The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans

The GTEx Consortium*

Understanding the functional consequences of genetic variation, and how it affects complex human disease and quantitative traits, remains a critical challenge for biomedicine. We present an analysis of RNA sequencing data from 1641 samples across 43 tissues from 175 individuals, generated as part of the pilot phase of the Genotype-Tissue Expression (GTEx) project. We describe the landscape of gene expression across tissues, catalog thousands of tissue-specific and shared regulatory expression quantitative trait loci (eQTL) variants, describe complex network relationships, and identify signals from genome-wide association studies explained by eQTLs. These findings provide a systematic understanding of the cellular and biological consequences of human genetic variation and of the heterogeneity of such effects among a diverse set of human tissues.

Over the past decade, there has been a marked increase in our understanding of the role of genetic variation in complex traits and human disease, especially via genome-wide association studies (GWAS) that have cataloged thousands of common genetic variants affecting human diseases and other traits (1-3). However, the molecular mechanisms by which this genetic variation predisposes individuals to disease are still poorly characterized, impeding the development of therapeutic interventions.

The majority of GWAS variants are noncoding, likely manifesting their effects via the regulation of gene expression (4, 5). Thus, characterization of the regulatory architecture of the human genome is essential, not only for understanding basic biology but also for interpreting GWAS loci. Expression quantitative trait locus (eQTL) analysis (6-8) is the most common approach used to dissect the effects of genetic variation on gene expression. However, comprehensive eQTL data from a range of human tissues are lacking, and eQTL databases are biased toward the most accessible tissues. Additionally, although many regulatory regions act in a tissue-specific manner (9, 10), it is unknown whether genetic variants in regulatory regions have tissue-specific effects as well. Complex diseases are often caused by the dysfunction of multiple tissues or cell types, such as pancreatic islets, adipose, and skeletal muscle for type 2 diabetes (11, 12), so it is not obvious a priori what the causal tissue(s) are for any given GWAS locus or disease. Hence, understanding the role of regulatory variants, and the tissues in which they act, is essential for the functional interpretation of GWAS loci and insights into disease etiology. The Genotype-Tissue Expression (GTEx) Project was designed to address this limitation by establishing a sample and data resource to enable studies of the relationship among genetic variation, gene expression, and other molecular phenotypes in multiple human tissues (13). To facilitate the collection of multiple different tissues per donor, the project obtains recently deceased donors through consented next-of-kin donation, from organ donation and rapid autopsy settings. The results described here were generated during the project’s pilot phase, prior to scaling up collection to 900 donors. All project data are made available at regular intervals to qualified researchers through dbGaP. Summary data are available on the GTEx Portal (http://gtexportal.org).

Study design

During the pilot, we recruited 237 postmortem donors, collecting an average of 28 tissue samples per donor spanning 54 distinct body sites (fig. S1 and tables S1 and S2). Blood-derived DNA samples were genotyped at approximately 4.3 million sites, with additional variants imputed using the 1000 Genomes phase I, resulting in ~6.8 million single-nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) of ~5% after quality control (tables S3 to S5) (14).

RNA was extracted from all tissues, but quality varied widely, with tissue site and sample specific ischemic time accounting for ~40% of the variance in RNA quality (fig. S2). To maximize statistical power, we prioritized RNA sequencing of samples from nine tissues that were most frequently collected and that routinely met minimum RNA quality criteria: adipose (subcutaneous), tibial artery, heart (left ventricle), lung, muscle (skeletal), tibial nerve, skin (Sun-exposed), thyroid, and whole blood (Table 1) (14).

We performed 76-base pair (bp) paired-end mRNA sequencing on a total of 1749 samples, of which 1641 samples from 43 sites, and 175 donors, constituted our final “pilot data freeze” reported on here (14). Median sequencing depth was 82.1 million mapped reads per sample (fig. S3A). The final data freeze included samples from 43 body sites: 29 solid-organ tissues, 11 brain sub-regions (with two duplicated regions), a whole-blood sample, and two cell lines derived from donor blood (EBV-transformed lymphoblastoid cell lines (LCLs)) and skin samples (cultured fibroblasts) (Table 1 and tables S1 and S2). Median sample size for the nine high-priority tissues was 105; median sample size for the other 34 sampled sites was 18.5.

Gene expression across tissues

We examined the patterns of expression of 53,934 transcribed genes across tissues [on the basis of Gencode V12 annotations] (14, 15). The number of biotypes [protein-coding genes, pseudogenes, and long noncoding RNAs (lncRNAs)] that were transcribed above a minimal threshold [reads per kilobase per million mapped reads (RPKM) > 0.1] was similar for most tissues (average of 20,940) (fig. S3B). Testis was an outlier, with substantially more transcribed regions detected than other tissues (31,240 on average), many of which are lncRNAs. Whole blood was also an outlier, exhibiting the fewest detected transcribed regions (17,160 on average).

Hierarchical clustering demonstrated that expression profiles accurately recapitulate tissue type, with blood samples forming the primary outgroup (Fig. 1A). The multiple brain regions cluster strongly together as a single unit, but among those the 11 individual sub-sampled regions are less distinct (Fig. 1A and fig. S4A). The most distinct brain region is the cerebellum (fig. S4A) (16), with preservation method having little impact on that signal (fig. S4B). The distribution of gene expression across tissues is described by Melé et al. (17), who show that tissue-specific transcription is typically dominated by a few genes that vary from tissue to tissue.

We quantified splicing events (splice junctions, exons, transcripts) by estimating exon inclusion levels, measured as PSI (percent spliced in) scores (14, 18). Clustering samples by PSI scores also largely, although less clearly, recapitulates tissue type. Samples from the brain, not blood, form the primary outgroup (Fig. 1B), which is divided into two groups: A group of 227 samples (from the cerebellum and cortex) forms an independent subcluster (cluster 1), and a smaller group of 97 samples (cluster 2, dominated by the remaining subregions) clusters closer to samples from the rest of the tissues.
Table 1. GTEx pilot samples. Characteristics of the 1641 RNA-sequenced samples included in the pilot data freeze. The second and third columns show the tissue abbreviation and color assigned to each tissue (used throughout). The nine tissues prioritized for sequencing are indicated by red stars. Boxes highlight two regions of the brain that were sampled in duplicate. A region each from the cerebellum and cortex (BRNCHA and BRNCTXA) was sampled on site during initial tissue collection, and again after the brain was received by the brain bank (BRNCHB and BRNCTXB). Cell lines included are an EBV-transformed lymphoblastoid cell line from blood (LCL) and a cultured primary fibroblast cell line from fresh skin (FIBRBLS). RIN, RNA integrity number.

<table>
<thead>
<tr>
<th>Tissue Site Detail</th>
<th>Abbreviation</th>
<th>Color</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Age</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adipose - Subcutaneous</td>
<td>ADPSBO</td>
<td></td>
<td>94</td>
<td>6.9</td>
<td>0.6</td>
<td>421.9</td>
<td>318.3</td>
<td>48.2</td>
<td></td>
</tr>
<tr>
<td>Adipose - Visceral (Omentum)</td>
<td>ADPVSC</td>
<td></td>
<td>19</td>
<td>7.2</td>
<td>0.6</td>
<td>401.4</td>
<td>303.1</td>
<td>46.9</td>
<td></td>
</tr>
<tr>
<td>Adrenal Gland</td>
<td>ADRNLG</td>
<td></td>
<td>12</td>
<td>8.3</td>
<td>0.8</td>
<td>172.8</td>
<td>107.2</td>
<td>51.4</td>
<td></td>
</tr>
<tr>
<td>Artery - Aorta</td>
<td>ARTACRT</td>
<td></td>
<td>24</td>
<td>7.8</td>
<td>0.8</td>
<td>280.5</td>
<td>178.3</td>
<td>50.8</td>
<td></td>
</tr>
<tr>
<td>Artery - Coronary</td>
<td>ARTCRN</td>
<td></td>
<td>9</td>
<td>7.5</td>
<td>0.7</td>
<td>312.4</td>
<td>307.9</td>
<td>55.3</td>
<td></td>
</tr>
<tr>
<td>Artery - Tibial</td>
<td>ARTTBL</td>
<td></td>
<td>112</td>
<td>7.5</td>
<td>0.7</td>
<td>486.7</td>
<td>350.0</td>
<td>47.9</td>
<td></td>
</tr>
<tr>
<td>Brain - Amygdala</td>
<td>BRNAMY</td>
<td></td>
<td>23</td>
<td>6.9</td>
<td>0.8</td>
<td>NA</td>
<td>NA</td>
<td>51.2</td>
<td></td>
</tr>
<tr>
<td>Brain - Anterior cingulate cortex (BA24)</td>
<td>BRNACC</td>
<td></td>
<td>17</td>
<td>7.0</td>
<td>0.7</td>
<td>NA</td>
<td>NA</td>
<td>51.4</td>
<td></td>
</tr>
<tr>
<td>Brain - Caudate (basal ganglia)</td>
<td>BRNCDT</td>
<td></td>
<td>36</td>
<td>7.5</td>
<td>0.8</td>
<td>NA</td>
<td>NA</td>
<td>52.5</td>
<td></td>
</tr>
<tr>
<td>Brain - Cerebellum [PAXgene]</td>
<td>BRNCHA</td>
<td></td>
<td>30</td>
<td>7.4</td>
<td>1.0</td>
<td>868.9</td>
<td>300.9</td>
<td>51.6</td>
<td></td>
</tr>
<tr>
<td>Brain - Cerebellar Hemisphere [Frozen]</td>
<td>BRNBC</td>
<td></td>
<td>24</td>
<td>7.6</td>
<td>1.1</td>
<td>NA</td>
<td>NA</td>
<td>49.7</td>
<td></td>
</tr>
<tr>
<td>Brain - Cortex [PAXgene]</td>
<td>BRNCTXA</td>
<td></td>
<td>23</td>
<td>7.1</td>
<td>0.9</td>
<td>837.2</td>
<td>280.3</td>
<td>51.3</td>
<td></td>
</tr>
<tr>
<td>Brain - Frontal Cortex (BA9) [Frozen]</td>
<td>BRNCTXB</td>
<td></td>
<td>24</td>
<td>7.5</td>
<td>0.9</td>
<td>NA</td>
<td>NA</td>
<td>55.1</td>
<td></td>
</tr>
<tr>
<td>Brain - Hippocampus</td>
<td>BRNHFP</td>
<td></td>
<td>24</td>
<td>6.9</td>
<td>0.7</td>
<td>NA</td>
<td>NA</td>
<td>51.1</td>
<td></td>
</tr>
<tr>
<td>Brain - Hypothalamus</td>
<td>BRNHPT</td>
<td></td>
<td>23</td>
<td>7.0</td>
<td>0.9</td>
<td>NA</td>
<td>NA</td>
<td>51.4</td>
<td></td>
</tr>
<tr>
<td>Brain - Nucleus accumbens (basal ganglia)</td>
<td>BRNCC</td>
<td></td>
<td>25</td>
<td>7.4</td>
<td>0.6</td>
<td>NA</td>
<td>NA</td>
<td>53.4</td>
<td></td>
</tr>
<tr>
<td>Brain - Putamen (basal ganglia)</td>
<td>BRNPTM</td>
<td></td>
<td>20</td>
<td>7.3</td>
<td>0.9</td>
<td>NA</td>
<td>NA</td>
<td>49.5</td>
<td></td>
</tr>
<tr>
<td>Brain - Spinal cord (cervical c-1)</td>
<td>BRNSCC</td>
<td></td>
<td>16</td>
<td>7.1</td>
<td>0.7</td>
<td>NA</td>
<td>NA</td>
<td>52.5</td>
<td></td>
</tr>
<tr>
<td>Brain - Substantia nigra</td>
<td>BRNSNG</td>
<td></td>
<td>25</td>
<td>6.8</td>
<td>0.7</td>
<td>NA</td>
<td>NA</td>
<td>53.8</td>
<td></td>
</tr>
<tr>
<td>Breast - Mammary Tissue</td>
<td>BREAST</td>
<td></td>
<td>21</td>
<td>7.0</td>
<td>0.7</td>
<td>645.6</td>
<td>425.4</td>
<td>50.3</td>
<td></td>
</tr>
<tr>
<td>Cells - EBV-transformed lymphocytes</td>
<td>LCL</td>
<td></td>
<td>39</td>
<td>9.9</td>
<td>0.2</td>
<td>59.7</td>
<td>502.0</td>
<td>46.2</td>
<td></td>
</tr>
<tr>
<td>Cells - Transformed fibroblasts</td>
<td>FIBRBLS</td>
<td></td>
<td>14</td>
<td>9.5</td>
<td>0.5</td>
<td>544.5</td>
<td>478.4</td>
<td>49.2</td>
<td></td>
</tr>
<tr>
<td>Colon - Transverse</td>
<td>CLNTRN</td>
<td></td>
<td>12</td>
<td>7.5</td>
<td>0.8</td>
<td>236.5</td>
<td>173.0</td>
<td>46.3</td>
<td></td>
</tr>
<tr>
<td>Esophagus - Mucosa</td>
<td>ESPMCS</td>
<td></td>
<td>18</td>
<td>8.6</td>
<td>0.7</td>
<td>330.9</td>
<td>219.6</td>
<td>51.7</td>
<td></td>
</tr>
<tr>
<td>Esophagus - Muscularis</td>
<td>ESPMSL</td>
<td></td>
<td>20</td>
<td>7.9</td>
<td>0.6</td>
<td>310.6</td>
<td>273.3</td>
<td>48.5</td>
<td></td>
</tr>
<tr>
<td>Fallopian Tube</td>
<td>FLPT</td>
<td></td>
<td>1</td>
<td>6.1</td>
<td></td>
<td>52.0</td>
<td></td>
<td>51.0</td>
<td></td>
</tr>
<tr>
<td>Heart - Atrial Appendage</td>
<td>HRTA</td>
<td></td>
<td>25</td>
<td>7.5</td>
<td>0.7</td>
<td>492.3</td>
<td>323.9</td>
<td>51.0</td>
<td></td>
</tr>
<tr>
<td>Heart - Left Ventricle</td>
<td>HRTLV</td>
<td></td>
<td>83</td>
<td>7.9</td>
<td>0.9</td>
<td>380.9</td>
<td>334.4</td>
<td>48.0</td>
<td></td>
</tr>
<tr>
<td>Kidney - Cortex</td>
<td>KDNCTX</td>
<td></td>
<td>3</td>
<td>6.8</td>
<td>0.4</td>
<td>583.0</td>
<td>585.0</td>
<td>56.3</td>
<td></td>
</tr>
<tr>
<td>Liver</td>
<td>LIVER</td>
<td></td>
<td>5</td>
<td>7.6</td>
<td>0.9</td>
<td>365.2</td>
<td>321.8</td>
<td>42.8</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>LUNG</td>
<td></td>
<td>119</td>
<td>7.6</td>
<td>0.9</td>
<td>447.0</td>
<td>380.4</td>
<td>48.9</td>
<td></td>
</tr>
<tr>
<td>Muscle - Skeletal</td>
<td>MSCLSK</td>
<td></td>
<td>138</td>
<td>7.9</td>
<td>0.6</td>
<td>486.1</td>
<td>358.1</td>
<td>49.2</td>
<td></td>
</tr>
<tr>
<td>Nerve - Tibial</td>
<td>NERVET</td>
<td></td>
<td>88</td>
<td>7.0</td>
<td>0.7</td>
<td>463.8</td>
<td>315.6</td>
<td>49.4</td>
<td></td>
</tr>
<tr>
<td>Ovary</td>
<td>OVARY</td>
<td></td>
<td>6</td>
<td>7.3</td>
<td>0.7</td>
<td>401.0</td>
<td>248.4</td>
<td>43.5</td>
<td></td>
</tr>
<tr>
<td>Pancreas</td>
<td>PNCRAES</td>
<td></td>
<td>19</td>
<td>6.8</td>
<td>0.7</td>
<td>200.4</td>
<td>115.3</td>
<td>49.3</td>
<td></td>
</tr>
<tr>
<td>Pituitary</td>
<td>PTTYR</td>
<td></td>
<td>13</td>
<td>7.3</td>
<td>0.6</td>
<td>841.3</td>
<td>305.5</td>
<td>51.9</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>PRSTTE</td>
<td></td>
<td>9</td>
<td>6.8</td>
<td>0.7</td>
<td>231.1</td>
<td>84.0</td>
<td>50.2</td>
<td></td>
</tr>
<tr>
<td>Skin - Not Sun Exposed (Suprapubic)</td>
<td>SKINS</td>
<td></td>
<td>23</td>
<td>7.2</td>
<td>0.7</td>
<td>557.2</td>
<td>388.2</td>
<td>48.7</td>
<td></td>
</tr>
<tr>
<td>Skin - Sun Exposed (Lower leg)</td>
<td>SKINS</td>
<td></td>
<td>96</td>
<td>7.0</td>
<td>0.7</td>
<td>498.5</td>
<td>364.0</td>
<td>49.0</td>
<td></td>
</tr>
<tr>
<td>Stomach</td>
<td>STMAC</td>
<td></td>
<td>12</td>
<td>7.4</td>
<td>0.9</td>
<td>250.0</td>
<td>131.6</td>
<td>47.8</td>
<td></td>
</tr>
<tr>
<td>Testis</td>
<td>TESTIS</td>
<td></td>
<td>14</td>
<td>7.0</td>
<td>0.9</td>
<td>293.6</td>
<td>236.3</td>
<td>52.4</td>
<td></td>
</tr>
<tr>
<td>Thyroid</td>
<td>THYROID</td>
<td></td>
<td>105</td>
<td>7.0</td>
<td>0.7</td>
<td>428.5</td>
<td>391.7</td>
<td>49.2</td>
<td></td>
</tr>
<tr>
<td>Uterus</td>
<td>UTERUS</td>
<td></td>
<td>7</td>
<td>7.4</td>
<td>1.0</td>
<td>313.3</td>
<td>184.3</td>
<td>48.7</td>
<td></td>
</tr>
<tr>
<td>Vagina</td>
<td>VAGINA</td>
<td></td>
<td>8</td>
<td>7.5</td>
<td>1.1</td>
<td>414.5</td>
<td>234.4</td>
<td>54.0</td>
<td></td>
</tr>
<tr>
<td>Whole Blood</td>
<td>WHLBLD</td>
<td></td>
<td>156</td>
<td>8.1</td>
<td>0.8</td>
<td>238.4</td>
<td>498.2</td>
<td>49.7</td>
<td></td>
</tr>
</tbody>
</table>

This is consistent with isof orm regulation playing a comparatively larger role in defining cellular specificity in the brain (18, 19). These analyses are extended in Melé et al. (17) to define tissue-specific splicing signatures and to investigate in depth the role of individual variation in splicing.

eQTL analyses: Single-tissue eQTL analysis

A primary goal of the GTEx project is to identify eQTLs for all genes for a range of human tissues. Past studies, hampered by the difficulties of obtaining human tissue samples, have typically examined no more than three tissues (8, 20). Although our pilot sample sizes are modest for eQTL discovery, the breadth of tissues provides an opportunity to assess differential eQTL discovery among tissues. Because of our small sample sizes, we primarily examined eQTLs that act in cis to the gene (cis-eQTLs; see box S1), as the expected effect size of trans-eQTLs (box S1) is too low to be efficiently detected at this time. We calculated cis-eQTLs separately for each of the nine tissues with sufficient sample sizes (>80 donors) for all SNPs within ±1 Mb of the transcriptional start site (TSS) of each gene (14). Significance correlations between genotypes and gene expression levels were determined by linear regression on quantile normalized gene-level expression values, after correction for known and inferred technical covariates (fig. S8) (14), using Matrix eQTL (21). To obtain gene-specific significance levels while correcting for testing multiple SNPs per gene, we computed permutation-adjusted P values for each gene for the most significant.
SNP per gene (14). We defined “eGenes” as genes with at least one SNP in cis significantly associated, at a false discovery rate (FDR) of ≤0.05, with expression differences of that gene (box S1) (14). A list of the significant SNP-gene pairs detected per eGene can be found on the GTEx Portal (http://gtexportal.org).

The number of eGenes ranged from 919 in heart to 2244 in thyroid, with a total of 6486 unique eGenes across the nine tissues (Fig. 2A). Rerunning the analysis on successively downsampled donor subsets from each tissue showed an approximately linear relationship between eGenes and sample size (slope of ~21 eGenes per sample; Fig. 2A). Interestingly, thyroid and nerve share a steeper slope with ~29 significant eGenes per sample, whereas muscle and blood share a shallower trajectory with ~15 eGenes per sample, whereas muscle and blood share a shallower trajectory with ~15 eGenes per sample, which may reflect the lower transcript complexity observed for these two tissues (17). The number of eGenes identified showed no signs of plateauing at current sample sizes in any of the tissues.

Consistent with previous work (20, 22), the majority of the significant cis-eQTLs clustered around the TSS of target genes in all nine tissues (fig. S9, A and B, and fig. S10). The eQTL signals also tended to show an upstream bias (fig. S10 and tables S6 and S7); an average of ~80% of significant eQTLs fell within ±100 kb around the TSS, and ~60% of all eQTLs were upstream despite testing a similar fraction of SNPs upstream and downstream of the TSS (tables S6 and S7) (23). A slight but distinct overrepresentation of nonsignificant eQTLs around the TSS (relative to other SNPs near the gene; red versus black line in figs. S8A and S11A) supports the existence of additional eGenes, which did not meet significance with current sample sizes.

To investigate the sensitivity and validity of our study, in particular because we used tissues from deceased donors, we compared the GTEx blood eQTLs to a previously reported eQTL study of whole blood samples in ~5300 individuals (7). Although many experimental and processing differences exist between these studies, a considerable fraction of GTEx eGenes (68%) were replicated in this study at FDR < 5% (14). Given the incomplete overlap in variants tested, we also compared our eQTLs against a smaller study of 911 blood samples taken from the Estonian Biobank, where we were able to apply a similar eQTL analysis pipeline to that used in GTEx and hence get better SNP coverage (fig. S11) (14). Notably, 98% of GTEx blood eQTLs at FDR < 5% showed consistent allelic direction with those eQTLs ($P < 10^{-200}$, binomial test) (14).

**Multitissue joint discovery of eQTLs: Tissue specificity and sharing of eQTLs**

The specificity or sharing of eQTLs among different tissues and cell types is of considerable biological interest (8, 22, 24), yielding insights into differential genetic regulation among tissues. Furthermore, cross-referencing tissue-specific eQTLs with disease genetic associations could help identify tissues most relevant to disease biology. We used the GTEx pilot data to examine eQTL sharing across multiple tissues and leveraged the large tissue range to discover weak but constitutively active eQTLs.

We investigated patterns of eQTL sharing using 22,286 genes (with RPKM > 0.1 in at least 10 samples) for each of the nine tissues with both a simple non-model-based analysis of every pair of tissues (22) and more sophisticated Bayesian models for joint analysis of...
all nine tissues (24, 25). Analyses focused on a ±100-kb window surrounding the TSS for each gene, which is smaller than used for the single-tissue analysis, as this is where we observe the highest eQTL density.

The non-model-based pairwise analysis method identifies significant SNP-gene pairs in a first tissue, and then uses the distribution of the $P$ values for these pairs in the second tissue to estimate $\pi_1$, the proportion of non-null associations in the second tissue (22, 26). Estimated values of $\pi_1$ ranged from 0.54 to 0.90 (Fig. 2B). For every pair of tissues, we observed a high level of sharing of eQTLs, and this pairwise sharing of eQTL among tissues roughly mirrored the tissue gene expression correlations (fig. S12).

To assess patterns of sharing beyond tissue pairs, we applied two Bayesian methods that assess for each of the 512 (29) possible null/alternative pairs, we applied two Bayesian methods that (fig. S12).

To assess patterns of sharing beyond tissue pairs, we applied two Bayesian methods that assess for each of the 512 (29) possible null/alternative pairs, we applied two Bayesian methods that (fig. S12).

We extended the model to (i) support the fact that not all tissues were sampled across all GTEx donors; (ii) calculate a gene-level FDR without requiring permutations (27); and (iii) include a fine mapping approach across multiple tissues (28). We also used a “SNP-based” approach (25), which assesses association of each SNP-gene pair separately, working directly with the $z$-statistics for each SNP-gene pair within each tissue [see also (29)]. We summarized the estimates of tissue specificity using marginal posterior probabilities for the number of tissues in which a randomly selected gene (gene-based model) or SNP-gene pair (SNP-based model) is active (Fig. 2C). Both approaches show a U-shaped pattern, with high tissue specificity (activity in a single tissue) or tissue ubiquity (activity in all nine tissues) more common than profiles involving only a few tissues, despite many more possible combinatorial patterns for intermediate specificity. Notably, both methods indicate that more than 50% of all detected eQTLs are common to all nine tissues. Reassuringly, both Bayesian methods produce pairwise tissue sharing probabilities that show agreement with the non-model-based analysis (fig. S13). Figure S14 illustrates the value of the multitissue analysis for an example in which the tissue specificity of an eQTL supports NDRG4 as a candidate to influence QT interval in the heart (30). Despite superficial similarity between patterns of tissue-tissue gene expression and eQTL sharing, extensive comparisons of eQTL evidence to expression levels indicate that the tissue-specific patterns are only modestly correlated with average tissue expression levels (fig. S15).

We also examined effect size estimates for eQTLs that are shared between tissues. The vast majority of shared eQTLs show consistent effect directions in different tissues. However, some SNPs showed apparent opposite effect directions in different tissues (fig. S16). A number of examples that we investigated using multi-SNP multitissue analyses (28) appeared to be due to pairs of SNPs, in LD with one another, that were separate eQTLs in different tissues (fig. S17) rather than being a single eQTL with opposite effects.

Beyond the biological interest of eQTL sharing among tissues, the availability of multiple tissues can also increase the power to detect eQTLs active in multiple tissues by combining information across tissues (24). To investigate, we conducted a permutation analysis by holding the expression data fixed and permuting the genotypes. In this manner, we identified the significant eGenes per tissue (as in the single-tissue analysis above, but for a smaller window near the TSS), or jointly for any combination of tissues by considering the minimum $P$ value across tissues under each permutation. With this approach, the number of eGenes identified for individual tissues was similar to the single-tissue eQTL analysis (Fig. 2A). Next, the minimum $P$ value for each gene across all nine tissues was used to test for eQTL evidence, and subjected to gene-level false discovery control. A total of 7425 eGenes with $FDR < 0.05$ were identified, representing a factor of 3 increase relative to the maximum number of significant eGenes for any single tissue. The Bayesian models, which leverage the high proportion of tissue-common eQTLs, increase the power to detect eQTLs for an individual tissue by borrowing strength from the remaining tissues. Thus, of the original 22,286 expressed genes, 10,030 showed a significant eQTL ($FDR < 0.05$) with the gene-based Bayesian multitissue model.
Allele-specific expression analysis

To complement eQTL analysis, we used allele-specific expression (ASE) of genes to indirectly estimate the overall effect of cis-regulatory variation on the expression of nearby genes. Individuals that are heterozygous for a cis-regulatory variant may differentially express each of the two alleles of the affected gene. In these cases, assuming no other causes for allele-specific expression, imbalanced allelic expression of a gene can be used to infer the effect of cis-regulation on the gene even without necessarily identifying the specific regulatory variant(s).

In addition to the subtle cis-regulatory effects described here, ASE of premature stop variants is informative of nonsense-mediated decay (3½), and consistent monoallelic expression of specific genes can be used to detect epigenetic silencing by imprinting.

ASE can be estimated from the abundance of RNA-seq reads supporting each of the gene’s alleles, measured at heterozygous sites within the transcribed region. We identified a median of 6385 sites per sample (across all tissues and all individuals) that were both heterozygous for a transcript SNP in the imputed genotypedata of the associated individual and well expressed in that sample (SNP covered by ≥ 30 RNA-seq reads) (fig. S18). Of these sites, a median of 390 sites per sample had significant ASE (binomial P < 0.005, or 144 sites with FDR 21%, after rigorous quality control of ASE sites and after downsampling to exactly 30 reads at each site) (Fig. 3A and table S8A) (14). To assess whether different tissues vary in terms of the overall effect of cis-regulatory variation in them, we estimated the proportion of those sites with significant ASE in each of the tissues (combining data from all individuals) (14) and found that the fraction of significant ASE sites varied widely across tissues, with a range of 1.7 to 3.7% (median 2.3%). Brain regions appeared depleted for allelic effects (2.0% ASE in brain versus 2.7% in other tissues combined; P < 2 × 10⁻⁶, Fisher’s exact test) (Fig. 3A), consistent with the hypothesis that brain-specific genes exhibit reduced genetic diversity (32). Our results generally support our eQTL findings; in thyroid, where we observed a high number of eQTLs (Fig. 2A), we also see an increased amount of ASE (Fig. 3B). Blood and LCLs have a higher proportion of allelic effects (3.6% versus 2.3% in others; P < 2 × 10⁻³).

![Fig. 3. Quantification of regulatory diversity by ASE.](http://gtexportal.org)

**Fig. 3. Quantification of regulatory diversity by ASE.** (A) Proportion of sites with significant ASE (P < 0.005) in each tissue (colored and labeled as in Table 1), with binomial confidence intervals. (B) Proportion of significant ASE sites for the nine tissues with eQTL data as a function of the proportion of eQTLs after regressing out the log of sample size. (C) Partitioning variation in allelic and total gene expression within and between individuals and tissues. We calculated pairwise Spearman rank correlations between all the samples using two metrics (D) and (E)). (D) Allelic ratios over sites (sampled to 30 reads each), which captures similarity in allelic effects that are a proxy for cis-regulatory variation. (E) Total read counts over the same sites, which captures similarity in total gene expression levels. The plots show the distributions of pairwise correlations for sample pairs that are from (1) different tissues and different individual, (2) different tissues within an individual, or (3) same tissues in different individuals. Gene expression levels are highly correlated within the same tissue (E3) (see Fig. 1A). However, allelic ratios show highest correlation among different tissues of the same individual that share the same genome (D2).
Fishers exact test (Fig. 3B), suggesting nongenetic sources of monoallelic expression possibly due to the more clonal nature of these cells (33).

To compare between-sample and between-tissue sharing of ASE with the overall similarity of gene expression, we constructed two Spearman rank correlation matrices between all pairs of samples. In one, we correlated ASE ratios, capturing sharing of cis-regulation; in the other, we correlated total coverage of the same sites, capturing similarity of overall gene expression levels, analyzing genes that are expressed in both tissues (both matrices used counts of reads covering heterozygous sites shared between the compared pair of samples) (34). The two matrices of tissue medians were highly correlated (Mantel test, \( r = 0.772, P < 0.0001 \); fig. S19), indicating that tissues with similar gene expression profiles also have a higher degree of sharing of genetic regulatory effects.

Interestingly, when partitioning the full sample-level matrix into correlations between samples from different tissues of the same individual, the same tissue across individuals, or different individuals from different tissues of the same individual, the proportion of shared ASE with the overall similarity of gene expression variability, gene expression levels and allelic ratios, are largely defined by independent factors.

A recent study provides tools to measure how relevant our eQTL findings can be at the genome-wide level across all tissues in the GTEx data set, including the tissues with sample sizes that were too small for eQTL analysis. Examining the significant eQTLs identified by single-tissue analysis for the nine main tissues, we looked at their ASE effects separately in each of the 42 tissues, calculating the odds ratio of significant versus nonsignificant ASE for eQTL heterozygote versus homozygote individuals (e.g., for thyroid eQTLs we looked for ASE in all 42 tissues, then for blood eQTLs and so on) (figs. S20 and S21). In addition to replication of eQTL signals in the discovery tissue, we can estimate how relevant our eQTL findings from nine tissues are to a wide variety of other tissues (independent of sample size or allele frequency) and then assess which tissue is the best proxy for capturing regulatory effects in another tissues of interest.

We found that the overall tissue specificity of the eQTL sets varies. For example, eQTLs identified in skeletal muscle were less active in other tissues. Some tissues, such as brain, were not well captured by any of the nine eQTL sets here. ASE analysis also allows quantification of genome-wide tissue sharing of cis-regulatory signals without relying on eQTL discovery, but instead by measuring how often a site with a significant ASE signal in one tissue is significant in another tissue of the same individual. The proportion of shared ASE effects between tissues pair within an individual varies between 36% and 58% (mean 46%), with slightly increased sharing observed between closely related tissues (conditioning on each site being measurable in both of the tissues) (fig. S22A). This relatively high degree of sharing is consistent with the eQTL results described above. If we relax the constraint of requiring a gene to be expressed in both tissues, then the proportion of shared ASE effects is substantially lower (0.65 to 39%, mean 11%; fig. S22B). This finding represents the total probability of detecting a regulatory effect in another tissue, and highlights a high degree of apparent tissue specificity that derives from the fact that a gene expressed in one tissue is often simply not expressed in another. This is particularly pronounced in brain, where the large proportion of genes showing tissue-specific isoform expression in that organ (17) drives a lower degree of overall sharing. Whole blood and skeletal muscle are partial outliers, relative to other tissues.
with lower sharing of eQTLs and lower average replication in other tissues (Fig. 2A, fig. S12, and figs. S20 to S22).

Although eQTL analysis and other association-based analyses are efficient for identifying common variants with phenotypic effects in populations, the genotype of those variants may not be a good indicator of the gene expression phenotype at an individual level. We investigated this using ASE analysis to estimate how consistent the allelic effect of single-tissue eQTL variants are across individuals that share the same genotype for the best associated eQTL variant per gene. We tested a subset of 606 eQTLs that had high read coverage and a large number of samples for transcript heterozygous sites in the eQTL discovery tissue. We identified 53 eQTLs with significant replication of the eQTL signal in ASE, as inferred by higher allelic imbalance in eQTL heterozygotes than in homozygotes (linear regression $P < 0.05$ after Bonferroni correction; Fig. 4, A and B). Further, for 22 of the 53 eQTL genes, individuals show variability in their allelic ratios that cannot be accounted for by the eQTL genotype or sampling error (Bonferroni corrected $P < 0.05$ from permutation of read counts; Fig. 4, C and D). These results highlight that common eQTLs do not explain all of the cis-regulatory effects in individuals and are inadequate predictors of allelic expression variance at the individual level. The higher variance in eQTL heterozygotes than in homozygotes (Mann-Whitney $P = 2.13 \times 10^{-7}$; Fig. 4B) suggests that part of this variance might originate from modification of the main eQTL effect by environmental or trans effects, or due to additional, independent regulatory variants ($P^{34}$).

**Analysis of splicing QTLs**

Beyond estimating overall levels of gene expression, RNA-seq data also allow for the quantification of expression levels of individual transcript isoforms, as well as components of these, such as exons, splice junctions, and untranslated regions. We refer to the quantitative variation of gene structure due to genetic variation as splicing QTLs (sQTLs) (Fig. 5A). To identify sQTLs, we used Altrans ($P^{35}$), a method that identifies SNPs that are associated with variation in the expression levels of exon junctions (sjQTLs; box S1) ($P^{14}$), and sQTLSeekeR ($P^{36}$), a method that identifies SNPs associated with the variation in the relative abundances of gene transcript isoforms (srQTLs; box S1) ($P^{14}$). Altrans identifies both novel and annotated splicing events, while sQTLSeekeR tests only annotated isoforms. Altrans, however, is restricted to changes in the usage of splice junctions, while sQTLSeekeR can in principle detect any variation in the relative abundance of transcript isoforms (fig. S23). Altrans was run using a ±1 Mb region around the TSS, while for sQTLSeekeR, we tested within the body of the gene ±5 kb ($P^{14}$).

We detected an average of ~1900 genes with Altrans, and ~250 with sQTLSeekeR, with at least one cis-sQTL per tissue (FDR = 0.05; table S9). The greater genomic window tested around the TSS for Altrans, and the capacity to detect novel splicing events, explains the much larger number of sQTLs detected by Altrans than by sQTLSeekeR (about 70% of Altrans sQTLs correspond to novel events, and only about 10% map within 5 kb of the TSS).

![Fig. 5. Splicing QTLs. (A) A splicing QTL that affects the relative usage of alternative splice isoforms for the tRNA methyltransferase 1 homolog gene (TRMT1). TRMT1 has three annotated isoforms, only two of which are abundant in skeletal muscle. The relative abundance of the two isoforms differs by genotype (number of individuals below each genotype), with heterozygotes showing an intermediate behavior. This SNP has not been detected as an eQTL. The right panel shows the exonic structure of the transcripts along with the location of the sQTL SNP (dotted line). (B) The relative proportions of the different types of splicing events detected by the two methods over the nine tested tissues (fig. S23). (C) Functional enrichment of sQTLs from Altrans and sQTLSeekeR. For the top-ranked SNPs associated with a given splicing event, we computed the relative frequency with which they map to different biologically determined ENCODE functional domains.]

Downloaded from http://science.sciencemag.org/ on July 16, 2017
annotated genes). Despite the difference, Alttrans and sQTLseekeR show strong complementarity, detecting variants that are associated with very different types of alternative splicing events. Most of the splice events detected by Alttrans (80%) are exon-skipping events, while 60% of those detected by sQTLseekeR correspond to complex splice events, such as mutually exclusive exons (Fig. 5B and fig. S23). If we consider only sQTLs associated with exon-skipping events, the overlap in sQTLs identified by both methods is substantial (36% of sQTLs are shared between Alttrans, P = 0.0044; figs. S23 and S25). An example of an sQTL that we identified with potential biological relevance is shown in Fig. 5A (further examples are in fig. S26).

Significant sQTLs show a high degree of sharing among tissue pairs, with tissue-specific sQTLs accounting only for 7 to 21% of the total depending on the tissue (figs. S27 and S28). The highest degree of sharing is between heart left ventricle and whole blood, whereas the two tissues that share the fewest sQTLs are whole blood and Sun-exposed skin. In general, sQTLs identified in whole blood are shared at lower levels with other tissues, as was observed for eQTLs (Fig. 2A and fig. S12, A and B). Although we observe the same enrichment of sQTLs around the TSS as seen for eQTLs, the sQTLs that are shared across multiple tissues tend to be closer to the TSS than those that are tissue-specific (fig. S28). On average, 20% of sQTLs associated with changes in exon junction abundance by Alttrans were also predicted to be eQTLs (P1 = 0.20, P1 = 0.14 to 0.27; table S10). An even larger fraction (48%) of sQTLs detected by sQTLseekeR associated with changes in relative abundances of gene transcript isoforms, were identified as eQTLs (P1 = 0.48, P1 = 0.13 to 0.70; table S10). This enrichment of eQTLs among sQTLs is larger than expected at FDR < 0.05, but a substantial fraction of sQTLs are unique and not detected by standard eQTL analysis.

**Functional annotation of eQTLs in noncoding regions**

Genetic variants affecting gene expression and splicing patterns have been shown to fall within regulatory elements, providing a potential molecular basis for their effects (37–39). To assess whether the eQTLs discovered across the nine tissues were enriched in regulatory regions, we used a set of regulatory annotations from the ENCODE project (20) and the Epigenomics Roadmap project (40), including regulator-bound locations, DNase I hypersensitivity sites, and maps of histone modifications for proximal and distal regulatory regions (14). For each tissue, we chose the top significant SNP per gene from the single-tissue eQTL analysis (14,431 eQTL SNPs). Discarding SNPs that were within annotated genes resulted in 4085 intergenic eQTL SNPs, which were compared to our regulatory annotations (14). Intergenic eQTL SNPs were enriched for transcription factor-bound sites, open chromatin, promoters, and enhancers (P = 4.3 × 10−18, 2.9 × 10−18, 1.7 × 10−10, and 0.003, respectively) relative to the density of these features within a 2.5-kb window of the tested eQTLs (fig. S29 and table S11). This enrichment was even more pronounced in a subset of 91 unambiguous intergenic eQTL SNPs (fig. S29). In contrast, we found sQTLs to be enriched within annotated splice junctions relative to other functional regions (Fig. 5C).

We also asked whether the specific SNP-gene links (“genetic links”) predicted by our eQTL analysis were supported by enhancer-gene links based on functional correlation (“functional links”) (14). We found that open chromatin regions containing an eQTL SNP show higher correlation to the corresponding gene, relative to open chromatin regions without an eQTL at similar distances and linked to genes also containing an independent eQTL (r = 0.037 versus 0.030, paired Mann-Whitney two-sided P = 4.3 × 10−14, SE = 0.0014, n = 32,168 eQTL-SNP pairs).

This demonstrates the value of eQTL information to establish functional links between regulatory elements and genes in the genome, an analysis that will increase in power as additional tissues and subjects are added to the data set.

**Gene network inference within tissues using cross-individual expression variation**

To provide a view of coordinated gene regulation arising from both cis and trans genetic effects and nongenetic sources, we inferred gene-gene coexpression networks within tissues. Studying each of the nine tissues in isolation and examining variation of gene expression across individuals, we linked pairs of genes that show correlated expression (top 1% of pairs by Pearson correlation) (14), revealing similar patterns of coexpression across tissue pairs (Fig. 6A). The median r1 statistic (fraction of true positive results) (26), estimating the total fraction of coexpressed gene pairs identified in one tissue that are replicated in a second tissue, is 0.44, ranging from 0.30 to 0.58 (Fig. 6A). Furthermore, the specific tissue pairs with higher overlap of coexpressed genes also have a greater overlap of cis-eQTLs than other tissue pairs, indicating a similar pattern of tissue relatedness underlying both results (P < 0.05 for correlation between similarity matrices of Figs. 2B and 6A). Thus, although coexpression networks primarily capture non-cis mechanisms, including trans regulation and environmental factors, the overall level of sharing and the specific patterns of tissue relatedness agree with the patterns observed from cis-eQTLs, which suggests that regulatory mechanisms beyond cis effects may be shared across tissues.

We also used a weighted gene coexpression network analysis (WGCNA) approach (41) to construct coexpression networks and extract gene modules for each tissue (Fig. 6B). The clustering of coexpressed genes into modules allows us to identify active biological processes across tissues (Fig. 6C). Modules enriched for common Gene Ontology (GO) biological processes were observed in all nine tissues (e.g., cell cycle, protein transport), while other biological processes were only seen in specific tissues (e.g., fatty-acyl-CoA metabolic process enrichment seen only in adipose, Benjamini-Hochberg corrected P = 2.6 × 10−5) (14). In addition to functional annotation, the identified coexpression networks also enabled us to search for potential transcriptional regulators of these modules (Fig. 6D). We found enrichment of transcription factor binding in promoter regions of genes in the same module using ENCODE chromatin immunoprecipitation (ChIP)-seq data (42), suggesting that some modules could be regulated by a large number of transcription factors (fig. S30A). Finally, we compared coexpression modules learned in each tissue individually to those learned in other tissues, based on common gene membership and correlation of first principal components of expression variation across individuals (fig. S30B). Many modules showed correlation between principal components to other modules across tissues but lacked common genes, demonstrating that similarity in patterns of variation is sometimes only visible at the module level.

**Gene network inference within individuals using cross-tissue expression variation**

The availability of multiple tissues per individual further allowed us to define modules of co-regulated genes within each individual, by correlating gene expression across tissues for the same individual (14). Merging modules across individuals produced 117 modules (fig. S6E), containing between 25 and 414 genes. Each of these modules corresponds to a multitissue expression pattern for a gene, enabling us to study changes in the regulatory program of genes that affect multiple tissues at a time.

We used these modules to identify instances of coordinated variation in multitissue expression patterns across individuals. For each gene, we estimated the proximity of each individual’s expression pattern to the median expression pattern of each module, corresponding to a “module membership score” (Fig. 6F and fig. S31A). We then calculated a module membership score for each individual (Fig. 6G) (14). Cases where depletion in membership of one module was accompanied by a corresponding increase in membership of a different module are called “module switching events.” The majority of genes showed conserved multitissue gene expression patterns among individuals, reuniting in the same module or switching between modules with similar expression patterns (correlation distance < 0.5). However, we identified 3965 genes (21%) that show switching between dissimilar modules (correlation distance > 0.5), which may have important biological consequences (fig. S31B). Using module membership scores as a quantitative trait, we searched for neighboring SNPs that are correlated with these module membership scores (Fig. 6H), which we call module-switching QTLs (modQTLs). Searching a window of 1Mb around each gene in cis, we found a total of 2102 modQTLs associated with statistically significant switching (FDR < 0.05) between dissimilar gene modules (correlation
distance > 0.5), suggesting that genetic variation for those genes leads to changes in multtissue regulatory programs. For example, the ZFP57 gene shows three distinct patterns of multitissue expression across individuals, and three nearby SNPs are associated with these changes, suggesting a genetic basis for these multitissue differences (Fig. 6H).

We compared the 2102 modQTLs with eQTLs discovered by both the single-tissue and multitissue eQTL analyses. At all correlation levels of stringency, the modQTLs capture 53% of tissue-specific eQTLs (calculated as the percentage of lead SNPs in LD $r^2$ > 0.8 with lead modQTLs) and 60% of multitissue eQTLs, and these two
together account for only 42% of modQTL SNPs. Hence, 58% of modQTLs are not discovered by other methods, which suggests that the approach has the potential to reveal novel genetic effects on the modulation of genes across tissues (fig. S31C).

**Personal transcriptomics and implications for human disease**

The in-depth analysis of multitissue transcriptomes enables both an understanding of the population-level properties of the transcriptome as well as individual level properties inferred from analyses of single transcriptomes or the transcriptomes of multiple tissues from a single individual. As is the case for personalized genomics, this individual-level transcriptome analysis is likely to become a crucial addition to the personalized assessment of an individual's biology and likely disease status.

**Impact of individual gene-disrupting variants on splicing and expression**

Assessing the functional impact of DNA sequence variants identified in whole-genome and exome sequencing studies remains a major challenge. Variants predicted to result in the truncation of proteins (splice, stop-gain SNVs, or frame-shift indels) may have large effects on biochemical function, but are also highly enriched for annotation artifacts (43). Errors in predicting the true functional impact of these and other variants can substantially reduce discovery power in common complex diseases and, more important, can affect disease diagnosis in clinical settings (44).

The GTEx multitissue expression data provide an opportunity to assess the real impact of protein-truncating variants (PTVs) on the human transcriptome.
transcriptome (see also (3I)). We combined exome sequencing and RNA-seq data from 173 GTEx individuals to assess the global properties of predicted high-confidence PTVs (14). PTVs are enriched in alternatively spliced exons, with just 38.4% of high-confidence PTVs having annotation support across all reported transcripts (Fig. 7A), and only 51 to 55% supported by the major transcript of at least one tissue (for all tissues with at least 10 samples). These numbers highlight the need for careful transcript-specific assessment of functional annotation for all classes of variation. Furthermore, if we require that a fixed percentage of the dominant isoform support for PTV prediction (Fig. 7B). This highlights the need for careful transcript sets across a broad array of tissues to enhance clinical interpretation for personal genome sequencing and disease studies.

An example with clinical ramifications is shown in fig. S32.

GWAS and eQTLs

The ultimate goal of the GTEx project is to provide a framework for biological interpretation of disease-related variants. To evaluate the relevance of the discovered eQTLs in disease mapping studies, we tested the eQTLs identified in each tissue for association with disease using the Wellcome Trust Case Control Consortium (WTCCC) studies of seven complex disorders (45) (see Fig. 8A). Using eQTLs identified in whole blood, we found an enrichment for top associations with autoimmune diseases (shown as a leftward shift from the null distribution in the Q-Q plot in Fig. 8A), for Crohn’s disease, rheumatoid arthritis, and type 1 diabetes among eQTLs (P < 2.2 × 10^{-16} for each disease), consistent with the utility of blood and lymphoblastoid cell lines (LCLs) in trait mapping for autoimmune disorders (46, 47). In contrast to the autoimmune disorders, we found no enrichment for association (no shift in P value distribution) with bipolar disorder or type 2 diabetes among the blood eQTLs (P > 0.05) (Fig. 8A). This tissue specificity of autoimmune enrichment results suggests that they are not due to the confounding effects that result from similar underlying genomic properties between eQTL and disease association regions. More generally, we observed trait-specific levels of enrichment for the WTCCC disease traits among the nine different single-tissue eQTL sets.

Remarkably, the use of eQTLs increased power to detect associations with hypertension (Fig. 8B). In particular, eQTLs in subcutaneous adipose were significantly enriched for multiple associations with hypertension relative to muscle, lung, thyroid, skin, heart, and tibial artery (P < 0.05, Kolmogorov-Smirnov test) (Fig. 8B). This is particularly noteworthy because the WTCCC GWAS of this disease did not yield any genome-wide significant associations, which suggests that larger sample sizes were required to identify highly significant SNP associations in the absence of functional data. Because the majority of GWAS-identified variants (~95%) lie in non-coding regions of the genome, we determined

**Fig. 9. A blood pressure-associated SNP is a significant eQTL in tibial artery, for ARHGAP42 and TMEM133.**

(A) The GWAS SNP, rs633185 in the intron of ARHGAP42, is associated with systolic blood pressure (P = 1.2 × 10^{-12}) and diastolic blood pressure (P = 2 × 10^{-15}). This GWAS SNP is in tight LD (r^2 = 0.93) with the most significant eQTL for ARHGAP42 in tibial artery, rs604723 (P = 1 × 10^{-8}), and is the most significant eQTL for TMEM133 in tibial artery (P = 2.7 × 10^{-8}). Tibial artery was the only significant tissue at FDR < 0.05 according to the single-tissue eQTL discovery method. (B) Average posterior probability of the most significant cis-eQTL, rs607562 for ARHGAP42 at FDR < 0.05 from the multitissue eQTL methods. (C) Similar plot for TMEM133. The most significant cis-eQTL for TMEM133 from the multitissue methods at FDR < 0.05 is the GWAS SNP, rs633185, in tibial artery.
which genome-wide significant trait associations ($P < 5 \times 10^{-8}$) reported to date are in LD with at least one GTEx-identified eQTL. We merged NCBI's Phenotype-Genotype Integrator (PhenGenI) (46) and the NHGRI GWAS catalog (49), yielding 10,129 genome-wide significant SNP associations with nearly 630 distinct complex traits. In total, 5195 “independent” SNPs were identified after LD pruning at $r^2 \geq 0.8$ and counting SNPs only once (14). Of these, 508 (~8%) are in strong LD ($r^2 > 0.8$) with a “best eQTL-per-gene” (at FDR < 0.05) from either the single-tissue or multitissue eQTL discovery analysis (table S12) in at least one of the nine tissues tested. For two-thirds of these (211 SNPs), no putative deleterious coding variants (nonsynonymous or splice variants) in the target gene product lie in LD ($r^2 > 0.8$) with the GWAS SNP; this finding suggests that regulatory effects may underlie the causal mechanism, although additional work is needed to prove causality. GWAS SNPs in LD with an eQTL show a factor of 2 higher representation in coding regions relative to all GWAS SNPs (11% versus 4.6%; table S13). Notably, about one-third of the eQTLs in LD with GWAS SNPs were detected only with methods that leverage the multitissue nature of GTEx data. Increasing both sample sizes and the range of tissues will likely increase the number of detected GWAS-eQTL loci.

Annotating a GWAS SNP with an eQTL can help to highlight candidate causal genes within a locus (i.e., the eQTL target gene). We found that proximity-based and eQTL-based gene assignments for GWAS SNPs were often discordant (47). A surprising proportion of trait-associated SNPs in LD ($r^2 > 0.8$) with a GTEx eQTL showed disagreement between the strongest eQTL-derived target gene and the genes that were physically proximal to the GWAS SNP (table S14). Of 190 GWAS loci ($P < 5 \times 10^{-8}$) where the lead SNP is an eQTL from the single-tissue analysis (FDR < 0.05) with only a single target gene, in 65 cases (~34%) this eQTL target gene differs from any of the genes that were closest to the SNPs in LD. These results were also observed when we restricted the target genes to protein-coding genes, when we pruned the GWAS SNPs for each trait examined ($r^2 > 0.8$), and when we used the eQTLs identified from the multitissue joint eQTL analysis.

In addition to prioritizing causal genes in GWAS loci, an eQTL catalog from multiple human tissues can highlight the relevant tissue(s) of action, evaluate the tissue specificity of GWAS loci, and characterize pleiotropic associations. We demonstrate the value of multitissue data to explore and resolve these issues for the GWAS intronic SNP, rs833185, located in ARHGAP42 (Fig. 9). This GWAS SNP is in high LD ($r^2 = 0.93$) with the best cis-eQTL (rs604723) targeting ARHGAP42, and the best cis-eQTL for a neighboring gene, TMEM133, in tibial artery. Evaluating eQTL significance in all nine tissues shows that although the eQTL's significance is indeed strongest in tibial artery, several other tissues may merit consideration in exploring the causal function of this locus, such as subcutaneous adipose and skin (Fig. 9, B and C, and fig. S33). This supports the need to explore the genetic basis of disease in the fuller context of a wide range of human tissues. The GTEx eQTLs may also be useful in highlighting the role of non-coding genes in disease risk and other complex traits (fig. S44). We found that ~13% of candidate genes proposed by GTEx eQTLs, and in LD to genome-wide significant GWAS SNPs, are non-coding genes (table S15).

Conclusions

We have described a large in-depth data set of multitissue human gene expression. We assessed the variability of the transcriptome among individuals in a large number of tissues at a resolution that provides unique insights into the diversity and regulation of gene expression among tissues. This analysis provides a unified view of genetic effects on gene expression across a broad range of tissue types, most of which have not been studied for eQTLs previously. We look forward to scaling up the resource to create a data set that will transform our understanding of how genetic variability influences different tissues and biological systems and ultimately complex diseases.

REFERENCES AND NOTES

14. See supplementary materials on Science Online.
15. J. Harrow et al., Genome Res. 22, 1760–1774 (2012).
The human transcriptome across tissues and individuals

Marta Melé,1,2,3 Pedro G. Ferreira,1,3,4,5 Ferran Reverter,1,6,7,8 David S. DeLuca,9 Jean Monlong,7,9 Michael Sammeth,1,7,10 Taylor R. Young,4 Jakob M Goldman,11,12 Dmitri D. Pervouchine,1,11,12 Timothy J. Sullivan,1,12,13,14 Rory Johnson,1,12 Ayellet V. Segré,1,12 Sarah Djebali,1,7,18,24 Anna M. Smith,51,52 Stephen A. Buia,52 Anita H. Undale,52 Laura A. Siminoff48,49,52

Transcriptional regulation and posttranscriptional processing underlie many cellular and organismal phenotypes. We used RNA sequence data generated by Genotype-Tissue Expression (GTEx) project to investigate the patterns of transcriptome variation across individuals and tissues. Tissues exhibit characteristic transcriptional signatures that show stability in postmortem samples. These signatures are dominated by a relatively small number of genes—which is most clearly seen in blood—though few are exclusive to a particular tissue and vary more across tissues than individuals. Genes exhibiting high interindividual expression variation include disease candidates associated with sex, ethnicity, and age. Primary transcription is the major driver of cellular specificity, with splicing playing mostly a complementary role; except for the brain, which exhibits a more divergent splicing program. Variation in splicing, despite its stochasticity, may play in contrast a comparatively greater role in defining individual phenotypes.

Gene expression is the key determinant of cellular phenotype, and genome-wide expression analysis has been a mainstay of genomics and biomedical research, providing insights into the molecular events underlying human biology and disease. Whereas expression data sets from tissues/primary cells (1, I) and individuals (3) have accumulated over recent years, only limited expression data sets have allowed analysis across tissues and individuals
The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans
The GTEx Consortium

*Science* 348 (6235), 648-660,
DOI: 10.1126/science.1262110

**Expression, genetic variation, and tissues**
Human genomes show extensive genetic variation across individuals, but we have only just started documenting the effects of this variation on the regulation of gene expression. Furthermore, only a few tissues have been examined per genetic variant. In order to examine how genetic expression varies among tissues within individuals, the Genotype-Tissue Expression (GTEx) Consortium collected 1641 postmortem samples covering 54 body sites from 175 individuals. They identified quantitative genetic traits that affect gene expression and determined which of these exhibit tissue-specific expression patterns. Melé et al. measured how transcription varies among tissues, and Rivas et al. looked at how truncated protein variants affect expression across tissues.

*Science*, this issue p. 648, p. 660, p. 666; see also p. 640