nature neuroscience

An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome

Bernard Ng[1,2], Charles C White[3], Hans-Ulrich Klein[3,4], Solveig K Sieberts[5], Cristin McCabe[3], Ellis Patrick[3], Jishu Xu[3], Lei Yu[6], Chris Gaiteri[6], David A Bennett[6], Sara Mostafavi[1,2,7,8] & Philip L De Jager[3,4,8]

We report a multi-omic resource generated by applying quantitative trait locus (xQTL) analyses to RNA sequence, DNA methylation and histone acetylation data from the dorsolateral prefrontal cortex of 411 older adults who have all three data types. We identify SNPs significantly associated with gene expression, DNA methylation and histone modification levels. Many of these SNPs influence multiple molecular features, and we demonstrate that SNP effects on RNA expression are fully mediated by epigenetic features in 9% of these loci. Further, we illustrate the utility of our new resource, xQTL Serve, by using it to prioritize the cell type(s) most affected by an xQTL. We also reanalyze published genome wide association studies using an xQTL-weighted analysis approach and identify 18 new schizophrenia and 2 new bipolar susceptibility variants, which is more than double the number of loci that can be discovered with a larger blood-based expression eQTL resource.

Genome wide association studies (GWAS) have identified thousands of SNPs that are associated with various human diseases[1]. However, most identified SNPs fall in the noncoding regions of the genome[2]. Connecting these regulatory changes to specific genes or to molecular pathways that may be implicated in human diseases is not straightforward. Suggestive evidence indicates that many more such SNPs exist, but they are difficult to detect due to their typically small effect sizes and the challenge of multiple-testing burden in genome-wide assessment of common genetic variation[3].

Expression quantitative trait locus (eQTL) analyses[4–6] have been very useful in understanding the functional consequences of trait- and disease-associated variants and in identifying genes that are likely to be affected by a risk allele. Recently, QTL analyses have been extended to other molecular phenotypes, such as DNA methylation (mQTL)[7,8] and histone modification (haQTL)[9]. Overall, SNPs associated with molecular phenotypes (collectively, xQTLs) are over-represented among SNPs that are linked to various traits and diseases[6,10], and previous studies have used eQTL hits to prioritize associations in GWAS, leading to improved detection sensitivity[11–13]. While a few data sets exist for brain tissue, large data sets measuring all three of these epigenomic and transcriptomic features have only recently been generated from the same brain region of each individual.

Here we present a Resource for the neuroscience community by performing xQTL analyses on a multi-omic data set that consists of RNA sequence (RNA-seq), DNA methylation and histone acetylation by chromatin immunoprecipitation and sequencing (H3K9Ac ChIP-seq) data derived from the dorsolateral prefrontal cortex (DLPFC) of up to 494 subjects (411 subjects having all three data types and genotypes available). Samples were collected at autopsy from participants of the Religious Orders Study (ROS) and the Rush Memory and Aging Project (MAP), which are two longitudinal studies of aging designed by the same group of investigators. These studies share the same sample and data collection procedures, which facilitate joint analyses[14,15]. At its heart, the Resource presents a list of SNPs associated with cortical gene expression, DNA methylation and/or histone modification levels that reflects the impact of genetic variation on the transcriptome and epigenome of aging brains. While our xQTLs replicated well in independent brain- and blood-derived QTL resources, a notable portion of xQTLs is specific to genes that are expressed only in older brains. Also, many SNPs influence multiple molecular features, with a small number having their impacts on gene expression mediated through epigenetics. Further, we apply a computational approach to prioritize the cell types that may be driving the tissue-level effect, a critical piece of information for designing follow-up molecular experiments in which an *in vitro* or *in vivo* target cell type needs to be selected. Finally, we illustrate the efficacy of an 'xQTL-weighted GWAS' approach that leverages our xQTL resource. We show that this approach increases the statistical power of GWAS, resulting in the detection of a number of new susceptibility variants for several diseases. All data used in this study are available from http://www.radc.rush.edu/, and the xQTL results and analysis scripts can be accessed through our online portal, xQTL Serve, at http://mostafavilab.stat.ubc.ca/xQTLServe/.

[1]Departments of Statistics and Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [2]Centre for Molecular Medicine and Therapeutics, Vancouver, British Columbia, Canada. [3]Broad Institute, Cambridge, Massachusetts, USA. [4]Center for Translational & Systems Neuroimmunology, Department of Neurology, Columbia University Medical Center, New York, New York, USA. [5]Sage Bionetworks, Seattle, Washington, USA. [6]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, USA. [7]Canadian Institute for Advanced Research, CIFAR Program in Child and Brain Development, Toronto, Ontario Canada. [8]These authors contributed equally to this work. Correspondence should be addressed to P.L.D.J. (pld2115@cumc.columbia.edu) or S.M. (saram@stat.ubc.ca).

Received 15 December 2016; accepted 2 August 2017; published online 4 September 2017; doi:10.1038/nn.4632

RESULTS

xQTL discovery

Genotype data[16] were generated from 2,093 individuals of European descent. Of these individuals, gene expression (RNA-seq; $n = 494$), DNA methylation[17] (450K Illumina array; $n = 468$) and histone modification data (H3K9Ac ChIP-seq; $n = 433$) were derived from postmortem frozen samples of a single cortical region, the dorsolateral prefrontal cortex (DLPFC) (**Fig. 1a**). 411 individuals had all four data types. Demographics of the analyzed individuals are summarized in **Supplementary Table 1**. Although some of these data have been previously published with respect to analysis of aging brain phenotypes (see **Supplementary Table 2**), here we report genome-wide xQTL analyses for these data sets for the first time. Genotype imputation was performed using BEAGLE[18] 3.3.2 with the 1000 Genomes reference panel[19], yielding 7,321,515 SNPs for analysis. For the molecular phenotype data, 13,484 expressed genes, 420,103 methylation sites and 26,384 acetylation peaks remained after quality control analyses (**Supplementary Figs. 1–3**). The effects of known and hidden confounding factors were removed from the molecular phenotype data using linear regression (Online Methods). Consistent with previous studies[20], we observed that accounting for hidden confounding factors greatly enhanced the statistical power of *cis* eQTL detection, and we confirmed that this observation held true for *cis* mQTL and *cis* haQTL detection (**Supplementary Fig. 4**).

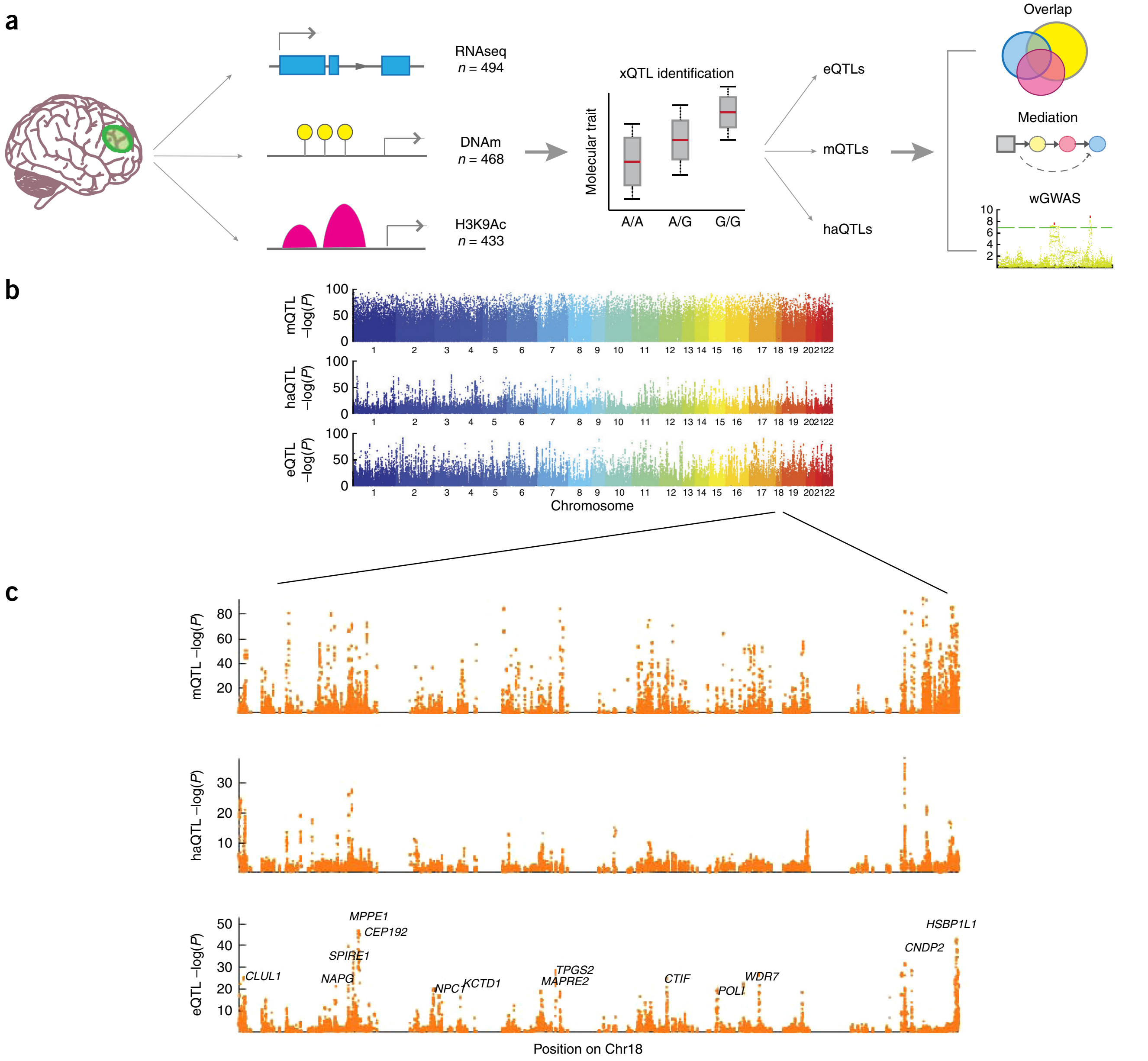

Figure 1 Overview of xQTL analysis. (**a**) Graphical summary of our data and analyses. We first associate genetic variation with each data type separately to establish our xQTL reference. We then use these xQTLs to assess whether a given SNP influences more than one data type, whether epigenomic features mediate the effects of SNPs on gene expression, and whether our xQTLs can be leveraged using a weighted GWAS (wGWAS) analysis approach to discover new susceptibility loci. (**b**) $-\log_{10}$ *P*-value of Spearman's correlation between SNPs and DNA methylation (mQTL), histone acetylation (haQTL) and gene expression (eQTL) vs. the SNPs' physical positions in the genome. Each dot represents the strongest association within a *cis* window for each SNP. (**c**) Zoomed-in Manhattan plot of chromosome 18 to illustrate *P*-value distribution of xQTLs at a higher resolution.

Table 1 Summary of xQTL associations

	No. associations (SNP–gene pairs)		No. features		No. SNPs	
	Tested	Significant	Tested	Significant	Tested	Significant
eQTLs (1 Mb)	60,456,556	405,429	12,979	3,388	6,442,864	313,467
mQTLs (5 kb)	9,939,236	693,696	412,152	56,973	2,358,873	383,920
haQTLs (1 Mb)	125,100,450	156,693	25,720	1,681	6,756,597	119,778

We employed Spearman's rank correlation to estimate the association strength between alleles of each SNP and gene expression (n = 494), DNA methylation (n = 468) and histone acetylation levels (n = 433). We refer to the measurement unit of each molecular phenotype data as a feature and a significant association between a SNP and a feature as an xQTL (that is, an xQTL is a SNP–feature pair). Based on the results of prior studies, we performed *cis* xQTL analysis between SNPs and each feature by defining a window of 1 Mb for eQTL analysis and haQTL analysis and a window of 5 kb for mQTL analysis[21,22]. The 1-Mb window for haQTL analysis was motivated by the possibility that SNPs in enhancer regions, which are far away, can indeed impact gene regulation through interaction with promoter regions (for example, chromatin looping). The much smaller window for the mQTL analysis was selected since the majority of *cis* mQTLs with the strongest correlation lie within a window of this size[22]. Also, the smaller window size helps reduce the multiple-testing burden, given the much larger number of DNA methylation features.

Using a Bonferroni corrected P-value threshold (α_{FWER} = 0.05, two-tailed), we found 3,388 genes associated with eQTL SNPs ($P < 8 \times 10^{-10}$), 56,973 CG dinucleotides linked to mQTL SNPs ($P < 5 \times 10^{-9}$) and 1,681 H3K9Ac peaks influenced by haQTL SNPs ($P < 4 \times 10^{-10}$) (**Fig. 1b**,**c** and **Table 1**). Among the eQTL genes, 133 of them correspond to long intergenic noncoding RNAs (lincRNAs), out of a total of 391 lincRNAs tested in the eQTL analysis. The complete lists of eQTLs, mQTLs and haQTLs are provided through the xQTL Serve webpage: http://mostafavilab.stat.ubc.ca/xQTLServe/.

Replication and cross-tissue comparisons

We evaluated the extent to which our xQTLs replicate brain eQTLs and mQTLs found in prior studies. We focused on eQTL and mQTL replication since relevant large-sample data sets are only available for these two xQTL types. Specifically, we assessed the replication rate of brain eQTLs discovered in the CommonMind[23] and Braineac[24] studies, and brain mQTLs in a fetal brain study[8], in our data set using the π_1 statistic[25], which estimates the proportion of these eQTLs or mQTLs that are also significant in our data set. π_1 values of the eQTLs are 0.91 and 0.56 for CommonMind and Braineac, respectively, and π_1 of mQTLs is 0.87 for the fetal brain study. All of these results are greater than their respective empirical null mean of 0.11 and 0.33 for eQTLs and mQTLs, respectively (P < 0.0001, one-tailed; see Online Methods). The lower replication rate of Braineac eQTLs compared to CommonMind eQTLs could be due to its smaller sample size. Also, the Braineac eQTLs were based on false discovery rate (FDR) correction whereas CommonMind eQTLs were defined using Bonferroni correction, and stronger associations captured by more stringent correction are more likely to replicate[26]. We also assessed the replication rate of our eQTLs in the CommonMind data and estimated a similar replication rate (π_1 = 0.90). For the mQTL replication analysis, we explored restricting our mQTL analysis to a 100-kb window and observed similar replication rate (π_1 = 0.87) on the fetal brain mQTLs[8], which suggests a 5-kb window already captures most of the stronger associations between SNP and DNA methylation.

For assessing cross-tissue replication, we used a large whole-blood eQTL data set from the DGN study[26] and two smaller eQTL data sets from the Immune Variation (ImmVar) study[27] that consist of monocyte and T cell data. π_1 values of these eQTLs in our data set are 0.63 (whole blood), 0.61 (monocytes) and 0.67 (T cells), which are greater than their empirical null mean of 0.10 (P < 0.0001 for all three data sets, one-tailed). Thus, a large proportion of blood eQTLs are present in our brain data. We also assessed the replication rate of our brain-derived eQTLs in the whole-blood DGN data set (**Fig. 2a**,**b**). When we considered SNP–gene pairs that could be tested in both studies, we observed a replication rate of 0.83 (**Fig. 2c**), which is greater than its empirical null mean of 0.30 (P < 0.0001, one-tailed). This higher replication rate may be due to the higher statistical power of the DGN study and the fact that cortical tissue consists of a large variety of cell types, which in aggregate express a large proportion of the transcriptome. Since blood contains a mixture of cell types, including immune cells, that share characteristics with those in brain, we further assessed the replication rate on three more tissues, namely subcutaneous adipose, visceral adipose and liver tissues from the GTEx study[28]. The replication rates were 0.51, 0.38 and 0.20, respectively, which are indeed lower than that of blood. Additional replication results for different tissues, window sizes and xQTL types are provided in **Supplementary Table 3**.

An important question to answer with our data is whether and which of the detected xQTLs are brain-specific. However, without tissue samples from the same individuals, distinguishing between subject-specific and tissue-specific effects is not possible. Nonetheless, based on the sparsity of 'population-specific' eQTLs[27] and a lower replication rate of eQTLs in blood compared to brain, a notable fraction of our eQTLs are likely tissue-specific. For example, when we considered only eQTLs that consist of the top SNP for each gene, we found that, of the 2,416 eQTLs discovered in our cortical tissue study that are testable in the whole-blood data set[26], 433 eQTLs (18%) had an unadjusted P > 0.05, indicating that this subset of brain eQTLs is unlikely to be present in blood (**Fig. 2b**). As an example, *NLRP1* RNA is expressed in both brain and blood (whole blood, monocytes and T cells), but its expression is associated only with brain-specific eQTL SNPs (**Fig. 2d**). *NLRP1* is a member of the inflammasome complex that is implicated in inflammatory response in both immune cells (in particular myeloid cells) and brain[29]. Notably, a few small-scale studies have linked polymorphisms in this gene with amyloid-β secretion and Alzheimer's disease[30]. In addition to the 2,416 eQTLs that were testable in both brain and blood, we identified 809 eQTL target genes from our brain eQTL analysis that were absent from the DGN blood eQTL analysis because these genes were not expressed in blood. As expected, this set of 809 brain-specific eQTL genes was enriched for brain-relevant functions (gene set enrichment analysis, FDR < 0.05, two-tailed) such as "neuronal system", "potassium channel components" and "neurotransmitter receptor binding."

Overall, the high cross-sample and cross-tissue replication rates suggest that a large number of SNPs that influence molecular phenotypes are shared across contexts. The degree of overlap between brain and blood eQTLs is high, with a π_1 of ~0.8. Nevertheless,

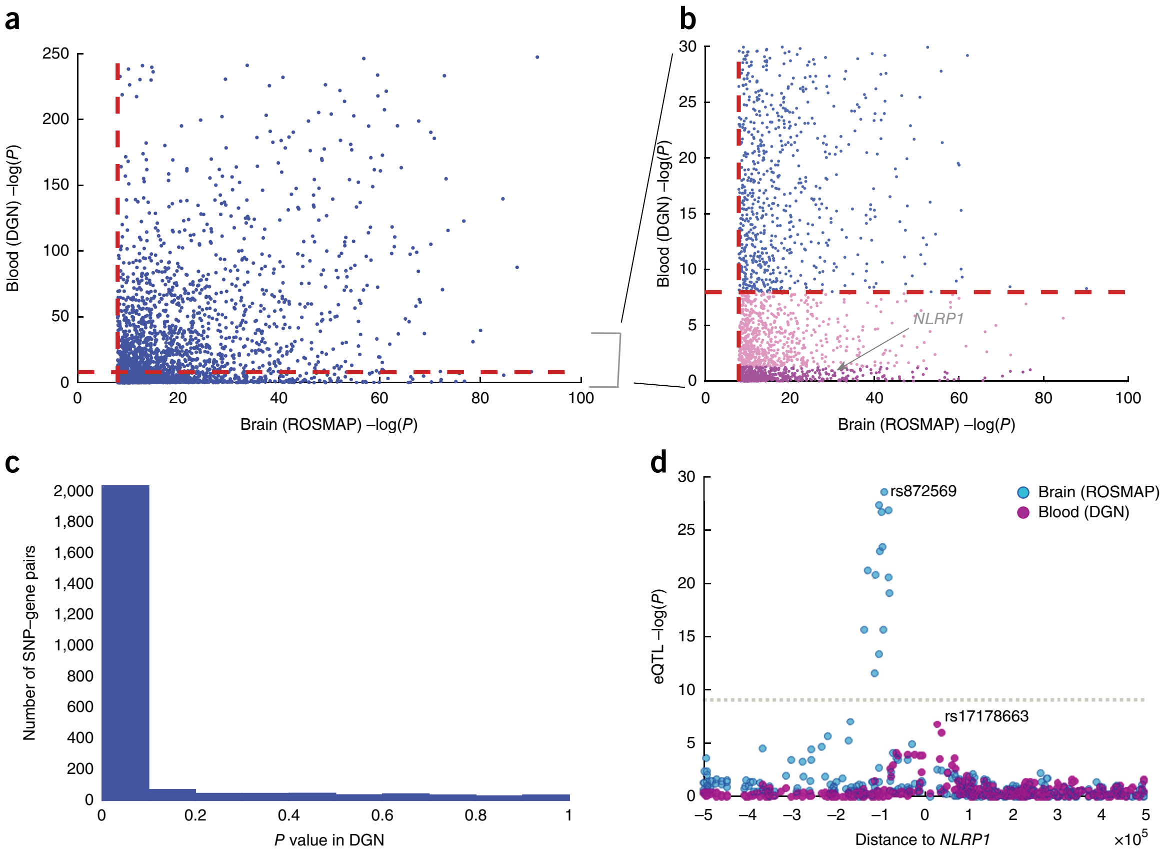

Figure 2 Cross-tissue replication analysis. (**a**) Scatter plot of $-\log_{10}$ *P*-values of associations between the lead brain eQTL SNPs and their associated genes in brain and blood. The dashed red lines denote the significance threshold ($\alpha_{\text{FWER}} = 0.05$ with Bonferroni correction). (**b**) $-\log_{10}$ *P*-value distribution of eQTLs that appear to be brain specific (pink and purple pink; the latter are specific to *NLRP1*) and those that appear to be non-tissue specific (blue). (**c**) Distribution of *P*-values from the DGN study restricted to brain eQTLs. Estimated replication rate (π_1 statistic) between blood and brain eQTLs is 0.83. (**d**) eQTL *P*-values at *NLRP1* locus. Each dot represents one SNP tested in either brain (ROSMAP) or blood (DGN). The *x* axis corresponds to the distance between each assessed *cis* SNP and *NLRP1*'s TSS, and the *y* axis corresponds to $-\log_{10}$ *P*-values for association between SNPs and *NLRP1* expression. The linkage disequilibrium between the lead SNP in blood and brain is $r^2 < 0.1$.

our results suggest that some eQTLs are tissue-specific, and more tissue-specific effects would likely emerge from analyses of purified cell populations.

Genetic architecture of xQTL SNPs and sharing across molecular phenotypes

We used epigenomic annotations derived by applying ChromHMM[31] to DLPFC tissue data to estimate the log odds of an xQTL SNP belonging to one of 15 chromatin states in comparison to all non-xQTL SNPs near our molecular features—that is, within 1-Mb, 5-kb and 1-Mb windows for eQTL, mQTL and haQTL analyses, respectively. eQTL SNPs were enriched mainly in promoters and transcribed regions (**Fig. 3a**), conforming to our understanding of how SNPs at transcription factor binding sites can affect protein–DNA interactions[32] and how SNPs in transcribed regions are known to affect mRNA processing and turnover[33]. haQTL SNPs were also largely enriched in promoter and transcribed regions, consistent with the role of H3K9Ac in transcriptional activation[34]. By contrast, mQTL SNPs were mainly enriched in bivalent regions (promoters and enhancers) and Polycomb-repressed regions, which matches previous findings that a large portion of mQTL SNPs resides in chromatin regions that are developmentally regulated[22]. Also, suppressed gene expression in Polycomb-repressed regions might partly explain why eQTL and haQTL SNPs derived from adult samples are scarce in these regions. Notably, xQTL SNPs that are shared across all three molecular phenotypes were mainly enriched close to the transcription start site (TSS), as well as in the 5′ and 3′ transcribed regions. With respect to transcribed sequences, we saw enrichment for all types of xQTLs in exons relative to introns (**Fig. 3b**), with this trend being most striking for mQTLs.

To quantify the degree to which an xQTL SNP influences more than one molecular phenotype, we first identified the list of xQTL SNPs for a 'discovery' phenotype and then estimated the π_1 statistics of the SNP–feature associations for a 'test' phenotype that share the same xQTL SNPs. Since an xQTL SNP might be tested for association with multiple *cis* features—for example, an mQTL SNP was, on average, tested for association with 18 gene expression levels—we needed to decide which SNP–feature associations to include in the π_1 estimation (Online Methods). In particular, we examined the distance between a discovery SNP and a test feature, and we found this distance to be a prime determinant of cross-phenotype sharing. For example, the most strongly associated eQTL gene for each mQTL SNP is often the gene closest to the mQTL SNP (**Fig. 3c** and **Supplementary Fig. 5**). On the basis of this observation, we estimated π_1 to be 0.41–0.63 when considering only the closest feature to each xQTL SNP (**Fig. 3d**). Also, we examined the effect of window size by restricting the haQTL analyses

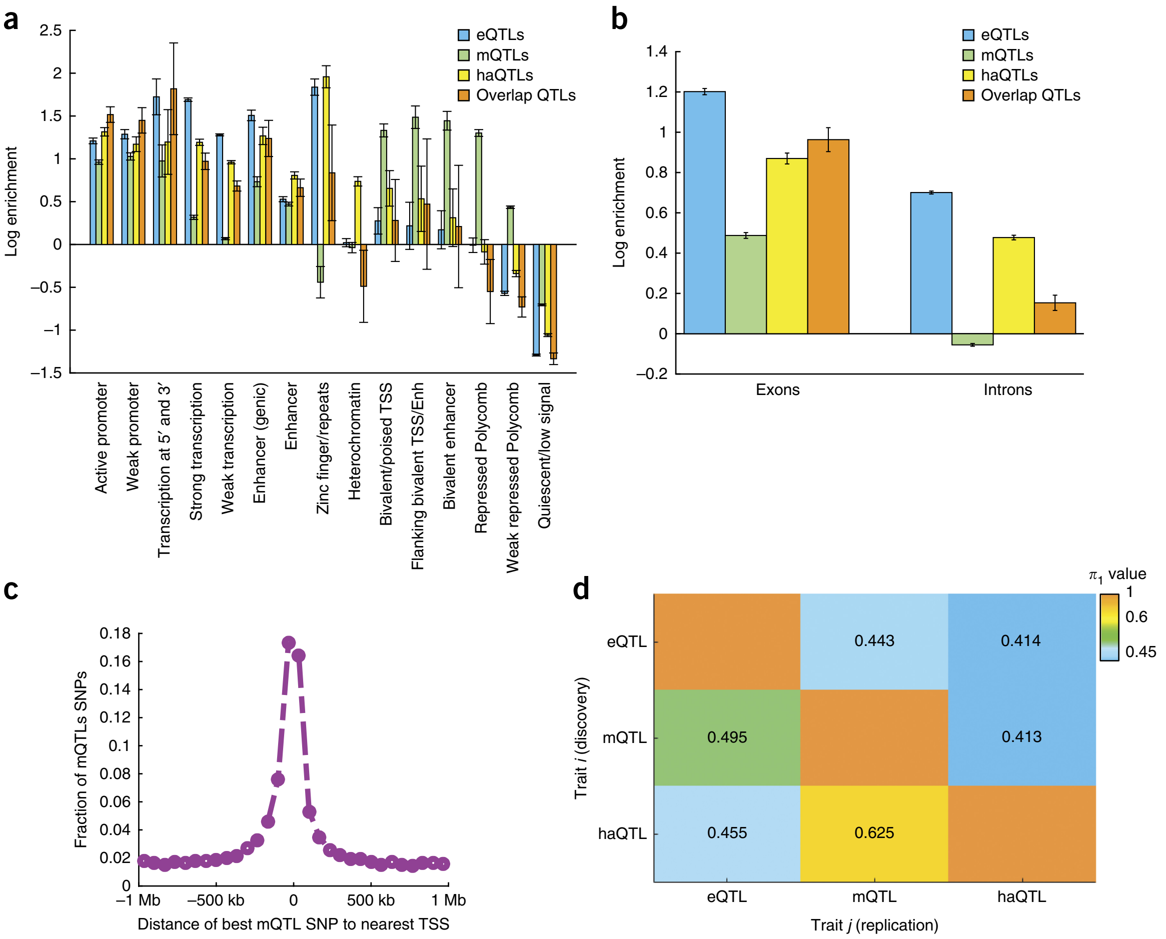

Figure 3 Genomic enrichment of xQTLs and their overlap. (**a**) Log odds ratio of xQTL SNP enrichment in 15 different chromatin states[31] as defined by the Roadmap Epigenomics project through applying ChromHMM to DLPFC samples from two cognitively unimpaired ROSMAP subjects. Error bars reflect s.d. (**b**) Log odds ratio of xQTL SNP enrichment in exons and introns. Error bars reflect s.d. (**c**) Distribution of distance between each lead mQTL SNP and its nearest TSS. (**d**) π_1 statistics for assessing xQTL sharing across the three molecular features. Each cell (i,j) corresponds to the proportion of xQTLs of trait j that share the same xQTL SNPs identified in trait i.

to 2-kb, 40-kb and 100-kb windows, as well as changing the eQTL and mQTL analysis window to 100 kb, and we found negligible differences in our estimates of xQTL sharing (**Supplementary Table 4**).

The availability of multi-omic data from the same individuals enabled us to go beyond 'overlap analyses' (**Fig. 4a**) and to investigate the cascading effect of genetic variation through the measured regulatory genomics layers. Specifically, we investigated whether the effect of a regulatory *cis* xQTL SNP is mechanistically mediated through its impact on epigenetic modification or gene expression using the casual inference test[35]. This analysis was performed on 10,897 xQTL SNPs (influencing 629 genes as based on the eQTL analysis) that were associated with all three molecular phenotypes, as only such SNPs satisfy the precondition for mediation analysis. With this analysis, we distinguished among three models for propagation of information from genetic variation: (i) independent effects of a SNP on *cis* gene expression and the *cis* epigenetic landscape (independent model), (ii) a propagation path from SNP to gene expression via epigenetic modifications (epigenetic mediation model), or (iii) a propagation path from SNP to the epigenome (namely DNA methylation) via gene expression (transcription mediation model) (**Fig. 4b**).

Using Bonferroni correction with the casual inference test ($n = 411$, two-tailed), 9% of the association sets conformed to the epigenetic mediation model, 3% conformed to the transcription mediation model, 85% conformed to the independent model and the remaining 3% could not be classified (**Fig. 4c** and **Supplementary Table 5**). As an example, an xQTL SNP (rs13015714) associated with celiac disease (GWAS $P < 10^{-8}$) was found to affect *IL1RL1* gene expression ($P < 10^{-11}$), DNA methylation ($P < 10^{-30}$) and histone modification ($P < 10^{-12}$), but the impact of this SNP on gene expression appeared to be fully mediated by epigenetic modifications (**Fig. 4d**,**e**), and thus this SNP conforms to the epigenetic mediation model. We also tested whether GWAS SNPs (downloaded from the GWAS catalog[1]) are preferentially enriched for any of these models but did not find any model-specific enrichment.

A large fraction of the shared xQTL SNPs appear to affect gene expression directly. This result could be explained by (i) epigenetic modification playing a passive role[21] whereby gene expression in fact lies upstream of epigenetic modification (3% based on the transcription mediation model), (ii) regulation of gene expression being dependent on a more complex combination of epigenetic marks that were not measured in our subjects, and (iii) artifactual decorrelation between the expression and epigenomic features due to technical or other factors. Thus, we should interpret the detected mediation as only a subset of true mediation—that is, these may be the most

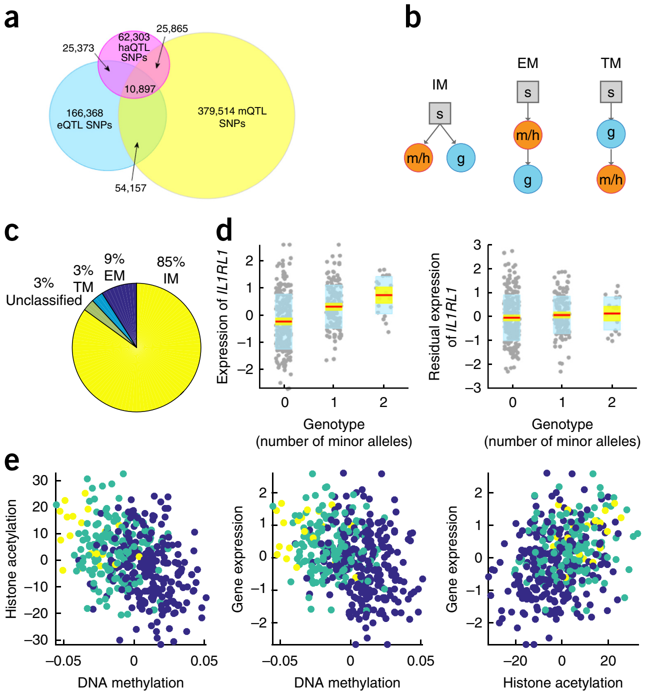

Figure 4 Epigenetic mediation of eQTLs. (**a**) Sharing of SNPs between eQTLs, mQTLs and haQTLs. 2,305,942 SNPs tested for all molecular phenotypes were considered. (**b**) Three models relating SNPs (s), epigenetic features (methylation and/or histone acetylation, m/h) and gene expression (g): (i) independent model (IM), wherein effects of SNPs on epigenetic features and transcripts are unrelated; (ii) epigenetic mediation model (EM), wherein epigenetic features mediate the effects of SNPs on gene expression; and (iii) transcription mediation model (TM), wherein the effects of SNPs on epigenetics are mediated through their effect on gene expression. The causal inference test was used for assessing mediation[35]. (**c**) Proportion of shared xQTL SNPs that are consistent with each model. (**d**) Expression level of *IL1RL1* vs. number of minor alleles present for rs13015714, which is a shared xQTL SNP that affects *IL1RL1* expression and nearby DNA methylation and histone acetylation levels. The red line corresponds to the mean. The yellow region corresponds to the 95% confidence interval of the mean. The edges of the blue region correspond to ± 1 s.d. The SNP effect disappears after regressing out the effect of the mQTL probes and haQTL peaks associated with rs13015714 from *IL1RL1* expression. (**e**) Association between *IL1RL1* expression and the levels of its associated methylation probes and acetylation peaks. Colors indicate the rs13015714 genotype: minor allele homozygotes (yellow), heterozygotes (green) or major allele homozygotes (blue).

robust subset of mediation events. Further work and more data may be needed to assess this issue more comprehensively. Indeed, when we separately included only DNA methylation or histone modification in the model, we identified a smaller subset of association sets for which an effect on gene expression was fully explained by the epigenetic features: 3% for DNA methylation and 6% for histone modification. Thus, a complementary (nonredundant) combination of DNA methylation and histone acetylation seems to be required to capture the mediation effect, and adding other nonredundant epigenetic features would likely further enhance detection of this type of functional propagation.

Enrichment of disease susceptibility SNPs among xQTL SNPs

Studies have shown that SNPs associated with eQTLs are more likely to influence complex traits and disease susceptibility[6,10]. Here we provide further support for this observation for eQTLs, mQTLs and haQTLs by performing an enrichment analysis on reported *P*-values of 16 GWAS data sets, including large-scale GWAS meta-analyses of Alzheimer's disease[36], schizophrenia[37], and type II diabetes[38] (Online Methods). Enrichment was assessed using stratified linkage disequilibrium score regression (LDSR)[39]. For all 12 GWAS studies (out of 16) with a minimum of 20,000 samples tested (**Fig. 5a** and **Supplementary Table 6**), we observed significant enrichment for the xQTL SNPs. We also repeated this analysis using a more stringent background model, wherein we considered enrichment of our xQTLs against a background set of SNPs falling in "generic" annotation categories as provided in the LDSR software[39]. Again, significant enrichment, albeit with lower effect size, was observed for many of the GWAS studies (**Fig. 5a** and **Supplementary Table 6**). Next we hypothesized that SNPs shared between xQTL types, which affect multiple molecular phenotypes, are more likely to affect downstream processes and could constitute a list of 'high confidence' functional SNPs. We therefore compared all xQTL SNPs shared across at least two molecular traits against those xQTLs only found for one molecular trait. We indeed observed enrichment for the shared xQTLs, but their enrichment was not always higher than the background xQTL SNPs—that is, the level of enrichment was somewhat trait dependent (**Supplementary Table 6**). To test the robustness of the results to window size, we repeated the analysis with 100-kb windows for all three xQTL types (**Supplementary Table 7**). The overall trend remained the same, with slightly higher enrichment observed.

The enrichment results are reassuring, and, as we describe later, we can use our list of xQTL SNPs to enhance susceptibility locus discovery in GWAS studies. Investigators can also confidently use our xQTL lists to annotate GWAS SNPs related to the brain or nervous system, which will accelerate the transition to functional studies. For example, we used our eQTLs to map the 21 SNPs (and correlated SNPs in linkage disequilibrium with $r^2 > 0.8$) reported in the IGAP Alzheimer's disease GWAS and identified four candidate Alzheimer's disease–associated genes that are absent from the reported gene list defined by proximity[36] (*MADD*, *MTCH2*, *PILRA* and *POLR2E*). The TSS of these eQTL-mapped genes were >100 kb, on average, from their respective Alzheimer's disease–associated SNPs. *MTCH2*, *PILRA* and *POLR2E* have also been found in recent eQTL mapping studies[40], demonstrating the robustness of our results. *MADD* has not been previously reported in this context but is a good candidate given that its expression correlates with neuronal cell death in Alzheimer's disease[41] and that it has also been reported to modulate Alzheimer's disease–related tau toxicity in a *Drosophila* model[42].

Accelerating the transition to functional studies in specific cell types

Selection of the relevant cell type to target for *in vitro* or *in vivo* follow-up functional studies is challenging because our xQTLs, like those identified in many other studies, rely on tissue profiles generated from a complex mixture of cell types. To help prioritize cell types for such follow-up efforts, we repeated the analyses relating each SNP to a given molecular feature but also included a variable that estimates the proportion of a cell type in the profiled tissue and an interaction term to identify those SNPs whose effects depend on the proportion of a target cell (Online Methods). This approach was recently validated using whole-blood data[43].

Using eQTL results as an example ($n = 494$, two-tailed), we examined the potential specificity of each lead eQTL SNP for five cell types that are abundant in the cortex: neurons, microglia, astrocytes, oligodendrocytes and endothelial cells. We found that assignment to a single cell type was ambiguous for most eQTLs (*P*-values available

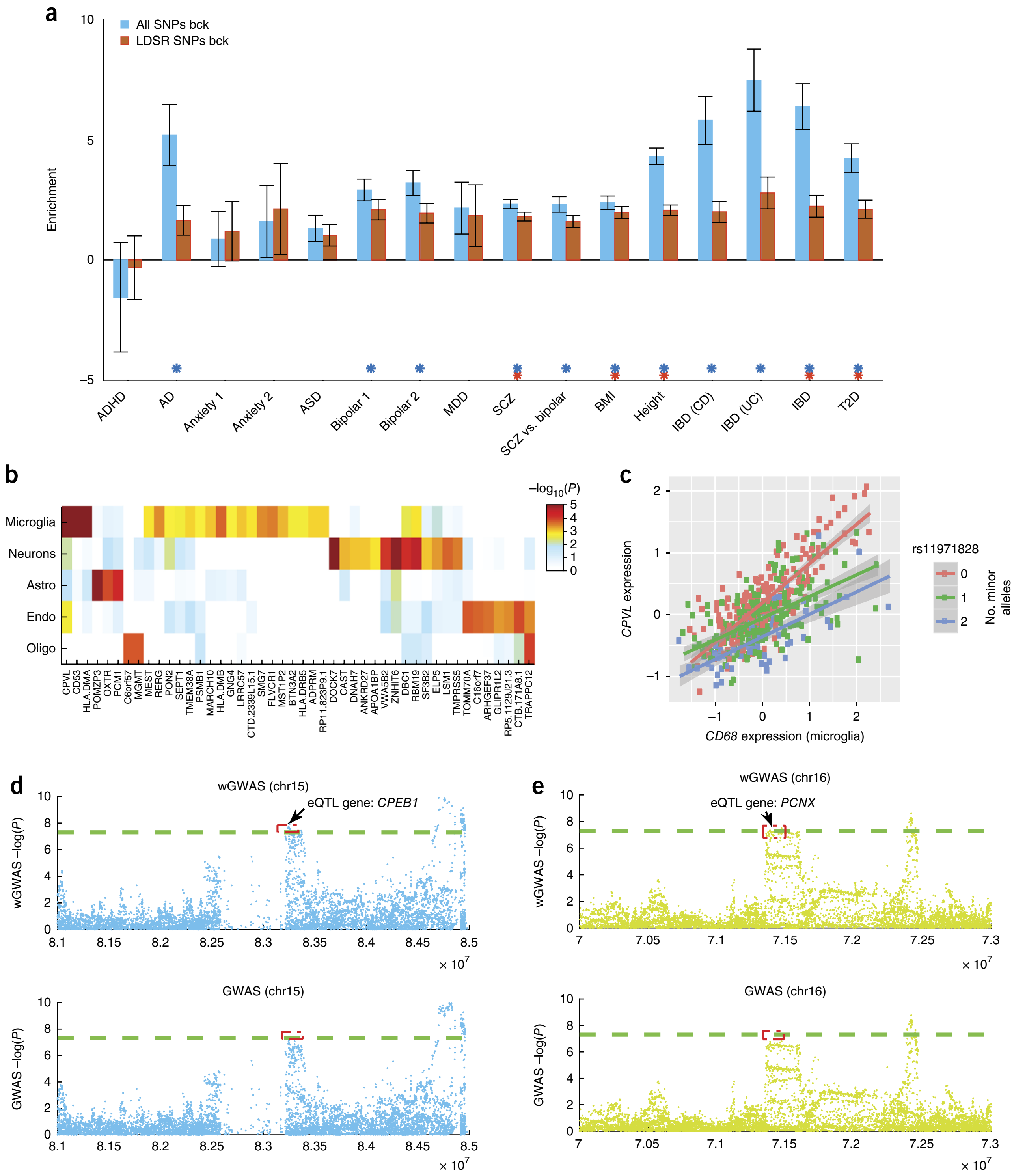

Figure 5 Application of the xQTL Resource for translational studies. (**a**) Enrichment of xQTL SNPs in published GWAS data sets as based on the LDSR model[39]. Enrichments are with respect to two sets of background SNPs: (i) all genome-wide SNPs and (ii) SNPs falling in generic functional sites previously defined by LDSR. Error bars reflect s.d. ADHD, attention deficit–hyperactivity disorder; AD, Alzheimer's disease; ASD, autism spectrum disorder; MDD, major depressive disorder; SCZ, schizophrenia; BMI, body mass index; IBD, inflammatory bowel disease; CD, Crohn's disease; UC, ulcerative colitis; T2D, type 2 diabetes; bck, background. (**b**) $-\log_{10}$ *P*-value of interaction test in quantifying cell specificity. Image shows 46 genes that survived FDR correction at a *q*-threshold of 0.2. Astro, astrocytes; Endo, endothelial cells; Oligo, oligodendrocytes. (**c**) Level of *CPVL* expression vs. a marker of microglial proportion (*CD68* gene). The gray shaded ares indicate 95% confidence intervals. *CPVL* expression increases with increasing proportion of microglia, particularly among major allele homozygotes (pink dots). (**d**,**e**) Zoomed-in Manhattan plot around the *PCNX* (**d**) and *CPEB1* (**e**) loci, showing the results of the published standard GWAS and our weighted GWAS (wGWAS). Each dot is one SNP. The dotted green line is the standard genome-wide significance threshold ($P < 5 \times 10^{-8}$).

at http://mostafavilab.stat.ubc.ca/xQTLServe). With a more lenient discovery strategy in which we thresholded the interaction term at FDR < 0.2, we found putative cell-type-specific effects in neurons (n = 13) and microglia (n = 22; **Fig. 5b**). In fact, in a minority of cases, our analysis returned an unambiguous cell type for the lead eQTL. For example, at FDR < 0.05, we identified six significant cell-specific eQTLs (one astrocytic, three microglial and one neuronal). An example is presented in **Figure 5c**. The *CPVL* locus harbors an eQTL effect (rs11971828) that is stronger in microglial cells. Even though only a small number of cell-specific eQTLs were identified with multiple-testing correction, our results can still be useful in prioritizing cell types for follow-up experiments based on the observation that suggestive cell-type-specific eQTL genes show clear cell type preferences. Many of these top cell-specific eQTL genes tend to conform to the expected function of the implicated cell. For example, the *MGMT* locus harbors an eQTL that ranks among the top three for oligodendrocyte specificity ($P = 1.5 \times 10^{-4}$). *MGMT* is known to function in oligodendrocytes, and its mutations are associated with oligodendrogliomas. These cell-specific results are intriguing but require molecular validation using purified cell populations from the cortex with matched genotypes to be confirmed.

xQTL-weighted GWAS for gene discovery efforts

Our large compendium of brain xQTLs can also be leveraged to accelerate gene discovery by boosting statistical power in GWAS. The simplest way of using our xQTL SNP list would be to restrict association analysis to our xQTL SNPs. However, such a strategy would miss other relevant SNPs that are not in our list (or were not tested in the *cis* xQTL analysis). Thus, we opted to use a weighted Bonferroni procedure[44], which permits all SNPs to be analyzed but weights their P-values by their potential phenotypic relevance. We refer to this approach as an "xQTL-weighted GWAS." Provided that the weights are non-negative and average to 1, strong control on family-wise error rate is guaranteed[44]. We employed a binary weighting scheme, in which P-values of xQTL SNPs were divided by w_1 and all other SNPs were divided by w_0 with $s = w_1/w_0 > 1$ (see Online Methods for s selection). Consistent with the standard GWAS convention, significance was declared at $P < 5 \times 10^{-8}$. To not over-count the number of significant hits due to correlations between SNPs, we applied PLINK[45] 1.9 on the 1000 Genomes phase 1 data[19] to remove SNPs among the significant hits that are in linkage disequilibrium with one another ($r^2 < 0.2$).

We compared five approaches: (i) no weighting, (ii) weighting xQTL SNPs found for any of the molecular phenotypes, (iii) weighting SNPs within predefined windows from the molecular features (1 Mb, 5 kb and 1 Mb for eQTL, mQTL and haQTL analyses, respectively) to account for distance bias, (iv) weighting generic functional SNP in the LDSR baseline model[39], and (v) weighting xQTL SNPs that are shared across any of the molecular phenotypes. Over the 19 GWAS data sets (Online Methods), weighting xQTL SNPs resulted in an equal or greater number of GWAS hits than no weighting, except for inflammatory bowel disease (**Supplementary Table 8**). For 8 of the 19 studies, the xQTL-weighted GWAS approach found at least two new independent loci (**Supplementary Table 8**). By contrast, weighting SNPs within predefined windows from the molecular features, as well as weighting SNPs in the LDSR baseline model, resulted in little change in detection sensitivity. Of note, the gain in sensitivity was not always the highest when we weighted the shared xQTL SNPs. Also, compared to weighting the DGN eQTL SNPs, weighting the union of all xQTL SNPs found in this study identified more additional independent susceptibility SNPs for a majority of the tested GWAS data sets, which demonstrates that additional signals are captured by mQTL and haQTL SNPs. In particular, weighting the xQTL SNPs found 22, 18 and 9 additional independent SNPs for schizophrenia, height and inflammatory bowel disease, respectively, compared to no weighting. In contrast, weighting the DGN eQTL SNPs found only 9, 3 and 2 additional independent SNPs. In fact, weighting just the eQTL SNPs in our data set identified 17 additional independent SNPs for schizophrenia, which illustrates the presence of eQTLs in our data that are enriched in brain diseases and not observed in blood.

Among the brain diseases that we examined, the largest detection gain was obtained with the schizophrenia data set[37], where 18 additional loci met genome-wide significance (excluding those near the MHC region) and were not in linkage disequilibrium ($r^2 < 0.2$) with the reported susceptibility SNPs[37]. Seven of these 18 SNPs were found to be associated with eQTLs (**Supplementary Table 8**), including rs57709857, which influences *LSM1*, a gene previously found in a Han Chinese schizophrenia study[46]. However, the *LSM1* locus had not reach genome-wide significance in individuals of European ancestry[47]. The list of eQTL genes also includes *PCNX* (associated with rs2189806), a gene encoding a member of the Notch signaling pathway that was reported to harbor a *de novo* copy number variant linked to autism spectrum disorder[48], and *CPEB1* (associated with rs1864699), which was recently implicated in experience-dependent neuronal development and circuit formation[49] (**Fig. 5d**,**e**). Thus, several of our new schizophrenia loci have some face validity, but further replication efforts are required to ensure that these are robust findings. In terms of the percentage increase in detection sensitivity, the largest gain was observed for bipolar disorder[50], where the standard GWAS approach identified one significant hit whereas the xQTL-weighted GWAS identified two additional independent loci.

DISCUSSION

Using one of the largest multi-omic data sets for brain tissue, we generated a list of xQTLs as a Resource for the neuroscience community to further investigate the interplay between the genome, epigenome and transcriptome in disease susceptibility. Our list of xQTLs replicates well in both brain and blood data sets, but it also contains xQTLs that appear unique to the older brain. Notable biological insights drawn from this Resource include significant sharing of xQTL SNPs across the measured molecular phenotypes. Also, the effects of some eQTL SNPs are fully mediated by our two epigenetic features, but further work and more data are needed to comprehensively assess the extent to which epigenomic features mediate eQTL effects. Overall, we create a large new reference with which investigators can functionally annotate their results; enhance their analyses, as illustrated by our xQTL-weighted GWAS approach; and guide functional studies, as in our cell type analysis. This Resource can be easily accessed through our portal, xQTL Serve.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the participants of ROS and MAP for their essential contributions and gift to these projects. This work has been supported by National Institute of Health (NIH) grants P330AG10161, U01 AG046152, R01AG16042, R01 AG036836, R01 AG015819, R01 AG017917 and R01 AG036547. The U01 AG046152 grant (to P.L.D.J. and D.A.B.) is a component of the AMP-AD Target Discovery and

Preclinical Validation Consortium, a program of the National Institute of Aging and the Foundation of the NIH.

AUTHOR CONTRIBUTIONS

Study design: S.M., B.N., P.L.D.J. Sample collection: D.A.B. Data generation and quality control analyses: B.N., C.M., H.-U.K., E.P., J.X., S.K.S., S.M., P.L.D.J. Analyses: B.N., C.C.W., C.G., S.M. Interpretation of results and critical review of the manuscript: B.N., C.M., H.-U.K., E.P., J.X., C.G., S.K.S., L.Y., D.A.B., S.M., P.L.D.J.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M.B. Annotating non-coding regions of the genome. *Nat. Rev. Genet.* **11**, 559–571 (2010).
3. Goldstein, D.B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
4. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
5. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
6. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
7. Gibbs, J.R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).
8. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* **19**, 48–54 (2016).
9. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
10. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
11. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
12. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J.D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
13. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
14. Bennett, D.A., Schneider, J.A., Arvanitakis, Z. & Wilson, R.S. Overview and findings from the religious orders study. *Curr. Alzheimer Res.* **9**, 628–645 (2012).
15. Bennett, D.A. *et al.* Overview and findings from the Rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646–663 (2012).
16. De Jager, P.L. *et al.* A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiol. Aging* **33**, 1017 (2012).
17. De Jager, P.L. *et al.* Alzheimer's disease: early alterations in brain DNA methylation at *ANK1*, *BIN1*, *RHBDF2* and other loci. *Nat. Neurosci.* **17**, 1156–1163 (2014).
18. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
19. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
20. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLOS Comput. Biol.* **6**, e1000770 (2010).
21. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
22. Do, C. *et al.* Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. *Am. J. Hum. Genet.* **98**, 934–955 (2016).
23. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
24. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **17**, 1418–1428 (2014).
25. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
26. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
27. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
28. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
29. Walsh, J.G., Muruve, D.A. & Power, C. Inflammasomes in the CNS. *Nat. Rev. Neurosci.* **15**, 84–97 (2014).
30. Pontillo, A., Catamo, E., Arosio, B., Mari, D. & Crovella, S. NALP1/NLRP1 genetic variants are associated with Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* **26**, 277–281 (2012).
31. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
32. Gaffney, D.J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012).
33. Johnson, A.D. *et al.* Polymorphisms affecting gene transcription and mRNA processing in pharmacogenetic candidate genes: detection through allelic expression imbalance in human target tissues. *Pharmacogenet. Genomics* **18**, 781–791 (2008).
34. Nishida, H. *et al.* Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. *Chromosome Res.* **14**, 203–211 (2006).
35. Millstein, J., Zhang, B., Zhu, J. & Schadt, E.E. Disentangling molecular relationships with a causal inference test. *BMC Genet.* **10**, 23 (2009).
36. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
37. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
38. Gaulton, K.J. *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
39. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
40. Karch, C.M., Ezerskiy, L.A., Bertelsen, S. & Goate, A.M. Alzheimer's disease risk polymorphisms regulate gene expression in the *ZCWPW1* and the *CELF1* loci. *PLoS One* **11**, e0148717 (2016).
41. Del Villar, K. & Miller, C.A. Down-regulation of DENN/MADD, a TNF receptor binding protein, correlates with neuronal cell death in Alzheimer's disease brain and hippocampal neurons. *Proc. Natl. Acad. Sci. USA* **101**, 4210–4215 (2004).
42. Dourlen, P. *et al.* Functional screening of Alzheimer risk loci identifies *PTK2B* as an *in vivo* modulator and early marker of Tau pathology. *Mol. Psychiatry* **22**, 874–883 (2017).
43. Westra, H.J. *et al.* Cell specific eQTL analysis without sorting cells. *PLoS Genet.* **11**, e1005223 (2015).
44. Roeder, K., Devlin, B. & Wasserman, L. Improving power in genome-wide association studies: weights tip the scale. *Genet. Epidemiol.* **31**, 741–747 (2007).
45. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
46. Shi, Y. *et al.* Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat. Genet.* **43**, 1224–1227 (2011).
47. Huang, L., Hu, F., Zeng, X., Gan, L. & Luo, X.J. Further evidence for the association between the LSM1 gene and schizophrenia. *Schizophr. Res.* **150**, 588–589 (2013).
48. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
49. Bestman, J.E. & Cline, H.T. The RNA binding protein CPEB regulates dendrite morphogenesis and neuronal circuit assembly *in vivo*. *Proc. Natl. Acad. Sci. USA* **105**, 20494–20499 (2008).
50. Ruderfer, D.M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol. Psychiatry* **19**, 1017–1024 (2014).

ONLINE METHODS

Experimental design and statistical analyses. Details on experimental design and reagents is described below and in the **Life Sciences Reporting Summary**. No power calculations were performed, but our sample sizes are similar to or larger than those of other brain QTL studies for RNA expression[23,24], DNA methylation[8] and ChIP-seq data[51]. Appropriate statistical tests have been specifically chosen for various analyses described below. For instance, we used Spearman's correlation for the xQTL analyses, which does not assume normality and equal variance. For assessing significance of the replication rate, we used the permutation test, which also does not assume normality. For testing genomic enrichment, which involves SNP counts as input, we used log odds. For testing mediation effects, we used the causal inference test, which take *P*-values from a set of associations. For testing trait and disease enrichment, we used LDSR and weighted GWAS, which take *P*-values as input.

Data acquisition, quality control, and normalization for known technical factors. The ROSMAP study is a longitudinal study in which all participants were healthy at enrollment. The sampled subjects thus represent a relatively random set of older individuals. By the time of death, 58% and 38% of participants were diagnosed with pathological and clinical Alzheimer's disease, respectively (**Supplementary Table 1**). These percentages are consistent with the Alzheimer's disease population prevalence. No randomization in subject selection was performed. For each -omic data type, data generation was attempted on all subjects with available frozen brain samples. The characteristics of the subjects in our analyses are similar to those of the overall ROS and MAP cohorts. A single person performed all of the dissections of the frozen tissues in isolating the gray matter for gene expression, DNA methylation and histone modification data generation to minimize technical variability in sample preparation. The individuals involved in collecting the autopsy samples and processing them during data generation were blinded to the phenotypic characteristics of the subjects.

Genotype data[16]. Genotyping of the ROS and MAP subjects was performed on the Affymetrix Genome-Wide HumanSNP Array6.0 ($n = 1,709$) and the Illumina OmniQuad Express platform ($n = 384$). DNA was extracted from whole blood, lymphocytes or frozen brain tissue as previously described[16]. To minimize population admixture, only self-declared non-Hispanic Caucasians were genotyped. At the sample level, samples with genotyping success rate < 95%, discordant genetically inferred and reported gender, or excess inter/intra-heterozygosity were excluded. At the probe level, genotyping data from both platforms were processed with same quality control (QC) metrics: Hardy-Weinberg equilibrium $P > 0.001$, genotype call rate < 0.95, mishap test $< 1 \times 10^{-9}$. QC was performed using version 1.08 of the PLINK software[52]. EIGENSTRAT[53] was used with the default setting to remove population outliers and to generate a genotype covariance matrix. The resultant data sets include 729,463 SNPs for 1,709 individuals (Affymetrix) and 624,668 SNPs for 384 individuals (OmniQuad). Dosages for all SNPs (>35 million) on the 1000 Genomes reference were imputed using version 3.3.2 version of the BEAGLE software[18] (1000 Genomes Project Consortium interim phase I haplotypes, 2011 phase 1b data freeze). Imputed SNPs were filtered based on minor allele frequency (MAF) > 0.01 and imputation INFO score > 0.3, resulting in 7,321,515 SNPs available for analysis.

Gene expression data[54]. Gene expression data were generated using RNA-seq from dorsolateral prefrontal cortex (DLPFC) of 540 individuals, at an average sequence depth of 90 million reads. Detailed description of data generation and processing will be described (S.M., C.G. *et al.*, unpublished data) and is summarized here.

Samples were submitted to the Broad Institute's Genomics Platform for transcriptome analysis following the dUTP protocol with poly(A) selection[55]. All samples were chosen to pass two initial quality filters: RNA integrity (RIN) score > 5 and quantity threshold of 5 µg. They were selected from a larger set of 724 samples. Sequencing was performed on the Illumina HiSeq with 101-bp paired-end reads and achieved coverage of 150 million reads of the first 12 samples. These 12 samples served as a deep coverage reference and included 2 males and 2 females each of unimpaired, mildly cognitively impaired and Alzheimer's disease status (**Supplementary Fig. 1**). The remaining samples were sequenced with a target coverage of 50 million reads. The mean coverage for the samples passing QC was 95 million reads (median 90 million reads) (**Supplementary Fig. 1**). The libraries were constructed and pooled according to the RIN scores such that similar RIN scores were pooled together (**Supplementary Fig. 1**). Varying RIN scores results in a larger spread of insert sizes during library construction and leads to uneven coverage distribution throughout the pool.

RNA-seq data were processed by our parallelized pipeline. This pipeline included trimming the beginning and ending bases from each read, identifying and trimming adaptor sequences from reads, detecting and removing rRNA reads, aligning reads to the reference genome using Bowtie[56] and quantifying transcript expression levels using RSEM[57]. Specifically, RNA-seq reads in FASTQ format were inspected using FASTQC program. Barcode and adaptor contamination and low quality regions (8 bp at the beginning and 7 bp at the end of each fastq read) were trimmed using the FASTX toolkit. To remove rRNA contamination, we aligned trimmed reads to rRNA reference (rRNA genes were downloaded from UCSC genome browser selecting the RepeatMask table) by BWA then extracted only paired unmapped reads for transcriptome alignment. rRNA-depleted reads were then mapped to the transcriptome reference (gencode v14) using the Trinity package with RSEM as the output option. Gene expression FPKM values were estimated using "rsem-calculate-expression" from RSEM.

Samples from 494 individuals were used in the eQTL analysis, which include those that had QCed genotype and passed the expression outlier test[6] ($D < 0.9$). To quantify the contribution of experimental and other confounding factors to the overall expression profiles, we performed a PCA on log-transformed FPKM values in all samples and computed the correlation between the top ten PCs and experimental factors (**Supplementary Fig. 2**). We observed significant correlations between many of these technical and confounding factors and top expression PCs (removal of which is described in the next section). With the log-transformed FPKM data, we used the COMBAT algorithm[58] to account for the effect of batch and linear regression to remove the effects of RIN, postmortem interval (PMI), sequencing depth, study index (ROS sample or MAP sample), genotyping PCs, age at death and sex. Finally, only highly expressed genes were kept (mean expression > $2\log_2$(FPKM)), resulting in 13,484 expressed genes for eQTL analysis. This FPKM-based threshold was determined through visual inspection of a histogram of mean expression values to approximately define two expression distributions: (i) no expression or very low expression and (ii) moderate to high expression.

DNA methylation data[17]. DNA methylation data were generated using the 450K Illumina array from DLPFC of 740 individuals. A detailed description of data acquisition and QC is previously published[17]. Briefly, methylation probes that coincided with common polymorphic sites were removed. Initial normalization of CpG probes to account for differences between type I and type II probes was performed using the BMIQ algorithm from the Watermelon package[59] and β-values were extracted for further analysis. The SNM approach[60] was then used to regress out the effects of batch, PMI, sex, age at death and a previously published estimate of proportion of neurons present in each sample[17]. In this study, samples from 468 individuals were analyzed, for whom gene expression data were also available. As described below, this decision was made to enable using gene expression data to estimate the proportions of the five main brain cell types. This correction for cell type proportions was done in addition to the regression approach for removing the effect of generic neuronal proportions based on DNAm marks[17].

Histone modification data[54]. Histone modification data were generated using H3K9Ac ChIP-seq from DLPFC of 714 individuals. Single-end reads were aligned by the BWA algorithm[61], and peaks were detected in each sample separately using the MACS2 algorithm[62] (using the broad peak option and a q cutoff of 0.001). A series of QC steps was employed to identify and remove low quality reads, and samples that did not reach (i) $\geq 15 \times 10^6$ unique reads, (ii) nonredundant fraction ≥ 0.3, (iii) cross- correlation ≥ 0.03, (iv) fraction of reads in peaks ≥ 0.05 and (v) ≥6,000 peaks were removed. Cross-correlation was defined as the maximum Pearson's correlation between the read coverage on the negative and positive strand after binning reads into 10-bp bins[63]. Cross-correlation was calculated after shifting the reads on the negative strand by s base pairs for $s = 0, 10, 20, \ldots, 1,000$, and the maximum cross-correlation was reported. In total, 669 samples passed quality control (**Supplementary Fig. 3**).

H3K9Ac domains were defined by calculating all genomic regions that were detected as a peak in at least 100 of the 669 samples (15%). Regions within 100 bp of each other were merged and very small regions of less than 100 bp were removed. Reads were then extended toward the 3′ end to the fragment size of the respective sample. The fragment size was estimated by the shift $s_{\max}$ that maximized the cross-correlation (mean $s_{\max} = 271$ bp). Finally, the number of extended reads in each H3K9Ac region was determined for each sample. Only uniquely mapped distinct reads were considered.

Quantified histone acetylation data were quantile normalized to account for variability in sequencing depth across individuals. Samples from 433 individuals for which gene expression data were available were used in our analysis.

Additional removal of known and hidden confounding factors. In addition to the data-specific QC and normalization described above, the effects of ancestry, cell type composition and 'hidden factors' were regressed out from the gene expression, DNA methylation and histone acetylation data. Variables representing ancestry were defined using the top three principal components of the genotype data. Cell type composition was estimated using gene expression levels of markers of main brain cell types: neurons (*ENO2*), oligodendrocytes (*OLIG2*, *MBP*, *CNP*), astrocytes (*GFAP*), microglia (*CD68*) and endothelial cells (*CD34*). Hidden confounding factors included top *N* PCs from the gene expression, DNA methylation and histone modification data (separately). PCA-based hidden factors typically capture variation in cell type proportions across individuals and other unmeasured confounding factors[20,64]. Following previous studies[26], for each molecular phenotype data, we varied *N* from 1 to 30 at a $\log_{10}$ scale and defined 'optimal' *N* as the value at which the number of significant hits in chromosome 18 saturated (**Supplementary Fig. 4**). We chose to assess performance on only chromosome 18 as opposed to all chromosomes to avoid overfitting. The optimal *N* was found to be approximately 10 for all three data types.

xQTL association analysis. Spearman's rank correlation was used to estimate the association strength between the alleles of each SNP and the three molecular phenotypes measured. For eQTL analysis ($n = 494$), we used SNPs that were up to 1 Mb upstream or downstream of the TSS of each gene. For mQTL analysis ($n = 468$), we used SNPs that were within 5 kb of each methylation site. For haQTL analysis ($n = 433$), we used SNPs that were within 1 Mb of each acetylation peak. The window sizes are informed by previous studies[21,22,65]. For each xQTL type, we declare an association as significant if its *P* was less than 0.05 after Bonferroni correction (two-tailed). Bonferroni threshold was determined separately for eQTL ($P < 8 \times 10^{-8}$), mQTL ($P < 5 \times 10^{-9}$) and haQTL ($P < 4 \times 10^{-10}$) analysis based on the number of tested associations.

Replication estimation with π_1 statistic. We performed replication analysis for eQTLs and mQTLs using previous brain-based studies[8,23,24,66], blood-based studies[26,27] and the GTEx study[28] to evaluate cross-sample replication and cross-tissue replication. Replication rates were estimated using the π_1 statistic[25], which provides an estimate of the proportion of xQTLs that are significant based on their *P*-value distribution. Only associations comprising the top SNP for each eQTL gene and each mQTL probe were included in the π_1 estimation, to avoid including many SNPs in linkage disequilibrium with each other in this analysis. For π_1 estimation, we used *P*-values from this study restricted to the eQTLs and mQTLs found in previous studies. That is, we used an existing reference eQTL or mQTL list and assessed the replication of those reported xQTLs in our data set. When possible, we also estimated π_1 in the other direction. Specifically, we assessed the replication rate of our eQTLs in a large DLPFC data set[23] and a large whole-blood data set[26]. To determine whether the replication rate was higher than chance level (one-tailed), we generated empirical null distributions by computing π_1 for 10^4 random *P*-value subsets of size *m*, where *m* is the number of eQTLs or mQTLs. Only *P*-values of associations that did not overlap with the eQTLs and mQTLs were used for null estimation.

Genomic annotations. To examine whether the xQTL SNPs are enriched in specific gene regions, we used genomic annotations from the ChromHMM resource[31], which comprise 15 categories: (1) active TSS (TssA), (2) flanking active TSS (TssAFlnk), (3) transcription at gene 5′ and 3′ (TxFlnk), (4) strong transcription (Tx), (5) weak transcription (TxWk), (6) genic enhancers (EnhG), (7) enhancers (Enh), (8) zinc finger genes and repeats (ZNF/Rpts), (9) heterochromatin (Het), (10) bivalent/poised TSS (TssBiv), (11) flanking bivalent TSS/enhancers (BivFlnk) (12) bivalent enhancers (EnhBiv), (13) repressed Polycomb (ReprPC), (14) weak repressed Polycomb (ReprPCWk) and (15) quiescent/low (Quies). We also used the knownGene table (GRCh37/hg19 assembly) provided on the UCSC genome browser website[67] to examine whether the xQTL SNPs are enriched in exons and introns. For each xQTL type, we computed the odds ratio of the xQTL SNPs being in each of the gene regions—that is, within predefined windows from the molecular features (1 Mb, 5 kb and 1 Mb for eQTL, mQTL and haQTL analyses, respectively)—versus all other tested SNPs. We further estimated the probability of observing an xQTL SNP at a certain distance from the TSS of its respective gene(s) by computing the number of xQTL SNPs at different distances away from TSS and dividing that by the number of tested SNPs to account for sampling biases.

Estimation of xQTL SNP sharing across molecular phenotypes. The π_1 statistic was employed to estimate the sharing of xQTL SNPs across molecular phenotypes. Using sharing between mQTLs and eQTLs as an example, with methylation and gene expression being the discovery and test phenotypes, respectively, we computed π_1 with *P*-values of the tested SNP–expression associations that consist of mQTL SNPs. This π_1 analysis provides an estimate of the proportion of SNP–expression associations that are significant when we restrict to associations comprising mQTL SNPs: that is, if most mQTL SNPs also drive gene expression, then the corresponding π_1 would be high. Since an mQTL SNP might be tested for association with expression levels of multiple genes, we had to decide which associations to include in the π_1 estimation. A lenient strategy would be to retain only the strongest association for each mQTL SNP, and a more stringent strategy would be to include all tested associations. With the lenient strategy, we estimated a cross-phenotype sharing (π_1) of ~0.83–0.97 for different pairs of phenotypes. With the more stringent strategy, we estimated a cross-phenotype sharing (π_1) of ~0.1–0.35. For the more stringent strategy, we note that as we decreased the allowable genomic distance between a discovery SNP and a tested feature, which by construction shrinks the coverage of xQTL SNPs, the cross-phenotype sharing increased (**Supplementary Fig. 5**).

The lenient strategy likely provides an over-estimate of π_1, since the retained associations were selected by their strength—that is, the smaller *P*-values kept. To tighten up our assessment of xQTL SNP sharing while not being overly stringent, we examined the distance between each discovery SNP and test feature, which we found to be a prime determinant of cross-phenotype sharing. For example, the strongest associated eQTL gene for each mQTL SNP is often the gene closest to the mQTL SNP. We also observed similar trends for other cross-phenotype comparisons (results not shown). Based on this observation, we modified our analysis to only consider the closest feature to each xQTL SNP.

Mediation analysis. We applied causal inference test (CIT)[35] to investigate whether the effect of a regulatory *cis* eQTL SNP is propagated through its impact on DNA methylation and/or histone modification (causal model), as well as whether the effect of an eQTL SNP on DNA methylation and/or histone modification is mediated through gene expression (reactive model). In brief, for the causal model, applying the CIT involves testing the following four associations: (i) an eQTL SNP is associated with the first PC of its associated histone acetylation peaks and methylation probes (that is, epigenome PC), (ii) this eQTL SNP is associated with expression of a gene, (iii) this eQTL SNP is associated with the epigenome PC conditioned on gene expression, and (iv) this eQTL SNP is independent of gene expression given epigenome PC. Testing the reactive model involves reversing the role of gene expression and epigenome PC. A *P*-value (two-tailed) was assigned to each set of associations using the intersection–union test[35]. Bonferroni correction was applied to account for the number of tested association sets *m*. We declared an association set as conforming to the causal model (or epigenetic mediation model) if pCausal < 0.05/*m* and pReact > 0.05/*m* and conforming to the reactive model (or transcription mediation model) if pCausal > 0.05/*m* and pReact < 0.05/*m*. An association set was declared as conforming to the independent model if pCausal > 0.05/*m* and pReact > 0.05/*m*. The remaining association sets were considered unclassified. The above analysis ($n = 411$) was performed on 20,916 association sets for the 10,897 xQTL SNPs that were associated with all three molecular phenotypes. We restricted analysis to these shared xQTL SNPs because only these SNPs would fulfill conditions (i) and (ii). The same analysis was also performed to assess the mediation of the shared xQTL SNPs through DNA methylation and histone acetylation separately. In this analysis, when multiple CpG probes (or acetylation peaks) were associated with a given xQTL SNP, we used their first PC to summarize their combination.

Disease enrichment analysis. We performed enrichment analysis on reported *P*-values of 16 GWAS data sets downloaded from the Psychiatric Genomics

Consortium website: https://www.med.unc.edu/pgc/results-and-downloads. Data from the following GWAS studies were used in the analysis:

- Attention deficit hyperactivity disorder (ADHD)[68]
- Alzheimer's disease (stage 1 data from IGAP)[36]
- Anxiety (case vs. control and factor score from ANGST)[69]
- Autism[70]
- Bipolar disorder[50,71]
- Major depressive disorder (MDD)[72]
- Schizophrenia[37,50]
- Body mass index (BMI)[73]
- Height[74]
- Crohn's disease[75]
- Ulcerative colitis[76]
- Inflammatory bowel disease[77]
- Diabetes[38]

Enrichment was assessed using stratified linkage disequilibrium score regression (LDSR) to estimate partitioned heritability[39]. For example, we labeled all xQTL SNPs as one category and SNPs in the LDSR baseline model as background. Significant enrichment was declared at an α of 0.05.

Cell type specificity analysis. We used a previous approach to estimate the cell specificity of an eQTL SNP, based on a statistical model that tests for an interaction effect between the SNP genotype and proportion of a cell type of interest[43]. Proportions of neurons, astrocytes, microglia, oligodendrocytes and endothelial cells were estimated with known cell type markers for these cells. Specifically, *ENO2* was used as the marker for neurons, *CD68* for microglia, *OLIG2* for oligodendrocytes, *GFAP* for astrocytes and *CD34* for endothelial cells. To reduce the number of tests, we only tested for cell specificity of the lead eQTL SNPs. That is, we tested (two-tailed) for cell specificity of lead SNPs that affected the expression levels of 3,388 genes with at least one significant eQTL SNP. In this analysis, we only corrected for known confounding factors, since regressing out the effect of hidden confounding factors would remove the effect of cell-specific expression[43].

xQTL-weighted GWAS. We used the weighted Bonferroni procedure[44] to prioritize xQTL SNPs in GWAS analysis. This procedure involves weighting *P*-values (or summary statistics) from a GWAS study by their potential relevance. Provided that the weights are non-negative and average to 1, strong control on family-wise error rate is guaranteed[44]. We used this approach with a simple binary weighting scheme on the 16 GWAS data sets listed in the previous section, as well as GWAS data sets pertaining to systolic and diastolic blood pressure[78] and BHRadj BMI, which were excluded from the LDSR enrichment analysis due to unavailability of some of the required summary statistics. Specifically, *P*-values of xQTL SNPs are weighted by w_1 and all other SNPs are weighted by w_0, where $w_1 = s/[1 + (s - 1)n_1/n]$ and $w_0 = 1/[1 + (s - 1)n_1/n]$ with $s = w_1/w_0$ ranging from 1 to 100. n_1 is the number of xQTL SNPs in our list and n is the number of SNPs in a study.

When only summary statistics—that is, GWAS *P*-values—are available, selecting the optimal s is nontrivial, since w_1 and w_0 do not depend on the *P*-values; that is, w_1 and w_0 depend only on s, the number of SNPs and the number of xQTL SNPs. Hence, we cannot 'train' w_1 and w_0 based on *P*-values. We thus instead proposed to divide the list of *P*-values into random half splits and use the following criterion: $J(s) = (D^1(s)/\pi_1^{\,1} + D^2(s)/\pi_1^{\,2}) \,/\, |D^1(s)/\pi_1^{\,1} - D^2(s)/\pi_1^{\,2}|$, where $D^i(s)$ is the number of SNPs in half split i with weighted $P < 5 \times 10^{-8}$ and $\pi_1^{\,i}$ is the estimated proportion of SNPs in half split i that are significant based on their unweighted *P*-values. The rationale behind the proposed criterion, $J(s)$, is twofold. First, if a given s is generalizable, then the detection rate should be similar for two half splits of randomly selected SNPs, as opposed to being large for one half but not the other. Second, among the s values that provide high reproducibility between splits, we should select the one that maximizes detection rate. We note that $|D^1(s) - D^2(s)|$ would not reflect reproducibility if the ground truth number of significant SNPs is different between the two splits. This complication is alleviated by dividing $D^i(s)$ by $\pi_1^{\,i}$. To determine the number of independent significant SNPs, we applied the PLINK[45] 1.9 pairwise linkage disequilibrium pruning function ($r^2 = 0.2$) on the 1000 Genomes phase 1 data[19].

Data availability. Data and samples from the ROSMAP study are available on the RADC Research Resource Sharing Hub at http://www.radc.rush.edu/ and https://www.synapse.org/#!Synapse:syn3219045. RNA-seq data are available at http://dx.doi.org/10.7303/syn3388564, ChIP-seq data at http://dx.doi.org/10.7303/syn4896408 and DNA methylation data at http://dx.doi.org/10.7303/syn3157275. The results of this study are available through the xQTL Serve website: http://mostafavilab.stat.ubc.ca/xQTLServe.

51. Shulha, H.P. *et al.* Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. *PLoS Biol.* **10**, e1001427 (2012).
52. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
53. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
54. Lim, A.S. *et al.* Diurnal and seasonal molecular rhythms in human neocortex and their relation to Alzheimer's disease. *Nat. Commun.* **8**, 14931 (2017).
55. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
56. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
57. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
58. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
59. Teschendorff, A.E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).
60. Mecham, B.H., Nelson, P.S. & Storey, J.D. Supervised normalization of microarrays. *Bioinformatics* **26**, 1308–1315 (2010).
61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
62. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
63. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
64. Mostafavi, S. *et al.* Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One* **8**, e68141 (2013).
65. Banovich, N.E. *et al.* Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* **10**, e1004663 (2014).
66. Sun, W. *et al.* Histone acetylome-wide association study of autism spectrum disorder. *Cell* **167**, 1385–1397 (2016).
67. Rosenbloom, K.R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).
68. Neale, B.M. *et al.* Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **49**, 884–897 (2010).
69. Otowa, T. *et al.* Meta-analysis of genome-wide association studies of anxiety disorders. *Mol. Psychiatry* **21**, 1391–1399 (2016).
70. Robinson, E.B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat. Genet.* **48**, 552–555 (2016).
71. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *ODZ4*. *Nat. Genet.* **43**, 977–983 (2011).
72. CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).
73. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
74. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
75. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
76. Anderson, C.A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
77. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
78. Ehret, G.B. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).

nature research

Corresponding author(s): Sara Mostafavi and Philip L. De Jager

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.	The sample size is adequate based on numerous previous papers that derive eQTLs and mQTLs. Indeed, using Bonferroni correction, our sample size yielded thousands of significant associations.

2. Data exclusions

Describe any data exclusions.	Except for analyses that require the presence of all three -omic datatypes, no samples were excluded in any of the analyses.

3. Replication

Describe whether the experimental findings were reliably reproduced.	The xQTLs replicated well in other published datasets.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.	NA

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.	NA

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	We have used both standard software packages (cited in the paper) and in-house scripts (availability of which is mentioned at the end of the Introduction).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	NA

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	NA

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	NA
b. Describe the method of cell line authentication used.	NA
c. Report whether the cell lines were tested for mycoplasma contamination.	NA
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	NA

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	NA

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Citation to previous study that describes the demographics of the subjects is provided (paragraph 2 of page 4). Number of subjects is clearly defined (paragraph 2 of page 4). Sex and age are treated as confounding factors and are regressed out before our analysis (Supplementary Information page 1-2).

nature research

Corresponding author(s): ____________________

□ Initial submission □ Revised version □ Final submission

ChIP-seq Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Data deposition

1. For all ChIP-seq data:

☒ a. Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ b. Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

2. Provide all necessary reviewer access links. *The entry may remain private before publication.*	https://www.synapse.org/#!Synapse:syn4896408
3. Provide a list of all files available in the database submission.	
4. If available, provide a link to an anonymized genome browser session (e.g. UCSC).	

Methodological details

5. Describe the experimental replicates.	n=714; data available for each individual in the study
6. Describe the sequencing depth for each experiment.	detailed provided on Supp Information: only samples with > 15x10^6 unique reads were utilized in the study
7. Describe the antibodies used for the ChIP-seq experiments.	NA - Described in previous publications
8. Describe the peak calling parameters.	Described in Supp Inf
9. Describe the methods used to ensure data quality.	Described in Supp Inf
10. Describe the software used to collect and analyze the ChIP-seq data.	Described in Supp Inf