Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals

Zhiguang Huo[1] Chi Song[2*] George Tseng[3*]

[1]Department of Biostatistics, University of Florida [2]Department of Biostatistics, The Ohio State University [3]Department of Biostatistics, University of Pittsburgh

Background

- Abundant transcriptomic data in public repositories.
- Meta-analysis to increase statistical power and reproducibility.
- Understand the interaction between the disease and different phenotypes.

Goal

1. Perform transcriptomic meta-analysis via combining p-values.
2. Discovery meta-analysis differential expression pattern (meta-pattern).

Background for meta-analysis

We consider three frequentists' hypothesis testing alternatives:

1. Biomarkers that are DE in all studies (HS_A):
   - $H_0: \vec{\theta} \in \bigcap\{\theta_s = 0\}$ vs $H_A: \vec{\theta} \in \bigcap\{\theta_s \neq 0\}$
2. Biomarkers that are DE one or more studies (HS_B):
   - $H_0: \vec{\theta} \in \bigcap\{\theta_s = 0\}$ vs $H_A: \vec{\theta} \in \bigcup\{\theta_s \neq 0\}$
3. Biomarkers that are DE in r or more studies (HS_r):
   - $H_0: \vec{\theta} \in \bigcap\{\theta_s = 0\}$ vs $H_A: \vec{\theta} \in \sum \mathbb{I}\{\theta_s \neq 0\} \geq r$

Problem: HS_A and HS_r are not complementary hypothesis testing setting.

Example for meta-pattern

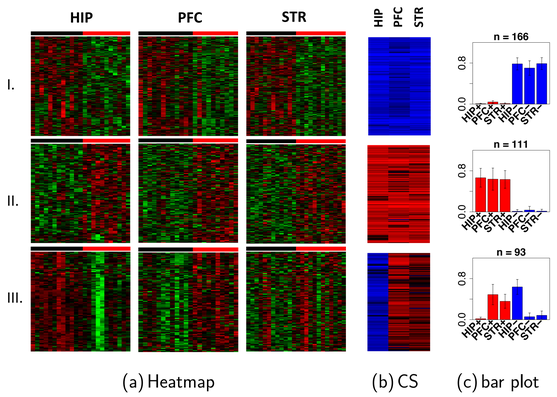

Figure 1: Three meta-pattern modules (on row) of biomarkers from HIV transgenic rats example. Each brain region (HIP, PFC or STR) represents a study.

Graphical model

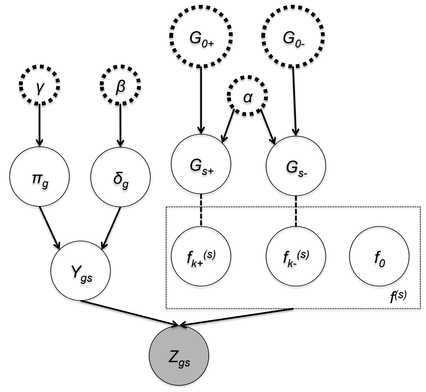

Figure 2: Graphical model representation of the Bayesian model.

Bayesian modeling

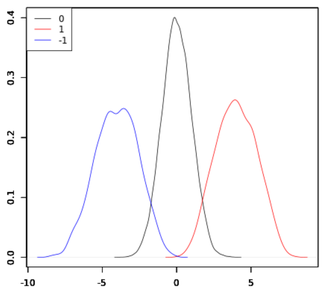

Figure 3: Z distribution.

Observed data:

- p_{gs} is one sided p-value
- $Z_{gs} = \Phi^{-1}(p_{gs})$.

Z mixture:

- Black: null.
- Red: positive DE.
- Blue: negative DE.

Generative process:

- $Y_{gs} \in \{-1, 0, 1\}$ is DE indicator:
  $f^{(s)}(Z_{gs}|Y_{gs}) = f_0^{(s)}(Z_{gs}) \cdot \mathbb{I}(Y_{gs} = 0)$
  $+ f_{+1}^{(s)}(Z_{gs}) \cdot \mathbb{I}(Y_{gs} = 1) + f_{-1}^{(s)}(Z_{gs}) \cdot \mathbb{I}(Y_{gs} = -1)$,
- $Y_{gs} \sim \text{Mult}\left(1, (1-\pi_g, \pi_g^+, \pi_g^-)\right) \cdot (0, 1, -1)$, where $\pi_g^+ = \pi_g \delta_g$, $\pi_g^- = \pi_g(1-\delta_g)$.
- $Z_{gs} \sim \text{N}(\mu_{gs}, 1)$
- $\mu_{gs} \sim \begin{cases} G_{s+} & \text{if } Y_{gs} = 1, \\ G_{s-} & \text{if } Y_{gs} = -1 \end{cases}$
- $G_{s+} \sim \text{DP}(G_{0+}, \alpha_+)$ and $G_{s-} \sim \text{DP}(G_{0-}, \alpha_-)$, where $G_{0\pm}$ is truncated normal distribution.

Priors:

- $\pi_g \sim \text{Beta}(\gamma, 1-\gamma)$;
- $\gamma \sim \text{UNIF}(0, 1)$;
- $\delta_g \sim \text{Beta}(\beta, \beta)$; $\beta = 1/2$

Bayesian computing

1. $\pi_g | Y_{gs} \sim$
   $\text{Beta}(\gamma/(G-\gamma) + Y_g^+ + Y_g^-, S - Y_g^+ - Y_g^- + 1)$,
   where $Y_g^+ = \sum_s \mathbb{I}(Y_{gs} = 1)$, $Y_g^- = \sum_s \mathbb{I}(Y_{gs} = -1)$.
2. $\delta_g | Y_{gs} \sim \text{Beta}(\beta + Y_g^+, \beta + Y_g^-)$.
3. Update Y_{gs}'s: First update C_{gs}'s s.t.
   $\Pr(C_{gs} = k | C_{-g,s}, Z_{gs}, \pi_g^{\pm}) \propto$
   $h_k^{(s)}(Z_{gs}|C_{-g,s})(\pi_g^+)^{\mathbb{I}(k>0)}(\pi_g^-)^{\mathbb{I}(k<0)}(1-\pi_g)^{\mathbb{I}(k=0)}$
   Set $Y_{gs} = \text{sgn}(C_{gs})$,
4. $\gamma \propto \Pi_{g=1}^G \text{dBeta}(\pi_g; \gamma, 1-\gamma)$

Decision making framework

- For meta-analysis purpose, we will declare differentially expressed genes which are in:
  - $\Omega_{\bar{A}}$: $\Omega_{\bar{A}}^1 = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) = S\}$.
  - Ω_B: $\Omega_B^1 = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) = 1\}$.
  - $\Omega_{\bar{r}}$: $\Omega_{\bar{r}}^1 = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) \geq r\}$.
- Compare our Bayesian FDR with FDR (Benjamini-Hochberg) from frequentists' perspective.

Simulation results

Table 1: FDR results, Nominal FDR 5%

	S	σ	$\mathcal{D}_{\bar{A}}$ BayesMP	$\mathcal{D}_{\bar{A}}$ maxP	$\mathcal{D}_B$ BayesMP	$\mathcal{D}_B$ Fisher	$\mathcal{D}_B$ AW	$\mathcal{D}_{\bar{r}}$ $(r = \lfloor S/2 \rfloor + 1)$ BayesMP	$\mathcal{D}_{\bar{r}}$ $(r = \lfloor S/2 \rfloor + 1)$ rOP
FDR	3	1	0.058 (0.008)	0.207 (0.014)	0.042 (0.004)	0.035 (0.005)	0.035 (0.004)	0.034 (0.004)	0.086 (0.007)
		2	0.058 (0.010)	0.198 (0.017)	0.047 (0.006)	0.035 (0.006)	0.036 (0.006)	0.037 (0.005)	0.080 (0.009)
		3	0.043 (0.016)	0.184 (0.025)	0.050 (0.009)	0.035 (0.008)	0.036 (0.009)	0.036 (0.009)	0.073 (0.014)
	5	1	0.075 (0.011)	0.361 (0.017)	0.043 (0.004)	0.034 (0.004)	0.034 (0.004)	0.037 (0.005)	0.130 (0.008)
		2	0.079 (0.019)	0.349 (0.022)	0.046 (0.006)	0.034 (0.005)	0.034 (0.005)	0.042 (0.006)	0.115 (0.010)
		3	0.062 (0.033)	0.330 (0.028)	0.050 (0.007)	0.034 (0.006)	0.034 (0.007)	0.042 (0.008)	0.099 (0.013)

Table 2: AUC results

	S	σ	$\mathcal{D}_{\bar{A}}$ BayesMP	$\mathcal{D}_{\bar{A}}$ maxP	$\mathcal{D}_B$ BayesMP	$\mathcal{D}_B$ Fisher	$\mathcal{D}_B$ AW	$\mathcal{D}_{\bar{r}}$ $(r = \lfloor S/2 \rfloor + 1)$ BayesMP	$\mathcal{D}_{\bar{r}}$ $(r = \lfloor S/2 \rfloor + 1)$ rOP
AUC	3	1	0.976 (0.003)	0.926 (0.003)	0.973 (0.002)	0.973 (0.002)	0.973 (0.002)	0.980 (0.002)	0.972 (0.003)
		2	0.906 (0.006)	0.876 (0.006)	0.880 (0.005)	0.878 (0.005)	0.876 (0.005)	0.902 (0.005)	0.873 (0.006)
		3	0.833 (0.008)	0.806 (0.008)	0.788 (0.006)	0.784 (0.006)	0.780 (0.006)	0.820 (0.007)	0.776 (0.008)
	5	1	0.974 (0.004)	0.920 (0.003)	0.978 (0.002)	0.978 (0.002)	0.979 (0.002)	0.985 (0.002)	0.979 (0.002)
		2	0.918 (0.007)	0.891 (0.006)	0.896 (0.004)	0.893 (0.004)	0.892 (0.004)	0.928 (0.004)	0.893 (0.004)
		3	0.866 (0.009)	0.833 (0.009)	0.812 (0.005)	0.805 (0.005)	0.800 (0.006)	0.859 (0.005)	0.801 (0.006)

Tight clustering to get meta-pattern

- $\vec{U}_{gs} = (\Pr(Y_{gs} = 1|Z), \Pr(Y_{gs} = -1|Z), \Pr(Y_{gs} = 0|Z))$
- Calculate the Cosine dissimilarity of $\vec{U}_{is}$ and $\vec{U}_{js}$, average over study index s.
- Apply tight clustering algorithm (Tseng and Wong, 2005) to the dissimilarity matrix.

Mouse metabolism data

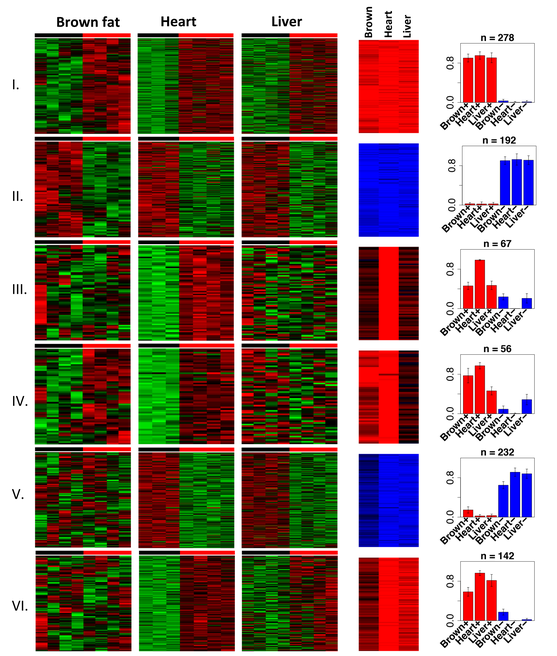

Figure 4: Six meta-pattern modules of biomarkers from the mouse metabolism example.

Table 3: Module related pathways

Target	pathway type	q value
module 1	KEGG LYSOSOME	$q = 2.8 \times 10^{-4}$
module 2	BIOCARTA AHSP PATHWAY	$q = 0.017$
module 3	DEFENSE RESPONSE	$q = 4.2 \times 10^{-8}$
module 4	BIOCARTA MCM PATHWAY	$q = 3.9 \times 10^{-3}$
module 5	none	
module 6	FC GAMMA R MEDIATED PHAGOCYTOSIS	$q = 0.067$

Implementation

BayesMP is implemented in R calling C++. The BayesMP package is publicly available at GitHub `https://github.com/Caleb-Huo/BayesMP`.