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
ShrinkBayes

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

Iowa State University

November 17, 2014

Outline

The problem

The workflow

Simulated data

The contenders
   edgeR
   baySeq
   ShrinkBayes

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

The workflow

Simulated data

The contenders
edgeR
baySeq
ShrinkBayes

The contest
ROC (receiver
operating
characteristic) curves
The results

Mock heterosis data

		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 problem

The workflow

Simulated data

The contenders
 edgeR
 baySeq
 ShrinkBayes

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

The workflow

Simulated data

The contenders
edgeR
baySeq
ShrinkBayes

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

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

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

Will Landau, Eric
Mittman

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}

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
ShrinkBayes

The contest
ROC (receiver operating characteristic) curves
The results

A single simulation (of 30)

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

A single simulation (of 30)

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

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

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

A single simulation (of 30)

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

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

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

A single simulation (of 30)

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

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

▶ where:
  ▶ Sample (library, column) $i = 1, \ldots, 3N$
  ▶ 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.

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

The problem

The workflow

Simulated data

The contenders
edgeR
baySeq
ShrinkBayes

The contest
ROC (receiver
operating
characteristic) curves
The results

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)

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 $p_{f,1}, \ p_{f,2}$

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

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	if...
$p_{f,1}/2$	$\widehat{\mu}_{f,3} < \widehat{\mu}_{f,1} \leq \widehat{\mu}_{f,2}$ or $\widehat{\mu}_{f,3} > \widehat{\mu}_{f,1} \geq \widehat{\mu}_{f,2}$
$p_{f,2}/2$	$\widehat{\mu}_{f,3} < \widehat{\mu}_{f,2} \leq \widehat{\mu}_{f,1}$ or $\widehat{\mu}_{f,3} > \widehat{\mu}_{f,2} \geq \widehat{\mu}_{f,1}$
1	$\widehat{\mu}_{f,1} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,2}$ or $\widehat{\mu}_{f,2} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,1}$

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
ShrinkBayes

The contest
ROC (receiver operating characteristic) curves
The results

baySeq

- Estimate main effects $\mu_{f,t}$ using 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	All $\mu_{f,t}$'s distinct

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:

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

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
ShrinkBayes

The contest
ROC (receiver operating characteristic) curves
The results

ShrinkBayes

▶ Built on inla (integrated nested Laplace approximation).

ShrinkBayes

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

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

ϕ_f	α_f	δ_f
parental mean	half parental difference	hybrid effect

▶ Use contrasts to calculate final posterior probabilities of heterosis:

Posterior probability	if...
0	$|\widehat{\delta_f}| < |\widehat{\alpha_f}|$, otherwise:
$P(\delta_f + \alpha_f > 0 \mid \text{data})$	$\widehat{\delta_f} > -\widehat{\alpha_f}$
$P(\delta_f - \alpha_f > 0 \mid \text{data})$	$\widehat{\delta_f} > \widehat{\alpha_f}$
$P(\delta_f - \alpha_f < 0 \mid \text{data})$	$\widehat{\delta_f} < \widehat{\alpha_f}$
$P(\delta_f + \alpha_f < 0 \mid \text{data})$	$\widehat{\delta_f} < -\widehat{\alpha_f}$

Outline

The problem

The workflow

Simulated data

The contenders
  edgeR
  baySeq
  ShrinkBayes

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

The workflow

Simulated data

The contenders
edgeR
baySeq
ShrinkBayes

The contest
ROC (receiver
operating
characteristic) curves
The results

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

▶ N_{true} heterosis features, N_{false} null features.

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

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

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

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

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

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

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

pval	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)

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

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

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

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

pval	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-p-value features to have heterosis.

pval	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,

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

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

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

Will Landau, Eric Mittman
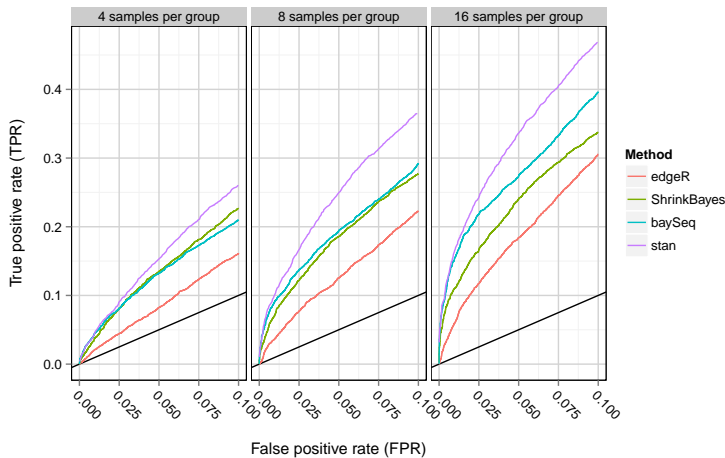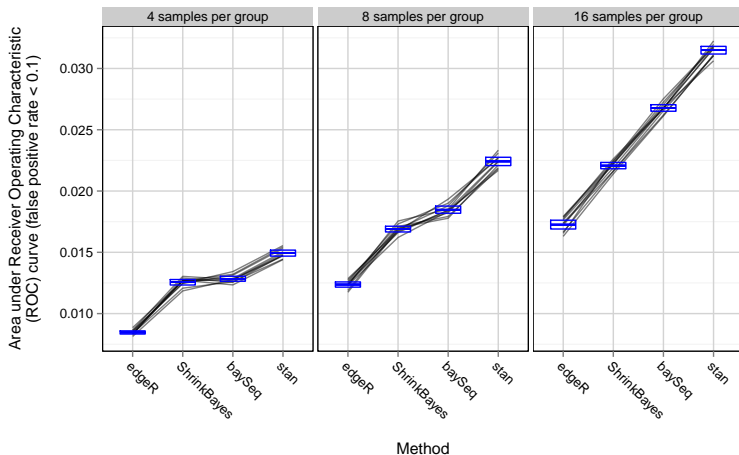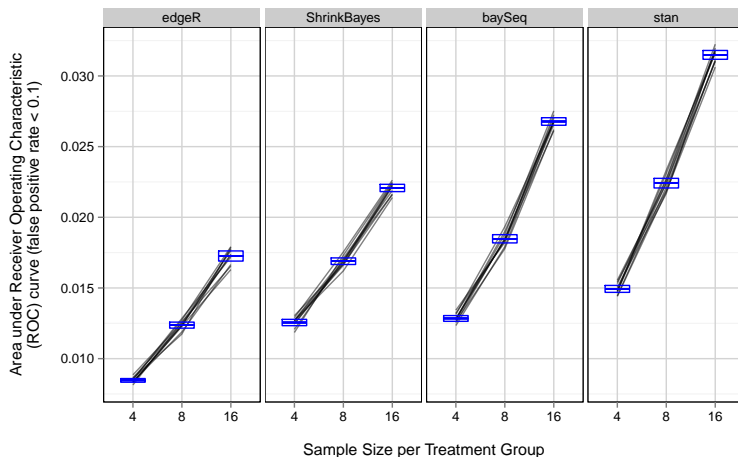
The problem

The workflow

Simulated data

The contenders
edgeR
baySeq
ShrinkBayes

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	$\cdots$	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

Areas under ROC curves

Areas under ROC curves