GTEx Release v6p

Laboratory Methods

Expression Data

- Affymetrix Expression Array
- Illumina TruSeq 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

Updated on 08/18/2016
Analysis information for V4 is available here
Analysis information for V6 is available here

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*
GENCODE Version	GENCODE v12	GENCODE v18	GENCODE v19

* Release v6 eQTLs are computed using RNA-seq data from release V5.

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. 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
- Chromosome names modified to match the reference genome used for alignment.

Collapsed Gene Model

Gene-level expression quantification was performed using RNA-SeQC (DeLuca, Bioinformatics, 2012) using a custom isoform collapsing procedure:

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.

Quantification

Gene-level quantifications: read counts and RPKM values were produced with RNA-SeQC, using the following read-level filters:

- Reads were uniquely mapped (corresponding to a mapping quality of 255 for TopHat 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 RPKM 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 Flux Capacitor (version 1.6).

eQTL Analysis

QC and Sample Exclusion Process

1. RNA-seq expression outliers were identified and excluded using a correlation-based statistic similar to that described in (Wright et al., 2014, Nat. Genet. 46(5):430-7). Briefly, for each tissue, read counts from each sample were normalized based on interquartile range 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 statistic di was then defined as the correlation between a sample i and the median of all samples from a given tissue. Samples with a di value outside of 1.5x the interquartile range were excluded.
2. Sex-specific expression outliers were removed.
3. Samples with less than 10 million mapped reads were removed.
4. 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 $150 \leq N < 250$, and 35 factors for $N \geq 250$, based on optimizing for the number of eGenes discovered.
- Genotyping array platform (Illumina OMNI 5M or 2.5M array).
- 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 RPKM in at least 10 individuals and ≥6 reads in at least 10 individuals.
- Expression values were quantile normalized to the average empirical distribution observed across samples.

- For each gene, expression values were inverse quantile normalized to a standard normal distribution across samples.

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).
- 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 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 frequencywas ≥ 0.01.
- 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 effect size 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).

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 44 tissues tested for the V6/V6p release.