

Biological Networks
Comparative Microarray Experiments,
Unwanted Variation
and
eQTL Analysis

October 3, 2014

Microarray (Expression) Arrays

Subject	Group/ Genotype	Gene				
		1	2	.	.	m
1	0					
2	0					
.	1					
.	0					
.	.					
.	.					
n	1					

y_{ig} = (log) expression of gene g in subject i

x_i = group indicator for subject i

Naive Analysis

$$y_g = X\beta_g + e_g \quad e_g \sim N(0, \sigma_g^2 I_n)$$

$$X = [1, x_g] \quad (n \times 2) \text{ matrix}$$

$$\beta_g = (\beta_{0g}, \beta_{1g})^T$$

$$\hat{\beta}_g = (X^T X)^{-1} X^T y_g \quad \text{least squares estimate}$$

$$s_g^2 \quad \text{pooled variance estimate}$$

t-tests

$$t_g = \frac{\hat{\beta}_{1g}}{s_g \sqrt{\nu_g}} \sim t_{d_g}, \quad g = 1, \dots, m$$

Multiple testing: Bonferroni correction or FDR control!

Typically m is very large; e.g. $O(10e3)$

Often n is quite small; e.g. $O(10e1)$ or $O(10e2)$

Hierarchical (Bayesian) Model

Smyth (2004) SAGMB 9, No.1, Article 39

Data Model

$$\hat{\beta}_{1g} \mid \beta_{1g}, \sigma_g^2 \sim N \left(\beta_{1g}, \frac{\sigma_g^2}{sxx_g} \right), \quad s_g^2 \mid \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2$$

Prior

$$\beta_{1g} \mid \sigma_g^2 \sim N(0, \nu_0 \sigma_g^2), \quad \sigma_g^{-2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2$$

Gamma prior with mean s_0^2

Posterior Analysis

Posterior mean of $\sigma_g^{-2} \mid s_g^2$ is $\tilde{s}_g^{-2}$ where

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

Moderated t-statistics:

$$\tilde{t}_g = \frac{\hat{\beta}_{1g}}{\tilde{s}_g \sqrt{\nu_g}} \sim t_{d_g + d_0}, \quad g = 1, \dots, m$$

if $\beta_{1g} = 0$

Mixture Model

Suppose $P(\beta_{1g} \neq 0) = p$ and

$$\beta_{1g} \mid \sigma_g^2, \beta_{1g} \neq 0 \sim N(0, \nu_0 \sigma_g^2)$$

Posterior (log) odds

$$B_g = \log \frac{P(\beta_{1g} \neq 0 \mid \hat{\beta}_{1g}, s_g^2)}{P(\beta_{1g} = 0 \mid \hat{\beta}_{1g}, s_g^2)}$$

monotone in $|\tilde{t}_g|$

Empirical Bayes

Estimate hyperparameters from marginal likelihood

Plug-in posterior mean of the variance and use the EM-algorithm to fit the mixture model

The LIMMA R package uses ad hoc estimates of the hyperparameters

Other similar approaches by Bar et al. (2010), Hwang and Liu (2010)

Unwanted Variation

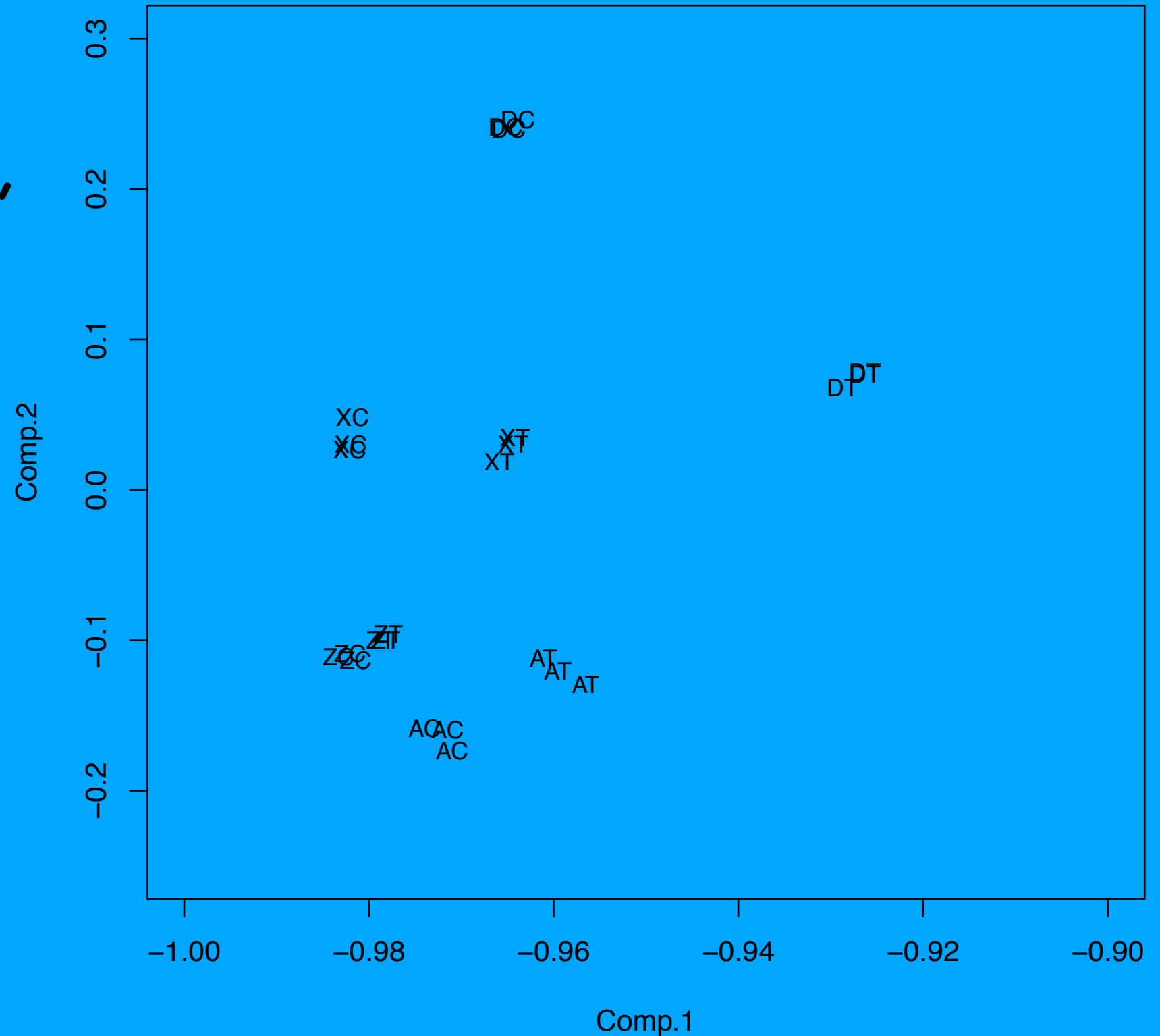
Jason Locasale

Division of Nutrition, Cornell

Gagnon-Bartsch and Speed (2012) Biostatistics

Lior Pachter's Blog

Correlations of Samples with PCs: Raw Data



Factor Analysis

$$y = \mu + \Lambda f + u$$

y ($n \times 1$) **multivariate response**

f ($k \times 1$) **vector of common factors**

Λ ($n \times k$) **factor loading matrix**

u ($n \times 1$) **specific/unique factors**

$$E(f) = 0 \quad V(f) = I$$

$$E(u) = 0 \quad V(u) = \text{diag}(\psi_i) \quad \text{or} \quad \psi I = D$$

$$C(f, u) = 0$$

Factor Analysis

The model implies

$$V(y) = \Lambda\Lambda^T + D = \Lambda GG^T \Lambda^T + D$$

for any orthogonal matrix, G

Constrain $\Lambda^t D^{-1} \Lambda$ to be diagonal

$$V(y_i) = \sum_{j=1}^k \lambda_{ij}^2 + \psi_i$$

communality + specific variance

λ_{ij}^2 is the extent to which y_i depends on the j th common factor

HEFT Model (Gao et al. 2013)

$$y_g = \mu + X\beta_g + \Lambda f_g + e_g, \quad g = 1, \dots, m$$

$n \times 1$

$$Y = \mu 1^T + XB + \Lambda F + E$$

$n \times m$

Columns of Y are standardized in advance

Means subtracted from rows of Y

$$e_g \sim N(0, \sigma^2 I_n) \quad \text{independently} \quad g = 1, \dots, m$$

$$\beta_g \sim N(0, I_2) \quad \text{independently} \quad g = 1, \dots, m$$

Lung airway dataset

n=118: 79 smokers, 37 non-smokers

m=7575 genes

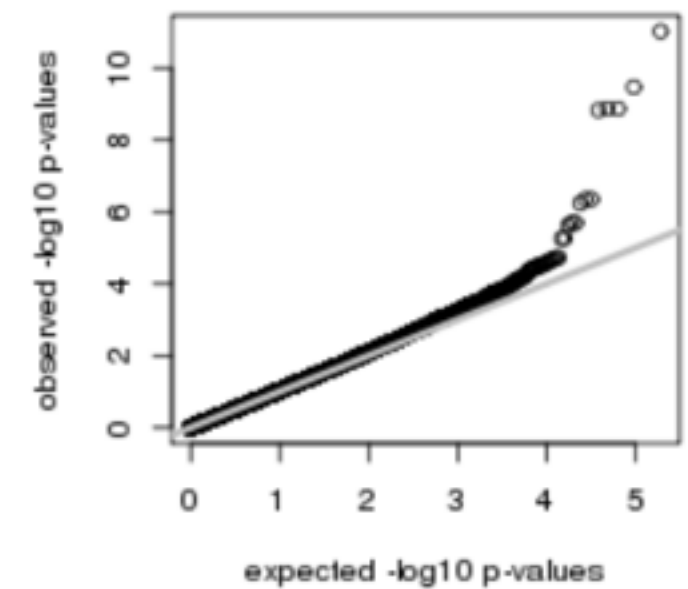
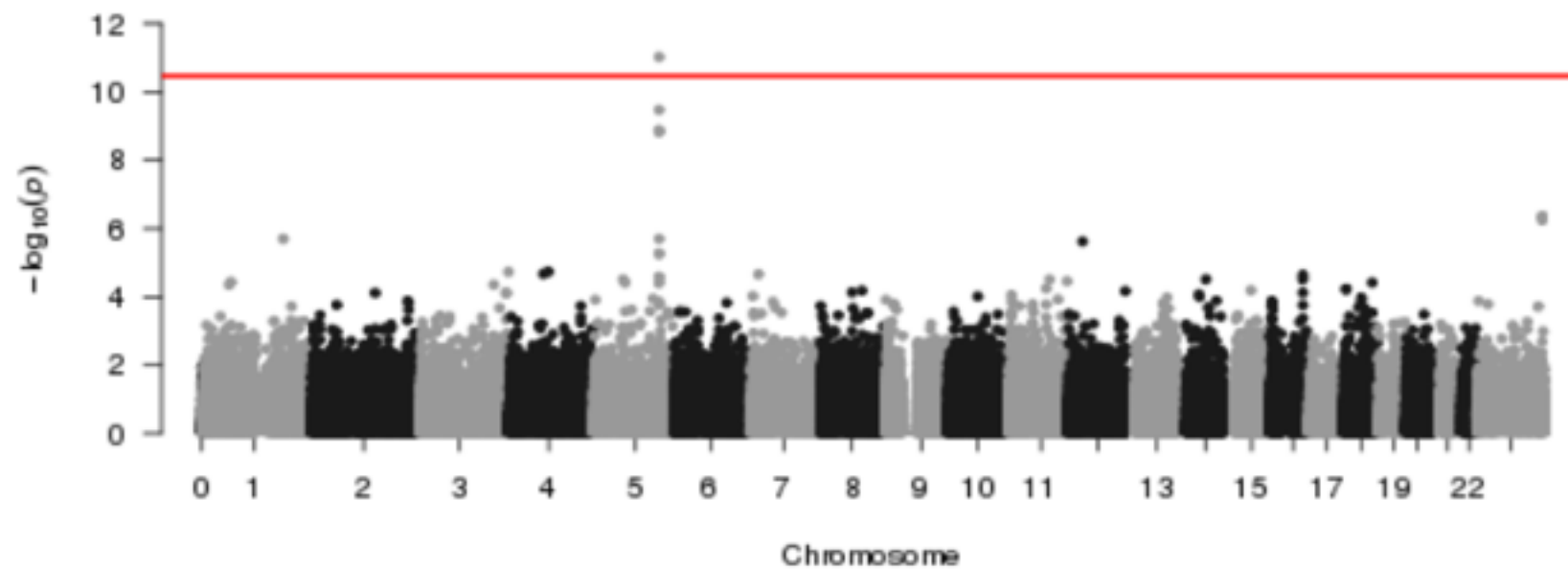
p=191,959 genotypes

96 non-duplicated pairs discovered using a Bonferroni
threshold $0.05/mp = 3.4 \text{ times } 10e-11$

61 of the 96 are cis

Manhattan and QQ plots

1 of 2 Extra Genes by LM: GeneID = 11005 , SPINK5



2 of 2 Extra Genes by LM: GeneID = 54059 , YBEY

