QRank: Quantile Regression for eQTL Analysis

Gen Li

Department of Biostatistics, Columbia University

March 27, 2017

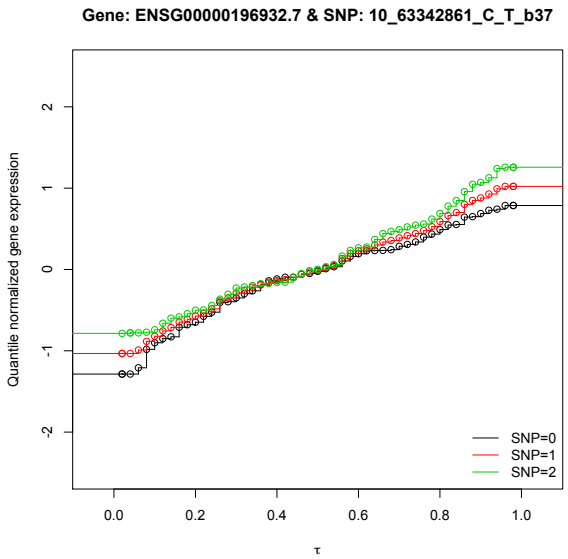
Motivation

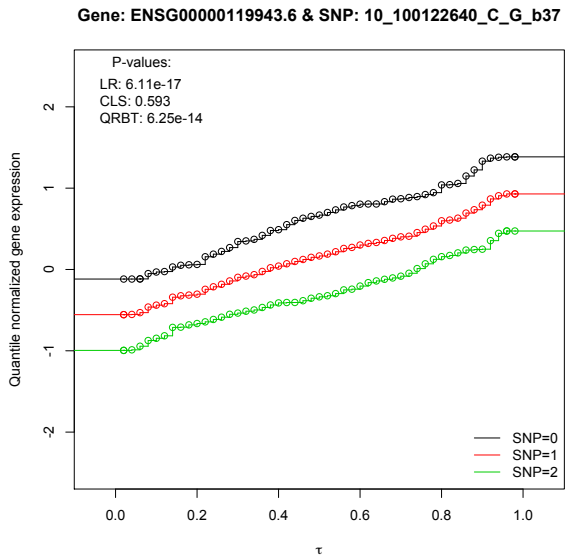
- Expression quantitative trait loci (eQTLs) analysis can elucidate genetic regulatory pathways of complex diseases
- Most eQTL studies focus on identifying mean effects
- Linear regression models are commonly used

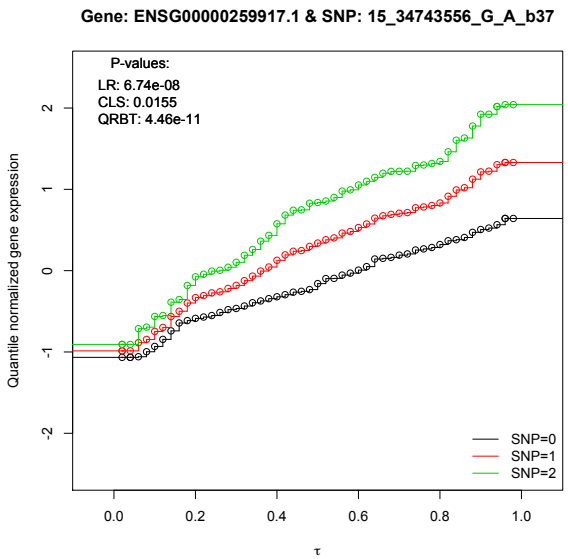
Motivation

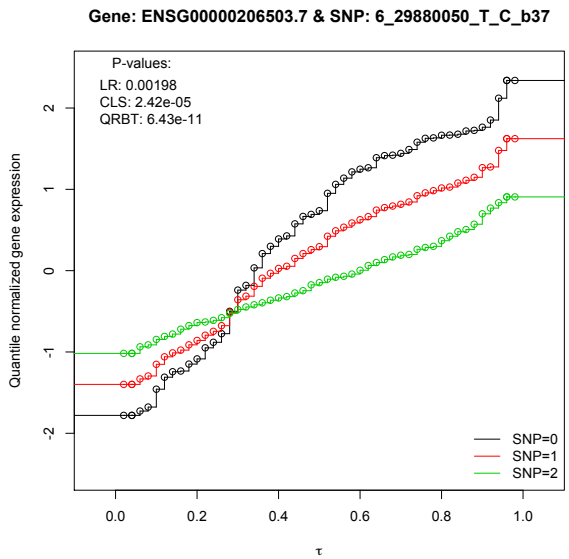
- Expression quantitative trait loci (eQTLs) analysis can elucidate genetic regulatory pathways of complex diseases
- Most eQTL studies focus on identifying mean effects
- Linear regression models are commonly used
- However, genetic variants may affect the entire distribution of gene expressions
- This distributional heterogeneity is understudied

Examples of Distributional Heterogeneity









Quantile Regression for eQTL Discovery

In a particular tissue with n samples, for $i \in \{1, 2, ..., n\}$,

- $Y_{i,k}$ is the expression level of the kth gene in the ith sample
- $x_{i,j}$ is the jth SNP within $\pm$1MB of the TSS of the gene in the ith sample
- $\boldsymbol{z}_i$ is a covariate vector for the ith sample

For each gene-SNP pair k and j, we assume the quantiles of $Y_{i,k}$ follow the model:

$$Q_{Y_{i,k}}(\tau|\boldsymbol{z}_i, x_{i,j}) = \boldsymbol{z}_i\boldsymbol{\alpha}_{jk,\tau} + x_{i,j}\beta_{jk,\tau},$$

- **Goal**: identify the k and j pairs for $\beta_{jk,\tau} \neq 0$ for any given $\tau \in (0, 1)$

QRank Test

Define the rank-score function for a fixed quantile τ as

$$S_{n,\tau} = n^{-1/2} \sum_{i=1}^{n} \rho_\tau\{y_{i,k} - \boldsymbol{z}_i \widehat{\boldsymbol{\alpha}}_{jk,\tau}\} x^*_{i,j}$$

- $\rho_\tau\{u\} = \tau - \mathbb{I}(u < 0)$ is an asymmetric sign function
- $\widehat{\boldsymbol{\alpha}}_{jk,\tau}$ is the estimated coefficient vector under the null $H_0 : \beta_{jk,\tau} = 0$
- $x^*_{i,j}$ is the genotype data adjusted by the covariates
- $S_{n,\tau}$ is close to zero if and only if $\beta_{jk,\tau} = 0$

QRank Test

- Test statistic at a fixed quantile τ:

$$T_{n,\tau} = S^2_{n,\tau}/V_n \longrightarrow \chi^2_1, \text{ as } n \to \infty$$

where V_n^{-1} is the variance of $S_{n,\tau}$ such that $V_n = n^{-1}\tau(1-\tau)\boldsymbol{x}_j^{*T}\boldsymbol{x}_j^*$

- Composite test statistic $(\tau_1, \cdots, \tau_\ell)$:

$$T_{n,\ell} = \boldsymbol{S}^T_{n,\ell}\boldsymbol{\Sigma}^{-1}_{n,\ell}\boldsymbol{S}_{n,\ell} \longrightarrow \chi^2_\ell, \text{ as } n \to \infty$$

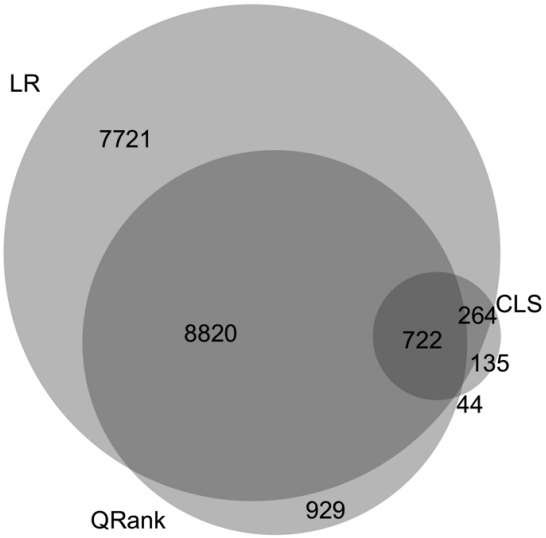
where $\boldsymbol{S}_{n,\ell} = (S_{n,\tau_1}, \cdots, S_{n,\tau_\ell})$ and $\boldsymbol{\Sigma}_{n,\ell}$ has an explicit expression

Numerical Results

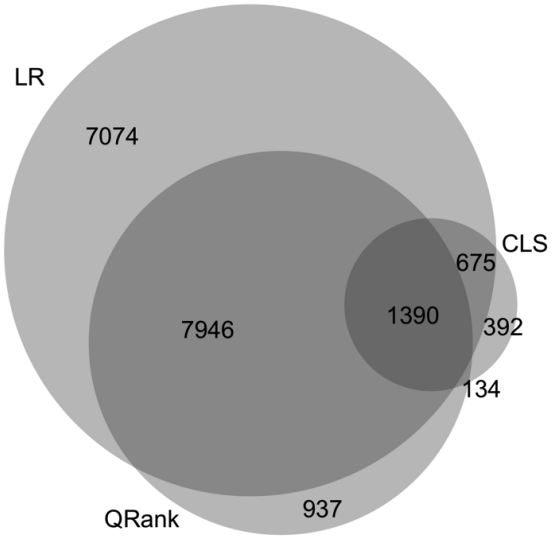
- Compare QRank (with 5 quantile levels) with LR and CLS[1]
- Type I error and power analysis in simulation
- GTEx v6 data analysis (in 4 tissues)
- Gene-level discoveries, tissue specificity, GWAS enrichment

[1] Brown et al., *Elife* (2014)

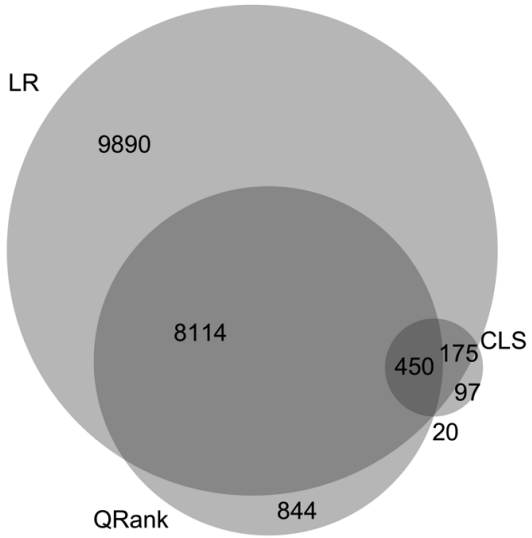
eQTL Discoveries (Unique Genes, FDR=0.05)



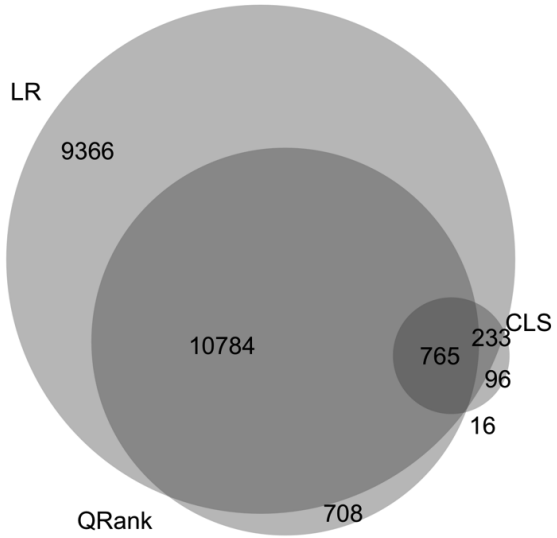
(a) Muscle Skeletal



(b) Whole Blood



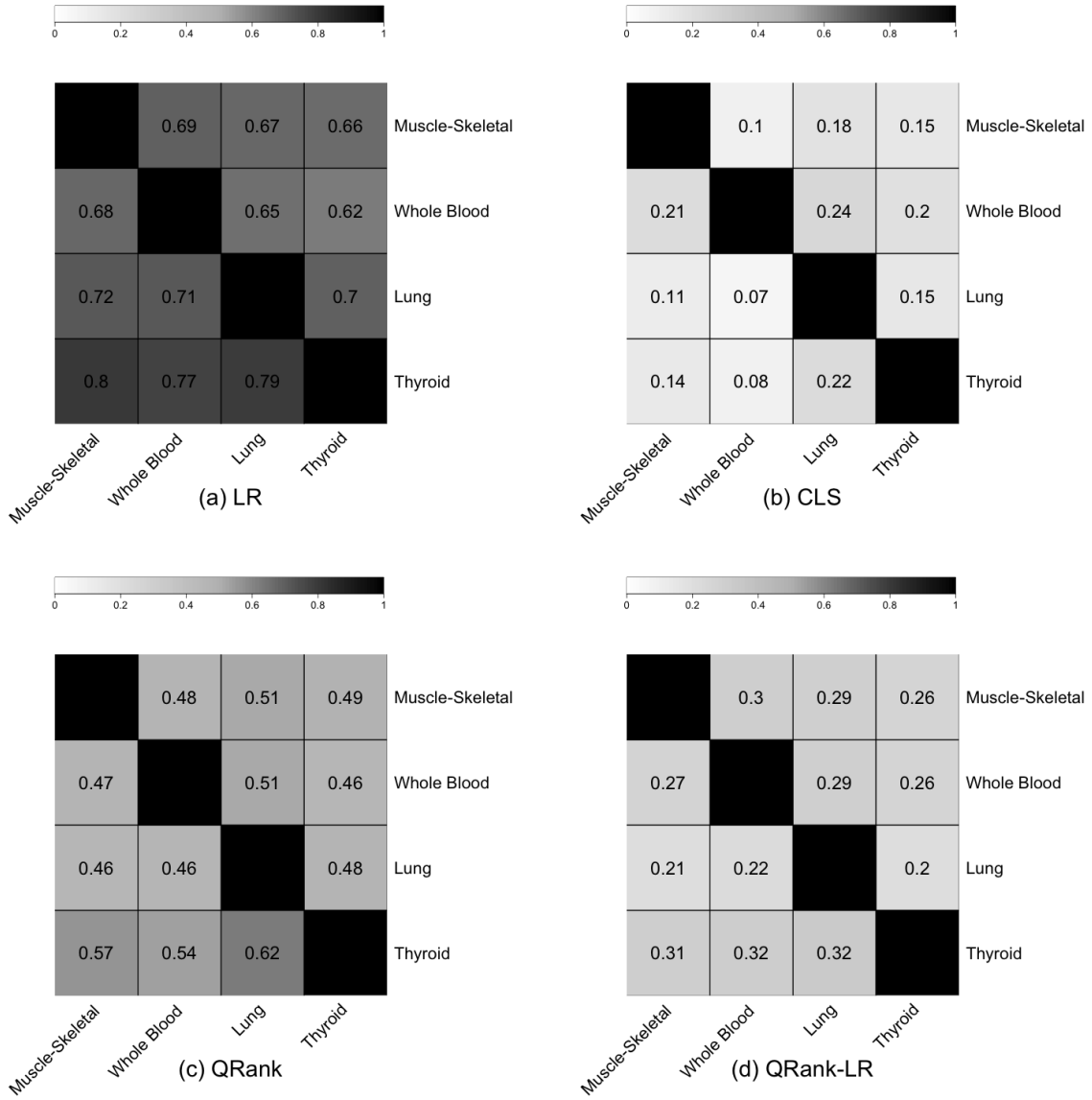
(c) Lung



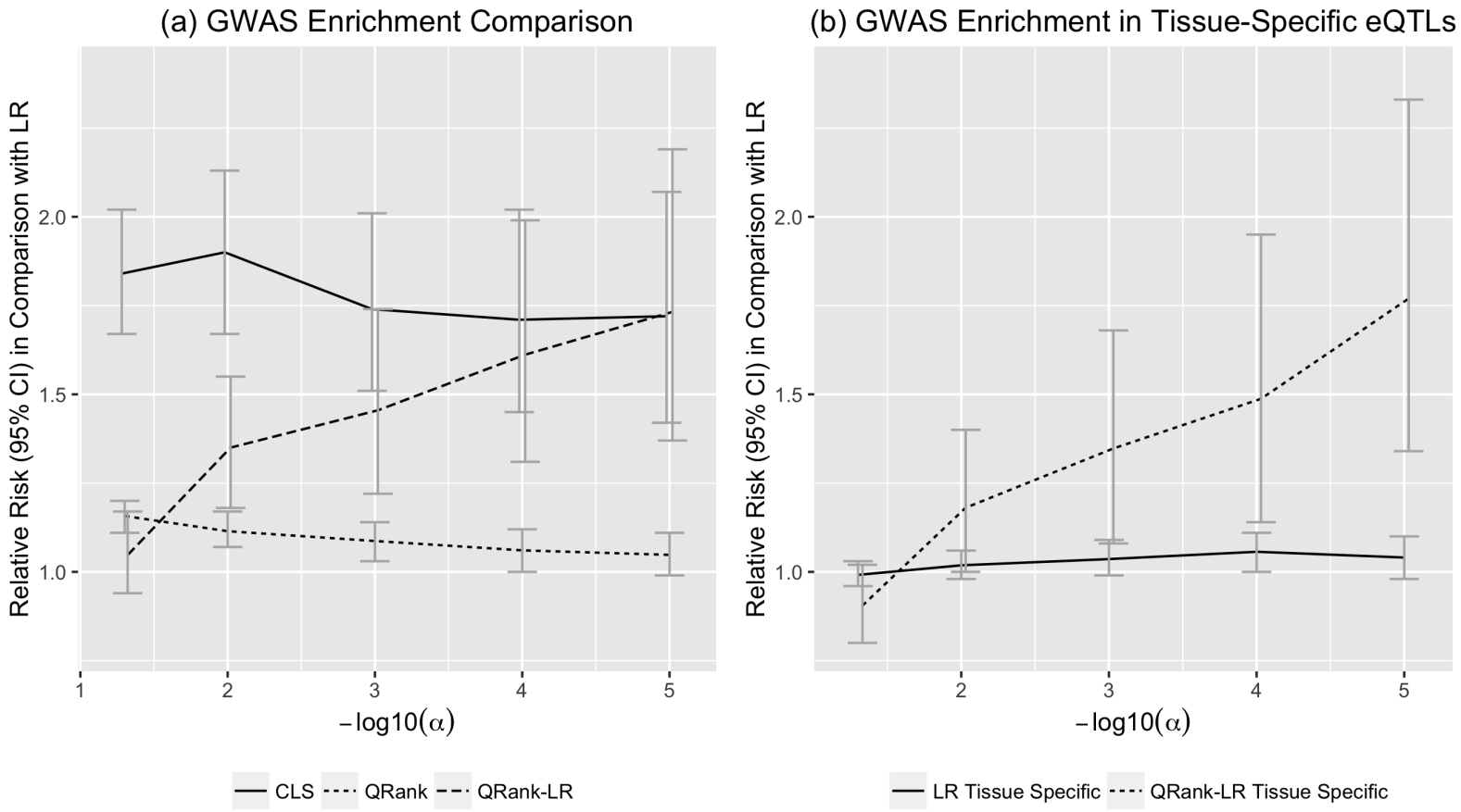
(d) Thyroid

Tissue Specificity

- eQTL sharing coefficient $\pi_{ij} = \mathbb{P}(\text{eQTL in tissue i} | \text{ eQTL in tissue j})$



GWAS Enrichment



Acknowledgement

Collaborators:
Xiaoyu Song
Zhenwei Zhou
Xianling Wang
Iuliana Ionita-Laza
Ying Wei

Funding:
NSF (DMS-120923)
NIH (R01HG008980)
Calderone Junior Faculty Award
MSPH, Columbia University

Song et al. “QRank: A novel quantile regression tool for eQTL discovery." *Bioinformatics*, 2017+
(`http://biorxiv.org/content/early/2016/08/17/070052`)
R package available at `https://github.com/cran/QRank`