For GTEx Release v6:

Laboratory Methods

Expression Data

- Affymetrix Expression Array
- Illumina TrueSeq RNA sequencing

Genotype Data

- Illumina OMNI 5M SNP Array (used for eQTL analysis)
- Illumina Human Exome SNP Array
- Whole exome sequencing (Agilent or ICE target capture, HiSeq 2000)
- Whole genome sequencing (HiSeq 2000 or HiSeq X)

Analysis Methods

Preprocessing

RNA-Seq Alignment

RNA-seq was performed using the Illumina TruSeq library construction protocol. This is a non-strand specific polyA+ selected library. The sequencing produced 76-bp paired end reads.

Alignment to the HG19 human genome was performed using Tophat v1.4.1 assisted by the GENCODE v19 transcriptome definition. In a post-processing step, unaligned reads are reintroduced into the bam. The final bam contains aligned and unaligned reads, and marked duplicates. It should be noted that Tophat produces multiple mappings for some reads, but in post-processing one read is flagged as the primary alignment.

GTEx dbGaP Release	V3 (Pilot Phase)	V4	V5[1]
GENCODE Version	GENCODE v12	GENCODE v18	GENCODE v19

Genotyping

DNA samples are sent to the Broad Institute Genetic Analysis Platform for genotyping, and are placed on 96-well plates using the Illumina HumanOmni5-4v1_B or HumanOmni2.5-8v1-1_B (Infinium 2.5 million Omni – 8 sample chip v1.1B) SNP arrays.

[1] Release V6 eQTLs are computed using RNA-seq data from release V5.

DNA isolated from blood samples collected from each GTEx donor was the primary source of DNA used for genotyping (> 360 ng of DNA input per sample). Omni genotypes were called using Illumina's GenTrain calling algorithm (Autocall version 1.6.2.2) and Birdsuite version 1.6, using the default cluster files, HumanOmni5-4v1-Multi_B.egt or HumanOmni2.5-8v1-1_B.egt.

After sample quality control (QC) for eQTL analysis (removing replicates, Klinefelter individuals, a chromosome 17 trisomy individual, and related individuals), we have 183 GTEx donors genotyped on Illumina's Human Omni 5M (4,276,680 hard call variants) from the pilot phase and 267 GTEx donors genotyped on Illumina's Human Omni 2.5M (2,378,075 hard call variants) from Option 1 phase, for a total of 450 donors with genotype data for eQTL analysis in release v6. Of these, 292 are males and 158 females. Our genotyping call rates per individual exceeded 98% for all samples, and all samples passed genetic fingerprint and gender concordance. All genotypes were aligned to chromosome positions from human genome build 37 (hg19). The Omni 2.5M portion of hard call genotypes from the Omni 2.5M or Omni 5M across all 450 donors were merged into one VCF. A series of standard QC steps were applied to the 450 donors' genotypes, including: excluding SNPs with genotyping call rate <95%, SNPs with differential missingness between Omni 2.5M and Omni 5M arrays, HWE P-value < 1E-06 (using European samples only), variants associated with plate batch, and variants with heterozygous haploid genotypes in non-pseudoautosomal regions of sex chromosomes in Males.

To increase eQTL discovery power, we imputed variants (SNPs and indels) from the 1000 Genomes Project Phase I version 3 reference panel (24 August 2012 release) into our QC'd array VCF of 450 GTEx donors using IMPUTE2 for imputation and SHAPEIT for prephasing.

For the final eQTL analysis, the following filters were applied to the imputed VCF: call rate < 95%, HWE P-value < 1E-6, monomorphic variants, indels of length >51 bp, and imputation quality score INFO<0.4. This yielded 11,555,102 genotyped and imputed autosomal variants.

Expression Quantification

Gene/Transcript Model

- Gencode Version 19
- Contig names modified to match the reference genome used for alignment

Procedure for collapsing transcript model into gene model

1. Primary source: gencode.v19
2. List exons as a set of intervals, discarding any labeled as 'retained_intron' and retaining only coding and linc rna.
3. Create a separate bin for other types of transcripts and process them independently.
4. Merge overlapping intervals.

5. Discard intervals associated with multiple genes (in V6, reads were only discarded if they overlapped between genes that are protein-coding or lincRNAs or between all other gene types, but not between these two categories).
6. Map intervals back to gene identifiers and output in GTF format.

Quantification

Gene level annotations are produced by taking the union of exons defined in GENCODE v19.

RPKMs are from the RNA-SeQC tool. **The RPKM values that are downloadable have not been normalized or corrected for any covariates**.

For gene/exon level read count and gene level RPKM values, we filter reads based on these requirements:

- Reads must have be uniquely mapped (for TopHat this corresponds to a mapping quality of 255).
- Reads must have proper pairs.
- Alignment distance must be <=6 (i.e. read must not contain more than six non-reference bases).
- Reads must be contained 100% within exon boundaries. Reads overlapping introns are not counted.

Exon

For exon read counts, if a read overlaps multiple exons, then a fractional value equal to the portion of the read contained within that exon is allotted.

Transcript

Transcript-level quantification is provided by Flux Capacitor, version 1.6

eQTL Analysis

QC and Sample Exclusion Process

1. D statistic outliers are removed. (D statistic described in: Wright et al., 2014, Nat. Genet. 46(5):430-7)
2. Gender-specific expression outliers are removed.
3. Samples with less than 10 million mapped reads are removed.
4. In the case of replicates, the samples with the greater number of reads are chosen.

Covariates

- 3 Genotyping PCs.
- Genotyping array platform (Illumina's OMNI 5M or 2.5M array).
- 15, 30 or 35 PEER factors: PEER factors were generated using the top 10,000 expressed genes per tissue that were normalized with the same procedure as for

the expression matrices (as described below). Number of PEERs used per tissue was determined by N, number of samples per tissue: For N < 150, we used 15 PEERs, for 150<=N<250, 30 PEERs, and for N >=250, 35 PEERs.
- Gender.

Expression

- RPKM data are used as produced by RNA-SeQC.
- Filter on >=10 individuals with >0.1 RPKM and raw read counts greater than 6.
- Quantile normalization was performed within each tissue to bring the expression profile of each sample onto the same scale.
- To protect from outliers, inverse quantile normalization was performed for each gene, mapping each set of expression values to a standard normal.

Genotypes

- Variants were imputed using 1000 Genomes Project Phase I, version 3.
  The following post-imputation genotype filters were applied:
  - Call Rate Threshold 95%.
  - Info score Threshold 0.4.
  - Minor Allele Frequency >= 1% (a tissue specific cutoff, as sample sets vary by tissue).
- Sex chromosomes were excluded from the current analysis.
- While the imputed array VCF contains genotype data for 450 GTEx donors, only 449 of the donors have RNA-seq data of good quality for eQTL analysis in release v6.

eQTL Analysis using Matrix eQTL

- Linear regression analysis using the covariates listed above was applied to gene-cis-SNP pairs using Matrix eQTL (Shabalin et al., Bioinf. 2012), assuming an additive model.
- **Cis window** was defined as +/- 1MB around the transcript start site (TSS).
- **Nominal eQTL p-values** were generated for each SNP-gene pair using a two-tailed t test, testing the alternative hypothesis that the beta (slope of the linear regression model) deviates from the null hypothesis of beta=0.
- The **effect size** of the eQTLs is defined as the slope ('beta') of the linear regression, and is computed as the effect of the alternative allele (ALT) relative to the reference allele (REF) in the human genome reference GRCh37/hg19 (i.e., the eQTL effect allele is the ALT allele).

Permutations with Matrix eQTL

- Based on the simplified assumption that we are primarily powered to detect one eQTL per gene (so-called "**eGenes**")

- A permutation regime is employed to correct for the many non-independent SNPs given for every gene due to linkage disequilibrium. Sample labels for the expression data were randomized. The randomized indexes were applied to the PEER factors and gender covariate, but not to the genotypes and genotyping PCs.
- The most significant nominal p-value, **min(P)**, of all variants tested in the cis window per gene, is used as the test statistic to generate the empirical distribution.
- For each gene, the empirical p-value is the fraction of permutations with the same or more significant min(P) than the observed min(P).
- The number of permutations performed for each gene depends on the significance level:
    - A minimum of 1,000 permutations are performed
    - A maximum of 10,000 permutations are performed
    - Between these ranges, permutation is terminated when there are 15 or more permuted values less than the test statistic being evaluated.
- Storey **FDR** is used to correct for multiple hypothesis.
    - Using the public R package with default values.
    - The correction is performed on the eGene empirical p-values.
    - eQTLs were chosen as significant for q-values <=5%.

Reporting all significant SNPs per eGene

- Since causal SNP(s) remain unknown for our eQTL calculations, it is useful to provide a list of all significant eQTL-associated SNPs for each eGene.
- To define a minimum significance threshold for SNPs per eGene in a given tissue, a **permutation threshold** was chosen as the empirical p-value of the gene that falls on the q-value = 0.05 threshold.
- For each eGene, **a gene-specific nominal p-value threshold** was determined as the nominal p-value of all permutations' min(P) per gene that corresponds to the permutation threshold defined above.
- SNPs with nominal eQTL p-values below or equal to this gene-specific nominal p-value threshold were included in the final list of significant eQTL SNPs per eGene.

Tissues for eQTL Analysis

In our experience, at least 70 samples are needed per tissue for proper eQTL discovery. eQTLs have only been generated for tissues that meet this threshold.