The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
Mittman

The problem
The workflow
Simulated data
The contenders
edgeR
ShrinkBayes
baySeq
The contest
Constructing ROC
curves
ROC results
FDR control
procedures
FDR control results

The heterosis problem: a comparison of Eric's method with `edgeR`, `baySeq`, and `ShrinkBayes`

Will Landau, Eric Mittman

Iowa State University

November 24, 2014

Outline

The problem

The workflow

Simulated data

The contenders
  edgeR
  ShrinkBayes
  baySeq

The contest
  Constructing ROC curves
  ROC results
  FDR control procedures
  FDR control results

The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
Mittman

The problem

The workflow

Simulated data

The contenders
edgeR
ShrinkBayes
baySeq

The contest
Constructing ROC
curves
ROC results
FDR control
procedures
FDR control results

Mock heterosis data

The heterosis problem: a comparison of Eric's method with edgeR, baySeq, and ShrinkBayes

Will Landau, Eric Mittman

The problem

The workflow

Simulated data

The contenders
edgeR
ShrinkBayes
baySeq

The contest
Constructing ROC curves
ROC results
FDR control procedures
FDR control results

	Parent (1)				Parent (2)				Hybrid (3)				Truth
HPH Feature 1	3	4	2	1	0	0	1	0	700	900	825	860	1
HPH Feature 2	0	1	1	0	2	7	5	18	50	501	400	90	1
Feature 3	100	225	0	15	300	106	200	400	70	279	100	123	0
LPH Feature 4	893	400	760	901	1000	513	760	580	5	5	6	7	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Feature 25000	10	13	6	4	902	912	999	825	819	761	800	465	0

Outline

The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
Mittman

Simulation workflow

- ▶ Simulate 30 datasets:
    - ▶ 10 datasets with 4 samples (libraries, columns, etc.) per group
    - ▶ 10 with 8 per group
    - ▶ 10 with 16 per group
- ▶ For each simulated dataset, test for heterosis with
    - ▶ empirical Bayes with STAN (Eric's method)
    - ▶ edgeR
    - ▶ baySeq
    - ▶ ShrinkBayes
- ▶ Compare methods with ROC curves

Outline

The problem

The workflow

Simulated data

The contenders
    edgeR
    ShrinkBayes
    baySeq

The contest
    Constructing ROC curves
    ROC results
    FDR control procedures
    FDR control results

The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
Mittman

The problem

The workflow

Simulated data

The contenders
edgeR
ShrinkBayes
baySeq

The contest
Constructing ROC
curves
ROC results
FDR control
procedures
FDR control results

Apply `edgeR` to real data to get simulation parameters

Normalization factors

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}

Main effects and dispersions

Parent (1)	Parent (2)	Hybrid (3)	Dispersion
$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,3}$	ψ_1
$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,3}$	ψ_2
...	...	...	...
$\mu_{27888,1}$	$\mu_{27888,2}$	$\mu_{27888,3}$	ψ_{27888}

Generating the data

- Does feature f truly have heterosis?

$$\text{truth}_f = I(\mu_{f,3} > \max(\mu_{f,1}, \mu_{f,2}) \text{ or}$$
$$\mu_{f,3} < \min(\mu_{f,1}, \mu_{f,2}))$$

- For a dataset with 4 libraries per group,

$$y_{f,i} \overset{\text{iid}}{\sim} NB\left(\exp\left(c_i + \mu_{f,t(i)}\right), \ \psi_f\right)$$

  - $t(i)$ is the group of library i.
  - Resimulate to increase the number of libraries per group.

- Remove extremely low-count features.

- Take a random subset of 25000 features from the remaining ones.

The heterosis problem: a comparison of Eric's method with edgeR, baySeq, and ShrinkBayes

Will Landau, Eric Mittman

The problem

The workflow

Simulated data

The contenders
edgeR
ShrinkBayes
baySeq

The contest
Constructing ROC curves
ROC results
FDR control procedures
FDR control results

Mock example data with 4 samples per treatment group

		Parent (1)				Parent (2)				Hybrid (3)			Truth
HPH	Feature 1	3	4	2	1	0	0	1	0	700	900	825	860	1
HPH	Feature 2	0	1	1	0	2	7	5	18	50	501	400	90	1
	Feature 3	100	225	0	15	300	106	200	400	70	279	100	123	0
LPH	Feature 4	893	400	760	901	1000	513	760	580	5	5	6	7	1
	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	Feature 25000	10	13	6	4	902	912	999	825	819	761	800	465	0

Outline

The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
Mittman

edgeR

- Fit a loglinear model to estimate main effects $\mu_{f,t}$
  - Feature $f = 1, \ldots, 25000$
  - Treatment group $t = 1$ (parent), 2 (parent), 3 (hybrid)
- Likelihood ratio tests to get p-values $p_{f,1}, \; p_{f,2}$

$$H_{0,1} : \mu_{f,3} = \mu_{f,1} \qquad H_{a,1} : \mu_{f,3} \neq \mu_{f,1}$$
$$H_{0,2} : \mu_{f,3} = \mu_{f,2} \qquad H_{a,2} : \mu_{f,3} \neq \mu_{f,2}$$

- Final p-value:

$$p_{f,\text{edgeR}} = \begin{cases} p_{f,1}/2 & \widehat{\mu}_{f,3} < \widehat{\mu}_{f,1} \leq \widehat{\mu}_{f,2} \text{ or } \widehat{\mu}_{f,3} > \widehat{\mu}_{f,1} \geq \widehat{\mu}_{f,2} \\ p_{f,2}/2 & \widehat{\mu}_{f,3} < \widehat{\mu}_{f,2} \leq \widehat{\mu}_{f,1} \text{ or } \widehat{\mu}_{f,3} > \widehat{\mu}_{f,2} \geq \widehat{\mu}_{f,1} \\ 1 & \widehat{\mu}_{f,1} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,2} \text{ or } \widehat{\mu}_{f,2} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,1} \end{cases}$$

ShrinkBayes

- ▶ Built on inla (integrated nested Laplace approximation).
- ▶ empirical Bayes with a zero-inflated NB likelihood and normal priors.
- ▶ I reparameterize

$$\phi_f = \frac{\mu_{f,1} + \mu_{f,2}}{2} \qquad \text{(parental mean)}$$

$$\alpha_f = \frac{\mu_{f,2} - \mu_{f,1}}{2} \qquad \text{(half parental difference)}$$

$$\delta_f = \mu_{f,3} - \frac{\mu_{f,1} + \mu_{f,2}}{2} \qquad \text{(hybrid effect)}$$

ShrinkBayes

The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
Mittman

The problem

The workflow

Simulated data

The contenders
edgeR
ShrinkBayes
baySeq

The contest
Constructing ROC
curves
ROC results
FDR control
procedures
FDR control results

$$\alpha_f = \text{ half parental difference}$$
$$\delta_f = \text{ hybrid effect}$$

▶ Use contrasts to calculate final posterior probabilities of heterosis:

$$p_{f,\text{ShrinkBayes}} = \begin{cases} 0 & \left|\widehat{\delta}_f\right| < |\widehat{\alpha}_f|, \text{ o.w.:} \\ P(\delta_f + \alpha_f > 0 \mid \text{data}) & \widehat{\delta}_f > -\widehat{\alpha}_f \\ P(\delta_f - \alpha_f > 0 \mid \text{data}) & \widehat{\delta}_f > \widehat{\alpha}_f \\ P(\delta_f - \alpha_f < 0 \mid \text{data}) & \widehat{\delta}_f < \widehat{\alpha}_f \\ P(\delta_f + \alpha_f < 0 \mid \text{data}) & \widehat{\delta}_f < -\widehat{\alpha}_f \end{cases}$$

baySeq

- Estimate main effects $\mu_{f,t}$ using edgeR.
- Calculate the posterior probability that each feature satisfies:

Model	Constraint
M_1	All $\mu_{f,t}$'s equal
M_2	$\mu_{f,1} = \mu_{f,2}$
M_3	$\mu_{f,1} = \mu_{f,3}$
M_4	$\mu_{f,2} = \mu_{f,3}$
M_5	All $\mu_{f,t}$'s distinct

- Final posterior probabilities of heterosis:

$$p_{f,\text{baySeq}} = \begin{cases} 0 & \widehat{\mu}_{f,1} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,2} \text{ or} \\ & \widehat{\mu}_{f,2} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,1} \\ P(M_2 \mid \text{data}) + \\ P(M_5 \mid \text{data}) & \text{otherwise} \end{cases}$$

Outline

The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
Mittman

Calculating false positive rate (FPR) and true positive rate (TPR)

- ► ROC = Receiver Operating Characteristic
- ► N_{true} heterosis features, N_{false} null features.
- ► Results of testing each feature for heterosis
    - ► For edgeR, the ranks are p-values.
    - ► For the other methods, take p(no heterosis | data) for each feature.

rank	0.802	0.935	0.539	0.001	$\cdots$	0.500	0.603
truth	0	0	1	1	$\cdots$	1	0

- ► Sort table by rank.

rank	0.000	0.001	0.005	0.006	$\cdots$	0.901	1.000
truth	1	1	0	1	$\cdots$	0	0

Calculating false positive rate (FPR) and true positive rate (TPR)

▶ In practice, we would declare the lowest-rank features to have heterosis.

rank	0.000	0.001	0.005	0.006	$\cdots$	0.901	1.000
truth	1	1	0	1	$\cdots$	0	0

▶ With 2 heterosis genes and 1 null gene,

$$FPR = \frac{1}{N_{false}} \qquad TPR = \frac{2}{N_{true}}$$

▶ Repeat for multiple cutoffs to get multiple (FPR, TPR) pairs.

Example ROC curves
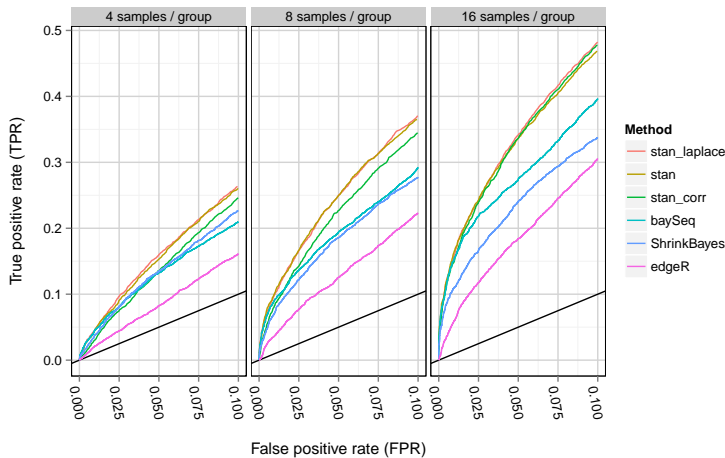
The heterosis problem: a comparison of Eric's method with edgeR, baySeq, and ShrinkBayes

Will Landau, Eric Mittman

The problem
The workflow
Simulated data
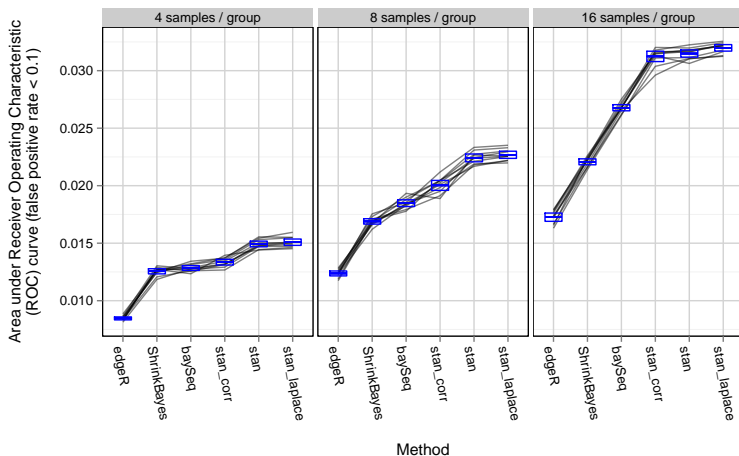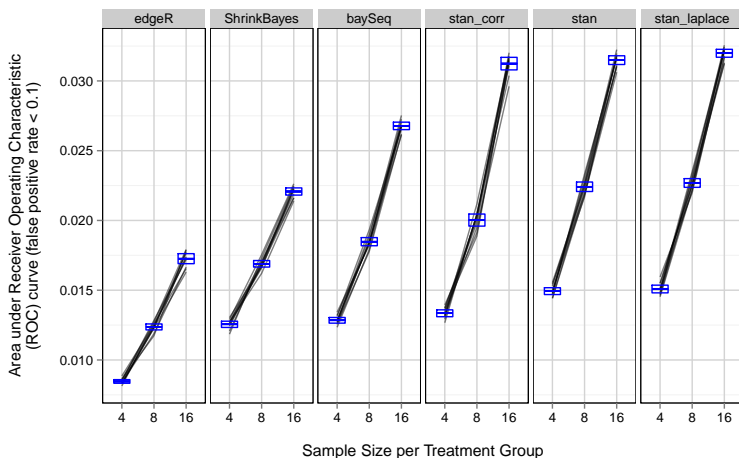The contenders
edgeR
ShrinkBayes
baySeq
The contest
Constructing ROC curves
ROC results
FDR control procedures
FDR control results

Areas under ROC curves

The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
Mittman

The problem

The workflow

Simulated data

The contenders
edgeR
ShrinkBayes
baySeq

The contest
Constructing ROC
curves
ROC results
FDR control
procedures
FDR control results

Areas under ROC curves

FDR control procedures

- Let $p_{(f)}$ ($f = 1, \ldots, 25000$) be the ordered ranks
  - For edgeR, the ranks are now q-values.
  - For the other methods, take p(no heterosis | data) for each feature.
- Let $I_{(f)} = I$(feature of $p_{(f)}$ is not heterosis)
- Method 1:

$$\overline{p}_m = \frac{1}{m} \sum_{f=1}^{m} p_{(f)} \qquad \overline{I}_m = \frac{1}{m} \sum_{f=1}^{m} I_{(f)}$$

  - To control FDR, we need $\overline{I}_m \leq \overline{p}_m$.
  - Plot $\overline{I}_m - \overline{p}_m$ vs $\overline{p}_m$ for multiple m.
- Method 2:

$$\overline{p}_m = \frac{1}{100} \sum_{f=m}^{m+100} p_{(f)} \qquad \overline{I}_m = \frac{1}{100} \sum_{f=m}^{m+100} I_{(f)}$$

The heterosis problem: a comparison of Eric's method with edgeR, baySeq, and ShrinkBayes

Will Landau, Eric Mittman

The problem

The workflow

Simulated data

The contenders
edgeR
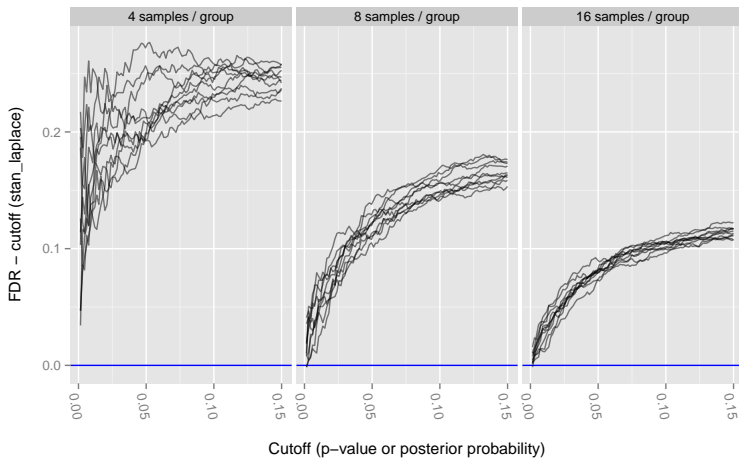ShrinkBayes
baySeq

The contest
Constructing ROC curves
ROC results
FDR control procedures
FDR control results

Method 1

Method 2

The heterosis
problem: a
comparison of
Eric's method with
edgeR, baySeq,
and ShrinkBayes

Will Landau, Eric
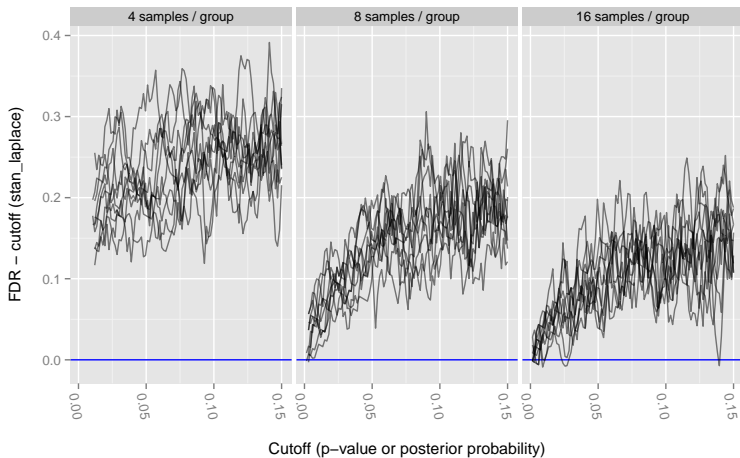Mittman

The problem

The workflow

Simulated data

The contenders
edgeR
ShrinkBayes
baySeq

The contest
Constructing ROC
curves
ROC results
FDR control
procedures
FDR control results