations, including cadherin 13 (CDH13). PloS one, 2012. 7(8): p. e42646. https://doi.org/10.1371/journal.pone.0042646 PMID: 22952603

49. Potkin S.G., et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. PloS one, 2009. 4(8): p. e6501. https://doi.org/10.1371/journal.pone.0006501 PMID: 19668339

50. Jun G., et al. A novel Alzheimer disease locus located near the gene encoding tau protein. Molecular psychiatry, 2015.

51. Ashok A., et al. Chronic cerebral hypoperfusion-induced impairment of Aβ clearance requires HB-EGF-dependent sequential activation of HIF1α and MMP9. Neurobiology of Disease, 2016. 95: p. 179–193. https://doi.org/10.1016/j.nbd.2016.07.013 PMID: 27431094

52. Severini C., et al. Bv8/prokineticin 2 is involved in Aβ-induced neurotoxicity. Scientific reports, 2015. 5.

53. Stoothoff W.H., et al. Axin negatively affects tau phosphorylation by glycogen synthase kinase 3β. Journal of neurochemistry, 2002. 83(4): p. 904–913. PMID: 12421363

54. Nalls M.A., et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nature genetics, 2014. 46(9): p. 989–993. https://doi.org/10.1038/ng.3043 PMID: 25064009

55. Purcell S.M., et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 2009. 460(7256): p. 748–752. https://doi.org/10.1038/nature08185 PMID: 19571811

56. Winham S., et al. Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. Molecular psychiatry, 2014. 19(9): p. 1010–1016. https://doi.org/10.1038/mp.2013.159 PMID: 24322204

57. Lei P., et al. Tau protein: relevance to Parkinson's disease. The international journal of biochemistry & cell biology, 2010. 42(11): p. 1775–1778.
58. Allen M., et al. Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels. Alzheimer's research & therapy, 2014. 6(4): p. 1.
59. Lu Q., et al. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics, 2016. 32(4): p. 542–548. https://doi.org/10.1093/bioinformatics/btv610 PMID: 26504140
60. Draber P., et al. SCIMP, a transmembrane adaptor protein involved in major histocompatibility complex class II signaling. Molecular and cellular biology, 2011. 31(22): p. 4550–4562. https://doi.org/10.1128/MCB.05817-11 PMID: 21930792
61. Talwar P., et al. Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease. BMC genomics, 2014. 15(1): p. 1.
62. Desikan R.S., et al. Polygenic overlap between C-reactive protein, plasma lipids and Alzheimer's disease. Circulation, 2015: p. CIRCULATIONAHA. 115.015489.
63. Sheu K.-F., et al. A DLST genotype associated with reduced risk for Alzheimer's disease. Neurology, 1999. 52(7): p. 1505–1505. PMID: 10227647
64. Davies G., et al. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N= 53 949). Molecular psychiatry, 2015. 20(2): p. 183–192. https://doi.org/10.1038/mp.2014.188 PMID: 25644384
65. Shin S.-Y., et al. An atlas of genetic influences on human blood metabolites. Nature genetics, 2014. 46(6): p. 543–550. https://doi.org/10.1038/ng.2982 PMID: 24816252
66. Jiang S., et al. Trafficking regulation of proteins in Alzheimer's disease. Molecular neurodegeneration, 2014. 9(1): p. 1.
67. Vardarajan B.N., et al. Identification of Alzheimer disease-associated variants in genes that regulate retromer function. Neurobiology of aging, 2012. 33(9): p. 2231. e15–2231. e30.
68. Jun G.R., et al. Transethnic genome-wide scan identifies novel Alzheimer's disease loci. Alzheimer's & Dementia, 2017.
69. Jonsson T., et al. Variant of TREM2 associated with the risk of Alzheimer's disease. New England Journal of Medicine, 2013. 368(2): p. 107–116. https://doi.org/10.1056/NEJMoa1211103 PMID: 23150908
70. Guerreiro R., et al. TREM2 variants in Alzheimer's disease. New England Journal of Medicine, 2013. 368(2): p. 117–127. https://doi.org/10.1056/NEJMoa1211851 PMID: 23150934
71. Gagliano S.A., et al. Genomics implicates adaptive and innate immunity in Alzheimer's and Parkinson's diseases. Annals of Clinical and Translational Neurology, 2016. 3(12): p. 924–933. https://doi.org/10.1002/acn3.369 PMID: 28097204
72. Huang K.-l. et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. *Nat Neurosci* advance online publication(2017).
73. Sarkar A.K., Ward L.D., and Kellis M., Functional enrichments of disease variants across thousands of independent loci in eight diseases. BioRxiv, 2016: p. 048066.
74. Finucane H., et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. bioRxiv, 2017: p. 103069.
75. Shi H., Kichaev G., and Pasaniuc B., Contrasting the genetic architecture of 30 complex traits from summary association data. The American Journal of Human Genetics, 2016. 99(1): p. 139–153. https://doi.org/10.1016/j.ajhg.2016.05.013 PMID: 27346688
76. Gamazon E.R., et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics, 2015. 47(9): p. 1091–1098. https://doi.org/10.1038/ng.3367 PMID: 26258848
77. Gusev A., et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics, 2016.
78. Mancuso N., et al. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American Journal of Human Genetics, 2017. 100(3): p. 473–487. https://doi.org/10.1016/j.ajhg.2017.01.031 PMID: 28238358
79. Ernst J. and Kellis M., Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nature biotechnology, 2015. 33(4): p. 364–376. https://doi.org/10.1038/nbt.3157 PMID: 25690853
80. Consortium T.F., A promoter-level mammalian expression atlas. Nature, 2014. 507(7493): p. 462–470. https://doi.org/10.1038/nature13182 PMID: 24670764
81. Abecasis G.R., et al. An integrated map of genetic variation from 1,092 human genomes. Nature, 2012. 491(7422): p. 56–65. https://doi.org/10.1038/nature11632 PMID: 23128226

82. Yang J., et al. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics, 2011. 88(1): p. 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011 PMID: 21167468

83. Willer C.J., Li Y., and Abecasis G.R., METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics, 2010. 26(17): p. 2190–1. https://doi.org/10.1093/bioinformatics/btq340 PMID: 20616382

84. Reimand J., et al. g: Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic acids research, 2016: p. gkw199.

85. Pruim R.J., et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics, 2010. 26(18): p. 2336–2337. https://doi.org/10.1093/bioinformatics/btq419 PMID: 20634204