eQTLs - expression quantitative trait loci

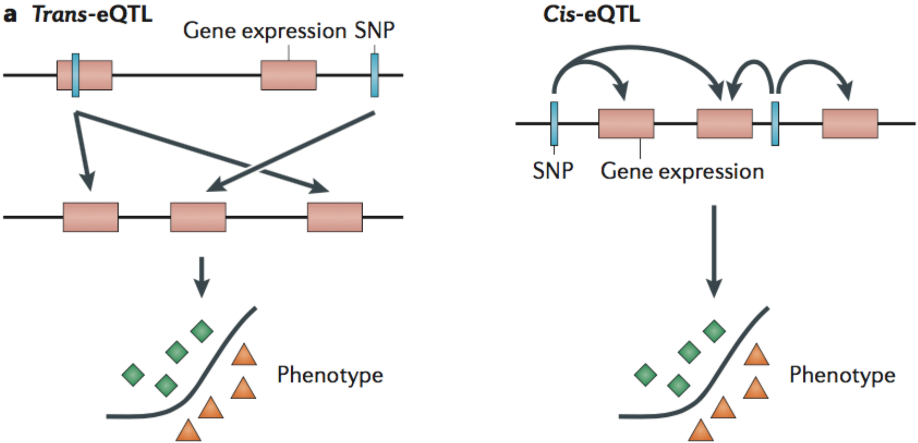
Mikhail Dozmorov

Spring 2018

eQTL analysis

- Understanding regulatory relationships between genetic variants and gene expression is important for deciphering biological mechanisms underlying a wide range of human diseases
- **eQTLs** - genomic loci that regulate expression level of mRNAs.
    - **cis-eQTL** - SNVs regulate the expression of the nearby genes
    - **trans-eQTLs** - SNVs regulare the expression of genes to which the SNVs do not directly relate
- Analysis of expression quantitative trait loci (eQTLs) analysis involves the identification of genetic variation associated with measures of quantitative gene expression

eQTLs


a *Trans*-eQTL
Gene expression
SNP
Phenotype
Cis-eQTL
SNP
Gene expression
Phenotype

Goal: Construct eQTL network using MultiVariate Linear Regression

- **Responses** - RNA expression levels $y_1, y_2, ..., y_Q$
- **Predictors** - SNV genotypes $x_1, x_2, ..., x_P$
- A straightforward approach - multivariate linear regression (MVL)

$$y_q = \sum_{p=1}^{P} x_p \beta_{pq} + \epsilon_q,\ \ q = 1, ..., Q$$

- **Goal:** construct the regulatory network - identify non-sero entries in the P x Q coefficient matrix $B = (\beta_{pq})$

eQTLs: Gene-environment interaction

$$H_0 : t_i = \beta_g g_i + \beta_e e_i + \mu + \epsilon_i$$

$$H_1 : t_i = \beta_g g_i + \beta_e e_i + \beta_{gXe} g_i e_i \mu + \epsilon_i$$

- t_i - notmalized total expression for individual i
- g_i is the genotype of the SNP encoded as $0, 1, 2$
- e_i is the environmental factor
- β_g, β_e, β_{gXe} are genetic, environment and interaction effect sizes, respectively
- μ is an intercept

Under the null the likelihood ratio $max_\beta P(t|\beta, H_1)/x_\beta P(t|\beta, H_0)$ is χ^2 distributed with one degree of freedom

eQTL problems

- Expression QTL analysis is known to be computationally intensive
- The issue is most pronounced for modern eQTL datasets, which have genotype measured over millions of SNPs and gene expression over tens of thousands of transcripts
- For such data, the eQTL analysis involves over ten billion tests

eQTL approaches

- Most eQTL studies perform separate testing for each transcript-SNP pair
- The association between expression and genotype can be tested for using linear regression and ANOVA models, as well as non-linear techniques including generalized linear and mixed models, Bayesian regression (Servin and Stephens, 2007), and models accounting for pedigree (Abecasis et al., 2001) and latent variables (Leek and Storey, 2007)
- Several methods has been developed to find groups of SNPs associated with expression of a single gene (Hoggart et al., 2008; Kao et al., 1999; Lee et al., 2008; Zeng, 1993)

Matrix eQTL R package

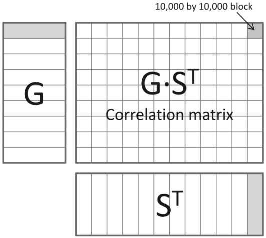
- Tests for association between each SNP and each transcript by modeling the effect of genotype as either additive linear (least squares model) or categorical (ANOVA model)
- Genotype-covariate interaction
- Covariates (e.g., gender, clinical variables, population structure and surrogate variables) can be included
- Proper treatment of relatedness of the samples
- FDR adjustment of p-values
- Distinguishing local (cis-) and distant (trans-) eQTLs
- Lightning fast

http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/

Shabalin, Andrey A. "Matrix EQTL: Ultra Fast EQTL Analysis via Large Matrix Operations." Bioinformatics (Oxford, England) 28, no. 10 (May 15, 2012): 1353–58. https://doi.org/10.1093/bioinformatics/bts163.

Matrix eQTL methods

- Deriving the common linear regression statistics from the sample correlation coefficient $r = cor(g, s)$, where g is the gene expression and s is the genotype
- Expressing correlation matrix as GS^T, where G is the gene expression matrix (genes x samples) and S is the genotype matrix (snps x samples)



http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/

Shabalin, Andrey A. "Matrix EQTL: Ultra Fast EQTL Analysis via Large Matrix Operations." Bioinformatics (Oxford, England) 28, no. 10 (May 15, 2012): 1353–58. https://doi.org/10.1093/bioinformatics/bts163.

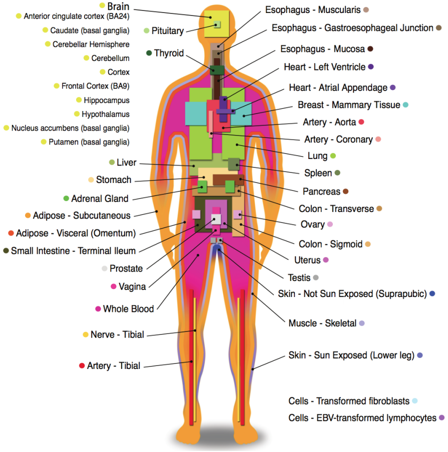
Other tools

- FastQTL - fast QTL mapping similar to Matrix eQTL. Permutation procedure to control for multiple testing modeled using beta distribution. Description of permutation techniques, from simple to adaptive early stopping to beta distribution approximation.

http://fastqtl.sourceforge.net/

The Genotype Tissue-Expression project

Atlas of gene expression and eQTLs in non-diseased human tissues from up to 960 recently deceased donors



- 53 tissue sites
- 11 distinct brain regions
- 2 cell lines
- WGS/WES (primary whole blood)
- RNA-seq

Overview of GTEx resources: open-access data

- Expression
  - Gene-level expression (TPM, counts)
  - Transcript-level expression (TPM, counts, isoform proportions)
  - Exon read counts
- QTLs
  - Single-tissue eQTLs (cis- and trans-)
  - Multi-tissue eQTLs
  - Future: splicing QTLs
- Histology images
- De-identified public access sample and subject metadata

All open-access data is available at https://gtexportal.org/home/

Overview of GTEx resources: protected data

- Sequence data:
  - RNA-seq (2x76 bp, unstranded, $>$50M reads/sample)
  - WGS (30x coverage) and WES (100x coverage)
  - Illumina Omni2.5/5 microarray genotypes (subset of 450 donors)
- Allele-specific expression (ASE)
- Full sample and subject metadata
- Future: eGTEx sequence data
  - ChIP-seq
  - WGBS-seq

All protected-access data is available at dbGaP, under accession phs000424

eGTEx: the Enhancing GTEx project

eGTEx data types

- Protein quantifications (x2)
- Methylation (WGBS)
- Histone modifications (ChIP-seq)
- Dnase-seq
- mmPCR-seq (deepASE)
- Somatic DNA-seq (deep exome seq)
- Analysis of telomere structure

eGTEx Project. "Enhancing GTEx by Bridging the Gaps between Genotype, Gene Expression, and Disease." Nature Genetics 49, no. 12 (December 2017): 1664–70. https://doi.org/10.1038/ng.3969.