nature genetics

# A survey of genetic human cortical gene expression

Amanda J Myers[1,2,10], J Raphael Gibbs[1,3,10], Jennifer A Webster[4,5,10], Kristen Rohrer[1], Alice Zhao[1], Lauren Marlowe[1], Mona Kaleem[1], Doris Leung[1], Leslie Bryden[1], Priti Nath[1], Victoria L Zismann[4,5], Keta Joshipura[4,5], Matthew J Huentelman[4,5], Diane Hu-Lince[4,5], Keith D Coon[4–6], David W Craig[4,5], John V Pearson[4,5], Peter Holmans[7], Christopher B Heward[8], Eric M Reiman[4,5,9], Dietrich Stephan[4,5,9] & John Hardy[1,3]



**It is widely assumed that genetic differences in gene expression underpin much of the difference among individuals and many of the quantitative traits of interest to geneticists. Despite this, there has been little work on genetic variability in human gene expression and almost none in the human brain, because tools for assessing this genetic variability have not been available. Now, with whole-genome SNP genotyping arrays and whole-transcriptome expression arrays, such experiments have become feasible. We have carried out whole-genome genotyping and expression analysis on a series of 193 neuropathologically normal human brain samples using the Affymetrix GeneChip Human Mapping 500K Array Set and Illumina HumanRefseq-8 Expression BeadChip platforms. Here we present data showing that 58% of the transcriptome is cortically expressed in at least 5% of our samples and that of these cortically expressed transcripts, 21% have expression profiles that correlate with their genotype. These genetic-expression effects should be useful in determining the underlying biology of associations with common diseases of the human brain and in guiding the analysis of the genomic regions involved in the control of normal gene expression.**

Large-scale assessments of the role of genetic variability in the control of gene expression have been attempted only recently. Two main approaches have been used: linkage-based analysis of gene expression in human lymphoblasts and multiple tissues from rat and mouse crosses, and association-based expression analyses in human lymphoblasts. These approaches have all shown that genetic variability is an important component in the regulation of gene expression[1–7].

Although these findings are encouraging, they are limited by the fact that the only human tissues that have been subject to extensive assay have been transformed lymphoblasts from individuals who did not receive any neurologic assessment. Very little has been done with other tissues because of their inaccessibility. However, it is well established that mRNA is stable postmortem in the human brain[8], and our and others' studies have shown that the apolipoprotein E (*APOE*) and microtubule-associated protein tau (*MAPT*) genes are subject to distortions in allelic expression[9–11]. Additionally, several studies using inbred mouse strains have mapped important expression quantitative trait loci (eQTL) in the mouse brain[1–3]. With this background, we developed a resource that allows the assessment of the genetic effects on normal human cortical gene expression. We isolated RNA and DNA from human cortical samples by standard protocols (see Methods) and carried out genotyping on the Affymetrix GeneChip Human Mapping 500K Array Set as previously described[12]. RNA expression was assessed using the Illumina HumanRefseq-8 Expression BeadChip system. We then treated the expression profile of each transcript as the sample phenotype and carried out a quantitative trait analysis on the genotype and expression data by linear regression to correlate allele dosage with expression. We analyzed samples for genetic relatedness and ethnic bias and outliers ($n = 3$ population outliers, $n = 5$ samples with some degree of relatedness, see **Supplementary Figs. 1** and **2** online and Methods) were excluded from our analysis. In addition, we corrected for several biological covariates (gender, age at death and cortical region) and several methodological covariates (day of expression hybridization, institute source of sample, postmortem interval and a covariate based on the total number of transcripts detected in each sample). See **Supplementary Table 1** online for covariate and sample statistics.

The Illumina HumanRefseq-8 chip probes 24,357 transcripts, of which we included 58% in our analyses because they were detected in at least 5% of our 193 samples. To avoid a possible bias introduced by poorer-quality samples, we added a methodological covariate based on a sample's detection for these transcripts. Additionally, SNPs with

[1]Laboratory of Neurogenetics, National Institute on Aging, Porter Neuroscience Building, National Institutes of Health Main Campus, Bethesda, Maryland 20892, USA. [2]Department of Psychiatry and Behavioral Sciences, Division of Neuroscience, Miller School of Medicine, University of Miami, Batchelors Children's Research Building, Room 609, 1580 10th Avenue, Miami, Florida 33136, USA. [3]Reta Lila Weston Institute and Departments of Molecular Neuroscience and Neurodegenerative Disease, Institute of Neurology, Queen Square, London WC1N 3BG, UK. [4]Neurogenomics Division, Translational Genomics Research Institute, Phoenix, Arizona 85004, USA. [5]Arizona Alzheimer's Consortium, Phoenix, Arizona 85006, USA. [6]Division of Thoracic Oncology Research, St. Joseph's Hospital and Medical Center, Phoenix, Arizona, 85013, USA. [7]Biostatistics and Bioinformatics Unit, Department of Psychological Medicine, Cardiff University, Cardiff, Wales CF14 4XN, UK. [8]Kronos Science Laboratory, Phoenix, Arizona 85016, USA. [9]Banner Alzheimer's Institute, Phoenix, Arizona 85006, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to A.J.M. (amyers@med.miami.edu).

Received 20 July; accepted 11 September; published online 4 November 2007; doi:10.1038/ng.2007.16

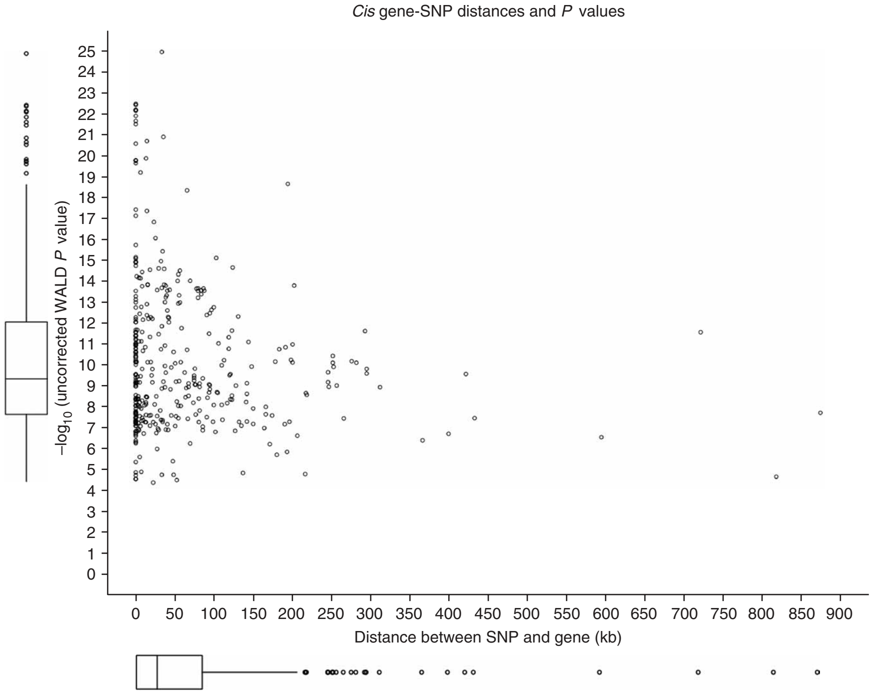


**Figure 1** Distance of *cis* effects. Only *cis* SNP-transcript pairs that were significant after correction for multiple testing, covariates and polymorphisms located in probes (see Methods) are plotted. See **Supplementary Table 2** for list of individual SNPs and transcripts. The *x* axis of the scatter plot is the distance between the SNP and the start or stop of the gene; for SNPs in the gene, the distance is given as zero. The *y* axis is the uncorrected and $-\log_{10}$ transformed WALD *P* values. Plot was created in R using the scatter plot function from the car package, which produces a scatter plot with box plots[19] included the axis margins. For each box plot, top bar is maximum observation, lower bar is minimum observation, top of box is upper or third quartile, bottom of box is lower or first quartile, middle bar is median value and circles are possible outliers.

call rates <90%, exact Hardy-Weinberg equilibrium *P* values <0.05 and minor allele frequencies <1% were filtered out during the analysis to eliminate noise from putative genotyping error. We assessed correlations among 366,140 SNPs on the Affymetrix platform and the expression of the 14,078 detected transcripts.

We divided our results into *cis* associations and *trans* associations. We defined *cis* associations as those that involved SNPs that were in the gene and within 1 Mb of either its 5′ or 3′ end. The mean and median of the sizes of these *cis* regions were both ~2.1 Mb. We defined *trans* associations as associations involving SNPs elsewhere in the genome. In this analysis, after using a permutation test correction (see Methods) and excluding results with a possible covariate effect, we found 433 SNP-transcript pairs (99 transcripts) that showed a significant (transcript-specific empirical *P* value ≤0.05) *cis* association and 16,701 SNP-transcript pairs (2,876 transcripts) that showed a significant *trans* association. Closer inspection of the positions of the *cis* SNP associations (**Fig. 1**) showed that most of these associations were near the gene, but in a few cases, effects were observable over long distances. Additionally, within ~70 kb of each associated transcript, there was an enrichment of *cis* associations over *trans* associations by ~3–3.5 fold (**Fig. 2**).

As we were looking at how DNA variation correlates with RNA expression, another possible confound was the presence of sequence variation within the transcript probe used on the Illumina expression chips. If such variation exists, the SNP may alter transcript binding in a way that is not biologically relevant. There was a polymorphism located within the transcript probe in 13% of our significant *cis* data and 5% of our significant *trans* data. **Table 1** shows a subset of our significant *cis* results from eight transcripts where there was no variation located within the transcript probe, where the transcript-specific empirical *P* values from 1,000 simulations were ≤0.05, the gene expression detection rate within the samples was ≥99%, the SNP

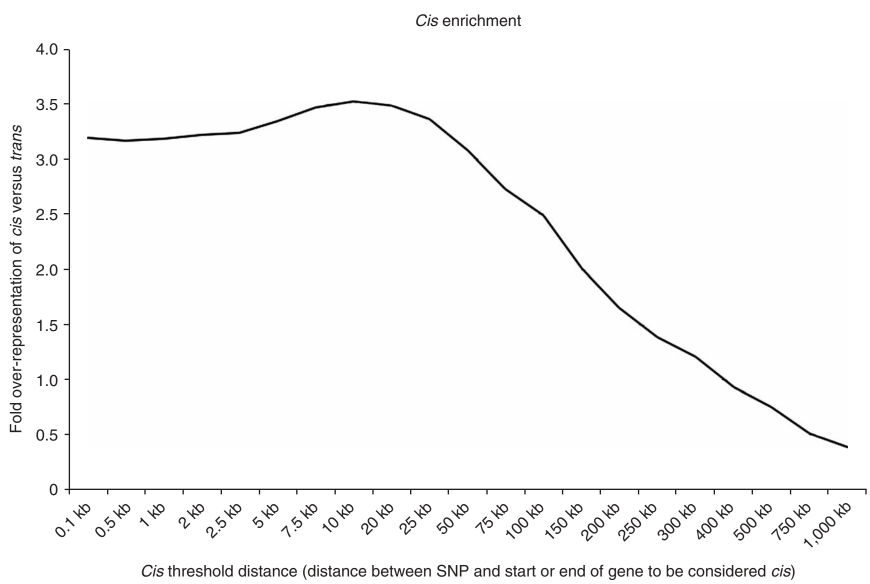


**Figure 2** Enrichment of *cis* associations over *trans* associations. Plotted is the distance between the SNP and the transcript (*x* axis) for *cis* SNPs against the average proportion of *cis* versus *trans* effects at that particular distance (*y* axis). Counts of *cis* versus *trans* effects were taken at 21 intervals from a distance of 0.1–1,000 kb from the transcript to the *cis* SNP (called the *cis* threshold distance, actual values used are marked on *x* axis). For each *cis* threshold distance, the number of identified *cis* SNP-transcript pairs was divided by the number of possible *cis* pairs for that distance. The same calculation was made for the *trans* pairs. The fold differences were then calculated by dividing the proportion of actual *cis* effects out of the total number of possible *cis* effects by the proportion of actual *trans* effects out of the total number of possible *trans* effects for a given distance from the transcript. As seen in the graph, there is an enrichment of *cis* effects at distances <1,000 kb from the gene, as expected, and the maximal overrepresentation of *cis* effects occurs at threshold distances <70 kb.

**Table 1 Subset listing of gene-SNP *cis* associated pairs**

| Gene | Ch | Start base | Stop base | SNP | SNP base | SNP loc | MAF | AA exp (s.d.) | AB exp (s.d.) | BB exp (s.d.) | pv1K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B3GTL | 13 | 30672131 | 30803656 | rs1005824 | 30714015 | Intron | 28% | 2.14 (0.17) | 2.02 (0.18) | 1.87 (0.23) | 0.001 |
| CHST7 | X | 46318135 | 46342781 | rs760697 | 46332287 | Intron | 45% | 2.37 (0.14) | 2.46 (0.14) | 2.56 (0.14) | <0.001 |
| HBS1L | 6 | 135323208 | 135417714 | rs1590975 | 135393780 | Intron | 49% | 2.17 (0.11) | 2.06 (0.12) | 1.97 (0.16) | <0.001 |
| HBS1L | 6 | 135323208 | 135417714 | rs2150681 | 135416924 | Intron | 49% | 2.17 (0.11) | 2.06 (0.12) | 1.97 (0.16) | <0.001 |
| HBS1L | 6 | 135323208 | 135417714 | rs4896128 | 135391448 | Intron | 35% | 2.13 (0.12) | 2.04 (0.13) | 1.92 (0.17) | 0.002 |
| HBS1L | 6 | 135323208 | 135417714 | rs6923765 | 135376868 | Intron | 49% | 2.17 (0.11) | 2.06 (0.12) | 1.97 (0.16) | <0.001 |
| HBS1L | 6 | 135323208 | 135417714 | rs7741515 | 135416060 | Intron | 49% | 2.17 (0.11) | 2.06 (0.12) | 1.97 (0.16) | 0.001 |
| KIF1B | 1 | 10193417 | 10364241 | rs10492972 | 10275698 | Intron | 33% | 2.39 (0.16) | 2.23 (0.18) | 1.83 (0.27) | <0.001 |
| KIF1B | 1 | 10193417 | 10364241 | rs12120042 | 10267911 | Intron | 35% | 2.40 (0.15) | 2.25 (0.18) | 1.86 (0.26) | <0.001 |
| KIF1B | 1 | 10193417 | 10364241 | rs12120191 | 10268358 | Intron | 35% | 2.40 (0.15) | 2.25 (0.18) | 1.86 (0.26) | <0.001 |
| KIF1B | 1 | 10193417 | 10364241 | rs1555849 | 10323188 | Intron | 33% | 2.39 (0.16) | 2.24 (0.18) | 1.85 (0.29) | <0.001 |
| KIF1B | 1 | 10193417 | 10364241 | rs3748577 | 10279992 | Intron | 33% | 2.39 (0.16) | 2.24 (0.18) | 1.83 (0.27) | <0.001 |
| KIF1B | 1 | 10193417 | 10364241 | rs3748578 | 10343504 | Intron | 31% | 2.36 (0.18) | 2.24 (0.20) | 1.88 (0.28) | <0.001 |
| KIF1B | 1 | 10193417 | 10364241 | rs946501 | 10232166 | Intron | 35% | 2.40 (0.15) | 2.25 (0.18) | 1.85 (0.27) | <0.001 |
| MAPT | 17 | 41327623 | 41461546 | rs17571739 | 41388780 | Intron | 23% | 2.28 (0.16) | 2.17 (0.17) | 2.03 (0.18) | 0.05 |
| PTD004 | 2 | 174645420 | 174821610 | rs10930638 | 174682841 | Intron | 45% | 1.92 (0.13) | 2.03 (0.14) | 2.14 (0.13) | <0.001 |
| PTD004 | 2 | 174645420 | 174821610 | rs10930654 | 174771758 | Intron | 48% | 2.12 (0.13) | 2.02 (0.14) | 1.91 (0.13) | <0.001 |
| PTD004 | 2 | 174645420 | 174821610 | rs11674895 | 174722208 | Intron | 49% | 2.12 (0.13) | 2.03 (0.13) | 1.91 (0.13) | <0.001 |
| PTD004 | 2 | 174645420 | 174821610 | rs4144329 | 174779123 | Intron | 48% | 2.12 (0.13) | 2.02 (0.14) | 1.91 (0.13) | <0.001 |
| PTD004 | 2 | 174645420 | 174821610 | rs4972643 | 174767946 | Intron | 49% | 2.12 (0.13) | 2.03 (0.14) | 1.91 (0.13) | <0.001 |
| PTD004 | 2 | 174645420 | 174821610 | rs6433464 | 174717017 | Intron | 48% | 2.12 (0.13) | 2.02 (0.14) | 1.91 (0.13) | <0.001 |
| SQSTM1 | 5 | 179180502 | 179197683 | rs10277 | 179197336 | Exon | 44% | 2.02 (0.23) | 1.93 (0.23) | 1.68 (0.22) | <0.001 |
| SQSTM1 | 5 | 179180502 | 179197683 | rs1065154 | 179197520 | Intron | 44% | 2.00 (0.21) | 1.94 (0.24) | 1.68 (0.22) | 0.006 |
| ZNF419 | 19 | 62690944 | 62697859 | rs2074074 | 62695684 | Intron | 28% | 2.38 (0.07) | 2.43 (0.06) | 2.48 (0.06) | <0.001 |
| ZNF419 | 19 | 62690944 | 62697859 | rs2360761 | 62699596 | 3′ | 28% | 2.38 (0.07) | 2.43 (0.06) | 2.47 (0.06) | <0.001 |
| ZNF419 | 19 | 62690944 | 62697859 | rs6510084 | 62694476 | Intron | 28% | 2.38 (0.07) | 2.43 (0.06) | 2.47 (0.06) | <0.001 |

Table shows chromosomal physical positions for the associated gene and SNP pairs (Ch, gene chromosomal location; start base, gene start; stop base, gene end; SNP base, SNP position; all relative to the published human sequence, build 36), the location of the SNP relative to the gene (SNP loc), the minor allele frequency for this SNP (MAF, based on these samples), genotype groups average expression with s.d. (AA, AB and BB Exp (s.d.)) and the empirical *P* values from 1,000 permutations (pv1K, see Methods). Subset listing was generated from the full list of *cis* associated gene-SNP pairs (full list is in **Supplementary Table 2**). Criteria for generation of subset include: no polymorphisms located within transcript probe, transcript specific empirical *P* value from 1,000 simulations $\leq 0.05$, gene expression detection rate within samples $\geq 99\%$, SNP call rate within portion of sample used $\geq 99\%$, number of minor homozygotes (BB genotype) $\geq 3$ and distance from SNP to gene $\leq 3$ kb. Genes are listed in alphabetical order.



call rates within the portion of the sample used were $\geq 99\%$, the number of minor homozygote samples was $\geq 3$ and the distance from the significant SNP to the gene was $\leq 3$ kb. **Supplementary Table 2** online lists all data for *cis* SNP-transcript pairs with transcript-specific, empirically significant *P* values in cases where there was no polymorphism located within the transcript probe and no covariate effects. **Supplementary Table 3** online lists all *cis* data for SNP-transcript pairs with transcript-specific, empirically significant *P* values where there was no covariate effect but there was a polymorphism located within the transcript probe. **Supplementary Table 4** online lists the subset of our *trans* pairs that met all the criteria we used to build **Table 1** for our *cis* results, with the exception that for this table, the SNP-transcript pair had to be mapped to distances greater than 1 Mb from the 5′ or 3′ end of the transcript and not within the transcript. Out of the 16,701 SNP-transcript pairs we found to be associated in *trans*, these 336 pairs (161 transcripts) represent our most probable *trans* results.

We assumed that the correlations we found between genotype and phenotype should be linear, such that expression would vary consistently with allele dosage; therefore, associations that follow the rules expAA > expAB > expBB or expBB > expAB > expAA, where A is the major allele at a locus and B is the minor allele at that locus, have *prima facie* biological plausibility. In contrast, cases where the heterozygote is either the highest or lowest expressor (for example, expAB > expAA > expBB, expAB > expBB > expAA, expAA > expBB > expAB or expBB > expAA > expAB) do not have *prima facie* biological plausibility. In our fully filtered (no covariate effects and no polymorphisms in probes) *cis* dataset, we found 1 instance out of 376 that did not follow these above rules (expAA > expAB > expBB or expBB > expAB > expAA). In addition, there were 10 instances where the heterozygote group had a higher expression level than the major homozygote group, but the expression level in the minor homozygote group was not reliably measured. For these 10 instances, because the expression level was unknown for the minor homozygote group, we could not determine whether the expression followed biologically plausible rules. In the *trans* dataset that contains our most likely effects (**Supplementary Table 4**), we found 6 SNP-transcript pairs with nonlinear expression-SNP correlations out of a total of 336 pairs. For all 336 pairs in the *trans* dataset, we had expression data for each possible genotype.

We have previously shown that within these samples, *MAPT* expression is affected by *MAPT* haplotype[11]. Analysis of our data from the genome-wide screen was consistent with our previous data on these samples: alleles that occurred on the major haplotype of *MAPT* (H1) were associated with higher Tau transcript expression (**Fig. 3**). This provides an internal positive control within our full-genome screen; by looking genome wide, we can find effects we have seen in candidate-gene analysis of our samples.

Comparing our screen to the previous eQTL screens carried out using human lymphoblasts[6,7] yielded few results in common. This was not surprising, considering the different sources and platforms for analysis. There were two results in common across the lymphoblast

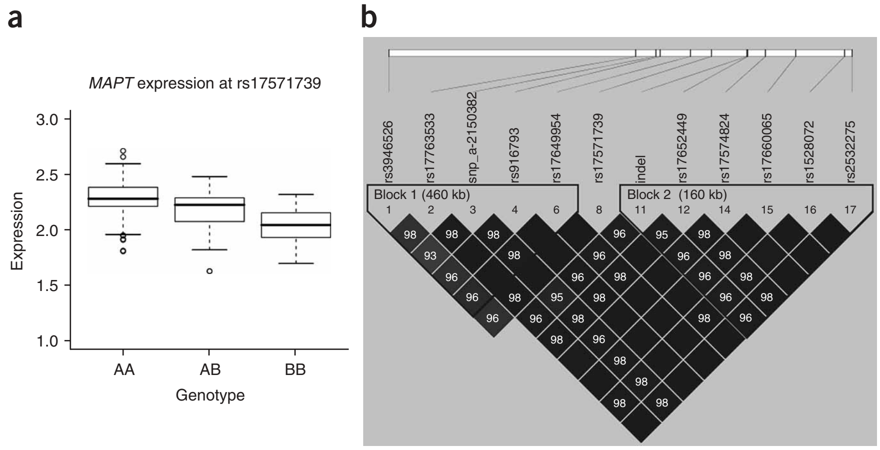


**Figure 3** *MAPT* result. (**a**) Box plot comparing the expression profiles of *MAPT* for the genotypes at rs17571739. rs17571739 was used because it was the most significantly associated SNP within 3 kb of *MAPT* (**Table 1**). The *x* axis represents the three genotype groups: AA (major homozygote), AB (heterozygote) and BB (minor homozygote). For this, SNP the major allele is A, and this allele falls on the previously defined high-expressing H1 *MAPT* haplotype[11]. Note that we could not detect subhaplotypes of H1 in our current screen. The genotype groups consist of the following numbers of samples: AA, $n = 113$; AB, $n = 66$ and BB, $n = 11$. The *y* axis is the expression level, which is the $\log_{10}$ value of the rank-invariant normalized intensity values. Plot was created in R using the box plot function from the graphics package, which produces a plot showing the five-number summaries for the three genotype groups in which the top bar is maximum observation, the lower bar is minimum observation, the top of box is upper or third quartile, the bottom of box is lower or first quartile, the middle bar is median value and the circles are possible outliers. (**b**) Linkage disequilibrium (LD) pattern of our dataset calculated using Haploview[20]. Black boxes with no numbers indicate $r^2 = 1$, which indicates that there is perfect LD and that each marker is a genetic surrogate for the other. For $r^2$ values $<1$, the $r^2$ value is given in white text in the box. As can be seen in the figure, all of the markers yielding empirically significant *cis P* values ($P < 0.05$) are in high LD with the major haplotypes of *MAPT* (H1 and H2), which can be delineated by a 238-base-pair insertion/deletion polymorphism in intron 9 of *MAPT* (indel on figure).

screens: eQTLs were found for transcripts encoding cystatin B (*CSTB*) and copine I (*CPNE1*) (refs 6,7). Within our screen, neither transcript was correlated with genotype. We found one transcript that was also reported to show a *cis* genotype-transcript association by Cheung and colleagues[6]. This transcript was *RPS26*, which encodes ribosomal protein S26, a ribosomal protein that is a component of the 40S subunit. This replication might reflect true results; however, further analysis will be needed to ensure that the same haplotypes are associated with expression in each screen. See **Figure 4** for the profiles of *RPS26* from our data. For our most likely *trans* associations (**Supplementary Table 4**), the only transcript that was in common with the other two screens was that encoding the binding protein of integrin β1 (*ITGB1BP1*), which was also found in the report by Cheung *et al.* (described in their series as ICAP-1A). However, each screen reported the *trans* effect with SNPs located on different chromosomes, indicating that each study was detecting a different *trans* effect from this transcript.

In this analysis we present associations between single SNPs and expression levels that suggest that genetic variability can contribute to the variability of transcript expression. The importance of these data is clear. First, when genetic associations are reported between SNPs and common neurologic or psychiatric diseases, one can use these data to predict relative mRNA expression levels at the locus, under the assumption that for common sporadic disease, the risk variants will be present in a considerable proportion of the control sample, for example, as with the H1 haplotype of *MAPT*. Second, they will provide the raw material for researchers to delineate the control of normal human cortical gene expression. Of course, it is likely that this analysis will underestimate the true contribution of SNP variation to gene expression, as the relationship between haplotypes at a locus and

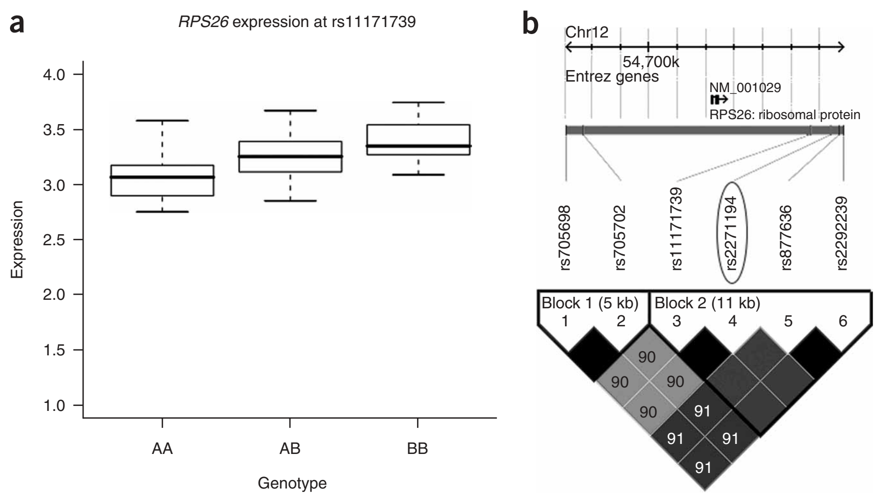


**Figure 4** *RPS26* result. Shown is the one transcript that we replicated from the previous genotype-expression screens[6,7]. (**a**,**b**) For RPS26, Cheung and colleagues[6] reported significant association with marker rs2271194, which is in complete LD with two out of our six associated SNPs, including rs11171739, which was the SNP that gave the strongest association in our screen and which is graphed in **a**. (**b**) LD plot using the CEPH HapMap data to compare the markers from the two studies for RPS26. The significant variant from Cheung *et al.* is circled. Because there is complete LD between our marker and the marker in Cheung's study, it is likely that both screens are picking up the same association for *RPS26*. Plots were created in R using the box plot function from the graphics package, which produces a plot showing the five-number summaries for the three genotype groups where top bar is maximum observation, lower bar is minimum observation, top of box is upper or third quartile, bottom of box is lower or first quartile, middle bar is median value and circles are possible outliers. Haplotype block plots were created using Haploview[20]. Black boxes with no numbers indicate $r^2 = 1$. For $r^2$ values $<1$, the $r^2$ value is given in white text in the box.

the gene expression is likely to be more complex. In addition, it is likely that different biological covariates and genomic duplications and deletions have important roles in the control of gene expression. Further analyses of these data and of data from other, similar datasets will be required to elucidate such complex interactions. To facilitate this analysis, we have made the data files used to generate the analysis for this paper available on our website (see Methods). Additional data and information is available through NCBIs Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE8919. Lastly, DNA from the samples used in this screen is available on request through the National Cell Repository for Alzheimer's Disease for fine mapping of particular effects.

## METHODS

**Samples.** We wrote to all the National Institute of Aging Alzheimer Centers and the Miami Brain Bank and asked for samples of 1 g of human cortex from control brain. We received 279 samples that met our criteria: first, they were self-defined as ethnically of European descent; second, they had no clinical history of stroke, cerebrovascular disease, Lewy bodies or co-morbidity with any other known neurological disease; third, they were assessed by a board-certified neurologist and, where available, they had a Braak and Braak score $<3$ (43% of controls used for this paper assessed) or a CERAD score indicating either sparse or no neuritic plaques (34% of controls used for this paper assessed); and fourth, they had an age at death $\geq 65$ years. 201 of those samples had both genotype and expression data, and 193 samples were used for analysis after excluding ethnic outliers and samples that were possibly related. Sample statistics are given in **Supplementary Table 1**.

**Genotyping and expression profiling.** 250 ng of DNA was hybridized to the Affymetrix GeneChip Human Mapping 500K Array Set as previously described[12]. Allele calls were determined using the Affymetrix BRLMM Analysis Tool. The resulting sample genotyping call rate had a mean of 97% and range of 90–99%.

250 ng of RNA was reverse transcribed into cRNA and biotin-UTP labeled using the Illumina TotalPrep RNA Amplification Kit (Ambion). We quantified cRNA by three replicate measurements using a nanodrop spectrophotometer. cRNA was hybridized to the Illumina HumanRefseq-8 Expression BeadChip using standard protocols (see URL in Methods for further details on chip design). We ran 6–8 chips (24–32 control samples) in parallel for each hybridization. Average detection scores across each expression chip were greater than 0.99. Transcripts that were detected in less than 5% of the series were excluded from our study. All expression profiles were extracted and rank-invariant normalized[13–15] using BeadStudio software (Illumina).

**Statistical analysis.** Before the analysis of the 366,140 SNPs and 14,078 gene transcripts, chromosome physical positions for each SNP and transcript were reannotated from NCBI's dbSNP and Entrez Gene based on Genome Build 36. We obtained information about the ethnic structure of our cohort using the program Structure[16,17] and removed ethnic outliers (**Supplementary Methods** online). After the three ethnic outliers were eliminated, we examined the degree of relatedness among the samples within our cohort by using the pairwise identity-by-state and identity-by-descent calculators available in the PLINK analysis toolset[18] and **Supplementary Methods**. Rank-invariant normalized expression data were $\log_{10}$ transformed, and missing data were encoded as missing, not as a zero level of expression. We excluded transcripts that were expressed in less than 5% of the series from the analysis. The following minimum SNP cut-off values were used during analysis: per sample call rate at least 90%, per SNP call rate at least 90%, per SNP minor allele frequency of at least 1%, and lack of significance ($P > 0.05$) for Hardy-Weinberg equilibrium tests. Categorical covariates were encoded and $\log_{10}$ transformed, again where missing values were indicated as such.

For our analysis, we used the PLINK analysis toolset (64-bit version) to carry out a one-degree-of-freedom allelic test of association. Briefly, the expression level of each transcript per sample was regressed on the number of minor alleles (0, 1 or 2) for the 366,140 SNPs that met the cut-off criteria to compute the effects of allele dosage on expression level. We analyzed transcripts one at a time and did not take into consideration interdependence among transcripts. Transcript-specific empirical $P$ values were calculated by permuting the sample identifiers (see multiple testing section below), and only those pairs with transcript-specific empirical $P$ values (1,000 permutations) $\leq 0.05$ were retained.

The analysis results were then separated into *cis* and *trans* significantly associated SNP-transcript pair sets. *Cis* SNPs were defined as SNPs within 1 Mb of the 5′ end of the transcript or 1 Mb of the 3′ end of the transcript and within the transcript. SNP-transcript associated pairs that had a methodological covariate (day of expression hybridization, institute source of sample, post-mortem interval and a covariate based on the total number of transcripts detected in each sample) or biological covariate (gender, age at death and cortical region) effect were then removed from the result set. For the assessment of covariate effects, we used a conservative approach in which any SNP-transcript pair covariate term with an uncorrected $P$ value $< 0.05$ was deemed to have an effect. To account for any potential confounding effect of SNPs located within the transcript hybridization probes on the Illumina ref-seq[8] chips, significant *cis* SNP-transcript effects were divided into pairs where there was no variant within the transcript probe (**Table 1** and **Supplementary Table 2**) and pairs where there was a variant in the transcript probe (**Supplementary Table 3**). Please see **Supplementary Methods** for further details. *Trans* SNP-transcript results were determined in the same fashion as the *cis* results but filtered to reduce the dataset from 16,701 SNP-transcript pairs to a more manageable 336 SNP-transcript pairs, which we believe are the most likely results within the larger dataset. The criteria for filtering the *trans* data were as follows: (i) no polymorphisms located within transcript probe, (ii) transcript-specific empirical $P$ value (1,000 permutations) $\leq 0.05$, (iii) gene expression detection rate within samples $\geq 99\%$, (iv) SNP call rate within portion of sample used $\geq 99\%$, (v) number of minor homozygotes (BB genotype) $\geq 3$ and (vi) distance between SNP and transcript greater than ±1 Mb of the gene.

This study used the high-performance computational capabilities of the Biowulf Linux cluster. We carried out permutation analysis on the Translational Genome Research Institute's IBM System Cluster 1350, which contains a total of 1,024 computing nodes and is housed on the Arizona State University campus.

**Statistical significance and corrections for multiple testing.** Multiple testing was corrected by simulation (**Supplementary Methods**). Uncorrected Wald $P$ values are given in the pvWALD columns in **Supplementary Tables 1–4**. All empirical $P$ values from 1,000 permutations are given in the pv1K columns on each table, and those from the 87 transcripts for which we carried out 100,000 replicates are shown on **Supplementary Table 2** in the pv100K column. Sidak multitranscript-corrected empirical $P$ values for the four transcripts where we carried out 1 million permutations and applied a Sidak correction for the effects of testing multiple transcripts are given in the pvSIDAK column on **Supplementary Table 2**.

**Data and biomaterial access.** The data files used to generate the analysis for this paper are available at http://labs.med.miami.edu/myers/. DNA from the samples used in this screen is available on request through the National Cell Repository for Alzheimer's Disease for fine mapping of particular effects http://ncrad.iu.edu. Details of Illumina chip design are available at http://www.illumina.com/pages.ilmn?ID=51.

**Accession codes.** National Center for Biotechnology Gene Expression Omnibus: Microarray data have been deposited with GEO accession code GSE8919.

*Note: Supplementary information is available on the Nature Genetics website.*

**ACKNOWLEDGMENTS**

We thank the individuals involved in this study and their families. Many data and biomaterials were collected from several sites funded by the National Institute on Aging (NIA) and National Alzheimer's Coordinating Center (NACC, grant #U01 AG016976). A.J. Myers (University of Miami, Department of Psychiatry) and J.A. Hardy (Reta Lila Weston Institute, University College London) collected and prepared the series. We thank M. Morrison-Bogorad, T. Phelps and W. Kukull for helping to coordinate this collection. The directors, pathologists and technicians involved include: R. Seemann (NIA); J.C. Troncoso and O. Pletnikova (Johns Hopkins Alzheimer's Disease Research Center, NIA grant AG05146); H. Vinters and J. Pomakian (University of California Los Angeles, NIA grant P50 AG16570);

C. Hulette and J.F. Ervin (The Kathleen Price Bryan Brain Bank, Duke University Medical Center, NIA grant AG05128, National Institute of Neurological Disorders and Stroke grant NS39764, National Institute of Mental Health MH60451, also funded by Glaxo Smith Kline); D. Horoupian and A. Salehi (Stanford University); J.P. Vonsattel and K. Mancevska (New York Brain Bank, Taub Institute, Columbia University); E.T. Hedley-Whyte and K. Fitch (Massachusetts General Hospital); R. Albin, L. Bain and E. Gombosi (University of Michigan, National Institutes of Health (NIH) grant P50-AG08671); W. Markesbery and S. Anderson (University of Kentucky, NIH grant AG05144); D.W. Dickson and N. Thomas (Mayo Clinic, Jacksonville); C.A. Miller, J. Tang and D. Diaz (University of Southern California); D. McKeel, J.C. Morris, E. Johnson Jr., V. Buckles and D. Carter (Washington University, St. Louis, Alzheimer's Disease Research Center (NIH grant P50AG05681); T. Montine and A. Schantz (University of Washington, Seattle, NIH grant P50 AG05136); J.Q. Trojanowski, V.M. Lee, V. Van Deerlin and T. Schuck (University of Pennsylvania School of Medicine, Alzheimer's Disease Research Center); A.C. McKee and C. Kubilus (Boston University Alzheimer's Disease Research Center, NIH grant P30-AG13846); J. Rogers, T.G. Beach and L.I. Sue (Sun Health Research Institute, Arizona, NIA grant P30 AG19610); B.H. Wainer and M. Gearing (Emory University); C.L. White III, R. Rosenberg, M. Howell and J. Reisch (University of Texas Southwestern Medical School); W. Ellis and M.A. Jarvis (University of California, Davis); D.A. Bennett, J.A. Schneider, K. Skish and W.T. Longman (Rush University Medical Center, Rush Alzheimer's Disease Center, NIH grant AG10161); and D.C. Mash, M.J. Basile and M. Tanaka (University of Miami/NPF Brain Endowment Bank). These studies were supported by Kronos Sciences Laboratory, the Verum Foundation, the Bisgrove charitable donation, the NIH Neuroscience Blueprint (U24NS051872), the ENDGAME Consortium (UO1HL084744) and the state of Arizona. We also thank E.B. Suh and J. Lowey at TGen for the use of the ASU-TGen cluster-based supercomputer. A.J.M. would like to thank the Johnnie B. Byrd Institute for support. None of the sponsors were involved in the design or conduct of the study, in the collection, analysis, and interpretation of the data, or in the preparation, review, or approval of the manuscript.

## AUTHOR CONTRIBUTIONS

A.J.M. conceived the experiment, supervised and performed the RNA screen and wrote the final manuscript. J.R.G. helped to refine the experiment and performed the final data analysis as well as helped to edit the manuscript. J.A.W. performed and supervised the DNA screen, carried out the permutation analysis and helped to edit the manuscript. K.R., A.Z., L.M., M.K., D.L., L.B. and P.N. helped to perform the RNA screen and APOE genotyping. V.L.Z., K.J., M.J.H., D.H.-L. and K.D.C. helped to perform the DNA screen. D.W.C. and J.V.P. performed initial analyses on the DNA screen. P.H. served as a statistical consultant for the final data analysis and helped edit the manuscript. C.B.H. helped to fund the study. E.M.R. and D.S. supervised the DNA portion of the screen as well as helped to fund the study. J.H. helped to conceive the experiment and wrote the first draft of the manuscript.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Hovatta, I. *et al.* DNA variation and brain-region specific expression profiles show different relationships between inbred mouse strains: implications for eQTL mapping studies. *Genome Biol.* **8**, R25 (2007).
2. Hovatta, I. *et al.* Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature* **438**, 662–666 (2005).
3. McClurg, P. *et al.* Genomewide association analysis in diverse inbred mice: power and population structure. *Genetics* **176**, 675–683 (2007).
4. Monks, S.A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
5. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
6. Cheung, V.G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
7. Stranger, B.E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).
8. Gilbert, J.M., Brown, B.A., Strocchi, P., Bird, E.D. & Marotta, C.A. The preparation of biologically active messenger RNA from human postmortem brain tissue. *J. Neurochem.* **36**, 976–984 (1981).
9. Lambert, J.C. *et al.* Distortion of allelic expression of apolipoprotein E in Alzheimer's disease. *Hum. Mol. Genet.* **6**, 2151–2154 (1997).
10. Bray, N.J. *et al.* Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes. *Hum. Mol. Genet.* **13**, 2885–2892 (2004).
11. Myers, A.J. *et al.* The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. *Neurobiol. Dis.* **25**, 561–570 (2007).
12. Coon, K.D. *et al.* A high density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J. Clin. Psychiatry* **68**, 613–618 (2007).
13. Workman, C. *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* **3**, research0048 (2002).
14. Schadt, E.E., Li, C., Ellis, B. & Wong, W.H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem. Suppl.* (Suppl. 37) 120–125 (2001).
15. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549–2557 (2001).
16. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
17. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
18. 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).
19. Tukey, J.W. *Exploratory Data Analysis* Section 2C (Addison-Wesley, Reading, Massachusetts, 1977).
20. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).