For release v4:

Laboratory Methods

Expression Data

- Affymetrix Expression Array
- Illumina TrueSeq RNA sequencing

Genotype Data

- Illumina OMNI 5M or 2.5M SNP Array (used for eQTL analysis)
- Illumina Human Exome SNP Array

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 v1.1.4 assisted by the GENCODE v12 transcriptome definition. In a post processing step, unaligned reads are reintroduced into the bam. The final bam contains aligned and unaligned reads, 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.

Genotyping

DNA samples that are sent to the Broad Institute Genetic Analysis Platform for genotyping, are placed on 96-well plates using the Illumina HumanOmni5-4v1_B SNP array. Omni genotypes are called using GenomeStudio v2010.3 with the calling algorithm/genotyping module version 1.8.4 using the default cluster file HumanOmni5-4v1-Multi_B.egt. Called genotypes are run through a standard QC pipeline and only samples passing a call rate threshold of 98%, and passing genetic fingerprint and gender concordance are passed. For the final eQTL analysis, the following filters were applied: call rate (< 95%), low HWE (pValue < 1E-6) or are monormorphic.

Expression Quantification

Gene/Transcript Model

- Gencode Version 18
- Contig names modified to match the reference genome used for alignment

Procedure for collapsing transcript model into gene model

1. Primary source: gencode.v18
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.
6. Map intervals back to gene identifiers and output in GTF format.

Quantification

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

- Reads must have be uniquely mapped (for tophat this is mapping quality > 3; == 255).
- Reads must have proper pairs.
- Alignment distance must be <=6.
- 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.
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.
- 15 Peer factors:
  The input to PEER are the post-normalization expression values described below.
- Gender.

Expression

- RPKM data are used as produced by RNA-SeQC.
- Filter on >=10 individuals having >0.1RPKM.
- Log and quantile normalize the expression values across all samples.
- Outlier correction: for each gene, rank values across samples then map to a standard normal.

Genotypes

- Imputation-based genotypes:
- Call Rate Threshold 95%.
- Info score Threshold 0.4.
- Minor Allele Frequency >= 5%.
- Sex chromosomes have been excluded excluded.

Permutations with Matrix eQTL

- Cis radius was defined as +/- 1mb from TSS
- Assume that we are powered only 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. Sample labels are randomized.
- The minimum p-value is used as the test statistic to generate the empirical distribution
- The number of permutations performed for each gene depends on the significance level
- A minimum of 1000 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

- The causal SNP(s) remain unknown for our eQTL calculations. Nonetheless, it is useful to provide the list of all eQTL-associated SNPs for each given eGene.
- For a given eGene, SNPs are filtered using a significance threshold that corresponds to the eGene empirical p-value for eGenes at the q-value = 0.05 threshold. We will call this empirical p-value the permutation threshold.

- For each gene, a gene-specific analytical p-value threshold is determined by using the permutation data to assess which analytical p-value would correspond to the chosen empirical p-value threshold
- Finally these gene-specific analytical p-value thresholds are used to which SNPs to include in the list of significant SNPs for a given eGene

Tissues for eQTL Analysis

In our experience, at least 60 samples are needed per tissue for proper eQTL discovery. eQTLs have only been generated for tissues that meet this threshold