


RESEARCH

Open Access



# Polymorphic mobile element insertions contribute to gene expression and alternative splicing in human tissues

Xiaolong Cao<sup>1†</sup>, Yeting Zhang<sup>1,2†</sup>, Lindsay M. Payer<sup>3</sup>, Hannah Lords<sup>1</sup>, Jared P. Steranka<sup>3</sup>, Kathleen H. Burns<sup>3</sup> and Jinchuan Xing<sup>1,2\*</sup> 

\* Correspondence: [xing@biology.rutgers.edu](mailto:xing@biology.rutgers.edu)

<sup>†</sup>Xiaolong Cao and Yeting Zhang contributed equally to this work.

<sup>1</sup>Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

<sup>2</sup>Human Genetic Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

Full list of author information is available at the end of the article

## Abstract

**Background:** Mobile elements are a major source of structural variants in the human genome, and some mobile elements can regulate gene expression and transcript splicing. However, the impact of polymorphic mobile element insertions (pMEIs) on gene expression and splicing in diverse human tissues has not been thoroughly studied. The multi-tissue gene expression and whole genome sequencing data generated by the Genotype-Tissue Expression (GTEx) project provide a great opportunity to systematically evaluate the role of pMEIs in regulating gene expression in human tissues.

**Results:** Using the GTEx whole genome sequencing data, we identify 20,545 high-quality pMEIs from 639 individuals. Coupling pMEI genotypes with gene expression profiles, we identify pMEI-associated expression quantitative trait loci (eQTLs) and splicing quantitative trait loci (sQTLs) in 48 tissues. Using joint analyses of pMEIs and other genomic variants, pMEIs are predicted to be the potential causal variant for 3522 eQTLs and 3717 sQTLs. The pMEI-associated eQTLs and sQTLs show a high level of tissue specificity, and these pMEIs are enriched in the proximity of affected genes and in regulatory elements. Using reporter assays, we confirm that several pMEIs associated with eQTLs and sQTLs can alter gene expression levels and isoform proportions, respectively.

**Conclusion:** Overall, our study shows that pMEIs are associated with thousands of gene expression and splicing variations, indicating that pMEIs could have a significant role in regulating tissue-specific gene expression and transcript splicing. Detailed mechanisms for the role of pMEIs in gene regulation in different tissues will be an important direction for future studies.

**Keywords:** Quantitative trait loci, Gene expression regulation, Alternative splicing, Transposable elements, Polymorphic mobile element insertions



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Introduction

Mobile genetic elements, or mobile elements (MEs), are segments of DNA that can move around and make copies of themselves within a genome [1]. At least 50% of the human genome is derived from MEs [2]. Three non-long terminal repeat (non-LTR) retrotransposons dominate the recent ME activity: the short interspersed element (SINE) *Alu* [3], the long interspersed element 1 (LINE1) [4], and the composite SVA (SINE-VNTR (variable-number tandem repeat)-*Alu*) [5, 6] element. LINE1 is an autonomous ME and encodes proteins that are required for the retrotransposition of LINE1 [7] and the non-autonomous *Alu* and SVA retrotransposons [8], as well as occasionally cellular RNAs [9]. Many diseases, including cancer [10] and psychiatric disorders [11], are associated with the activities of MEs [12, 13]. In addition to causing genomic structural changes, MEs can also alter mRNA splicing [14] and gene expression levels [15, 16] via a wide variety of mechanisms, including acting as promoters [17], enhancers [18], splicing sites [19], and terminators for transcription [20] and affecting chromatin looping [21].

The activities of MEs create new insertional mutations in the genome, leading to thousands of polymorphisms among human individuals and populations [22–24]. The effects of polymorphic mobile element insertions (pMEIs) on gene expression have been studied in the transformed B lymphocytes cell lines (LCLs) of the 1000 Genomes Project (1KGP) [25–28] and in human induced pluripotent stem cells [28]. Together, several hundred pMEI loci were identified as expression quantitative trait loci (eQTLs). However, the full extent of the impact of pMEIs on human gene expression in diverse tissues has not been extensively examined.

The Genotype-Tissue Expression (GTEx) project provides a public resource to study tissue-specific gene expression and regulation [29–31]. In the v7 release, GTEx provides 11,668 high-depth RNA sequencing (RNA-seq) datasets from 51 tissues and 2 cell lines of 714 donors. More than 600 of the donors have also been subjected to high-depth whole genome sequencing (WGS). This rich dataset makes it possible to assess the impact of different types of genomic variants on gene expression. For example, studies have reported the impact of structural variants [32], rare variants [33], and short tandem repeats [34] on gene expression variation. However, the role of pMEIs in gene regulation and alternative splicing, especially for pMEIs not annotated in the reference genome, has not been fully evaluated. Given that thousands of common pMEIs exist in human populations, pMEIs might explain a large proportion of gene expression variation among humans. With the large GTEx dataset, we systematically identified pMEIs in each donor and examined the impact of common pMEIs on gene expression and splicing.

## Results

### Detection of pMEIs in GTEx individuals

We obtained WGS data from the GTEx v7 release. Using the Mobile Element Locator Tool (MELT) [35], we identified MEs that are present in the sequenced individuals but absent in the reference genome, as well as MEs that are present in the reference genome but absent in a subset of sequenced individuals. We refer to these two types of ME polymorphisms as non-reference MEIs (nrMEIs) and reference MEIs (rMEIs) in the following text, respectively. We identified a total of 80,057 candidate nrMEI and rMEI loci in 639 individuals, including 638 GTEx individuals and the HuRef sample (Table 1). Overall,

**Table 1** Overview of pMEIs in the MELT call set, eQTL, and sQTL analyses

ME type	MELT call set			eQTL			sQTL		
	Raw	HQ	Common	All	Causal	Highest	All	Causal	Highest
nrAlu	62,864	13,870	2157	1451	562	147	1071	539	191
nrL1	11,159	2130	246	177	81	23	126	71	18
nrSVA	1877	558	69	61	32	12	51	27	13
rAlu	3837	3687	968	671	253	84	444	202	106
rL1	192	188	59	42	15	7	28	18	8
rSVA	128	112	21	20	13	9	14	9	6
<b>Total</b>	<b>80,057</b>	<b>20,545</b>	<b>3520</b>	<b>2422</b>	<b>956</b>	<b>282</b>	<b>1734</b>	<b>866</b>	<b>342</b>

MELT call set: raw—all pMEI loci identified by MELT; HQ—high-quality loci after quality control; common—pMEIs used for the eQTL and sQTL analysis

eQTL/sQTL analysis: all—unique pMEIs in eQTL/sQTL analysis (FDR < 10%); causal—unique pMEIs identified as the causal variant; highest—unique pMEIs identified as the causal variant with the highest causal probability

99.5% of sites have no-call rates < 25%, demonstrating the high quality of the sequenced genomes.

The initial candidate ME loci were further filtered based on quality scores, no-call rates, and other criteria (see the “Methods” for details). After filtering, 20,545 high-quality loci were selected for further analysis. Most pMEIs have allele frequency < 0.05, especially nrMEIs (Additional file 1: Fig. S1a, S1b). Because the human reference genome is based on only a small number of individuals, pMEIs present in the reference genome (rMEIs) should be more common than pMEIs absent in the reference genome (nrMEIs). As expected, overall, rMEIs have higher allele frequencies than nrMEIs (Additional file 1: Fig. S1a, S1b). The number of loci with pMEI present in an individual is correlated with their self-reported ancestry. In general, the number of loci with nrMEI and rMEI present in African individuals is larger than in non-African individuals (Additional file 1: Fig. S1c). We define common pMEIs as those with allele frequency between 0.05 and 0.95. Overall, 3076 nrMEIs and 1662 rMEIs are common, which are 18.58% and 41.68% of the high-quality nrMEI and rMEI call sets, respectively. After further quality control, a total of 3520 common pMEIs were selected for the following analyses (Table 1).

### Identification of pMEI-associated eQTLs

Next, we determined the effect of pMEIs on nearby gene expression by identifying pMEI-associated *cis*-eQTLs. The GTEx v7 release includes expression data of 56,202 genes, including 19,820 protein-coding genes and 36,382 non-coding genes (Table 2). We selected 46 tissues and 2 cell lines with expression data in more than 70 individuals

**Table 2** Summary of genes

Gene	Total	Expressed	eQTLs	ME causal	ME highest causal
Protein-coding	19,820	19,064	4243	1062	294
Non-coding	36,382	19,111	2099	526	139
<b>Total</b>	<b>56,202</b>	<b>38,175</b>	<b>6342</b>	<b>1588</b>	<b>433</b>

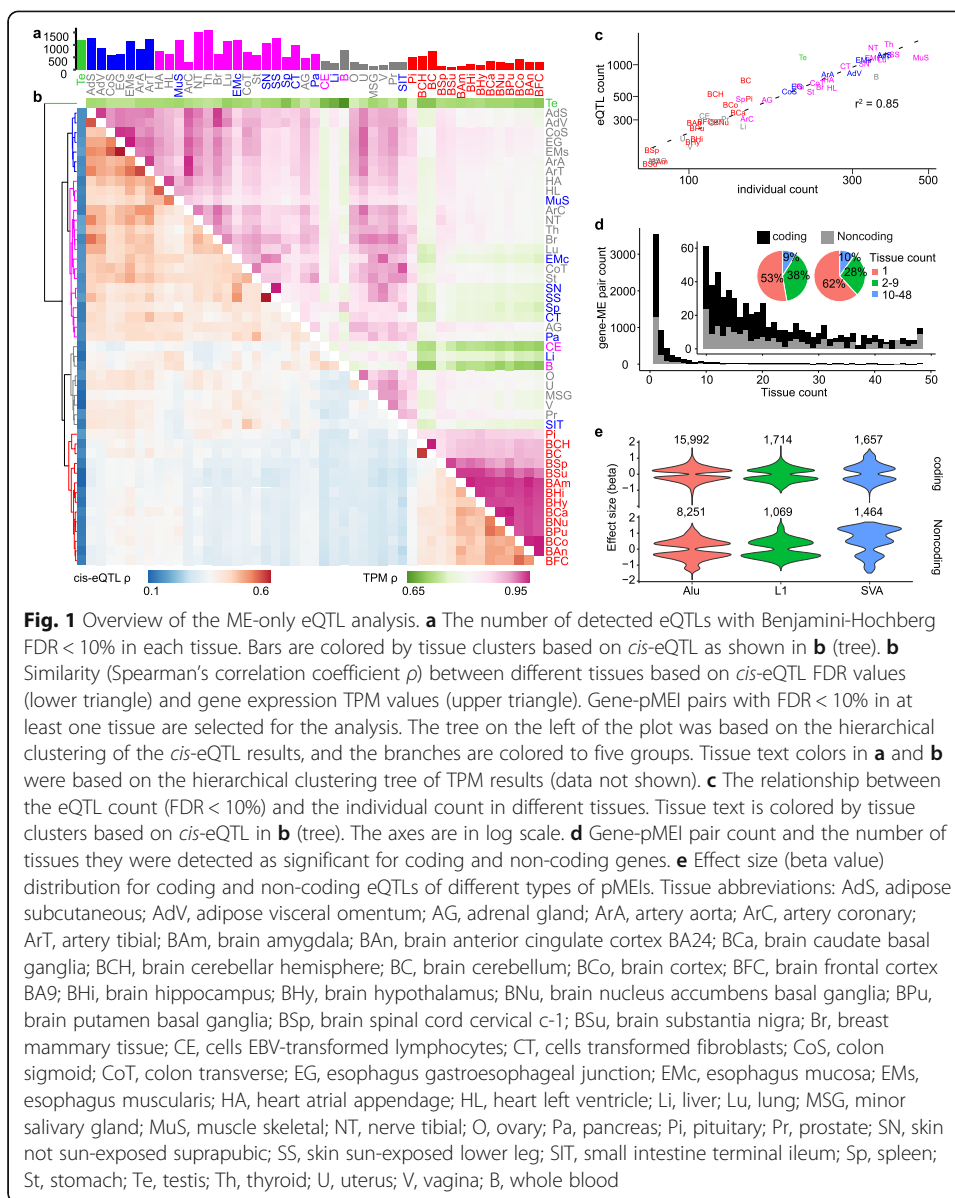
Expressed—genes used in the eQTL analysis of at least one tissue

eQTLs—number of unique genes in the ME-only eQTL analysis with FDR < 10%; ME causal and ME highest causal—unique genes with pMEIs predicted as a causal variant or a causal variant with the highest probability, respectively

for the analysis (ranging from 78 to 481 individuals per tissue or cell line) (Additional file 2: Table S1). We will refer to both tissues and cell lines as tissues for simplicity in the following text. After excluding low-expressed genes from the analysis, the average number of tested protein-coding genes in each tissue is 16,461 with a standard deviation (SD) of 598 (see the “Methods” section for detail). For non-coding genes, the testis is an outlier with 14,970 expressed genes. The average number of expressed non-coding genes in tissues other than testis is 7294 with an SD of 826.

We performed *cis*-eQTL mapping with Matrix eQTL [36] in each tissue. Here, we define an eQTL as a unique combination of tissue-gene-variant. Among all tissues, we identified 30,147 eQTLs with 6342 distinct genes, 2422 distinct pMEIs, and 8204 distinct gene-ME pairs with a false discovery rate (FDR) < 10%. pMEIs that are eQTLs showed strong enrichment near the transcription start site (TSS) of the affected genes, although some eQTL-pMEIs are much further away from the affected genes (Additional file 1: Fig. S2a). In comparison, there is no enrichment of pMEIs at TSS among all tested tissue-gene-pMEI combinations (Additional file 1: Fig. S2b). Next, we define an eGene as a tissue-gene pair that was identified in the eQTL analysis with an FDR < 10%, while an eVariant as a tissue-variant pair with an FDR < 10%. Because an eGene can be influenced by multiple variants and an eVariant may have an impact on multiple genes, the numbers of eGenes (24,109) and eVariants (17,230) are smaller than the total number of eQTLs. The number of eQTLs (FDR < 10%) per tissue ranges from 118 to 1609, and the sample size is strongly correlated with the number of detected eQTLs ( $r^2 = 0.85$ , Fig. 1a, c, Additional file 2: Table S1). This strong correlation was also observed in similar studies [30, 31, 34]. The correlation is even stronger ( $r^2 = 0.92$ ) when we added the number of expressed genes as a covariate in the linear regression analysis of the number of eQTLs. For eQTLs, most gene-ME pairs were identified in only one tissue, accounting for 53% of coding gene-ME pairs and 62% of non-coding gene-ME pairs (Fig. 1d, Additional file 2: Table S1). The higher tissue specificity of non-coding gene eQTLs could be explained by the fact that non-coding genes more frequently have tissue-specific expression patterns.

To determine if closely related tissues show similar eQTL profiles, we evaluated the eQTL correlations among different tissues for ME-gene pairs using Spearman's correlation ( $\rho$ ). The Spearman's correlation of the expression level of these eQTL genes (calculated as transcript per million (TPM)) was also calculated to determine the impact of similarities of gene expression on eQTL identification. As shown in Fig. 1b, tissues from different brain regions were clustered together by eQTL correlations and eQTL gene expression levels. Testis (Te) showed the highest difference with other tissues in both eQTLs and gene expression levels. Highly similar tissues, such as skin sun-exposed (SS) and skin not sun-exposed (SN), brain cerebellum (BC) and brain cerebellum hemisphere (BCH), are highly similar in the eQTL significance ( $\rho > 0.6$ ) and the gene expression level ( $\rho > 0.95$ ). However, whole blood and EBV-transformed lymphocytes (B and CE) showed lower gene expression correlation with other tissues ( $\rho < 0.8$  in general) than other tissue pairs, suggesting a different expression pattern in blood and cell line samples. It is also obvious that the correlations are higher for gene expression than for eQTLs. This is partial because gene expression values can be more accurately determined and normalized than eQTL significance values.



To determine if the presence/absence of an ME has a directional impact on the gene expression, we examined the association between the direction of the gene expression change (positive or negative beta value) and the presence or absence of an ME. We observed several significant differences (e.g., SVA for non-coding genes, Fig. 1e, Additional file 3: Table S2). However, because some pMEIs are eQTLs in multiple tissues and/or affect multiple genes, these pMEIs can potentially bias the result, especially for small datasets such as SVAs. To control for this bias, we selected a single best eQTL for each pMEI for the testing. Using one eQTL per pMEI locus, we observed no statistically significant difference in the direction of the effect for any of the three types of pMEIs (*Alu*, *L1*, *SVA*), for either coding or non-coding genes (Additional file 3: Table S2). This result suggests that for common pMEI eQTLs, the ME-specific sequence is a less important factor affecting the nearby gene expression than the presence/absence of an ME. We also compared the correlation of the direction of the effect

(i.e., the sign of the beta value) among tissue pairs. Overall, the direction of the effect for MEs is highly consistent among tissue pairs, with an apparent exception of the testis (Additional file 1: Fig. S3). Excluding the testis, the effect direction among tissue pairs is consistent for  $98.6 \pm 1.7\%$  of eQTLs.

#### **Fine-mapping of causal pMEIs for eGenes**

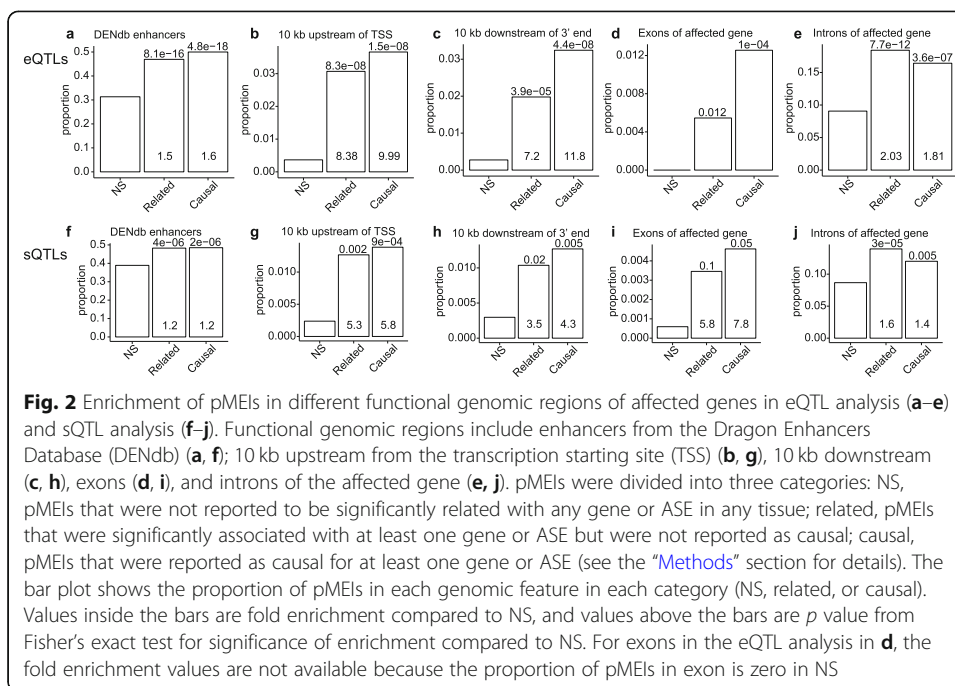
Due to the linkage disequilibrium among genetic variants, several tightly linked variants can be identified as eQTLs along with the causal variant. To determine whether the pMEIs identified in the eQTL analysis are the causal variants, we applied a fine-mapping approach for each eQTL locus. To do this, we gathered the single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) from GTEx individuals and selected a total of 6,334,405 high-quality common variants, including 5,837,891 SNPs and 496,514 indels. For the 6342 unique eGenes identified in the ME-only eQTL analysis, we performed joint analyses for pMEIs and these common variants to identify all variants associated with an eGene in each tissue. Then, we applied a fine-mapping method for each of the 24,109 eGenes to identify the contributions of MEs in altering gene expression. Overall, pMEIs were included in the causal variant set for 13.98% of eGenes, ranging from 10.69% in the sun-exposed lower leg skin to 25.33% in the hippocampus. pMEIs were detected as the highest probability causal variant for 4.55% of tested eGenes (2.67–9.18% among tissues) (Table 1, Additional file 2: Table S1). This is slightly more frequent than the 3.5% (2.4–4.4% among tissues) detected for structural variants in a previous study [32].

#### **Enrichment of eQTL-pMEIs with functional genomic elements**

To explore the potential molecular mechanisms by which MEs influence gene expression, we examined the enrichment of pMEIs relative to functional genomic elements. We grouped the 3520 common pMEIs into three categories: not an eVariant for any gene (NS), identified as an eVariant but not a causal eVariant (related), and identified as a causal eVariant (causal). Compared to the NS set, pMEIs that are eVariants (related and causal) are significantly enriched in enhancers, 10 kb upstream or downstream of affected genes, and exons and introns of affected genes (Fig. 2a–e). This is consistent with the observation that eQTL-pMEIs are enriched near the TSS of genes (Additional file 1: Fig. S2). Importantly, pMEIs in the “causal” category are more significantly enriched in functional regions than “related” pMEIs in all categories except in introns. This enrichment suggests pMEIs in the causal set are more likely to be the true functional variant for the gene expression change. Only a small portion of pMEIs are in the exon of genes, and all of them are detected as eQTLs and showed a stronger enrichment in the causal set (Fig. 2e). Given the size of the pMEIs, it is expected that the exonic pMEIs will have a strong impact on the gene expression level. The enrichment of pMEIs in functional elements is similar to structural variants in general, as structural variants impacting gene expression are also enriched in enhancers, promoters, and regions close to the affected genes [32].

#### **Identification of pMEI-associated sQTLs**

We next investigated the impact of pMEIs on alternative splicing of genes. We analyzed splicing quantitative trait loci (sQTLs) similarly to eQTLs, except we used percent



splicing (PSI) scores of alternative splicing events (ASEs) instead of TPM of genes (see the “Methods” section for a full definition of the ASEs). When determining ASEs, genes sharing one or more exons were grouped together as a gene cluster. We will refer to these gene clusters as genes in the sQTL analysis for simplicity. There are 165,882 ASEs from 17,015 genes (Table 3). About half of the events occur inside the gene, these include alternative 3’/5’ splicing site (A3/A5), mutually exclusive exons (MX), retained intron (RI), and skipped exon (SE). The other half occur at the edge of a gene, including alternative first/last exons (AF/AL) (Table 3). We detected a total of 21,529 sQTLs with 7184 distinct splicing events from 2992 genes with FDR < 10%. Similar proportions

**Table 3** Summary of alternative splicing events

ASEs	Total events (genes)	Events in sQTL (genes)	ME causal (genes)	ME highest causal (genes)
A3	14,918 (7419)	537 (456)	165 (154)	50 (49)
A5	14,197 (7144)	576 (484)	185 (165)	55 (53)
AF	70,352 (9036)	3063 (1332)	994 (533)	253 (172)
AL	18,369 (5103)	887 (513)	314 (198)	103 (72)
MX	4803 (2681)	210 (179)	71 (61)	21 (18)
RI	5718 (3237)	219 (178)	78 (66)	25 (23)
SE	37,525 (12,232)	1692 (1267)	494 (418)	154 (135)
<b>Total</b>	<b>165,882 (17,015)</b>	<b>7184 (2992)</b>	<b>2301 (1231)</b>	<b>661 (435)</b>

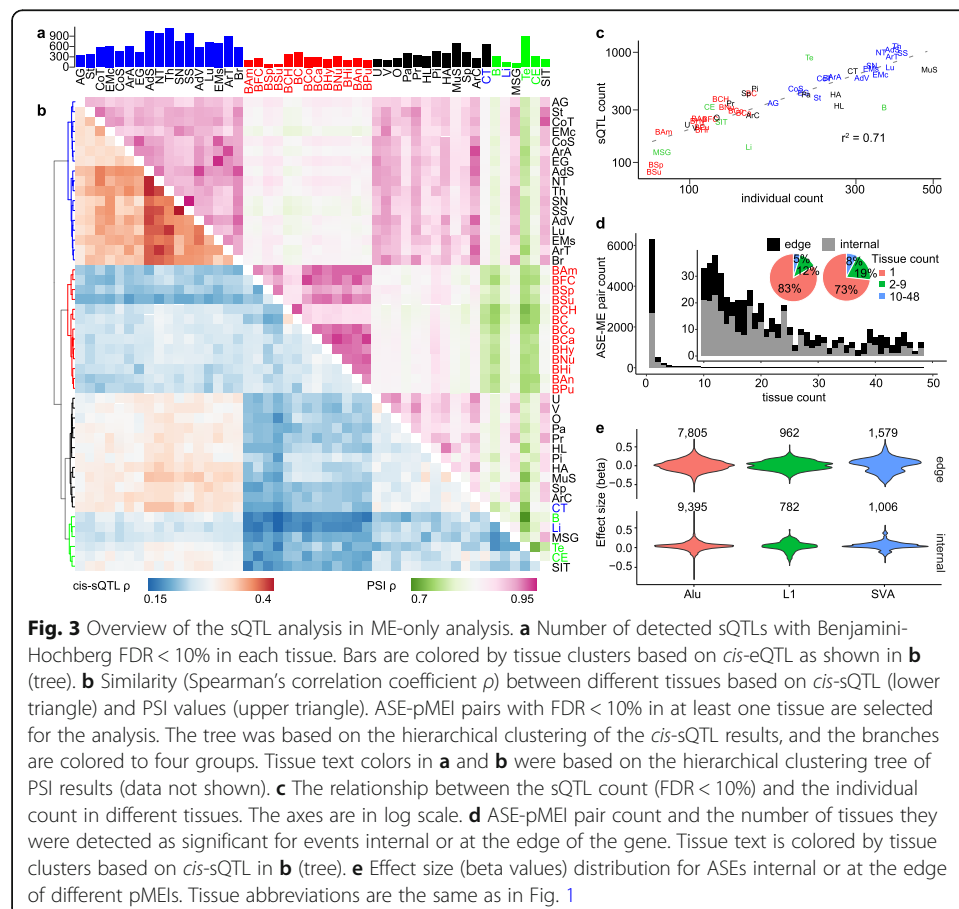
Events in sQTL—number of unique ASEs in the ME-only sQTL analysis with FDR < 10%

ME causal and ME highest causal—number of unique ASEs with pMEIs predicted as a causal variant or a causal variant with the highest probability, respectively

The numbers in the parentheses are the number of genes/gene clusters of the corresponding ASEs. Genes sharing the same exons were merged into gene clusters by SUPPA when calculating PSI scores. Because some genes have multiple ASEs, the overall gene count is not the sum of the gene count in different ASEs

ASEs alternative splicing events, A3/A5 alternative 3’/5’ splice site, AF/AL alternative first/last exon, MX mutually exclusive exon, RI retained exon, SE skipping exon

of sQTLs are from events internal (11,183) and at the edge (10,346) of the gene. The numbers of detected sQTLs with events internal or at the gene edge are proportional to the total number of possible events in these regions, indicating weak or no selective preference. The number of sQTLs in each tissue ranges from 81 to 1120 (Fig. 3a, Additional file 4: Table S3). Similar to eQTLs, the number of sQTLs is highly correlated with the number of donors for each tissue ( $r^2 = 0.71$ , Fig. 3c). Eighty-three percent internal and 73% of gene edge splicing event-pMEI pairs are only detected in one tissue, suggesting the impact of pMEIs on gene splicing is highly tissue-specific (Fig. 3d). Of note, sQTL analysis uses the transcript level PSI information, which is noisier than the gene level TPM used in the eQTL analysis. Therefore, the higher tissue specificity of sQTLs than eQTLs may also be partly due to the lower power and higher level of false negatives in the sQTL analysis. Although sQTLs appear highly tissue-specific, we did identify similarities among related tissues (e.g., brain regions) based on sQTL significance and PSI scores for ASEs (Fig. 3b). We also observed high agreement in the direction of effect by the pMEIs (Additional file 1: Fig. S3), similar to eQTLs. Overall, tissues show more variance based on gene alternative splicing (PSI values) than gene expression levels (TPM values), and the similarity of sQTL and PSI metrics are less than eQTL and TPM metrics. The effect size for sQTLs can be either positive or negative (Fig. 3e), but values of beta are much smaller than eQTLs due to the small variation of PSI values (0–1).



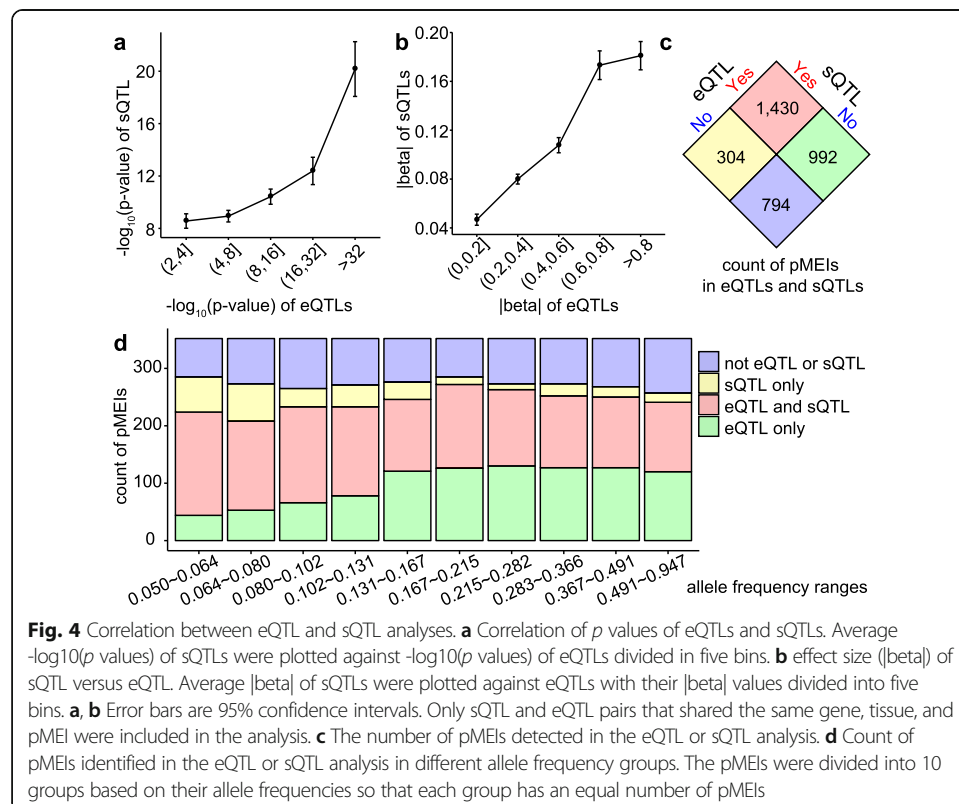


Next, we applied the fine-mapping strategies to identify the causal pMEI-sQTLs. pMEIs were identified as causal for 17.26% (13.10–35.11% among tissues) and as the highest probability causal for 4.33% (2.05–7.38%) of ASEs. The same as eQTLs, pMEIs detected as sQTLs (related) or identified as causal variants for at least one ASE (causal) are significantly enriched in enhancer regions and regions close to the affected genes (Fig. 2f–j). However, the enrichment and significance of pMEIs are lower compared to eQTLs, likely because of the noisier measurement of PSI values than TPM values for eQTL analysis.

To determine if pMEIs affect the expression and splicing of genes simultaneously, we identified genes with both eQTLs and sQTLs. Both the significance and the effect size for eQTLs and sQTLs are positively correlated, indicating that a pMEI that influences the expression of a gene is also likely to impact the alternative splicing and isoform abundance of that gene (Fig. 4a, b). Although ~40% of pMEIs were identified in both eQTL and sQTL analyses, some pMEIs were only identified in one, indicating either a specific functional impact of some pMEIs or different sensitivities of the two analyses (Fig. 4c). pMEIs detected only in sQTL analysis tend to have lower allele frequencies than pMEIs only in the eQTL analysis (Fig. 4d).

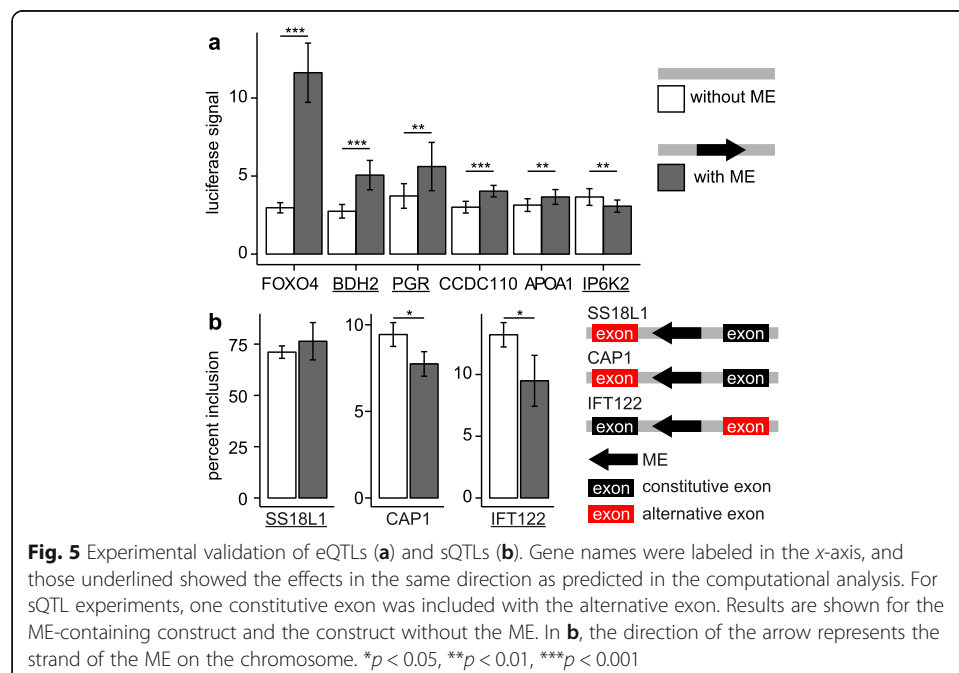
#### Experimental validation of eQTLs and sQTLs

To experimentally verify the predicted impact of specific pMEIs on gene expression and splicing, we evaluated selected loci in ectopic reporter assays (see the “Methods” section for details). We selected loci for validation based on the requirements of ectopic



reporter assays (e.g., pMEI size, sequence availability), the supportive evidence from the eQTL/sQTL analysis, and the importance of the associated genes. For pMEIs predicted to be causal in the eQTL analysis, we selected six loci for experimental validation. All six tested ME loci showed a significant difference in the gene expression between the presence and absence of the pMEI ( $p < 0.05$ , unpaired 2-tailed  $t$  test) (Fig. 5a). The presence of the pMEI resulted in the upregulation of luciferase expression in five cases, with only one locus, *IP6K2*, where the presence of the pMEI reduced the luciferase expression relative to the pre-insertion allele. These results indicate that pMEI in their genomic context can alter transcription levels, supporting their role as eQTLs. Three pMEIs have the same direction of effect (i.e., either up- or downregulation in the presence of the ME) in the reporter assay as predicted computationally for the closest eGenes: *BDH2*, *PGR*, and *IP6K2* (Fig. 5a). Because all three pMEIs are eQTLs in multiple tissues and in all tissues the pMEIs have the same predicted direction of effect, these pMEIs are likely to regulate gene expression across tissue types using a similar mechanism.

Next, we performed an experimental validation of pMEI sQTLs using ectopic reporter assays. We focused on pMEIs within genes and near differentially incorporated exons to enable evaluation with a minigene reporter. We evaluated three pMEI loci for sQTLs and identified significant effects of the ME at two of the three loci ( $p < 0.05$ , unpaired 2-tailed  $t$  test) (Fig. 5b). In both cases, the presence of the ME resulted in less incorporation of the alternatively incorporated exon. We compared these results to the effects predicted for the ME-containing allele in our sQTL analysis. For *IFT122*, we predicted that the presence of the ME would decrease the exon inclusion in all tissues with sQTLs, and this prediction agrees with our ectopic assay. However, for *CAP1*, the predicted effect of the presence of the ME on splicing did not agree with the experimental result. Altogether, these data confirm that pMEI can alter gene expression levels



and isoform proportions largely consistent with the predicted effects in our QTL analysis.

## Discussion

MEs play important roles in gene regulation and have the capacity of creating new gene regulatory networks [18, 37, 38]. However, most previous studies on the impact of pMEIs on gene expression focused only on eQTLs and on LCLs from the 1KGP project [26, 32, 39]. The GTEx project provides an excellent opportunity to study the impact of pMEI on gene expression and alternative splicing across human tissues [30]. Although a previous study from the GTEx Consortium included some pMEIs that are present in the reference genome (rMEIs), the study did not consider non-reference pMEIs, and it is based on the smaller GTEx v6 release (147 individuals, 13 tissues) [32]. In this study, we identified both pMEIs that are present in the reference genome (rMEIs) and absent in the reference genome (nrMEIs) in more than 600 individuals. Combining the genotypes of common pMEIs with the GTEx RNA-seq data, we examined the impact of pMEIs on gene expression and gene splicing comprehensively in 48 tissues.

The high-depth WGS from the GTEx project (mean coverage about 40-fold) resulted in sensitive pMEI identification and accurate genotyping [35]. We identified a total of 20,545 pMEI loci from 639 individuals, including 16,558 non-reference pMEIs and 3987 reference pMEIs. The total number of pMEIs in our study is about ten times more than the 2051 reference pMEIs identified in the previous study [32]. The number is also higher than the 17,934 pMEIs identified in phase 3 of the 1KGP from 2504 individuals, which was based on low-coverage WGS (mean coverage 7.4-fold) [27]. Less than half of the pMEI loci (8456) were identified both in this project and in the 1KGP. Recent studies have been able to continually annotate new pMEIs as techniques improve and individuals from diverse populations are considered [24, 40]. In addition, because of the repetitive nature of MEs, many pMEIs are missed by the current short-read sequencing technology [41]. Therefore, when more diverse populations are included and long-read sequencing technologies are used, we expect a lot more pMEIs will be identified.

We identified 6342 genes with expression levels correlated with 2422 pMEIs and the 2992 transcript splicing events correlated with 1734 pMEIs in at least one of the 48 tested tissues. The number of pMEIs identified as eQTLs (2422) in our study is much higher compared to previous studies [26, 28, 32, 39]; the GTEx structure variation study reported 265 rMEIs at FDR < 10% [32], and the 1KGP study identified 235 pMEIs at FDR ≤ 5% [26]. We also identified a large number of pMEIs as the potential causal variant for eQTLs (956) and sQTLs (866). This highlights the value of both the large number of pMEIs identified from the high-coverage WGS data and the many tissues examined in our study. The numbers of detected eQTLs and sQTLs in each tissue were highly correlated with the sample size of each tissue ( $r^2 = 0.85$  and  $0.71$  for eQTLs and sQTLs, respectively) (Fig. 1c, Fig. 3c). Because the power of the QTL (eQTL and sQTL) analysis is closely related with the sample size, this linear relationship indicates that the sample size is still too small in most tissues. It is likely that many additional QTLs were not detected due to the small sample size in many tissues. In GTEx v8, there is no significant sign of eGenes/sGenes showing plateauing at a sample size of 600 [31],

suggesting more than 600 samples are needed to reach sufficient power to identify all eQTLs and sQTLs.

A previous study showed that different analysis methods can produce very different eQTL results, even with the same raw dataset [28]. It is also known that different ME identification programs have different sensitivity and specificity. In addition, the pMEI selection can further introduce discrepancies among studies [42]. For example, our analyses focused on a small set of common pMEIs, which accounts for only 17% of the high-quality call set. To assess the consistency of our eQTL analysis with other studies, we compared the eQTLs identified in LCLs with an eQTL study of LCLs from 1KGP samples [26]. Our dataset contains 113 individual-derived LCLs, which is much smaller than the 445 LCLs in the 1KGP study [43]. With  $FDR < 5\%$  as a cutoff, we identified 255 pMEI-associated eQTLs in GTEx LCLs. Despite that the differences in sequencing protocols, sample composition, and data processing, 67 of these eQTLs were also identified in the study of 1KGP eQTLs (Additional file 5: Table S4). This result suggests that many of the pMEI-associated eQTLs are strong eQTLs that show consistent signal in individuals from different populations.

The significance of pMEI-associated eQTLs and sQTLs is similar in related tissues (Fig. 1b, Fig. 3b). Except for the testis, tissue pairs also show strong consistency in the direction of the effect for the pMEI in eQTLs and sQTLs (Additional file 1: Fig. S3). Our results agree with a previous study showing that the testis is unique in gene expression compared to other tissues [44]. The overall high consistency of the direction of effects for eQTLs and sQTLs among tissues suggests that when a pMEI is affecting gene expression or splicing in multiple tissues, similar regulators are involved. However, because gene expression and alternative splicing patterns are also correlated among related tissues (Fig. 1b, Fig. 3b), the similarity of eQTLs and sQTLs could also be attributed to the correlated gene expression/splicing patterns among related tissues.

Although the QTL analyses can detect the association of pMEIs with gene expression and splicing changes, they do not provide information on the molecular mechanisms for the effect. By examining the enrichment of pMEIs, we found that pMEIs in regions close to genes (intron, exon, 10 kb upstream or downstream) are more likely to correlate with gene expression and alternative splicing (Fig. 2, Additional file 1: Fig. S2). These pMEIs likely affect *cis*-elements (e.g., promoter, splicing sites) of the associated genes. However, not all pMEIs identified in the eQTL and sQTL analyses are near genes. Many of these pMEIs are far from the associated genes. These pMEIs may impact gene regulation through several mechanisms, such as serving as distal enhancers [45, 46] or altering chromatin looping structure [21, 37]. An interesting observation is that the effects of pMEIs on expression and splicing were highly correlated for some genes (Fig. 4). This may be because the regulation of gene expression was isoform-specific; the pMEI altered the transcript level of specific isoforms and is then detected as both an eQTL and an sQTL. pMEIs with eQTL/sQTL signals are also highly enriched in enhancer regions (Fig. 2a, f). Because enhancers are key regulators for tissue-specific gene expression [47], this enrichment suggests that pMEIs could play a role in regulating tissue-specific expression and splicing.

In addition to the enrichment analysis, we also experimentally validated the predicted impact of several pMEIs using ectopic reporter assays. Such reporter assays are beneficial as several loci can be evaluated quickly to confirm computational predictions.

However, while we have included as much of the endogenous locus as technically feasible, the ectopic assay does not capture the full genomic context of the pMEI. Therefore, locus-dependent or tissue-specific effects may not be recapitulated in the reporter system. Further, the cloned pMEI loci were limited to the subset of pMEIs we could evaluate. In the end, our experiments did validate the predicted effect of most of the tested pMEIs. To fully assess the functional impact of pMEIs, large-scale functional validation, including validation at the endogenous locus, will be needed in the future.

## Conclusions

Overall, our study showed that pMEIs are associated with thousands of gene expression and splicing variations in different tissues. Given the majority of pMEI-associated eQTLs/sQTLs are tissue-specific and pMEIs are enriched in the enhancer regions, pMEIs could have a significant role in regulating tissue-specific gene expression/splicing. Detailed mechanisms for the role of pMEIs in gene regulation in different tissues will be an important direction for future studies.

## Methods

### pMEI identification and filtering

WGS data from the GTEx project v7 release were downloaded from dbGaP (phs000424.v7.p2). Of the 650 individuals in the v7 release, 12 were excluded from the analysis because of issues during the dbGaP retrieval or the read mapping. WGS data from a reference sample HuRef (<https://www.coriell.org/1/HuRef>) was also included for quality control purposes. HuRef DNA sample was purchased from Coriell (NS12911, Camden, NJ, USA), and WGS was performed by Novogene (Sacramento, CA, USA) on the Illumina HiSeq platform using a PCR-free library and the pair-end 150-bp sequencing format.

The Mobile Element Locator Tool (MELT, version 2.1.5) [35] was used to identify pMEIs using the WGS data from the 639 individuals (638 GTEx individuals and HuRef). Briefly, WGS reads were aligned to the human reference genome GRCh38 with a decoy sequence used in the 1KGP [48] using the Burrows-Wheeler Aligner (BWA, ver. 0.7.15) [49]. Output files were sorted and indexed with SAMtools (ver. 1.7) [50]. To identify pMEIs that are not present in the reference genome (nrMEIs), MELT (ver. 2.1.5) was run in the “MELT-SPLIT” mode under the default setting. The “MELT-SPLIT” mode includes five steps: Preprocess, IndivAnalysis, GroupAnalysis, Genotype, and MakeVCF. To identify pMEIs that are present in the reference genome but absent in the sequenced individuals (rMEIs), MELT was run in the “MELT-Deletion” mode which includes two steps: Genotype and Merge. The ME reference files for *Alu*, LINE1, and SVA were downloaded within the MELT program. The final output is three files for nrMEIs and three for rMEIs in the VCF format.

The call sets were filtered to reduce false positives and to focus on common variants. For nrMEIs, loci with < 25% no-call rate, MELT ASSESS score  $\geq 3$ , VCF filter column with “PASS” or “rSD,” and split reads > 2 were kept. For rMEIs, sites with < 25% no-call rate were kept. For both nrMEIs and rMEIs, only loci with allele frequencies between 0.05 and 0.95 in the dataset were kept. Hardy-Weinberg equilibrium test was performed for each locus using individuals with “European” in the race description. Loci

with a  $p$  value  $< 10^{-10}$  were considered low-quality and were excluded from the analysis. The genomic coordinates of the loci were then lifted over from the human reference genome version GRCh38 to GRCh37/hg19 using CrossMap (ver. 0.2.7) [51]. Because of the known low-quality calls on the Y chromosome, only the loci from the autosomes and X chromosomes were used for the downstream analysis.

### **cis-eQTL mapping**

Matrix eQTL (ver. 2.3) was used to identify the association between genotypes and gene expression with a linear regression method [36]. Two genotype files were prepared: one file with only pMEIs for the ME-only analysis and one file with pMEIs plus common SNPs and indels for the joint analysis. The SNP and indel genotypes were obtained from the GTEx project (phs000424.v7.p2, GTEx\_Analysis\_20160115\_v7\_WholeGenomeSeq\_635Ind\_PASS\_AB02\_GQ20\_HETX\_MISS15\_PLINKQC.PIR.vcf). SNPs and indels were filtered to remove sites with  $> 25\%$  no-call rate or with Hardy-Weinberg equilibrium test  $p$  value  $< 10^{-10}$  in “European” individuals as described above.

Gene expression data were downloaded from the GTEx website (<https://gtexportal.org/home/datasets>, GTEx\_Analysis\_2016-01-15\_v7\_RNASeQCv1.1.8\_gene\_tpm.gct.gz, and GTEx\_Analysis\_2016-01-15\_v7\_RNASeQCv1.1.8\_gene\_reads.gct.gz). Normalized expression data of genes in each tissue were generated following the official GTEx QTL pipeline to reduce the effect of technical bias (<https://github.com/broadinstitute/gtex-pipeline/tree/master/qtl>). Briefly, in each tissue, a gene was kept if it has a transcript per million (TPM)  $\geq 0.1$  and a raw read count  $\geq 6$  in  $\geq 20\%$  samples. Read counts among samples were normalized with the method described by [52] to obtain the trimmed mean of  $M$  (TMM) values. Then, TMM values of each gene were inverse normal transformed across the samples in each tissue.

The covariates for each tissue were downloaded from the GTEx website (<https://gtexportal.org/home/datasets>, GTEx\_Analysis\_v7\_eQTL\_covariates.tar.gz). The covariates include sex, three genotyping principal components, sequencing platform, and probabilistic estimation of expression residuals (PEER) factors based on the number of individuals ( $N$ ) in each tissue type (15, 30, and 35 PEERs for  $N < 150$ ,  $150 \leq N < 250$ , and  $N \geq 250$ , respectively) [30, 53]. Input files for Matrix eQTL were generated with Python scripts for each tissue, and Matrix eQTL was run with a window of 1 million bp (Mb) on either side of each gene. The  $p$  value cutoffs ( $-p$ ) were set at 1 for the ME-only analysis and 0.05 for the joint analysis. For the ME-only analysis, all genes were used as input and only eQTLs with FDR  $< 10\%$  by the Benjamini-Hochberg method were used for further analysis. For the joint analysis, in each tissue, only genes reported in ME-only analysis with FDR  $< 10\%$  were used as input. From both eQTL analyses, a gene whose expression level showed an association with a variant with FDR  $< 10\%$  in a given tissue is defined as an eGene. Protein-coding genes and non-coding genes were defined based on GENCODE gene models.

### **cis-sQTL mapping**

TPM value for each transcript and transcript model for each gene were downloaded from the GTEx website (<https://gtexportal.org/home/datasets>, GTEx\_Analysis\_2016-01-15\_v7\_RSEMv1.2.22\_transcript\_tpm.txt.gz, and gencode.v19.transcripts.patched\_contigs.gtf). ASEs were determined using SUPPA2 [54], with “-pool-genes” option enabled to group

genes together if they are on the same genomic strand and share at least one exon. Seven types of ASEs were calculated: skipping exon (SE), alternative 5' splice sites (A5), alternative 3' splice sites (A3), mutually exclusive exons (MX), retained intron (RI), alternative first exons (AF), and alternative last exons (AL). Then, the percent spliced in (PSI) values were calculated by SUPPA2 based on the TPM values of transcripts in each sample. Similar to the eQTL analysis, sex, three genotyping principal components, sequencing platform, and PEER factors were included as covariates. PEER factors of different tissues were calculated by r-peer with PSI values [53]. The number of PEER factors was set based on the number of individuals in each tissue type, the same as in the eQTL analysis. ASEs with empty values were excluded from the r-peer analysis. The cutoff for significant sQTLs was set at 10% FDR.

#### **Fine-mapping of causal variants for each eGene and ASE**

CAVIAR (ver. 2.1) [55] was used to identify causal variants in the associated region for each eGene. CAVIAR takes a linkage disequilibrium file and a *z*-score file as inputs and reports a list of possible causal variants and the posterior probabilities of input variants being causal. pMEIs in the ME-only analysis and the 100 most significant SNPs/indels in the joint analysis were chosen for each FDR-controlled eGene in the ME-only analysis. The signed *r* values for the linkage disequilibrium file were calculated with PLINK (version 1.90), and the *t*-statistic values in Matrix eQTL output were used as the *z*-score. For each eGene, CAVIAR was run under the default setting (rho-prob 0.95, gamma 0.01, causal 1).

To identify causal *cis*-sQTL variants, similar analyses were performed as the eQTL analysis using CAVIAR (ver. 2.1). pMEIs in the ME-only analysis and the 100 most significant SNPs/indels in the joint analysis were chosen for each FDR-controlled ASE in the ME-only analysis. Here, ASEs were used in place of eGenes, and PSI values were used in place of gene expression levels.

#### **Enrichment analysis of pMEIs**

Fisher's exact test was performed to check the enrichment of pMEIs in different regions of the affected genes. To test for enrichment in the eQTL analysis, common pMEIs were grouped into three categories based on their effect on gene expression: pMEIs not correlated with any gene (NS), correlated with at least one gene but not causal (related), and being causal for at least one gene (causal). For pMEIs grouped as NS and related, the affected gene of a pMEI is defined as the gene with the smallest FDR value by Matrix eQTL. For causal pMEIs, the affected gene is defined as the gene with a pMEI as the causal variant and with the smallest eQTL FDR value. pMEIs that are not within the 1-Mb window of any gene were excluded from the analysis. Functional genomic regions include enhancers from the Dragon Enhancers Database (DENdb, <https://www.cbrc.kaust.edu.sa/dendb/src/enhancers.csv.zip>) [56], 10 kb upstream of the transcription starting site (TSS), 10 kb downstream of the affected gene, and exons and introns of the affected gene. For each category, the number of pMEIs in different genomic functional groups was counted, and Fisher's exact test was performed to determine the enrichment of pMEIs in those genomic regions in the related and causal categories relative to the NS category.

The enrichment analysis for sQTLs was performed similarly. For pMEIs grouped as NS and related, the affected ASE of a pMEI is defined as the ASE with the smallest FDR value; for pMEIs grouped as causal, the affected ASE is the ASE with a pMEI as the causal variant and with the smallest FDR value. The affected gene is the gene containing the affected ASE. If ASE includes transcripts from more than one gene, the longest gene among the overlapping genes was used to define the genomic functional groups. pMEIs that are not within the 1-Mb window of any ASE were excluded from the analysis.

#### Dual-luciferase reporter assay for eQTLs

The effects of six representative pMEIs on gene expression were tested using a standard luciferase enhancer assay. For loci where the pMEI was predicted as causal for multiple eGenes, the gene closest to the pMEI location was selected. About 300 bp of each genomic locus encompassing the pMEI insertion site was cloned into a modified pGL4.26 vector (Payer LM, et al.: Alu insertion variants alter gene transcript 722 levels, submitted) using Gateway cloning (Invitrogen). The locus was amplified from 1KGP individuals using the primers listed in Additional file 6: Table S5. For each locus, two independent clones were generated with the pMEI present and two clones without the pMEI. The orientation of the locus and the pMEI relative to the eGene was maintained relative to the luciferase reporter gene. All constructs were verified by Sanger sequencing. The firefly luciferase vectors were each co-transfected with a Renilla plasmid (pRL, Promega) into 293T cells using Fugene HD (Promega). 293T cells were selected because they are easy to transfect and are frequently used in ectopic reporter assays. After 48 h, luciferase levels were measured using the Dual-Glo Luciferase Assay System (Promega) and the GloMax-Multi Detection System (Promega). Firefly and Renilla levels were normalized to the background in wells with no transfected plasmids, and a ratio of firefly to Renilla levels in each well accounted for any differences in transfection efficiency. Results were graphed as relative luciferase units for each construct, and an unpaired 2-tailed *t* test was performed for each locus.

#### Ectopic minigene reporter assay for sQTLs

The effects of four representative pMEIs on alternative splicing were experimentally evaluated with an ectopic minigene reporter assay as previously described [14]. Briefly, for each locus, a genomic fragment surrounding the pMEI and nearby exons was cloned into an intron between rat insulin exons in the pSpliceExpress vector (Addgene) [57] using Gateway cloning (Invitrogen). The region was amplified using primers listed in Additional file 6: Table S5 from the DNA of 1KGP individuals. Two constructs were generated for each evaluated locus: one with the pMEI present and one without the pMEI. Two independent clones were isolated for each construct and verified by Sanger sequencing. The plasmids were transfected (Fugene HD, Promega) into 293T cells, and after 24 h, RNA was extracted (Quick RNA MicroPrep Kit, Zymo Research) and reverse transcribed to cDNA (iScript cDNA Synthesis Kit, BioRad). RT-PCR was performed with primers that bind within the rat insulin exons (Ins1: 5'-CAGCACCTTTGTGGTTCTCA-3' and Ins2: 5'-AGAGCAGATGCTGGTGCAG-3'). For the *IFT122* locus, to enable a sensitive quantification of the rare alternative exon, we increased the



specificity by repeating the RT-PCR with a primer in the constitutive exon from this locus (5'-AAAGTAAAGATCGAGCGGCC-3' paired with Ins2). For each locus, the relative quantification of alternatively spliced RNA isoforms was performed on ethidium bromide-stained agarose gels with band intensities normalized for DNA fragment length. Two transfections were performed for each independent clone of each construct, resulting in four data points for each type of construct (i.e., with or without the pMEI) for each locus. Quantification is graphed as percent of transcripts that include the alternative exon, and unpaired *t* tests compared the percent inclusion when the pMEI was present versus absent at each locus.

### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13059-020-02101-4>.

**Additional file 1: Figure S1.** Overview of high confidence pMEIs in GTEx individuals. **Figure S2.** Enrichment of pMEIs around Transcription Starting Sites (TSSs) of genes. **Figure S3.** Agreement of the effect direction between a pair of tissues for eQTLs (lower-left) and sQTLs (upper-right).

**Additional file 2: Table S1.** Number of samples, expressed genes, and eQTLs for each tissue.

**Additional file 3: Table S2.** Impact of MEs on the direction of gene expression change.

**Additional file 4: Table S3.** Number of samples and sQTLs for each tissue.

**Additional file 5: Table S4.** LCL eQTLs in the current study and an 1KGP study (Spirito et al. 2019).

**Additional file 6: Table S5.** Primers for eQTL and sQTL cloning.

**Additional file 7.** ME-only eQTLs.

**Additional file 8.** ME-only sQTLs.

**Additional file 9.** Review history.

### Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from dbGaP accession number phv00169064.v7.p2 in February 2018. We gratefully acknowledge access to the HPC facilities and support of the computational STEM and bioinformatics scientists from the Office of Advanced Research Computing (OARC) at Rutgers University. We further acknowledge the critical work made possible through access to the Perceval Linux cluster operated by OARC under NIH 1S10OD012346-01A1. This work also used resources from the Rutgers Discovery Informatics Institute (the Caliburn Linux cluster), which are supported by Rutgers and the state of New Jersey.

### Peer review information

Yixin Yao was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Review history

The review history is available as Additional file 9.

### Authors' contributions

XC and JX designed the research. The analysis was performed primarily by XC and YZ, and some by HL and JX. The validation experiments were performed by LP, JS, and KB. XC, YZ, LP, and JX wrote the draft of the manuscript. All authors read and approved the final version of the manuscript.

### Funding

This study was supported by the startup fund from the Human Genetics Institute of New Jersey to JX.

### Availability of data and materials

The VCF files of individual pMEI genotypes are available under the dbGaP project "Impact of Mobile Element Insertions on Human Transcriptome Variation" (Study Accession: phs002030) [58]. The raw results of the eQTL and sQTL analyses are listed in Additional file 7 and Additional file 8.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

**Competing interests**

The authors declare no competing interests.

**Author details**

<sup>1</sup>Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. <sup>2</sup>Human Genetic Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. <sup>3</sup>Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

Received: 18 May 2020 Accepted: 14 July 2020

Published online: 27 July 2020

**References**

- Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS, et al. Ten things you should know about transposable elements. *Genome Biol.* 2018;19:199.
- de Koning APJ, Gu WJ, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 2011;7:e1002384.
- Batzer MA, Deininger PL. A human-specific subfamily of Alu-sequences. *Genomics.* 1991;9:481–7.
- Brouha B, Schustak J, Badge RM, Lutz-Prigget S, Farley AH, Moran JV, Kazazian HH. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A.* 2003;100:5280–5.
- Wang H, Xing J, Grover D, Hedges DJ, Han KD, Walker JA, Batzer MA. SVA elements: a hominid-specific retroposon family. *J Mol Biol.* 2005;354:994–1007.
- Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet.* 2003;73:1444–51.
- Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol.* 2001;21:1429–39.
- Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Lower J, Stratling WH, Lower R, Schumann GG. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* 2012;40:1666–83.
- Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet.* 2000;24:363–7.
- Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell.* 2002;110:327–38.
- Guffanti G, Gaudi S, Fallon JH, Sobell J, Potkin SG, Pato C, Macchiardi F. Transposable elements and psychiatric disorders. *Am J Med Genet Part B-Neuropsychiatric Genet.* 2014;165:201–16.
- Payer LM, Burns KH. Transposable elements in human genetic disease. *Nat Rev Genet.* 2019;20:760–72.
- Kazazian HH Jr, Moran JV. Mobile DNA in health and disease. *N Engl J Med.* 2017;377:361–70.
- Payer LM, Steranka JP, Ardeljan D, Walker J, Fitzgerald KC, Calabresi PA, Cooper TA, Burns KH. Alu insertion variants alter mRNA splicing. *Nucleic Acids Res.* 2019;47:421–31.
- Chen LL, Yang L. ALU alternative regulation for gene expression. *Trends Cell Biol.* 2017;27:480–90.
- Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet.* 2017;18:71–86.
- Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 2003;19:68–72.
- Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science.* 2016;351:1083–7.
- Gal-Mark N, Schwartz S, Ast G. Alternative splicing of Alu exons - two arms are better than one. *Nucleic Acids Res.* 2008;36:2012–23.
- Conley AB, Jordan IK. Cell type-specific termination of transcription by transposable element sequences. *Mob DNA.* 2012;3:15.
- Diehl AG, Ouyang N, Boyle AP. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat Commun.* 2020;11:1796.
- Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, Jorde LB. Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.* 2009;19:1516–26.
- Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP, et al. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* 2011;7:e1002236.
- Loh JW, Ha H, Lin T, Sun N, Burns KH, Xing J. Integrated mobile element scanning (ME-Scan) method for identifying multiple types of polymorphic mobile element insertions. *Mob DNA.* 2020;11:12.
- Wang L, Norris ET, Jordan IK. Human retrotransposon insertion polymorphisms are associated with health and disease via gene regulatory phenotypes. *Front Microbiol.* 2017;8:1418.
- Spirito G, Mangoni D, Sanges R, Gustincich S. Impact of polymorphic transposable elements on transcription in lymphoblastoid cell lines from public data. *Bmc Bioinformatics.* 2019;20:495.
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015;526:75.
- Goubert C, Zavallos NA, Feschotte C. Contribution of unfixed transposable element insertions to human regulatory variation. *Philos Trans R Soc Lond Ser B Biol Sci.* 2020;375:20190331.
- Ardlie KG, DeLuca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science.* 2015;348:648–60.
- Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, Mohammadi P, Park Y, Parsana P, Segrè AV, et al. Genetic effects on gene expression across human tissues. *Nature.* 2017;550:204–13.
- Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, Kasela S, Kim-Hellmuth S, Liang Y, Oliva M, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* 2019:787903.

32. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Consortium GT, et al. The impact of structural variation on human gene expression. *Nat Genet.* 2017;49:692–9.
33. Li X, Kim Y, Sang EKT, Davis JR, Damani FN, Hiang CC, Hess GT, Zappala Z, Strober BJ, Scott AJ, et al. The impact of rare variation on gene expression across tissues. *Nature.* 2017;550:239.
34. Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M. The impact of short tandem repeat variation on gene expression. *Nat Genet.* 2019;51:1652.
35. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Genomes Project C, Devine SE. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* 2017;27:1916–29.
36. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics.* 2012;28:1353–8.
37. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* 2014;24:1963–76.
38. Buzdin AA, Prassolov V, Garazha AV. Friends-enemies: endogenous retroviruses are major transcriptional regulators of human DNA. *Front Chem.* 2017;5:35.
39. Wang L, Rishishwar L, Marino-Ramirez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Res.* 2017;45:2318–28.
40. Witherspoon DJ, Zhang Y, Xing J, Watkins WS, Ha H, Batzer MA, Jorde LB. Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. *Genome Res.* 2013;23:1170–81.
41. Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, Moran JV, Mills RE. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* 2020;48:1146–63.
42. Vendrell-Mir P, Barteri F, Merenciano M, Gonzalez J, Casacuberta JM, Castanera R. A benchmark of transposon insertion detection tools using real data. *Mob DNA.* 2019;10:53.
43. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* 2013;501:506–11.
44. Djureinovic D, Fagerberg L, Hallstrom B, Danielsson A, Lindskog C, Uhlen M, Ponten F. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Mol Hum Reprod.* 2014;20:476–88.
45. Chuong EB, Rumi MA, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet.* 2013;45:325–9.
46. Garcia-Perez JL, Widmann TJ, Adams IR. The impact of transposable elements on mammalian development. *Development.* 2016;143:4101–14.
47. Ko JY, Oh S, Yoo KH. Functional enhancers as master regulators of tissue-specific gene regulation and cancer development. *Mol Cells.* 2017;40:169–77.
48. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, GA MV, Abecasis GR. A global reference for human genetic variation. *Nature.* 2015;526:68–74.
49. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754–60.
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford).* 2009;25:2078–9.
51. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics.* 2014;30:1006–7.
52. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010;11:R25.
53. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc.* 2012;7:500–7.
54. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyraas E. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* 2018;19:40.
55. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. *Genetics.* 2014;198:497–508.
56. Ashoor H, Klefogiannis D, Radovanovic A, Bajic VB. DENdb: database of integrated human enhancers. *Database (Oxford).* 2015;2015:bav085.
57. Kishore S, Khanna A, Stamm S. Rapid generation of splicing reporters with pSpliceExpress. *Gene.* 2008;427:104–10.
58. Cao X, Zhang Y, Payer LM, Lords H, Steranka JP, Burns KH, Xing J. Impact of mobile element insertions on human transcriptome variation. *Datasets. dbGAP.* 2020. [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs002030.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002030.v1.p1). Accessed 10 July 2020.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

