GTEx Release v7

Laboratory Methods

Expression Data

- Illumina TruSeq RNA sequencing
- Affymetrix Human Gene 1.1 ST Expression Array (V3; 837 samples)

Genotype Data

- Whole genome sequencing (HiSeq X; first batch on HiSeq 2000)
- Whole exome sequencing (Agilent or ICE target capture, HiSeq 2000)
- Illumina OMNI 5M Array or 2.5M SNP Array
- Illumina Human Exome SNP Array

Analysis Methods

Updated on 09/05/2017
Analysis information for V4 is available here
Analysis information for V6 is available here
Analysis information for V6p is available here

RNA-seq was performed using the Illumina TruSeq library construction protocol (non-stranded, polyA+ selection) .

Total RNA was quantified using the Quant-iTTM RiboGreen®RNA Assay Kit and normalized to 5 ng per µL. An aliquot of 200 ng for each sample was transferred into library preparation, which was an automated variant of the Illumina Tru SeqTM RNA sample preparation protocol (Revision A, 2010). This method used oligo dT beads to select mRNA from the total RNA sample followed by heat fragmentation and cDNA synthesis from the RNA template. The resultant cDNA then went through library preparation (end repair, base 'A' addition, adapter ligation, and enrichment) using Broad Institute-designed indexed adapters substituted in for multiplexing. After enrichment, the libraries were quantified with qPCR using the KAPA Library Quantification Kit for Illumina Sequencing Platforms and then pooled equimolarly. The entire process was performed in 96-well plates and all pipetting was performed by either Agilent Bravo or Hamilton Starlet liquid handlers with electronic tracking throughout the process in real-time, including reagent lot numbers, specific automation used, time stamps for each process step, and automatic registration.

Pooled libraries were normalized to 2 nM and denatured using 0.1 N NaOH prior to sequencing. Flow cell cluster amplification and sequencing were performed according to the manufacturer's protocols using either the HiSeq 2000 or HiSeq 2500. Sequencing generated 76bp paired-end reads and an eight-base index barcode read, and was run with a coverage goal of 50M reads (the median achieved was ~82M total reads).

Preprocessing

RNA-seq Alignment

Alignment to the human reference genome hg19/GRCh37 was performed using STAR v2.4.2a, based on the GENCODE v19 annotation. Unaligned reads were kept in the final BAM file. Among multi-mapping reads, one read is flagged as the primary alignment by STAR. The alignment pipeline is available at https://github.com/broadinstitute/gtex-pipeline/tree/master/rnaseq

GTEx DbGaP release	V3 (Pilot Phase)	V6p	V7
GENCODE Version	v12	v19	v19

Genotyping

Whole genome sequencing (WGS) was performed by the Broad Institute's Genomics Platform on DNA samples from 652 GTEx donors at an average coverage of 30X. Of these, 68 samples were sequenced on Illumina HiSeq 2000 using 101-bp paired-end reads, and 584 samples on Illumina HiSeq X using 151-bp paired-end reads. DNA isolated from blood samples collected from each GTEx donor was the primary source of DNA used for genotyping (~100 ng of DNA input per sample).

WGS BAMs were processed through a pipeline based on Picard (http://picard.sourceforge.net/), using base quality score recalibration and local realignment at known indels. The WGS reads were aligned to the human reference genome build hg19/GRCh37 with BWA-MEM (http://bio-bwa.sourceforge.net). Joint variant calling was performed across all WGS samples using GATK's HaplotypeCaller v3.4. Genotyping call rates per individual exceeded 98% for all samples, and all samples passed genetic fingerprint and sex concordance checks. To increase variant calling quality, genotype posterior probabilities were calculated for all calls based on allele frequency in 1000 Genomes Project Phase 3 version 1, and genotype quality (GQ) scores were updated accordingly, using GATK's CalculateGenotypePosteriors.

To detect samples with large deletions or duplications for exclusion from eQTL analysis, we applied the Genome STRiP module that detects Long Copy Number Variation (>1Mb) to all WGS samples. 16 donors with >10Mb CNVs, or >1Mb CNVs associated with known syndromes based on literature searches, were excluded from the analysis freeze.

Analysis freeze for eQTL analyses: After sample quality control (QC) (removing replicates, individuals with large chromosomal abnormalities, such as Klinefelter individuals and a chromosome 17 trisomy individual, and one individual from a related pair), 635 donors (406 males and 229 females) with genotype data remained in release v7. Multi-allelic sites were split into bi-allelic sites using Hail (https://hail.is/). Multiple variant-level QC steps were applied to all bi-allelic sites in the resulting WGS VCF of 635 donors, including: Variant Quality Score Recalibration (VQSR) filtering; removal of variants in low-complexity regions or with inbreeding coefficient <= -0.3; assigning calls with allelic imbalance (AB) <0.2 or >0.8 and/or GQ <20 to missing; excluding variants with genotyping call rate <85%, with HWE P-value < 1E-06 (using European samples only and females for chrX), associated with library construction batch or sequencing technology (P < 1E-08), or with heterozygous haploid genotypes in

non-pseudoautosomal regions of sex chromosomes in males. This yielded a total of 10,526,813 variants at MAF ≥ 1%.

Expression Quantification

Transcript Model

GENCODE 19 (http://www.gencodegenes.org/releases/19.html), with chromosome names modified to match the hg19 chromosome names.

Collapsed Gene Model

Gene-level expression quantification was based on the GENCODE 19 annotation, collapsed to a single transcript model for each gene using a custom isoform collapsing procedure, comprising the following steps:

1. Exons associated with transcripts annotated as “retained_intron” and “read_through” were excluded.
2. Exon intervals overlapping within a gene were merged.
3. The intersections of exon intervals overlapping between genes were excluded.
4. The remaining exon intervals were mapped to their respective gene identifier and stored in GTF format.

Code for generating the collapsed model is available at https://github.com/broadinstitute/gtex-pipeline/tree/master/gene_model .

Quantification

Gene-level quantifications: read counts and TPM values were produced with RNA-SeQC v1.1.8 (DeLuca et al., Bioinformatics, 2012), using the following read-level filters:

- Reads were uniquely mapped (corresponding to a mapping quality of 255 for START BAMs).
- Reads were aligned in proper pairs.
- The read alignment distance was <=6 (i.e., alignments must not contain more than six non-reference bases).
- Reads were fully contained within exon boundaries. Reads overlapping introns were not counted.

These filters were applied using the “-strictMode” flag in RNA-SeQC.

The TPM values that are downloadable have not been normalized or corrected for any covariates.

Exon-level quantifications: for exon-level read counts, if a read overlapped multiple exons, then a fractional value equal to the portion of the read contained within that exon was allotted.

Transcript-level quantifications were calculated using RSEM v1.2.22.

eQTL Analysis

QC and Sample Exclusion Process

1. RNA-seq expression outliers were identified and excluded using a multidimensional extension of the statistic described in (Wright et al., Nat. Genet. 2014). Briefly, for each tissue, read counts from each sample were normalized using size factors calculated with DESeq2 and log-transformed with an offset of 1; genes with a log-transformed value >1 in >10% of samples were selected, and the resulting read counts were centered and unit-normalized. The resulting matrix was then hierarchically clustered (based on average and cosine distance), and a chi2 p-value was calculated based on Mahalanobis distance. Clusters with ≥60% samples with Bonferroni-corrected p-values <0.05 were marked as outliers, and their samples were excluded.
2. Samples with <10 million mapped reads were removed.
3. For samples with replicates, the replicate with the greatest number of reads was selected.

Covariates

- Top 3 genotyping principal components.
- A set of covariates identified using the Probabilistic Estimation of Expression Residuals (PEER) method (Stegle et al., PLoS Comp. Biol., 2010), calculated for the normalized expression matrices (described below). The number of PEER factors was determined as function of sample size (N): 15 factors for N<150, 30 factors for 150≤ N<250, 45 factors for 250≤ N<350, and 60 factors for N≥350, as a result of optimizing for the number of eGenes discovered.
- Genotyping platform (Illumina HiSeq 2000 or HiSeq X).
- Sex.

Expression

Gene expression values for all samples from a given tissue were normalized using the following procedure:

- Genes were selected based on expression thresholds of >0.1 TPM in at least 20% of samples and ≥6 reads in at least 20% of samples.
- Expression values were normalized between samples using TMM as implemented in edgeR (Robinson & Oshlack, Genome Biology, 2010).
- For each gene, expression values were normalized across samples using an inverse normal transform.

Genotypes

The genotype data used for eQTL analyses in release V7 was based on WGS from 635 donors (of these 620 donors have RNA-seq data available in V7). The variant QC filtering steps applied are described above; only variants with MAF ≥ 1% were considered.

eQTL Analysis using FastQTL

cis-eQTL mapping was performed using FastQTL (Ongen et al., Bioinformatics, 2016), using the covariates described above.

- **Nominal p-values** were generated for each variant-gene pair by testing the alternative hypothesis that the slope of a linear regression model between genotype and expression deviates from 0.
- The **mapping window** was defined as 1 megabase up- and downstream of the transcription start site.
- For each tissue, variants in the VCF were selected based on the following thresholds: the minor allele was observed in at least 10 samples, and the **minor allele frequency** was ≥ 1%.
- The adaptive **permutations** mode was used with the setting "--permute 1000 10000".
- Beta distribution-adjusted **empirical p-values** from FastQTL were used to calculate q-values (Storey & Tibshirani, PNAS, 2003), and a false discovery rate (FDR) threshold of ≤0.05 was applied to identify genes with a significant eQTL ("eGenes").
- The **normalized effect size (NES)** of the eQTLs is defined as the slope 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).

Note: NES are computed in a normalized space where magnitude has no direct biological interpretation. The version of FastQTL used for GTEx analyses is available at https://github.com/francois-a/fastqtl.

Allelic Fold-Change

Log allelic fold-change (aFC), a measure of cis-eQTL effect size, is defined as the log-ratio between the expression of the haplotype carrying the alternative eVariant allele to the one carrying the reference allele.

aFC is calculated using the approach described in Mohammadi et al., Genome Research, 2017 . Briefly, the model assumes an additive model of expression in which the total expression of a gene in a given genotype group is the sum of the expression of the two haplotypes: e(genotype) = 2e_r, e_r + e_a, 2e_a, for reference homozygotes, heterozygotes, and alternate homozygotes, respectively, where e_r is the expression of the haplotype carrying the reference allele, and e_a the expression of the haplotype carrying the alternative allele. The allelic fold change k is defined as: e_a = k e_r where $0 < k < \infty$; aFC is represented in log2 scale as s = log2 k, and is capped at 100-fold to avoid outliers ($|s| < \log_2 100$).

Currently, the aFC of the top variant of each eGene is available in the eGene table.

Identification of all significant variant-gene pairs

To identify the list of all significant variant-gene pairs associated with eGenes, a genome-wide empirical p-value threshold, p_t, was defined as the empirical p-value of the gene closest to the 0.05 FDR threshold. p_t was then used to calculate a nominal p-value threshold for each gene based on the beta distribution model (from FastQTL) of the minimum p-value distribution $f(p_{min})$ obtained from the permutations for the gene. Specifically, the nominal threshold was calculated as $F^{-1}(p_t)$, where F^{-1} is the inverse cumulative distribution. For each gene, variants with a nominal p-value below the gene-level threshold were considered significant and included in the final list of variant-gene pairs.

Tissues for eQTL Analysis

A threshold of at least 70 samples per tissue was determined to provide sufficient statistical power for eQTL discovery, resulting in a set of 48 tissues tested for the V7 release.