Ancestry-eGenes : Differentially Expressed Genes by Ancestry and Genotype

Tae Hyun Kim

Department of Statistics, The University of Chicago

1. Introduction
    - Motivation
    - Goal
2. Data
    - GTEx and 1000GP
    - Ancestry Inference
3. Model and Methods
    - Full Model
    - Testing
4. Results
    - Ancestry-eGenes
    - Interpretation

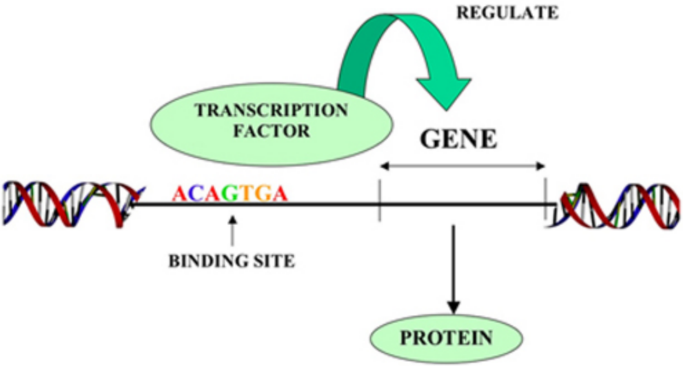
Outline

1. Introduction
   - Motivation
   - Goal
2. Data
   - GTEx and 1000GP
   - Ancestry Inference
3. Model and Methods
   - Full Model
   - Testing
4. Results
   - Ancestry-eGenes
   - Interpretation

eGenes and eQTLs

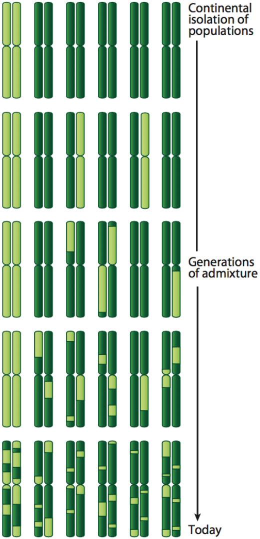
- GWAS : found genetic variants that affect certain phenotypes
  $\Rightarrow$ But how?
  - variants in coding area : has a role in protein synthesis
  - variants in non-coding area :
    genotypes $\Rightarrow$ **gene expression** $\Rightarrow$ **phenotypes**
- gene expression : measures how active a gene is
  ex) the amount of protein a gene produces.
- GTEx: Gene Tissue Expression Project
  - **eQTLs** (expression quantitative trait loci) : variants that affect the gene expression
  - **eGenes** : genes that have at least 1 eQTL

Gene expression regulation



http://2013.igem.org/Team:XMU Software/Project/promoter

Population Admixture



- two or more populations interbreed
  - ex) African American, Latino American
- analyzing admixed population : brings together different genomes and naturally control for environmental confounders

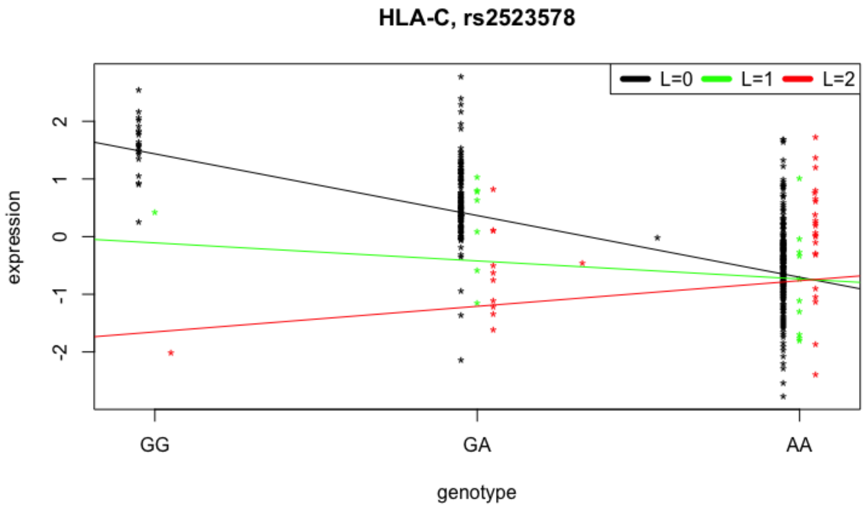
Goal

Find **ancestry-eQTLs** and **ancestry-eGenes**

- ancestry-eQTLs : SNPs that affect gene expression level differently based on the gene's local ancestry
- ancestry-eGenes : Genes that have at least 1 ancestry-eQTLs in its cis-region (start site - 1Mb, end site + 1Mb)

This project only focuses on muscle-skeletal tissue, but the method is applicable to all other tissues.

Goal



Outline

1. Introduction
   - Motivation
   - Goal
2. Data
   - GTEx and 1000GP
   - Ancestry Inference
3. Model and Methods
   - Full Model
   - Testing
4. Results
   - Ancestry-eGenes
   - Interpretation

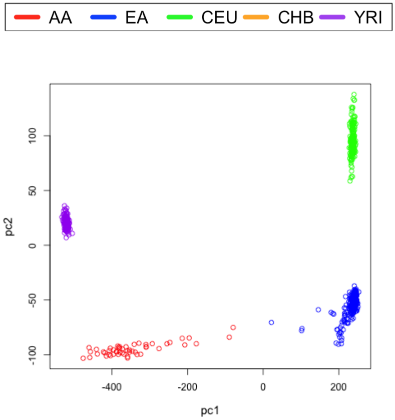
GTEx

- Subjects
  - 51 self-reported African Americans
  - 305 self-reported European Americans
    $\Rightarrow$ 4 of them turn out to be admixed
  - final sample : 55 African Americans, 301 European Americans
- Expression (muscle skeletal tissue)
  - pre-processed expression level data from GTEx Portal
  - truncated for having at least 0.1 RPKM
  - normalized, log-transformed, corrected for technical artifacts
  - total 22,248 genes in autosomal chromosomes
- Genotypes
  - Illumina OMNI 5M SNP array + imputation
  - minor allele frequency $>$5%
  - total 6,954,165 SNPs

Pure population data

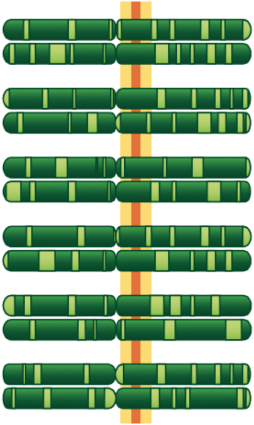
- **1000GP** has haplotype data for pure populations including CEU, YRI, CHB etc.
- Population code
  - CEU: Central Europeans in Utah
  - YRI: Yoruban in Ibadan Nigeria (West Africa)
  - CHB: Han Chinese from Beijing
- With this information
  (1) verify the population information of GTEx subjects through PCA
  (2) compute pure population allele frequency

Principal Component Analysis



- 1st principal component:
  left: African
  right: European
  ⇒ can compute global ancestry
- 4 blue outliers

Definition of Local Ancestry



- Local ancestry of a gene: local ancestry at the locus that is closest to the gene
- L=2 : two African chromosomes
- L=1 : heterozygous
- L=0 : two European chromosomes

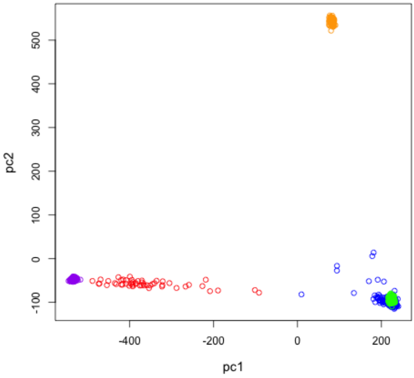
LAMP

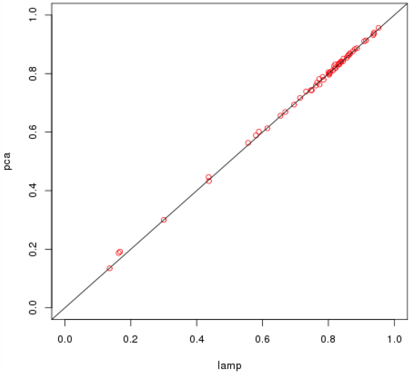
- Window-based approach
- Assumes that each window comes from the same ancestry
- Allows overlapping windows, adaptive window size
- Achieves $\sim$ 98% accuracy for distinguishing YRI/CEU
- Sample Output

Global Ancestry

Compare

- average local ancestry inferred from LAMP
- the distance to YRI divided by the distance between CEU and YRI





Outline

1. Introduction
   - Motivation
   - Goal
2. Data
   - GTEx and 1000GP
   - Ancestry Inference
3. Model and Methods
   - Full Model
   - Testing
4. Results
   - Ancestry-eGenes
   - Interpretation

Model

$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + G_s\lambda_{ks} + LG_{ks}\theta_{ks} + X\nu_{ks}$$

$$H_0 : \theta_{ks} = 0$$

n: sample size $= 356$

k: gene index, s: SNP index

y_k: gene expression level of gene k

μ_{ks}: intercept of this model

$\mathbb{1}_{EA}$: mean gene expression level of European Americans

A: global ancestry vector of length n

L_k: local ancestry vector of length n for gene k

G_s: genotype vector of length n for SNP s

LG_{ks}: interaction vector, element-wise product of L_k and G_s

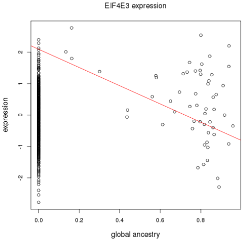
$\alpha, \beta, \gamma, \lambda, \theta$: scalar coefficients

Joint Modeling

$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + G_s\lambda_{ks} + LG_{ks}\theta_{ks} + X\nu_{ks}$$

Is it statistically valid to use the same model for both African Americans and European Americans?

- Indicator variable $\mathbb{1}_{EA}$
  - $E(y_k) = \mu_k + \mathbb{1}_{EA}\ \alpha_k + A\beta_k + L_k\gamma_k$
  - Is assigning 0 to European Americans' global ancestry okay?
  - $\Rightarrow$ No. Separate mean term for the expression of European Americans



- Constant variance test
  - $y_k = \mu_k + \mathbb{1}_{EA}\alpha_k + A\beta_k + L_k\gamma_k + \epsilon_k$
  - $\epsilon_{k,AA} \sim \mathcal{N}(0, \sigma^2_{k,AA}), \epsilon_{k,EA} \sim \mathcal{N}(0, \sigma^2_{k,EA}),\ H_0\text{:}\ \sigma_{k,AA} = \sigma_{k,EA}$
  - p-values are very close to uniform

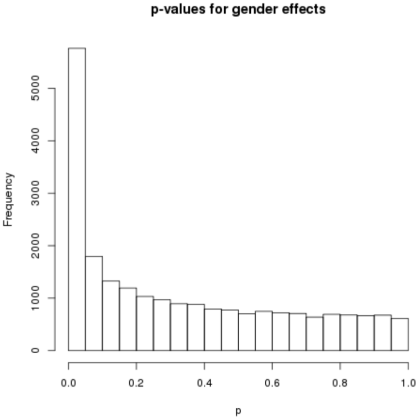
Covariates X

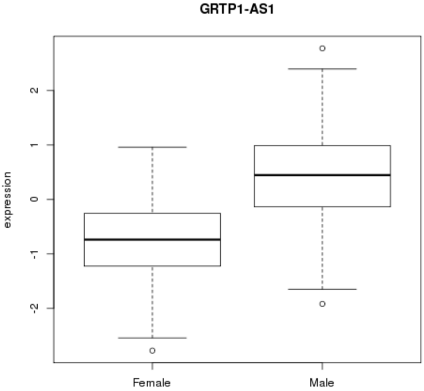
$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + X\nu_{ks}$$

- Gender
- Age
- The first two principal components of expression level matrix

Covariates X

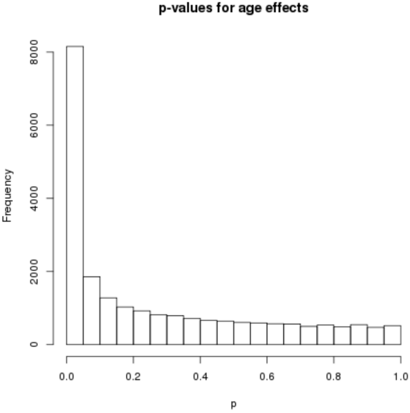
- Gender
- Age
- The first two principal components of expression level matrix

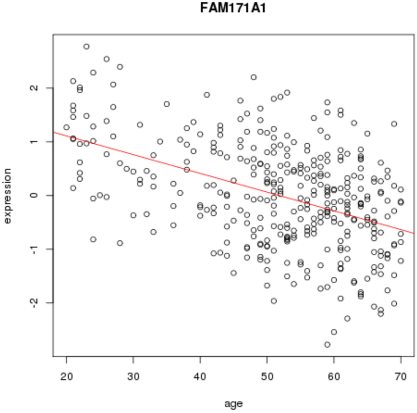

p-values for gender effects


GRTP1-AS1

Covariates X

- Gender
- Age
- The first two principal components of expression level matrix

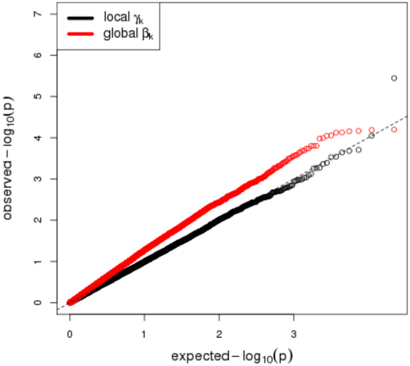


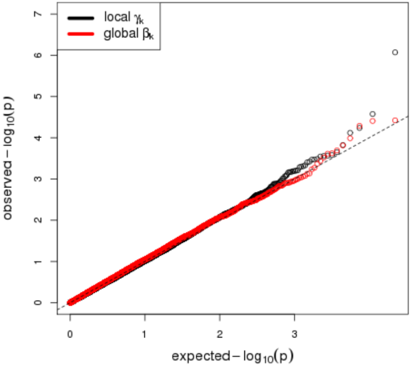


Covariates X

- Gender
- Age
- The first two principal components of expression level matrix

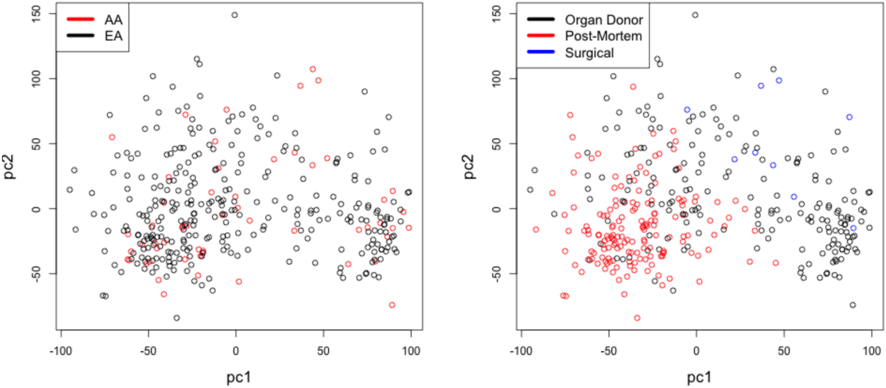
$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + X\nu_{ks}$$





Covariates X

- Gender
- Age
- The first two principal components of expression level matrix



Test Statistic

$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + G_s\lambda_{ks} + LG_{ks}\theta_{ks} + X\nu_{ks}$$

- Goal : find gene k where for at least one SNP s, $\theta_{ks} \neq 0$
- Problem
  - correlations due to linkage disequilibrium
  - appropriate type I error control
- Use gene-specific test statistic
  - get T_{ks} = t-value testing the correlation of interaction and expression
  - get $t_k = max_s(|T_{ks}|)$
  - t_k : evidence that gene k has at least one ancestry-eQTL

Permutation Test

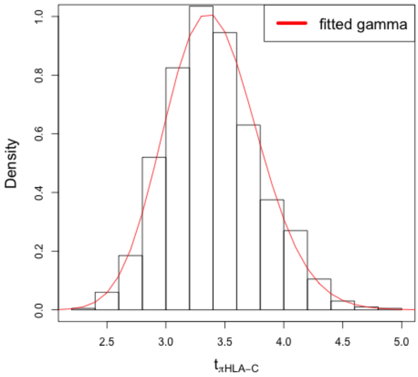
T_{ks} = t-value testing the correlation of interaction and expression
$t_k = max_s(|T_{ks}|)$

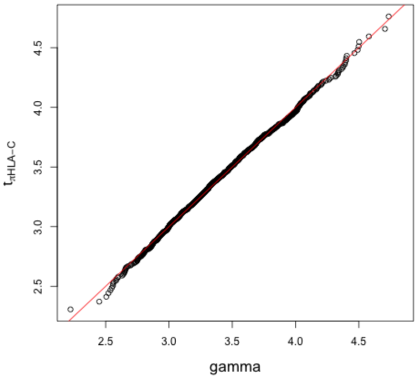
How do I get p-value of t_k?

- Estimate null distribution of t_k through permutation test
- Compute $t_{\pi,k}$ for each gene for permutations π_1, .., π_{1000}
- Which variables should be permuted?
  - SNPs are correlated due to LD : must be preserved
  - don't permute genotype information
    $\Rightarrow$ genotype, interaction term
  - permute everything else
    $\Rightarrow$ expression, covariates, mean term, ancestry
- Compare t_k to the estimated null distribution of $t_{\pi,k}$

Null Distribution of t_k

$$t_k = max_s(|T_{ks}|)$$
$$p_k = Pr(x > t_k; x \sim \Gamma(\alpha_k, \beta_k))$$

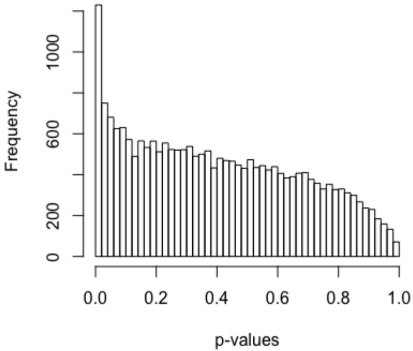


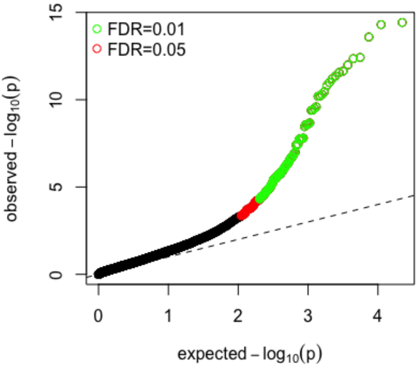


Outline

1. Introduction
   - Motivation
   - Goal
2. Data
   - GTEx and 1000GP
   - Ancestry Inference
3. Model and Methods
   - Full Model
   - Testing
4. Results
   - Ancestry-eGenes
   - Interpretation

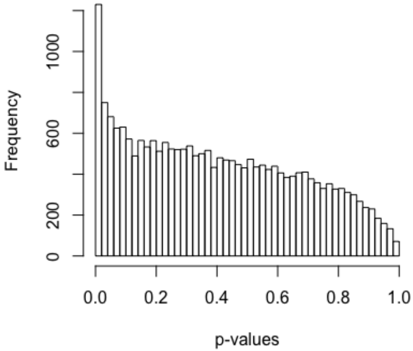
Distribution of *p*-values

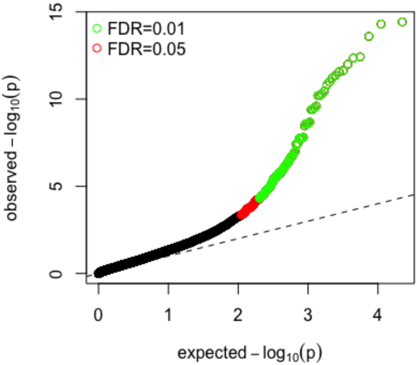




- unexplained confounder
- gamma misfitting
- real signals

Distribution of *p*-values





- found 201 genes under FDR=0.05 (red)
- found 110 genes under FDR=0.01 (green)

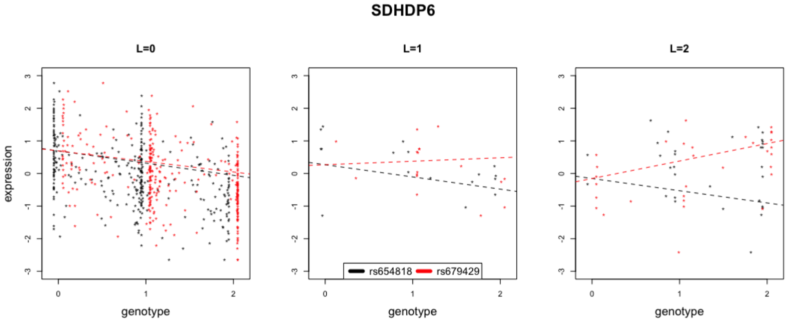
Functional Analysis

- DAVID: extract biological functions from the gene list
- Returned enrichment of keyword 'MHC'
- Fisher exact test
  - odds ratio 12.23
  - p-value 8.56e-13
- MHC (major histocompatibility complex) region contains HLA (human leukocyte antigens) genes.
  - responsible for making receptors for pathogens
  - play a large role in immunity

	MHC	non-MHC
Signals (FDR <0.05)	14	187
Non-Signals	165	21852

LD causing interaction

- Europeans have higher LD than Africans
- This can lead to an interaction between genotype and ancestry confounded with an eQTL



Other explanations for interaction

- Africans and Europeans have different transcription factors
- Africans and Europeans have different binding sites
- A SNP interacts with another SNP that only exists in one of the populations

Summary

- Found ancestry-eGenes that have at least one SNP (ancestry-eQTL) in their cis-region that have different effects on the expression level based on local ancestry
- Significant enrichment of these genes in the MHC region
- Suggests immunological difference between Africans and Europeans
- Possible explanation : linkage disequilibrium
- Other molecular mechanisms to be studied further