Mediation test

Scientific background

The goal of the current anaysis is to estimate the contribution of DNA methylation to phenotype differences (e.g., tissue, species, etc.) in gene expression. The general idea is that methylation is highly correlated with gene expression and is likely a mediating factor for phenotype difference in gene expression. In other words, the total effect of phenotype on expression is consisted of direct effect not due to methylation and indirect effect due to methylation. Consider a hypothetical example. Say we know A causes B, and this relationship is likely due to C which causes B and is caused by A (A -> C -> B). The effect size estimating A-> B is the total effect. Now, say we control for the effect of C on B, the we can estimate the direct effect of A on B (A|C -> B). Total effect equals the sum of direct and indirect effect. Hence, we can estimate indirect effect by taking the difference between the total and the direct effect. The magnitude of the indirect effect provides an estimate for the mediating effect of C on the relationship between A and B. For example, in a completely mediation case, indirect effects are signficantly small, possibly close to zero. And in a partial mediation case, indirect effects are significantly small and can even be at the same magnitude as the direct effects. We use `ash` to compare direct versus indirect effects and to obtain a robust estimate of effect size differences, by accounting for standard error of both effect sizes.

We use `ash` to compare direct versus indirect effects and to obtain a robust estimate of effect size differences, by accounting for standard error of both effect sizes.

In this document, we assume the phenotype is species, although one can readily substitute other phenotypes for species (so long as it's two-group comparison).

Approach

We use linear models to perform the following analysis:

Step 1. Estimate the effect of species on gene expression levels
Step 2. Regression methylation levels out from expression levels. Step 3. Estimate the effect of species on gene expression levels after controlling for methylation levels
Step 4. Compute the covariance of the differences of the effect sizes in Step 1 versus Step 2
Step 5. Use `ash` to estimate false discovery rate

Data

Gene expression levels and methylation levels have been preprocessed and transformed to logarithm scale. For gene g, we observe gene expression level Y_g and DNA methylation level M_g for species s in sample i. We assume that

1. DNA methylation level M_g is fixed, i.e., not a random variable.
2. $E(Y_g) = \mu_g$ and $Var(Y_g) = \sigma_g^2 \mathbf{I}_N$, where $Y^g \sim N(\mu_g, \sigma_g^2 \mathbf{I}_N)$. In other words, sample gene expression levels are independent both within and between species.

Methods

Notations

$s = 1, 2$: two species, human and chimpanzee

$N = n_1 + n_2$: Total sample size N equals n_1 human samples plus n_2 chimp samples.

$R = (R_H, R_C)'$: length-N RIN scores.

$Y_g = (Y_{gH}, Y_{gC})'$: length-N transcriptional expression vector for gene g.

$M_g = (M_{gH}, M_{gC})'$: length-N methylation level vector for gene g.

Estimate the unconditional effet of species on expression

We fitting a linear model including species and RIN score as covariates

$$E(Y_g) = X_1\beta_{g1}$$

where $\beta_{g1} = (\beta_{g1}^I, \beta_{g1}^S, \beta_{g1}^R)'$ denote intercept, species and RIN socre effect, respectively. X_1 is the $N \times 3$ design matrix

$$X_1 = \begin{bmatrix} \mathbf{1}_n & \mathbf{1}_n & R_H \\ \mathbf{1}_n & -\mathbf{1}_n & R_C \end{bmatrix}.$$

It follows that the least-square estimator $\hat{\beta}_{g1} = (X'X)^{-1}X'Y_g$, and

$$Var(\hat{\beta}_{g1}) = \sigma_g^2(X_1'X_1)^{-1}.$$

Estimate the conditional effect of species on expression

We estimate the conditional effect of species on expression after controlling for methylation levels in two steps. First, we fit an analysis of covariance model with methylation levels as the covariate.

$$E(Y_g) = X_{2g}\beta_{g2} = (1_N, M_g)\beta_{g2} \qquad \text{(Eq. 2)}$$

where $\beta_{g2} = (\beta_{g2}^I, \beta_{g2}^M)'$ correspond to intercept, species, RIN score effect, and methylation effect, respectively. The least-square solution $\hat{\beta}_{g2} = (X_{2g}'X_{2g})^{-1}X_{2g}'Y_g$, and

$$Var(\hat{\beta}_{g2}) = \sigma_g^2(X_{2g}'X_{2g})^{-1}.$$

The estimated methylation effect $\hat{\beta}_{g2} = C'\beta_{g2}$ where $C' = (0, 1)$. And

$$Var(\hat{\beta}_{g2}^M) = Var(C'\hat{\beta}_{g2}) = \sigma_g^2 C'(X_{2g}'X_{2g})^{-1}C.$$

The second step estimates the species effect on expression after accounting for the effect of methylation levels on expression. We fit a linear model

$$E(Y_g - M_g C'\hat{\beta}_{g2}) = X\beta_{g3}$$

where $\beta_{g3} = (\beta_{g3}^I, \beta_{g3}^S, \beta_{g3}^R)'$ corresponds to intercept, species effect, and RIN score effect on gene expression after controlling for methylation levels. The least-square solution $\hat{\beta}_{g3}$ is $(X_1'X_1)^{-1}X_1'(Y_g - M_g C'\hat{\beta}_{g2})$. Subsitute $\hat{\beta}_{g2}^M$ and denote $A_1 = (X_1'X_1)^{-1}X_1'$, $A_{2g} = (X_{2g}'X_{2g})^{-1}X_{2g}'$, then

$$\begin{aligned}\hat{\beta}_{g3} &= (X_1'X_1)^{-1}X_1'(Y_g - M_gC'\hat{\beta}_{g2}) \\ &= \hat{\beta}_{g1} - A_1M_gC'\hat{\beta}_{g2},\end{aligned}$$

and

$$\begin{aligned}Var(\hat{\beta}_{g3}) &= Var(\hat{\beta}_{g1}) + A_1M_gC'Var(\hat{\beta}_{g2})CM_g'A_1' - 2Cov(\hat{\beta}_{g1}, A_1M_gC'\hat{\beta}_{g2}) \\ &= Var(\hat{\beta}_{g1}) + A_1M_gC'Var(\hat{\beta}_{g2})CM_g'A_1' - 2\sigma_g^2A_1A_{2g}'CM_g'A_1'\end{aligned}$$

Covariance of the difference between direct and indirect species effects

First, we compute the covariance between $\hat{\beta}_{g1}^S$ and $\hat{\beta}_{g3}^S$,

$$\begin{aligned}Cov(\hat{\beta}_{g1}, \hat{\beta}_{g3}) &= Cov(\hat{\beta}_{g1}, \hat{\beta}_{g1} - A_1M_gC'\hat{\beta}_{g2}) \\ &= Var(\hat{\beta}_{g1}) - Cov(A_1Y_g, A_1M_gC'A_2Y_g) \\ &= Var(\hat{\beta}_{g1}) - \sigma_g^2A_1A_2'CM_g'A_1'\end{aligned}$$

Putting the above results together, we can compute the standard error for the difference of effect size

$$Cov(\hat{\beta}_{g1}^S - \hat{\beta}_{g3}^S) = Var(\hat{\beta}_{g1}^S) + Var(\hat{\beta}_{g3}^S) - 2Cov(\hat{\beta}_{g1}^S, \hat{\beta}_{g3}^S).$$