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

Will Landau, Eric Mittman

Iowa State University

November 15, 2014

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

Will Landau, Eric Mittman

The problem

.... .........................

imulated data

The contenders
edgeR
baySeq

#### Outline

#### The problem

The workflow

Simulated data

#### The contender

edgeR baySeq ShrinkBay

#### The contest

ROC (receiver operating characteristic) 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
baySeq

The contest

ROC (receiver operating characteristic) curve: The results

#### Mock heterosis data

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

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

Will Landau, Eric Mittman

The problem

#### Outline

The problem

The workflow

Simulated data

The contenders

edgeR baySeq ShrinkBay

The contest

ROC (receiver operating characteristic) curves

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

The heterosis

Will Landau, Eric Mittman

The problem

The workflow

Simulated data

The contenders edgeR baySeq

The contest

ROC (receiver operating characteristic) curves

The results

#### 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

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
baySeq

#### 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

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
baySeq

The contest

ROC (receiver operating

#### 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

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

Will Landau, Eric Mittman

The workflow

edgeR

#### Outline

#### Simulated data

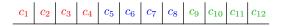
The heterosis problem: a comparison of Eric's method with edgeR, baySeq, and ShrinkBayes

Will Landau, Eric Mittman

Simulated data

# Apply edgeR to real data to get simulation parameters

Normalization factors



Main effects and dispersions

Parent (1)	Parent (2)	Hybrid (3)	Dispersion
$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,3}$	$\phi_1$
$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,3}$	$\phi_2$
$\mu_{27888,1}$	$\mu_{27888,2}$	$\mu_{27888,3}$	$\phi_{27888}$

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

Will Landau, Eric Mittman

l he problem

The working

Simulated data

he contenders

edgeR baySeq ShrinkBayes

The contest

ROC (receiver
operating
characteristic) curve

The results

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

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

Will Landau, Eric Mittman

The problem

i ne worktio

Simulated data

The contenders

edgeR

baySeq

ShrinkBayes
The contest

ROC (receiver operating characteristic) curves The results

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

$$y_{f,i} \stackrel{\text{iid}}{\sim} NB \left( \exp \left( c_{\lceil 4i/N \rceil} + \mu_{f,\lceil i/N \rceil} \right), \ \phi_f \right)$$

- where:
  - ▶ Sample (library, column) i = 1, ..., 3N
  - $\triangleright$  N =samples per treatment group (4, 8, or 16)

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

Will Landau, Eric Mittman

The problem

The working

Simulated data

The contenders

edgeR

baySeq

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

$$y_{f,i} \stackrel{\mathsf{iid}}{\sim} \mathsf{NB} \left( \exp \left( c_{\lceil 4i/N \rceil} + \mu_{f,\lceil i/N \rceil} \right), \ \phi_f \right)$$

- where:
  - ▶ Sample (library, column) i = 1, ..., 3N
  - ightharpoonup N =samples per treatment group (4, 8, or 16)
- Remove extremely low-count features.

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
baySeq

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

$$y_{f,i} \stackrel{\text{iid}}{\sim} NB \left( \exp \left( c_{\lceil 4i/N \rceil} + \mu_{f,\lceil i/N \rceil} \right), \ \phi_f \right)$$

- where:
  - ▶ Sample (library, column) i = 1, ..., 3N
  - ightharpoonup N =samples per treatment group (4, 8, or 16)
- ▶ 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

i ne worktio

Simulated data

The contenders
edgeR
baySeq

# 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

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

Will Landau, Eric Mittman

Simulated data

#### Outline

#### The contenders

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

Will Landau, Eric Mittman

#### The contenders

edgeR

- Fit a loglinear model to estimate main effects  $\mu_{f,t}$ 
  - Feature f = 1, ..., 25000
  - ▶ Treatment group t = 1 (parent), 2 (parent), 3 (hybrid)

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, ..., 25000
  - ▶ Treatment group t = 1 (parent), 2 (parent), 3 (hybrid)
- ▶ Likelihood ratio tests to get p-values  $p_{f,1}$ ,  $p_{f,2}$

$$H_{0,1}: \mu_{f,3} = \mu_{f,1}$$

$$H_{a,1}: \mu_{f,3} \neq \mu_{f,1}$$

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

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

Will Landau, Eric Mittman

The problem

Simulated data

The contenders
edgeR
baySeq
ShrinkPayor

The contest

ROC (receiver operating characteristic) curves

The results

▶ Fit a loglinear model to estimate main effects  $\mu_{f,t}$ 

- Feature f = 1, ..., 25000
- ▶ Treatment group t = 1 (parent), 2 (parent), 3 (hybrid)
- ▶ Likelihood ratio tests to get p-values  $p_{f,1}$ ,  $p_{f,2}$

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

Final p-value	if
$p_{f,1}/2$	$\widehat{\mu}_{f,3} < \widehat{\mu}_{f,1} \le \widehat{\mu}_{f,2} \text{ or } \widehat{\mu}_{f,3} > \widehat{\mu}_{f,1} \ge \widehat{\mu}_{f,2}$
$p_{f,2}/2$	$\mu_{f,3} < \mu_{f,2} \le \mu_{f,1} \text{ or } \mu_{f,3} > \mu_{f,2} \ge \mu_{f,1}$
1	$ \begin{vmatrix} \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} \\ \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{vmatrix} $

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

Will Landau, Eric Mittman

The problem

THE WOLKHOW

imulated data

The contenders
edgeR
baySeq

The contest

ROC (receiver operating characteristic) curves

The results

### baySeq

▶ Estimate main effects  $\mu_{f,t}$  using edgeR.

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

Will Landau, Eric Mittman

The problem

.... ..........

Simulated data

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

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

Will Landau, Eric Mittman

The problem

imulated data

I he contenders
edgeR
baySeq
ShrinkBayes

- ▶ 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:

Posterior probability	if
Posterior probability $0$ $P(M_3 \mid \text{data}) + P(M_5 \mid \text{data})$	$\widehat{\mu}_{f,1} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,2}$ or
	$\widehat{\mu}_{f,2} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,1}$
$P(M_3 \mid data) + P(M_5 \mid data)$	otherwise

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

Will Landau, Eric Mittman

The problem

imulated data

edgeR
baySeq
ShrinkBayes

#### ShrinkBayes

▶ Built on inla (integrated nested Laplace approximation).

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

Will Landau, Eric Mittman

The problem

Simulated data

edgeR baySeq

#### ShrinkBayes

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

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

Will Landau, Eric Mittman

i ne probiem

Simulated data

edgeR baySeq

edgeR baySeq

## ShrinkBayes

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

$$\begin{split} \phi_f &= \frac{\mu_{f,1} + \mu_{f,2}}{2} \qquad \text{(parental mean)} \\ \alpha_f &= \frac{\mu_{f,2} - \mu_{f,1}}{2} \qquad \text{(half parental difference)} \\ \delta_f &= \mu_{f,3} - \frac{\mu_{f,1} + \mu_{f,2}}{2} \qquad \text{(hybrid effect)} \end{split}$$

$\phi_{f}$	$\alpha_{f}$	$\delta_f$
parental mean	half parental difference	hybrid effect

Use contrasts to calculate final posterior probabilities of heterosis:

Posterior probability	if
0	$ \delta_f  <  \alpha_f $ , otherwise:
$egin{aligned} P(\delta_f + lpha_f > 0 \mid data) \ P(\delta_f - lpha_f > 0 \mid data) \ P(\delta_f - lpha_f < 0 \mid data) \end{aligned}$	$\delta_f > -\alpha_f$
$P(\delta_f - \alpha_f > 0 \mid data)$	$\delta_f > \alpha_f$
$P(\delta_f - \alpha_f < 0 \mid data)$	$\delta_f < \alpha_f$
$P(\delta_f + \alpha_f < 0 \mid data)$	$\delta_f < -\alpha_f$

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

Will Landau, Eric Mittman

The problem

The contenders

edgeR baySeq ShrinkBayes

#### Outline

The problem

The workflow

Simulated data

The contender

edgeR baySeq ShrinkBa

#### The contest

ROC (receiver operating characteristic) curves

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

Will Landau, Eric Mittman

The problem

The worknow

Simulated data

The contenders
edgeR
baySeq

The contest

ROC (receiver operating characteristic) curves The results

 $\triangleright$   $N_{\text{true}}$  heterosis features,  $N_{\text{false}}$  null features.

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

Will Landau, Eric Mittman

The problem

Simulated data

I he contenders edgeR baySeq

- $\triangleright$   $N_{\text{true}}$  heterosis features,  $N_{\text{false}}$  null features.
- ► Results of testing each feature for heterosis (25000 columns here):

ļ	oval	0.802	0.935	0.539	0.001	 0.500	0.603
t	truth	0	0	1	1	 1	0

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

Will Landau, Eric Mittman

The problem

ormulated data

I he contenders
edgeR

The contest

ROC (receiver operating characteristic) curves The results

- $\triangleright$   $N_{\text{true}}$  heterosis features,  $N_{\text{false}}$  null features.
- Results of testing each feature for heterosis (25000 columns here):

pval	0.802	0.935	0.539	0.001	 0.500	0.603
truth	0	0	1	1	 1	0

Sort table by p-value (or other binary classifier)

pval	0.000	0.001	0.005	0.006		0.901	1.000
truth	1	1	0	1		0	0

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

Will Landau, Eric Mittman

i ne problem

Simulated data

edgeR
baySeq

The contest
ROC (receiver operating

characteristic) curves
The results

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

Will Landau, Eric Mittman

The problem

... ......

Simulated data

edgeR baySeq

ShrinkBayes

ROC (receiver operating characteristic) curves.

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

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

Will Landau, Eric Mittman

The problem

imulated data

edgeR baySeq

The contest

ROC (receiver operating characteristic) curves

The results

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

pval	0.000	0.001	0.005	0.006		0.901	1.000
truth	1	1	0	1		0	0

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

Will Landau, Eric Mittman

i ne problem

THE WORKHOW

Simulated data

edgeR eaySeq

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

pval	0.000	0.001	0.005	0.006	 0.901	1.000
truth	1	1	0	1	 0	0

▶ With 2 heterosis genes and 1 null gene,

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

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

Will Landau, Eric Mittman

The problem

THE WOLKHOW

he contenders

dgeR aySeq

The contest

ROC (receiver
operating
characteristic) curves

ロト 4回 ト 4 重 ト 4 重 ト 1 重 の9 0

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

pval	0.000	0.001	0.005	0.006		0.901	1.000
truth	1	1	0	1	• • •	0	0

▶ With 2 heterosis genes and 1 null gene,

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

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

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

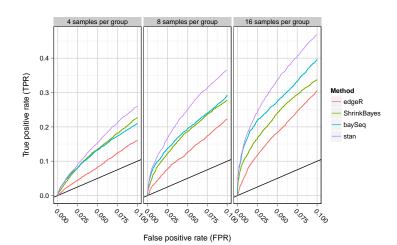
Will Landau, Eric Mittman

The problem

The contenders

edgeR DaySeq

### Example ROC curves



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

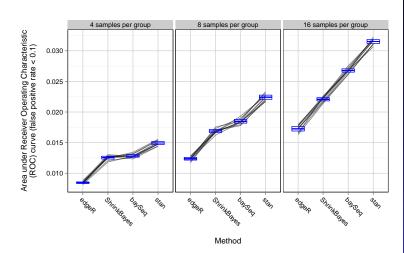
Will Landau, Eric Mittman

The problem

Simulated data

edgeR
baySeq
ShrinkBayes

#### Areas under ROC curves

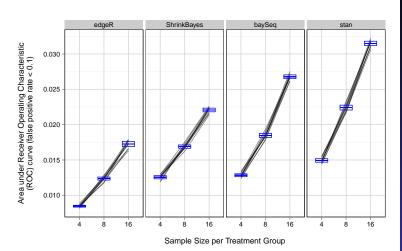


The heterosis problem: a comparison of Eric's method with edgeR, baySeq, and ShrinkBayes

Will Landau, Eric Mittman

The 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 WOLKHOW

Simulated data

The contenders
edgeR
baySeq