Tissue-Specific Genetic Control of Splicing: Implications for the Study of Complex Traits

Erin L. Heinzen[1]ᵊ, Dongliang Ge[1]ᵊ, Kenneth D. Cronin[1], Jessica M. Maia[1], Kevin V. Shianna[1], Willow N. Gabriel[1], Kathleen A. Welsh-Bohmer[2], Christine M. Hulette[2], Thomas N. Denny[3], David B. Goldstein[1]*

1 Institute for Genome Sciences & Policy, Center for Human Genome Variation, Duke University, Durham, North Carolina, United States of America, **2** Joseph and Kathleen Bryan Alzheimer's Disease Research Center, Duke University, Durham, North Carolina, United States of America, **3** Human Vaccine Institute, Duke University, Durham, North Carolina, United States of America

Numerous genome-wide screens for polymorphisms that influence gene expression have provided key insights into the genetic control of transcription. Despite this work, the relevance of specific polymorphisms to in vivo expression and splicing remains unclear. We carried out the first genome-wide screen, to our knowledge, for SNPs that associate with alternative splicing and gene expression in human primary cells, evaluating 93 autopsy-collected cortical brain tissue samples with no defined neuropsychiatric condition and 80 peripheral blood mononucleated cell samples collected from living healthy donors. We identified 23 high confidence associations with total expression and 80 with alternative splicing as reflected by expression levels of specific exons. Fewer than 50% of the implicated SNPs however show effects in both tissue types, reflecting strong evidence for distinct genetic control of splicing and expression in the two tissue types. The data generated here also suggest the possibility that splicing effects may be responsible for up to 13 out of 84 reported genome-wide significant associations with human traits. These results emphasize the importance of establishing a database of polymorphisms affecting splicing and expression in primary tissue types and suggest that splicing effects may be of more phenotypic significance than overall gene expression changes.

Citation: Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, et al. (2008) Tissue-specific genetic control of splicing: Implications for the study of complex traits. PLoS Biol 6(12): e1000001. doi:10.1371/journal.pbio.1000001

Introduction

The release of the HapMap data in 2003 and the availability of immortalized cell lines from HapMap participants initiated a new era of research investigating how SNPs affect how genes are expressed at the mRNA level. In 2005, two landmark publications evaluated how SNPs affect overall transcription in immortalized cell line samples collected from unrelated individuals [1,2]. Since those initial publications, the work has advanced with additional studies using more sophisticated microarrays, larger more diverse sample sets, and with studies of heritability of transcript and exon-level expression [3–6].

The work to date however has been limited in scope, largely focusing on the control of overall gene expression in immortalized cells, which may not be representative of in vivo patterns in specific cellular populations [7]. Only two genome-wide studies have focused on human primary cells [8,9], and most studies have considered only overall expression with no attempt to identify polymorphisms that have their effects primarily on alternative splicing.

Here we extend the previous body of work by studying the genetic control of both exon-level and whole-transcript level variation in expression in two primary cell types, including peripheral blood mononucleated cells (PBMCs) and cortical brain tissue from a set of control individuals, combined with parallel genome-wide genotyping of these samples. The implementation of identical genome-wide screens in two primary tissue types has allowed us to identify polymorphisms with clear effects on both overall expression and splicing, and to show that these effects are often tissue specific. We have also established an easy-to-use database that allows users to assess whether a given polymorphism is associated with any local changes in expression and have shown that these data suggest possible underlying causes of several published disease associations.

Results/Discussion

Exon-level microarrays were used to quantify expression levels of fully annotated coding sequences, EST-predicted exons, and bioinformatically predicted exons across the genome. These data allow direct inferences about expression levels of specific exons. By averaging sets of exons it is also possible to estimate expression levels for transcript species (Figure 1, top panel). While the exon-level expression data do not allow inference about the representation of specific (full)

Academic Editor: Edison Liu, Genome Institute of Singapore, Singapore

Received May 19, 2008; **Accepted** November 12, 2008; **Published** December 23, 2008

Abbreviations: eQTL, expression quantitative trait locus; LD, linkage disequilibrium; MAF, minor allele frequency; PBMC, peripheral blood mononucleated cell; sQTL, splicing quantitative trait locus

* To whom correspondence should be addressed. E-mail: d.goldstein@duke.edu

ᵊ These authors contributed equally to this work.

Author Summary

Although humans have a relatively small complement of genes, the proteins encoded by those genes and their biologic function are far more complex. The increased complexity is achieved in part through processes that create different messages from the same gene sequence (alternative splicing) and that regulate the expression of those messages in a tissue-specific fashion. These processes expand the functional capacity of the human genome, but also can create predisposition to disease when these processes go awry. In this study, we investigated how single nucleotide polymorphisms influence both overall gene expression and alternative splicing in two important cell types (brain and blood) highly relevant to human disease. Extensive and tissue-specific regulation of gene expression and alternative splicing were observed in the two tissue types, and some of these polymorphisms were shown to be connected to other polymorphsims that have been recently implicated in human diseases through genome-wide association studies. Most of these connections appeared to relate to alternative splicing as opposed to overall expression changes, suggesting that changes in splicing patterns may be more consequential for disease than those affecting only expression. These data emphasize the importance of comprehensive studies into genetic regulation of gene expression in all human tissue types in order to help understand how genetic variation influences risk of common diseases.

transcripts resulting from a given alternative splicing event, they do reflect splicing events in how they influence the proportion of transcripts with and without a given exon (Figure 1, middle and bottom panels).

We first used a principal component analysis to evaluate overall variation in exon-level expression data, and found that tissue source is the most important determinant of that variation (Figure 2). We have therefore implemented genome-wide screens for SNPs controlling gene expression and splicing (referred to as expression quantitative trait locus [eQTL] and splicing quantitative trait locus [sQTL], respectively) separately in the tissue types.

Our screen for (*cis*-acting) polymorphisms controlling expression and splicing evaluated SNPs in or near (within 100 kb) either the target gene or exon. We limited this screen to SNPs with a minor allele frequency (MAF) > 0.04 in our sample sets (requiring at least six alleles to be present in the tissue type investigated). The screen for *cis*-acting sites controlling overall expression and those regulating exon expression levels required approximately ten and 85 million tests, respectively. On average 40 SNPs were considered for each of the ~22,000 genes, including ~12 transcripts per gene and ~four exons per transcript. Thus, thresholds for study-wide significance were 5×10^{-9} for transcript level associations and 6×10^{-10} for exon-level associations. We identified 584 study-wide significant eQTLs meeting the MAF requirements, but many of these were associated with one another and therefore appeared to reflect the same causal eQTL. We used stepwise regression to eliminate associated SNPs, separately evaluating the two tissue types, and identified 81 independent eQTLs. Significant associations that overlapped between the two tissue types were merged, resulting in 77 transcript level associations. Associations were separated into high confidence (Table 1) and low confidence (Table S2), depending on whether the transcripts were core transcripts, which indicates the highest level of confidence.

For exon-level assessments 5,357 significant associations were identified in the two tissue types combined and 1,554 remained after removal of associated SNPs. We also removed associations where the probeset contained the associated SNP or a SNP in high linkage disequilibrium (LD, $r^2 > 0.5$), leaving 985 associations. Significant associations that overlapped between the two tissue types were merged, resulting in 929 unique exon-level associations. We also identified a subset of these as high confidence (see Table 2 and Table S3) on the basis of the following criteria: (1) $p < 10^{-12}$, (2) no reported SNPs within the regions covered by the associated probesets, and (3) no suggested cross-hybridization of the associating probeset.

For all high confidence associations identified, we evaluated how often the expression effects of SNPs were observed in the other tissue, and found that 74% of eQTLs and 51% of sQTLs appeared to act exclusively in one tissue or the other. These data clearly indicate a significant role of tissue-specific genetic regulation.

To confirm the accuracy of the exon array technology, and in particular the conclusion of tissue specificity, we selected a subset of sQTLs to evaluate with quantitative real-time PCR (qRT-PCR). Events were selected to replicate the detected event as closely as possible, and also to establish that tissue specificity was not the result of low resolution of the array when exons are differentially expressed in a tissue type. We found a highly significant correlation between measurements using both technologies (Figure 3) with an overall associated *p*-value comparing the two methodologies (linear regression) of 1×10^{-35}. Importantly, we found clear replication of tissue specificity.

We also evaluated our associations for overlap with previously reported expression QTLs. We confirmed associations with a previously reported eQTL in *LRAP* (renamed *ERAP2*) [1], and also SNP regulation of *RPS26* expression [2,8]. While previous reports document an overall transcript change [2,8,9], we identified the effects of the SNP to be localized to specific exons in the *RPS26* transcript. This discrepancy is probably due to microarray platform differences. We also confirmed in our PBMC sample set several previously reported sQTLs established in HapMap cell lines, including sQTLs in *ULK4, PARP2, C17orf57,* and others [5].

Taking the full list of high confidence sQTLs we evaluated how often LD extends into or surpasses regions known to be important in splicing (Figure 4) [10]. We found that 78% of study-wide significant sQTLs, or their extended regions of SNPs in high LD ($r^2 > 0.2$), were located near the exons they regulated for at least one transcript containing the exon. The remaining approximately 22% likely reflect unknown exons not screened for in the array, and also possibly novel regulatory regions that regulate splicing outside of these well-documented regions.

Amongst all the study-wide significant sQTLs, only two of them are themselves located in a consensus splicing sequence for the relevant exons, rs10814567 in *POLRE1* and rs7770794 in *PIP3-E*. We note, however, that the probeset screening for the associated exon in *POLRE1* contains a SNP in perfect LD with the consensus site SNP and therefore may be the result of poor hybridization to the target. Given that most common polymorphisms are now known, it is surprising that there are so few cases where a candidate polymorphism responsible for a splicing change is in the consensus sequence, although this

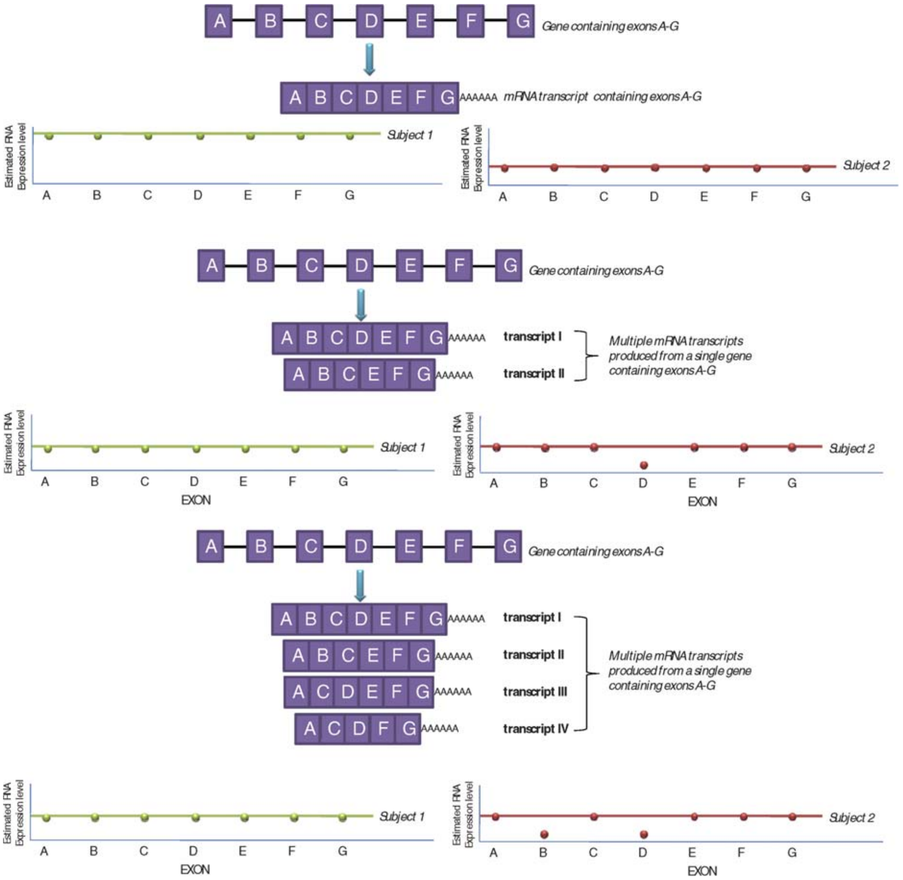

Gene containing exons A-G
mRNA transcript containing exons A-G
transcript I
transcript II
transcript III
transcript IV
Multiple mRNA transcripts produced from a single gene containing exons A-G
Estimated RNA Expression level
Subject 1
Subject 2
EXON
A
B
C
D
E
F
G

Figure 1. Idealized Representation of How Overall Expression and Alternative Splicing Events Are Reflected in the Exon Array Data

Top panel: In this study, all exon-level data were normalized across all exons and individuals. Transcript-level expression was reported for each transcript interrogated on the array by averaging (PLIER method) exon expression levels for all exons contained in a transcript (annotation details can be found at http://www.affymetrix.com). Subject 1 in this example has a higher overall transcript expression level (indicated by green line representing an average of exons A–G within subject 1) compared to subject 2 (red line). In this example, all exons contained in the transcript were expressed at approximately equal levels, suggesting that this transcript does not have alternative splice variants in either subject. Middle panel: An example of the detection of alternative splicing in which multiple transcripts are produced from a single gene through unique combination of the coding regions. Exon D in subject 2 appears to be expressed at lower quantities when compared to the other exons in the transcript (exon D expression levels lie below the average transcript line), indicating that this exon may be spliced out of the transcript in this subject (i.e., higher expression of transcript isoform II). Bottom panel: A scenario where we cannot definitively establish the combinatorial assembly of exons using these data. In this example, subject 2 has lower expression of both exons B and D. We cannot conclude that this subject has a higher proportion of transcript IV expressed compared to subject 1 or if transcript II and III are expressed at higher levels. Despite this shortcoming in situations of large heterogeneity of transcript isoforms produced with multiple alternative splicing events, these data provide a clear indicator of alternative exon composition within transcripts for a given individual. This study was specifically focused on the *cis*-acting genetic regulation of overall expression and splicing. Therefore, we were interested in identifying groups of splicing and expression patterns unique to individuals with the same genotype at certain commonly variant loci in and surrounding the transcript or exon. In cases where the expression of multiple exons was under genetic regulation, we declared it a splicing event if $<$40% of all exons contained in the transcript were associated with high confidence with the genotype. If $>$40% of exons were implicated, this was considered to be an overall expression change.

doi:10.1371/journal.pbio.1000001.g001

scarcity may be due to low primary representation of these SNPs on the array. To further evaluate the role of polymorphisms in consensus sequences, we identified all known polymorphisms in the conserved region located at the exon boundaries of the close to 300,000 core exons measured on the Affymetrix array (three basepairs into the exon and eight into the intron, Ensembl database, National Center for Biotechnology Information [NCBI] Build 36 hg18) and assessed how these influenced the expression levels of neighboring exons. A total of 2,078 SNPs were identified with an MAF $>$ 0.1, of which 1,011 were represented by a proxy on the Illumina genotyping chip ($r^2 = 1$ with the splicing SNP in Centre d'Etude du Polymorphisme Humain from Utah (CEU) HapMap samples). For both tissue types, fewer than 7% of consensus site SNPs associated with relevant exon expression levels (Table S4). While it is likely that some associations are missed because of unknown exons not included on the array, this number was surprisingly low given the common conception that disruption of this highly conserved region would very likely disrupt exon assembly. We emphasize that this analysis only evaluates systematically the effects of common SNPs in consensus regions at the exon boundaries and note that rare variation may produce profoundly different effects.

We also assessed transcript-level associations for proximity of LD regions associated with the eQTL to promoter regions (within 10 kb upstream of the transcript start) or in the 3′UTR regions of transcripts, key regions involved in tran-

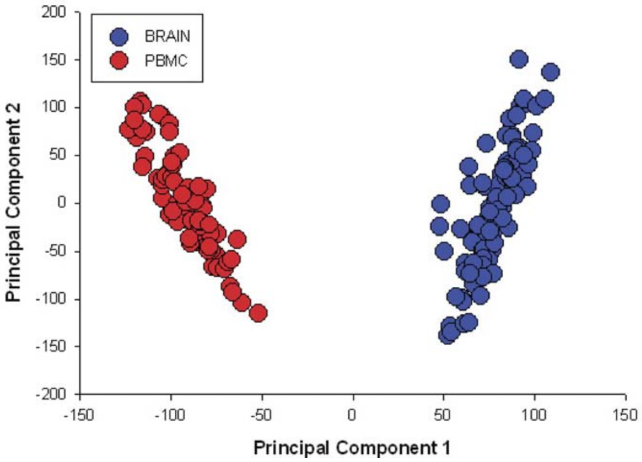


Figure 2. Principal Components Analysis of All Exon Expression Level Data for Both Brain and PBMC Samples

The differentiated pattern of expression suggests the need for tissue specific evaluations of alternative splicing and expression and demonstrates the added benefit of studying genetic regulation of splicing and expression in two unique and important cellular populations. A similar profile was observed for the transcript level expression values in the two tissue types suggesting the same level of tissue specificity for splicing.
doi:10.1371/journal.pbio.1000001.g002

scription and stability of mRNA transcripts [11]. Twenty-one out of 23 high confidence eQTLs or SNPS in the LD region ($r^2 > 0.2$) were found to be located in or extending beyond these relevant regions involved in the steady state expression level of mRNA transcripts.

One motivation for the current project is to facilitate rapid evaluation of whether polymorphisms implicated in human disease influence gene expression or splicing in relevant tissue types. We have therefore established a user interface called SNPExpress, which permits rapid interrogation of the localized effects of common SNPs on exon and transcript level expression (Figure 5). This resource is freely available at: http://people.genome.duke.edu/~dg48/SNPExpress/.

As of April 2008, >60 genome-wide associations studies were published identifying SNPs with convincing associations to complex human traits. While the association of these SNPs to the study phenotype is secure, how these polymorphisms (or variants associated with them) confer their effects is largely unknown. Of these published genome-wide association scans, 41 papers document genome-wide significant findings for 50 different traits (84 variants). Interestingly, outside of identifying nonsynonymous coding SNPs, only six claim to have identified a functional molecular-level consequence that may contribute to the phenotype, all of which are expression changes at the mRNA transcript level [12–18].

To test the utility of the SNPExpress database, we evaluated the 84 variants (Table S5) for localized associations within the transcript/exon containing this SNP or transcripts/exons within 100 kb of the SNP and determined thirteen to have a strong ($p < 1 \times 10^{-5}$) effect on an exon or transcript-level expression level (Figure 5, top panel; Table 3). Of these, rs11171739, associated with type 1 diabetes [12], was found through the use of an Illumina proxy to have an association with the exon-level expression of the *RPS26* gene. In a follow-up analysis, a SNP in LD with rs11171739 (rs2292239 $r^2 = 0.71$

Table 1. High-Confidence *cis*-Acting SNPs That Were Shown in This Study to Influence Transcript Level Expression

Tissue Type	Transcript ID[a]	Gene	Chromosome	SNP Association (eQTL)	MAF[b]	*p*-Value	*p*-Value in Other Tissue Type[c]
Brain	3456313	*ATP5G2*	12	rs1971762	0.32	3.88 e^{-17}	NS
Brain	3391724	*TMPRSS5*	11	rs1318296	0.16	7.04 e^{-11}	NS
Brain	3319613	*RPL27A*	11	rs2073687	0.19	1.23 e^{-09}	0.003
Brain	2501317	*LOC654433*	2	rs12620738	0.5	1.92 e^{-09}	3.76 e^{-09}
Brain	3509645	*SOHLH2*	13	rs943895	0.47	4.45 e^{-09}	NS
PBMC	3382061	*XRRA1*	11	rs4944950	0.19	4.48 e^{-25}	1.91 e^{-12}
PBMC	3757329	*JUP*	17	rs11079013	0.08	4.47 e^{-24}	NS
PBMC	3333247	*FADS2*	11	rs968567	0.18	6.58 e^{-22}	NS
PBMC	3757399	*NT5C3L*	17	rs1046403	0.23	1.64 e^{-20}	NS
PBMC	2821347	*ERAP2*	5	rs2549782	0.48	7.82 e^{-20}	6.10 e^{-15}
PBMC	3960174	*LGALS2*	22	rs2235338	0.43	1.28 e^{-15}	NS
PBMC	3315549	*PSMD13*	11	rs7128029	0.25	1.80 e^{-15}	6.73 e^{-06}
PBMC	2502686	*MARCO*	2	rs4491733	0.24	1.31 e^{-12}	NS
PBMC	3299504	*ACTA2*	10	rs1926196	0.49	7.27 e^{-12}	NS
PBMC	3545130	*VASH1*	14	rs10483877	0.28	3.12 e^{-11}	NS
PBMC	3921391	*WRB*	21	rs2836999	0.33	3.90 e^{-10}	NS
PBMC	3337835	*IGHMBP2*	11	rs604524	0.28	4.51 e^{-10}	NS
PBMC	2351687	*CHI3L2*	1	rs654997	0.08	4.79 e^{-10}	NS
PBMC	3935486	*S100B*	21	rs2070435	0.39	6.26 e^{-10}	NS
PBMC	3318989	*ZNF215*	11	rs1491823	0.32	8.50 e^{-10}	NS
PBMC	3261923	*AS3MT*	10	rs9527	0.18	1.07 e^{-09}	0.003
PBMC	2438093	*C1orf85*	1	rs2274226	0.43	1.54 e^{-09}	NS
PBMC	2371065	*LAMC1*	1	rs10752893	0.4	4.63 e^{-09}	NS

[a] Identifiers can be linked to genomic regions at https://www.affymetrix.com/site/login/login.affx.
[b] MAF observed in the study samples.
[c] This column defines whether or not the association was observed in the other tissue type studied (brain/PBMC), and if so the uncorrected *p*-value is given.
ID, identifier; NS, not significant.
doi:10.1371/journal.pbio.1000001.t001

Table 2. High-Confidence *cis*-Acting SNPs That Were Shown in This Study to Influence Exon-Level Expression

Tissue Type	Probeset ID[a]	Exon ID[a]	Transcript ID[a]	Exon Gene/ Transcript ID	Chromosome	SNP Association (sQTL)	MAF[b]	*p*-Value	*p*-Value in Other Tissue Type[c]
Brain	2412599	59726	2412529	*NRD1*	1	rs10888734	0.19	3.01 e^{-13}	1.78 e^{-11}
Brain	2452744	84104	2452724	*PM20D1*	1	rs708727	0.2	1.74 e^{-14}	NS
Brain	2480976	102015	2480961	*TACSTD1*	2	rs4953495	0.23	1.16 e^{-13}	NS
Brain	2665484	218169	2665472	*EFHB*	3	rs4103004	0.11	8.08 e^{-15}	9.22 e^{-06}
Brain	2670619	221387	2670481	*ULK4*	3	rs1052501	0.11	1.64 e^{-14}	8.26 e^{-14}
Brain	2676700	224952	2676671	*TKT*	3	rs3736151	0.22	2.66 e^{-17}	5.35 e^{-12}
Brain	2895678	362200	2895650	*SIRT5*	6	rs2804919	0.19	8.89 e^{-16}	1.88 e^{-13}
Brain	3391746	669898	3391724	*TMPRSS5*	11	rs1318296	0.27	7.00 e^{-12}	NS
Brain	3779866	907351	3779817	*CEP192*	18	rs1786263	0.34	3.79 e^{-12}	0.014
Brain	3847876	948510	3847873	*SLC25A23*	19	rs173229	0.37	3.60 e^{-13}	NS
Brain	3872278	962393	3872274	*VN1R1*	19	rs11084499	0.16	7.72 e^{-17}	8.88 e^{-08}
PBMC	2395130	49060	2395123	*UTS2*	1	rs161811	0.43	8.60 e^{-24}	NS
PBMC	2395127	49059	2395123	*UTS2*	1	rs161811	0.05	3.34 e^{-18}	NS
PBMC	2395129	49059	2395123	*UTS2*	1	rs161811	0.46	1.10 e^{-14}	NS
PBMC	2395134	49062	2395123	*UTS2*	1	rs161811	0.46	5.66 e^{-13}	NS
PBMC	2492088	109042	2492064	*JMJD1A*	2	rs2367575	0.35	2.61 e^{-30}	3.48 e^{-23}
PBMC	2522629	128190	2522616	*CFLAR*	2	rs4487072	0.16	3.94 e^{-12}	NS
PBMC	2545738	142708	2545681	*GTF3C2*	2	rs4803	0.26	9.14 e^{-12}	0.001
PBMC	2553702	147836	2553682	*C2orf63*	2	rs7349405	0.36	9.96 e^{-14}	NS
PBMC	2553712	147841	2553682	*C2orf63*	2	rs4435493	0.36	8.78 e^{-12}	NS
PBMC	2562840	153481	2562821	*AK095433*	2	rs1437614	0.04	8.61 e^{-12}	3.72 e^{-06}
PBMC	2845747	331076	2845699	*SLC12A7*	5	rs4580814	0.26	1.04 e^{-13}	NS
PBMC	2845703	331042	2845699	*SLC12A7*	5	rs4580814	0.26	1.04 e^{-13}	NS
PBMC	2845743	331074	2845699	*SLC12A7*	5	rs4580814	0.26	4.17 e^{-12}	NS
PBMC	2845737	331068	2845699	*SLC12A7*	5	rs4580814	0.05	4.69 e^{-12}	NS
PBMC	2856062	337718	2856044	*EMB*	5	rs1039797	0.45	9.91 e^{-12}	NS
PBMC	2859687	339993	2859667	*CENPK*	5	rs380327	0.44	1.99 e^{-19}	NS
PBMC	2859711	340012	2859667	*CENPK*	5	rs6897886	0.26	2.19 e^{-13}	NS
PBMC	2859712	340012	2859667	*CENPK*	5	rs380327	0.26	2.49 e^{-13}	NS
PBMC	2890747	359199	2890741	*SCGB3A1*	5	rs2453176	0.04	1.69 e^{-17}	NS
PBMC	2898863	364257	2898746	*LRRC16*	6	rs17317669	0.36	6.86 e^{-12}	NS
PBMC	2952341	396999	2952323	*MDGA1*	6	rs6938061	0.33	1.96 e^{-12}	NS
PBMC	3031645	446399	3031624	*TMEM176A*	7	rs7806458	0.11	5.22 e^{-18}	NS
PBMC	3079177	476145	3079172	*TMEM176B*	7	rs7806458	0.12	2.44 e^{-24}	NS
PBMC	3079179	476147	3079172	*TMEM176B*	7	rs7806458	0.33	2.49 e^{-24}	NS
PBMC	3125968	505623	3125915	*MTUS1*	8	rs17125630	0.43	2.77 e^{-14}	NS
PBMC	3125964	505621	3125915	*MTUS1*	8	rs17125630	0.26	3.63 e^{-13}	NS
PBMC	3158566	526211	3158516	*CPSF1*	8	rs3817681	0.26	7.94 e^{-13}	0.006
PBMC	3158552	526204	3158516	*CPSF1*	8	rs3817681	0.37	4.73 e^{-12}	0.007
PBMC	3210531	557982	3210497	*PRUNE2*	9	rs561970	0.33	1.47 e^{-14}	NS
PBMC	3220415	564156	3220384	*EDG2*	9	rs1411424	0.26	7.65 e^{-14}	NS
PBMC	3220416	564156	3220384	*EDG2*	9	rs1411424	0.04	5.04 e^{-12}	NS
PBMC	3269681	594934	3269662	*BCCIP*	10	rs10794030	0.22	2.15 e^{-14}	1.11 e^{-04}
PBMC	3293248	609845	3293244	*SAR1A*	10	rs870801	0.43	1.73 e^{-13}	1.27 e^{-08}
PBMC	3299591	613763	3299585	*LIPA*	10	rs2243547	0.07	7.43 e^{-14}	NS
PBMC	3308539	619413	3308489	*KIAA1598*	10	rs10787735	0.28	4.21 e^{-14}	0.034
PBMC	3308522	619400	3308489	*KIAA1598*	10	rs10787735	0.33	3.66 e^{-12}	NS
PBMC	3449119	705224	3449068	*TMTC1*	12	rs4931215	0.38	1.69 e^{-12}	NS
PBMC	3568684	779134	3568667	*MAX*	14	rs1271582	0.33	6.48 e^{-15}	1.34 e^{-04}
PBMC	3629245	816715	3629243	*RBPMS2*	15	rs7174486	0.17	2.27 e^{-20}	0.008
PBMC	3633384	819218	3633347	*MAN2C1*	15	rs4886699	0.49	7.80 e^{-13}	3.21 e^{-06}
PBMC	3633383	819217	3633347	*MAN2C1*	15	rs4886699	0.49	7.86 e^{-15}	8.34 e^{-04}
PBMC	3633393	819221	3633347	*MAN2C1*	15	rs4886699	0.41	7.73 e^{-16}	8.44 e^{-04}
PBMC	3633379	819215	3633347	*MAN2C1*	15	rs4886699	0.41	2.76 e^{-15}	0.001
PBMC	3633376	819213	3633347	*MAN2C1*	15	rs4886699	0.33	2.05 e^{-18}	0.005
PBMC	3633387	819219	3633347	*MAN2C1*	15	rs8028182	0.45	4.50 e^{-12}	NS
PBMC	3675130	844504	3675116	*TMEM8*	16	rs3830160	0.45	1.61 e^{-12}	0.005
PBMC	3675125	844502	3675116	*TMEM8*	16	rs3830160	0.07	1.25 e^{-13}	NS
PBMC	3675121	844499	3675116	*TMEM8*	16	rs3830160	0.41	4.57 e^{-13}	NS
PBMC	3705422	862698	3705412	*LOC400566*	17	rs6565724	0.04	4.41 e^{-12}	3.19 e^{-07}
PBMC	3710287	865492	3710277	*C17orf48*	17	rs440655	0.33	3.34 e^{-12}	1.49 e^{-06}
PBMC	3724617	874037	3724591	*C17orf57*	17	rs2175290	0.26	4.20 e^{-19}	0.045
PBMC	3726602	875229	3726569	*SPATA20*	17	rs989128	0.38	2.22 e^{-17}	9.50 e^{-07}
PBMC	3726601	875229	3726569	*SPATA20*	17	rs989128	0.41	8.65 e^{-18}	1.42 e^{-06}
PBMC	3726604	875231	3726569	*SPATA20*	17	rs989128	0.43	1.83 e^{-13}	2.05 e^{-05}
PBMC	3726584	875222	3726569	*SPATA20*	17	rs989128	0.49	7.12 e^{-13}	3.72 e^{-05}
PBMC	3726597	875228	3726569	*SPATA20*	17	rs989128	0.33	3.49 e^{-12}	1.11 e^{-04}
PBMC	3726603	875230	3726569	*SPATA20*	17	rs989128	0.33	7.31 e^{-15}	2.30 e^{-04}

Table 2. Continued.

Tissue Type	Probeset ID[a]	Exon ID[a]	Transcript ID[a]	Exon Gene/ Transcript ID	Chromosome	SNP Association (sQTL)	MAF[b]	*p*-Value	*p*-Value in Other Tissue Type[c]
PBMC	3726583	875221	3726569	*SPATA20*	17	rs989128	0.48	$2.18\ e^{-12}$	$9.73\ e^{-04}$
PBMC	3849702	949589	3849688	*ZNF266*	19	rs10401135	0.33	$4.40\ e^{-16}$	$3.64\ e^{-04}$
PBMC	3849706	949590	3849688	*ZNF266*	19	rs10401135	0.28	$6.93\ e^{-12}$	NS
PBMC	3859901	955509	3859899	*TMEM149*	19	rs2871921	0.23	$1.91\ e^{-17}$	$2.66\ e^{-05}$
PBMC	3870669	961516	3870611	*LILRB3*	19	rs103294	0.49	$4.89\ e^{-12}$	NS
PBMC	3894744	976106	3894727	*SIRPB1*	20	rs11696842	0.3	$2.01\ e^{-25}$	0.005
PBMC	3894745	976106	3894727	*SIRPB1*	20	rs11696842	0.22	$1.35\ e^{-20}$	0.03
PBMC	3894746	976107	3894727	*SIRPB1*	20	rs11696842	0.49	$8.10\ e^{-22}$	NS
PBMC	3894747	976107	3894727	*SIRPB1*	20	rs11696842	0.38	$2.47\ e^{-21}$	NS
PBMC	3894748	976108	3894727	*SIRPB1*	20	rs11696842	0.18	$1.96\ e^{-18}$	NS
PBMC	3947314	1007909	3947310	*C22orf32*	22	rs1801311	0.38	$9.72\ e^{-14}$	0.009
PBMC	3947319	1007912	3947310	*C22orf32*	22	rs1801311	0.34	$3.05\ e^{-12}$	0.023

[a]Identifiers can be linked to genomic regions at https://www.affymetrix.com/site/login/login.affx.
[b]MAF observed in the study samples.
[c]This column defines whether or not the association was observed in the other tissue type, and if so the uncorrected *p*-value is given.
ID, identifier; NS, not significant.
doi:10.1371/journal.pbio.1000001.t002

with rs11171739) was found to be more highly associated with type 1 diabetes [19]. Interestingly, a SNP located upstream of both of these SNPs, rs10876864, was found in our dataset to have the strongest association with a splicing event in *RPS26*. This observation extends what was recently reported by Schadt et al. [9] by specifically identifying *RPS26* splicing as responsible for the expression association with the implicated polymorphism in type 1 diabetes. More generally, however, these results illustrate the exceptional difficulty of moving from phenotypic associations to underlying biological mechanisms. While the strong splicing association with *RPS26* makes it a convincing candidate for being responsible for the diabetes risk, the originally reported polymorphism is located in the *ERBB3* gene, which also has been suggested as having direct relevance in type 1 diabetes [19]. Fortunately, the effects may in this case be resolvable because rs10876864 has a stronger association with the splicing change than does rs2292239. The association between these polymorphisms, while high, is not complete and it should be possible to resolve which SNP is the more likely causal variant by testing whether the originally identified polymorphism has a stronger association with type 1 diabetes than does the polymorphism more strongly associated with the splicing change.

This scan of genome-wide associations for effects on expression changes also identified splicing effects of rs6678677, a SNP originally identified as a risk factor in rheumatoid arthritis and later a contributor to type 1 diabetes predisposition, in the *PTPN22* gene [12,19–21]. We note that the associated SNP is located directly in the region targeted by the associated probeset, which may result in a false positive association, however the SNP effects were not observed in the brain tissue despite a similar expression level in the minor allele homozygotes. Additional work is needed to confirm this association.

Other e- or sQTLs that overlapped with associations in genome-wide association studies were found for SNPs previously implicated in ankylosing sponylitis, asthma, celiac disease, Crohn disease, HDL cholesterol, lupus, multiple sclerosis rheumatoid arthritis, and type 1 diabetes. It is unclear why the majority of the splicing/expression associations we have found are for SNPs originally implicated in autoimmune diseases. Although this result could reflect a particular importance of splicing variation in autoimmunity, two other possibilities seem more plausible. First, the imbalance could be the result of a methodological bias in evaluating a tissue type clearly relevant to immune system function (PBMCs). While brain tissue was included, little progress has been made identifying common variants that influence brain-specific phenotypes in genome-wide studies. Interestingly, for each e- or sQTL the association in the immune system relevant PBMCs in all cases was stronger compared to that observed in the brain tissue samples (Table 3), which argues for the importance of assessing expression and splicing effects in tissue types most relevant to the disease under study.

The second possible explanation for the clear excess of candidate mechanisms in the case of autoimmune diseases is more fundamental and relates to the growing recognition of the importance of rare variants in common disease [22–24]. It is generally assumed that when a common SNP is associated with disease in a genome-wide study, that it, or some other common variant in LD with it, is responsible for the association. It is theoretically possible, however, that many of the associations observed are not due to single common variants, but rather due to a constellation of more rare disease-causing variants that happen to occur, by chance, more frequently along with one of the common alleles at given SNP as opposed to the other. In such a case, the signal of association credited to a common SNP is actually a synthetic association resulting from the contributions of multiple rare SNPs. In such cases a screen for a common SNP associated with an underlying biological effect (such as expression or splicing) is not likely to identify a causal site. Our failure to identify any good strong candidate SNPs controlling expression or splicing associated with disease implicated SNPs in conditions other than autoimmune conditions could reflect a difference in the importance of

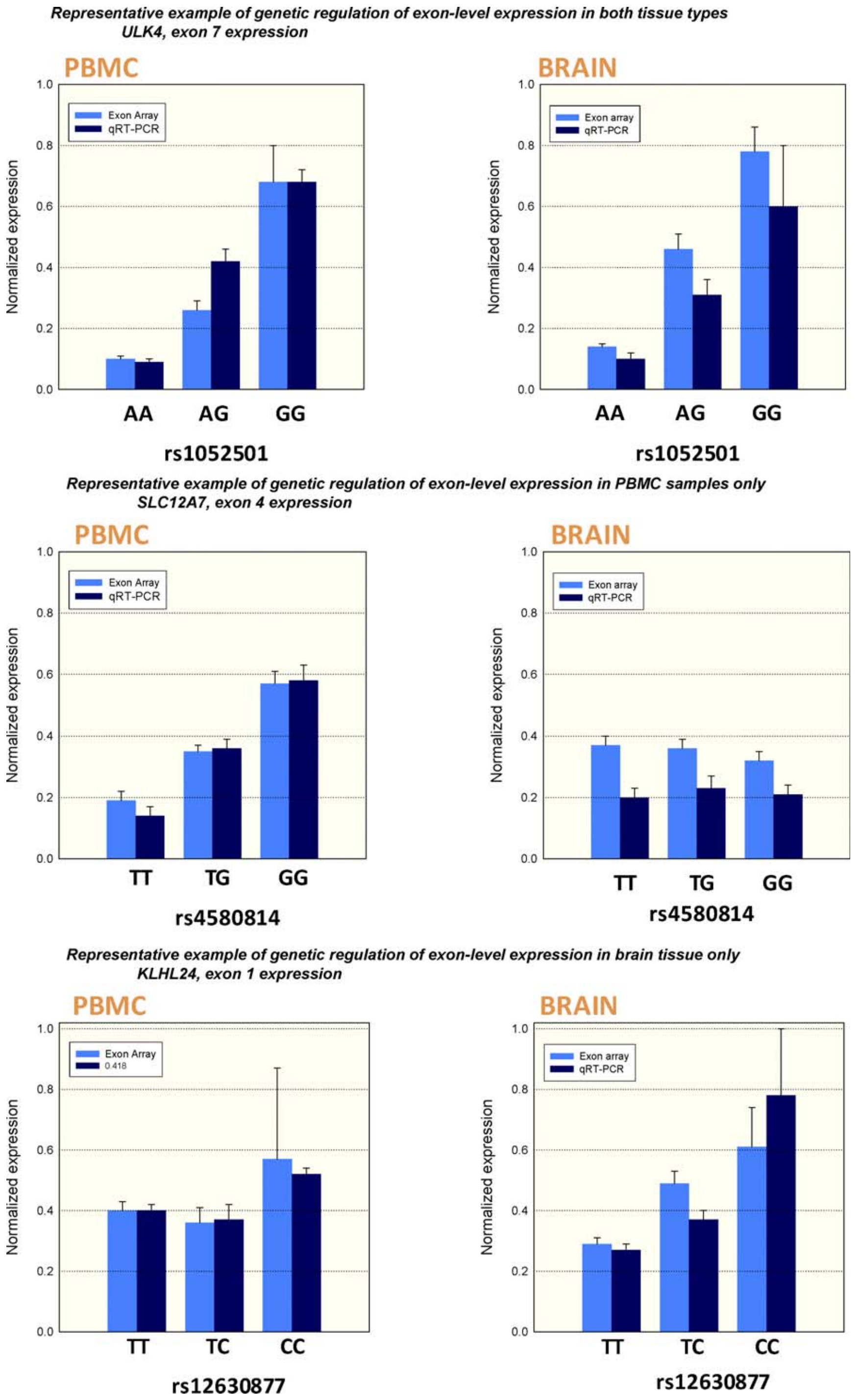


Figure 3. Quantitative Real-Time PCR Confirmation of Selected Genetically Regulated Exon-Level Expression Changes in the Two Tissue Types

Three representative scenarios are presented. The top panel shows an sQTL that was present in both brain and PBMCs. The middle panel and bottom panel show sQTLs unique to a particular tissue type, providing unequivocal evidence for tissue specific genetic regulation.

doi:10.1371/journal.pbio.1000001.g003

common variants in autoimmune disease versus other diseases. Such a difference in the role of common variants could be an indirect consequence of selection [25] related to infectious disease, which has created predispositions to autoimmune conditions. In short, outside of autoimmunity, it is possible that many of the reported associations are synthetic, due to multiple rare variants, and therefore the reason that no clear expression or splicing effects have been consistently identified at these loci.

A key challenge in human disease genomics is establishing appropriate resources to elucidate the underlying biological causes of polymorphisms that are associated with disease. As demonstrated here, one key element in this effort is the development of appropriate databases that describe the relationship between polymorphisms and patterns of gene

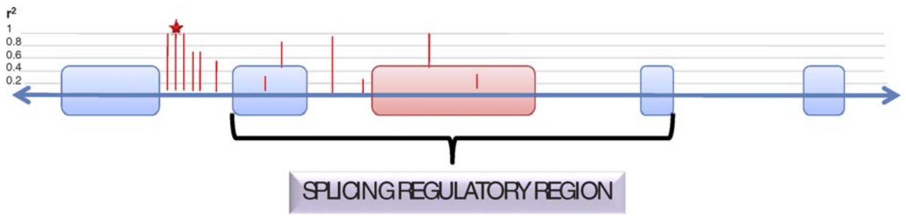


Figure 4. Methodological Details Evaluating the Proximity of a Detected sQTL and Its Region of LD to a Splicing Regulatory Region

Red box represents an exon whose expression is correlated with the SNP indicated by the starred red bar. All SNPs in LD with this sQTL are shown by red bars and the height of the bar indicates the level of correlation (r^2) with the starred SNP. We assessed how often the range of LD for a given sQTL (defined by $r^2 > 0.2$ with the sQTL) extended into or surpassed the splicing regulatory region. This analysis was performed by evaluating all mRNA transcripts containing the exon regulated by the sQTL. The splicing regulatory region was defined as the genomic region from the start of the exon located upstream of the associated exon through the stop site of the downstream exon interrogated. If the exon was part of multiple transcripts the region including the most distal and proximal neighboring exons was defined the splicing regulatory region. If the affected exon was located at the beginning or end of the transcript then the range was truncated at the start or stop of the transcript, respectively. Finally, if a single SNP associated with more than one exon in a single transcript they were considered as a single entry in this analysis (i.e., sQTL LD needed only come in close proximity of one of the affected exons to be counted as a positive entry).

doi:10.1371/journal.pbio.1000001.g004

expression and splicing in multiple human primary tissue types. As the field transitions to the study of rare variants it will be critical to supplement these datasets with complete DNA resequencing data to comprehensively characterize the full spectrum of genetic regulation of expression.

Methods

Samples. Brain tissue samples (frontal cortex) from neurologically healthy control individuals were obtained from the National Neuro-AIDS Tissue Consortium (NNTC), the Kathleen Price Bryan Brain Bank (KPBBB) at Duke University, and the Oregon Brain Bank. PBMCs from healthy living participants were purchased from Seracare Bioservices, Cellular Technology Ltd., and also provided by the Duke Human Vaccine Institute. All samples used in this analysis were of European ancestry. Sample demographics are included in Table S1. PBMCs obtained from living participants were completely de-identified and obtained according to standards set forth by the Duke University Institutional Review Board.

Genome-wide expression and genotyping. Affymetrix Human ST 1.0 exon arrays were used to assess exon and transcript expression levels for all samples used in the study. Genome-wide genotyping was performed using Illumina Human Hap550K chips. DNA and RNA were extracted using standard Qiagen protocols. Exon array sample preparation from total RNA was conducted based on standard Affymetrix protocols.

Exon array data were evaluated using a series of quality control steps defined by Affymetrix for uniform hybridization intensity, abnormal background signals, and sample outliers. The data were normalized across all samples for a tissue type on an exon and transcript level (four separate normalizations) per Affymetrix PLIER protocol with a sketch-quantile normalization procedure (Affymetrix Expression Console). This algorithm also removed undetectable signals for the dataset using a screen for signals below a group of antigenomic probesets. Principal component analysis (PCA) was performed to look secondarily to identify sample outliers using Partek Genomics Suite. Individual sample positions on top principal component (PC) axes were exported and the effects postmortem interval (applicable only to brain tissue analyses), age, gender, and sample source/processing day were tested for significance using STATA/IC 10.0. All of the four covariates were deemed to impact the sources of variability on both an exon and transcript level, and were therefore included in subsequent genetic association analyses as covariates in linear regression models. A combined normalization on both brain tissue samples and PBMCs also was performed on both the exon and transcript level. Principal components analyses were performed for a combined normalization in order to demonstrate the unique expression patterns in cortical tissue and PBMCs.

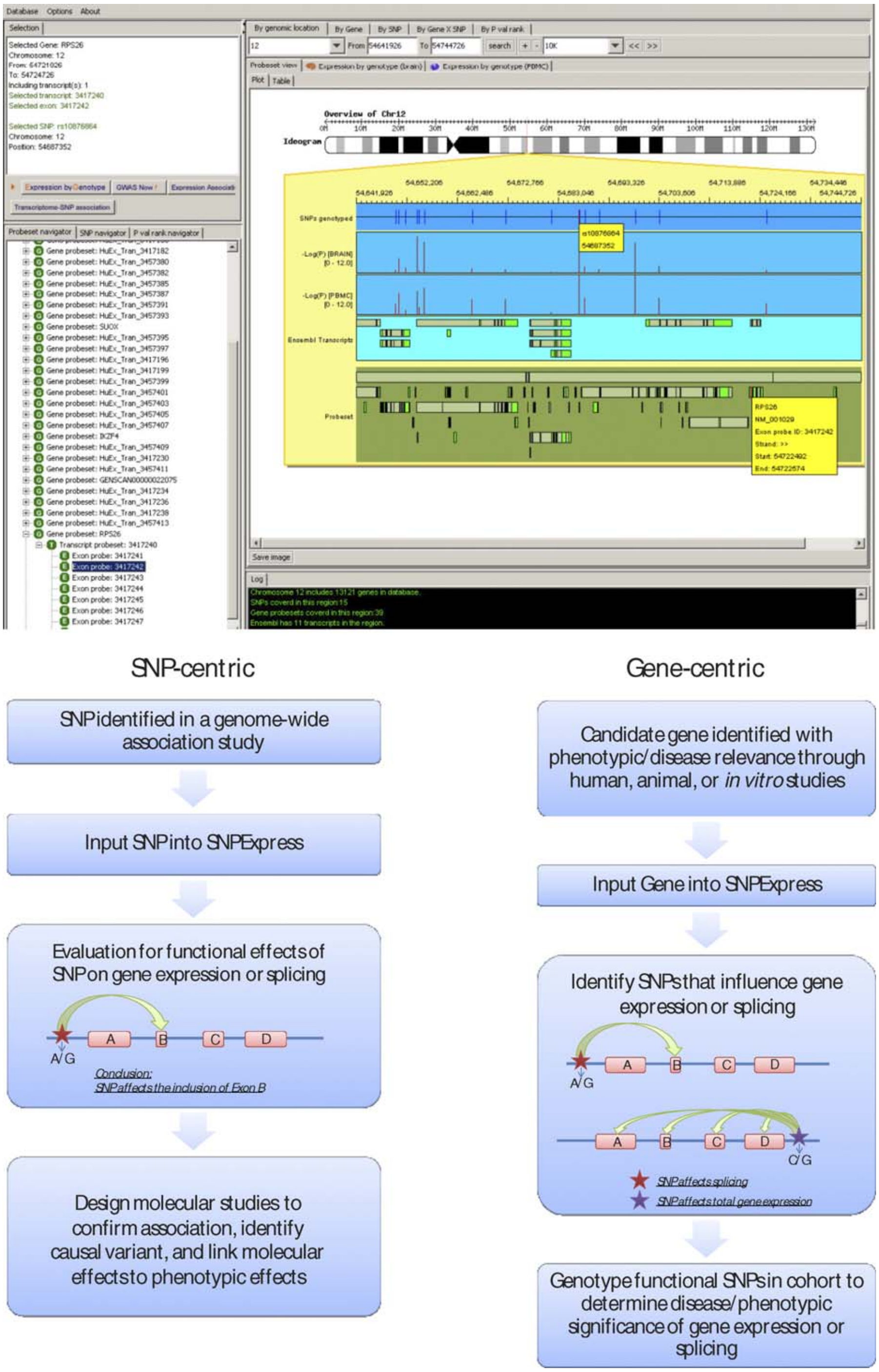
Genotyping quality was assessed using previously published methods [13]. Briefly, all SNPs that we called with a genotyping frequency of >99% across individuals (1% rule) were included in the analysis. All participants were also required to have a genotyping success rate of >99% for all SNPs that passed the 1% rule. Finally, each study-wide significant SNP identified in this analysis was manually evaluated in the Illumina Bead Studio files for genotyping quality/accuracy.

Quantitative real-time PCR. Taqman-based real-time PCR was used to confirm exon-level expression changes. Primers and fluorescently-labeled probes were custom designed for specific detection of exon-level expression. The follow primers/probes were used: ULK4, TCTCGTCCTAAAGCTTCTTCAGATT; ULK4, CTTTTCTGAGGATCTCTTTGAAGT; ULK4.PROBE, VIC- ATTAATTTGCTTGATGGGTT; SLC12A7F, ATCCTGGGCGTCATCCTCT; SLC12A7R, CACATGGCCACGATGAGG; SLC12A7.PROBE, VIC-CTGGTGTCCTGGAGTCCT; KLHL24F, TGGTACTAATATTGGGACGCAGAC; KLHL24R, CGCTTAGTTGCTGGGGAATC; KLHL24.PROBE, VIC- TAAACAGAGAGGATCTTGGG.

FAM-labeled β-actin was used as an internal control in a multiplex reaction. Assays were performed according to standard methods (900-nM primer and 250-nM probe in 20-μl reaction mix, Applied Biosystems). Fluorescence outputs were quantified in real time using a 7900HT Fast Real Time PCR System and the data were analyzed using SDS software v.2.2.2 (Applied Biosystems).

Statistical analyses. To screen for *cis*-acting genetic regulation of splicing/expression genetic association analyses were conducted to search for *cis*-acting SNPs that regulate exon-level and gene-level expression. Specifically, associations were limited to SNPs lying in or 100 kb surrounding the region of the transcripts or exons. Linear regression incorporating all the covariates was performed using PLINK genome-wide association analysis toolkit (http://pngu.mgh.harvard.edu/~purcell/plink/) [26]. To control for the possibility of spurious associations resulting from population stratification, we used a modified EIGENSTRAT method [13,27]. A total of four separate analyses were conducted, including PBMC transcript level, PBMC exon level, brain transcript level, and brain exon level. Thresholds for significance at each level were calculated based on the total number of association tests conducted within the four separate analyses.

Initially, all SNPs were excluded from analysis if the minor allele was not present at least six times in the sample group, translating to an MAF cutoff of 0.04. All significant observations that had *p*-values below the threshold and that met the MAF cutoff requirement were exported and the list was evaluated based on the following criteria: (1) Associations that were present on the exon or transcript list that were actually due to the opposing level were moved to the appropriate list. Specifically, a study-wide significant transcript level association was reported if the *p*-value achieved the threshold requirements, the affected transcript contained >2 exons, and >40% of exons contained in a transcript were significantly associated at exon-level study-wide significance level. Consistent with those rules, transcripts containing <2 exons associating with genotype that were study-wide significant and/or <40% of exons within a transcript were affected were removed and allowed on the exon-level list if they achieved exon-level study-wide significance. (2) Using stepwise linear regression (STATA/IC 10.0) associations were removed that were redundant due to LD between SNPs. Specifically, following inclusion of the most significant SNP-probeset association, only SNPs that contributed significantly above and beyond the initial association at a *p*-value of $<10^{-6}$ were considered as a separate association. (3) Finally, if the sQTL or a SNP in LD ($r^2 > 0.5$) was located in a probeset for the



Selected Gene: RPS26
Chromosome: 12
From: 54721026
To: 54724726
Including transcript(s): 1
Selected transcript: 3417240
Selected exon: 3417242

Selected SNP: rs10876864
Chromosome: 12
Position: 54687352

```
Chromosome 12 includes 13121 genes in database.
SNPs coverd in this region:15
Gene probesets coverd in this region:39
Ensembl has 11 transcripts in the region.
```

SNP-centric

SNP identified in a genome-wide association study

Input SNP into SNPExpress

Evaluation for functional effects of SNP on gene expression or splicing

Conclusion: SNP affects the inclusion of Exon B

Design molecular studies to confirm association, identify causal variant, and link molecular effects to phenotypic effects

Gene-centric

Candidate gene identified with phenotypic/disease relevance through human, animal, or *in vitro* studies

Input Gene into SNPExpress

Identify SNPs that influence gene expression or splicing

Genotype functional SNPs in cohort to determine disease/phenotypic significance of gene expression or splicing

Figure 5. SNPExpress Database

Top panel: Output showing an example of an association between rs10876864 and a splicing change in *RPS26*. Software permits the input of an SNP, a gene, or a genomic region for comprehensive interrogation of associations between SNPs and exon/transcript expression levels in the regions surrounding the SNP. The blue frame indicates the SNPs genotyped on the chip, the two lighter blue frames correspond to the –log *p*-values for the brain and PBMC samples, the turquoise panel contains all Ensembl transcripts, and the bottom green panel shows all of the exons/transcripts screened for on the array. Bottom panel: This database can be applied in a SNP-centric or gene-centric approach for determining the functional and phenotypic consequences of genetic variation on the transcriptome.

doi:10.1371/journal.pbio.1000001.g005

exon-level associations, it was excluded from any list of significant associations. Step 3 was not applied to transcript level associations as these expression levels were determined over a range of exons thereby reducing the contribution of SNPs to the expression levels. Post hoc evaluation of the transcript level associations we identified were manually inspected for effects of SNPs or SNPs in LD contributing to the association by exporting the raw values, eliminating probesets that contained a SNP, and recalculating the expression level. None of the reported associations could be accounted for by a SNP in a probeset measuring the transcript.

We also assessed whether associations were present in the other sample type. The *p*-value for declaring an effect in the other tissue

Table 3. Overlap between e- and sQTLs and Genome-Wide Significant Associations with Complex Human Traits

Trait	Variant	Gene/Region	GWAS *p*-Value	Reference	Illumina Proxy	GWAS SNP-Illumina Proxy r^2	Affected Gene/ Level of Expression Change	Illumina SNP Expression Change *p*-Value/Group[a]
Ankylosing sponylitis	rs30187	*ARTS1*	3.4 e^{-10}	[28]	rs30187	—	*ARTS1, exon*	0.011/brain, 2.98 e^{-7}/PBMC
Asthma	rs7216389[b]	*17q21*	9 e^{-11}	[17]	rs7216389	—	*GSDML, exon*	1.11 e^{-6}/brain, 1.27 e^{-10}/PBMC
Celiac disease	rs9357152	*HLA-DQA1/B1*	5.2 e^{-14}	[29]	rs9357152	—	*HLA-DQB2, exon*[c]	NS/brain, 5.93 e^{-10}/PBMC
Celiac disease	rs9275141	*HLA-DQA1/B1*	3.9 e^{-16}	[29]	rs9275141	—	*HLA-DQB2, exon*[c]	0.001/brain, 7.2 e^{-10}/PBMC
Crohn disease	rs9858542	*3p21*	3.58 e^{-8}	[12]	rs3197999	0.955	*APEH, exon*	6.73 e^{-15}/brain, 3.82 e^{-20}/PBMC
HDL cholesterol	rs2338104	*MMAB*	3.4 e^{-8}	[30]	rs2058804	1	*UBE3B, exon*	5.82 e^{-10}/brain, 4.65 e^{-11}/PBMC
Lupus	rs9275572	*HLA region, 6p21*	2.8 e^{-12}	[31]	rs9275572[d]	—	*HLA-DQB2, exon*	0.031/brain, 4.59 e^{-11}/PBMC
Lupus	rs10239340	*IRF5/TNPO3*	7 e^{-16}	[31]	rs10239340	—	*ITF5, exon*[e], [6]	3.61 e^{-11}/brain, 1.97 e^{-12}/PBMC
Multiple sclerosis	rs3135388	*HLA-DRA*	8.94 e^{-81}	[32]	rs9271366[f]	0.96	*HLA-DRB5, gene*	1.81 e^{-6}/brain, 4.67 e^{-24}/PBMC
Rheumatoid arthritis	rs6679677	*PTPN22*	5.5 e^{-25}	[12]	rs2476601	1	*PTPN22, exon*[g]	NS/brain, 6.16 e^{-29}/PBMC
Rheumatoid arthritis	rs6457617	*MHC*	5.18 e^{-75}	[12]	rs6457617[d]	—	*HLA-DQB2, exon*	NS/brain, 3.87 e^{-8}/PBMC
Type 1 diabetes	rs11171739	*12q13*	9.7 e^{-11}	[12]	rs10876864	0.862	*RPS26, exon*[h]	7.07 e^{-21}/brain, 6.92 e^{-37}/PBMC
Type 1 diabetes	rs9270986	*HLA-DRB1*	2.3 e^{-122}	[12]	rs9271366[f]	0.913	*HLA-DRB5, gene*	1.81 e^{-6}/brain, 4.67 e^{-24}/PBMC
Type 1 diabetes	rs6679677	*PTPN22*	5.4 e^{-26}	[17]	rs2476601	1	*PTPN22,exon*[g]	NS/brain, 6.16 e^{-29}/PBMC

[a] The lowest *p*-value is reported if multiple probesets associated with the Illumina proxy SNP.
[b] This SNP previously was shown to influence the overall expression of *ORMDL3*.
[c] rs9357152 and rs9275141 both associate with the expression of an *HLA-DQB2* exon, r^2 of these two SNPs in this study, and in HapMap data are ~0.25.
[d] rs9275572 and rs6457617 both associate with the expression of an *HLA-DQB2* exon, r^2 of these two SNPs in this study, and in HapMap data are ~0.64.
[e] The genetic regulation of this splicing event was previously reported (prior to the identification of this SNP in a genome-wide association study) [6].
[f] The same Illumina proxy was identified for the association between type 1 diabetes and multiple sclerosis.
[g] The control of this SNP in regulating exon-level expression may be misrepresented by the presence of the associating SNP being present in the probeset.
[h] Genetic regulation of overall *RPS26* expression has been reported previously [2,8,9].
NS, not significant.
doi:10.1371/journal.pbio.1000001.t003

was based on a p-value cutoff of 0.05. Directionality was confirmed to be the same in the two tissue types for all overlapping associations. All uncorrected p-values for observations in the other tissue types are provided in the relevant tables.

See Figure 4 for methodological details regarding the screen for proximity of sQTLs to affected exons.

To screen for effects of consensus site SNPs, we first identified all known SNPs located in highly conserved consensus site regions at the exon-intron boundary including all SNPs eight basepairs into an intron and three basepairs into an exon (Ensembl database). A total of 2,078 common consensus site SNPs were identified, of which 1,011 had proxies ($r^2 = 1$) with one or more SNPs on the Illumina Human Hap550K chip allowing assessments of their effects in the present dataset. Specifically, consensus site SNPs, or their proxies, were assessed for significant associations (uncorrected $p < 0.05$) with the expression level of the immediate exon or exons located up- or downstream for all transcripts containing that exon.

Supporting Information

Table S1. Sample Demographics

Found at doi:10.1371/journal.pbio.1000001.st001 (48 KB RTF).

Table S2. Lower Confidence *cis*-Acting SNPs That Were Shown in This Study to Influence Transcript Level Expression

Found at doi:10.1371/journal.pbio.1000001.st002 (321 KB RTF).

Table S3. Lower Confidence *cis*-Acting SNPs That Were Shown in This Study to Influence Exon-Level Expression

Found at doi:10.1371/journal.pbio.1000001.st003 (4.3 MB RTF).

Table S4. Table of Associations between Consensus Site SNPs and Adjacent Exons

Found at doi:10.1371/journal.pbio.1000001.st004 (285 KB RTF).

Table S5. A List of Genome-wide Association Studies Interrogated for Significant Associations That Affect Expression and Splicing in the SNPExpress Database.

Full references are provided below this table.

Found at doi:10.1371/journal.pbio.1000001.st005 (459 KB RTF).

Acknowledgments

Author contributions. ELH, DG, and DBG conceived and designed the experiments. ELH, KDC, and KVS performed the experiments. ELH, DG, JMM, KVS, and DBG analyzed the data. DG, JMM, WNG, KAW-B, CMH, and TND contributed reagents/materials/analysis tools. ELH wrote the paper.

Funding. This work was funded in part by the Institute for Genome Sciences and Policy at Duke University, the Center for HIV-AIDS Vaccine Immunology, the Bryan Alzheimer's Disease Research Center (NIA, AG028377), and a pilot grant from the Kathleen Price Bryan Alzheimer's Disease Research Center. The tissue specimens used in this study were kindly provided by the National NeuroAIDS Tissue Consortium (NNTC), the Kathleen Price Bryan Brain Bank (KPBBB) at Duke University, Layton Aging & Alzheimer's Disease Center at Oregon Health and Science University, and the Duke Human Vaccine Institute.

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

References

1. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365–1369.
2. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78. doi:10.1371/journal.pgen.0010078
3. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nat Genet 39: 1202–1207.
4. Hull J, Campino S, Rowlands K, Chan MS, Copley RR, et al. (2007) Identification of common genetic variation that modulates alternative splicing. PLoS Genet 3: e99. doi:10.1371/journal.pgen.0030099
5. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, et al. (2008) Genome-wide analysis of transcript isoform variation in humans. Nat Genet 40: 225–231.
6. Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, et al. (2008) Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet 82: 631–640.
7. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, et al. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21: 650–659.
8. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, et al. (2007) A survey of genetic human cortical gene expression. Nat Genet 39: 1494–1499.
9. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6: e107. doi:10.1371/journal.pbio.0060107
10. Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3: 285–298.
11. Shyu AB, Wilkinson MF, van Hoof A (2008) Messenger RNA regulation: to translate or to degrade. Embo J 27: 471–481.
12. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
13. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317: 944–947.
14. Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, et al. (2008) Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N Engl J Med 358: 900–909.
15. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40: 189–197.
16. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, et al. (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet 3: e58. doi:10.1371/journal.pgen.0030058
17. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448: 470–473.
18. Thorleifsson G, Magnusson KP, Sulem P, Walters GB, Gudbjartsson DF, et al. (2007) Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. Science 317: 1397–1400.
19. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39: 857–864.
20. Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, et al. (2005) Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. Am J Hum Genet 77: 1044–1060.
21. Bottini N, Musumeci L, Alonso A, Rahmouni S, Nika K, et al. (2004) A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. Nat Genet 36: 337–338.
22. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, et al. (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455: 232–236.
23. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539–543.
24. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, et al. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358: 667–675.
25. Gibson G, Goldstein DB (2007) Human genetics: the hidden text of genome-wide associations. Curr Biol 17: R929–R932.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
27. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
28. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39: 1329–1337.
29. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 39: 827–829.
30. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161–169.
31. Harley JB, Alarcon-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, et al. (2008) Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. Nat Genet 40: 204–210.
32. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357: 851–862.