DEXSeq paper discussion

L. Collado-Torres

December 10th, 2012
1 Background

2 DEXSeq paper

3 Results
Gene Expression

High-Throughput Sequencing \(^2\)

2Source: Metzker, Sequencing technologies — the next generation, 2010, Nat Rev Genet
Alignment (Mapping) ³

³Source: Trapnell et al, How to map billions of short reads onto genomes, 2009, Nat Biotech
What can we find? 4

4Source: Sorek and Cossart, Prokaryotic transcriptomics a new view on regulation, physiology and pathogenicity, 2010, Nat Rev Genet
What can we find? 5

Source: Sorek and Cossart, Prokaryotic transcriptomics a new view on regulation, physiology and pathogenicity, 2010, Nat Rev Genet
What can we find?  

6

---

6Source: Sorek and Cossart, Prokaryotic transcriptomics a new view on regulation, physiology and pathogenicity, 2010, Nat Rev Genet
What can we find? 

Source: Sorek and Cossart, Prokaryotic transcriptomics a new view on regulation, physiology and pathogenicity, 2010, Nat Rev Genet
Main ideas

Compare two or more conditions of interest to find the DE exons (DEX).

- Focus on DE: assume a transcript inventory
- Account for biological variation
- Use GLMs
- Fine tuning to make it fast, control for false positives, and when possible increase power
Simplifying the exome: *counting bins* \(^8\)

---

\(^8\)Source: Anders, Reyes, Huber; Detecting differential usage of exons from RNA-seq data, 2012, Genome Research
Model

Using count data and assume it follows a negative binomial distribution

\[ K_{ijl} \sim NB \ (\text{mean} = s_j \mu_{ijl}, \text{dispersion} = \alpha_{il}) \] (1)

- counting bin \( l \)
- gene \( i \)
- sample \( j = 1, \ldots, m \)
- size factor \( s_j \): needed because each sample is sequenced at a different \textit{depth}
- \( \alpha_{il} \) is the dispersion parameter
Poisson vs NB

Poisson GLM
- Outcome \( Y \sim Poisson(\mu) \)
- Link function: \( \log \mu = x' \beta \)
- Variance function \( \text{Var}(Y) = \text{Var}(\mu) = \alpha \mu \) where \( \alpha = 1 \). \( \alpha \neq 1 \) is the quasi-likelihood approach.

Negative Binomial Model: Gamma-Poisson mixture construction
- Assume unobserved r.v. \( E \) where \( E \sim Gamma(\theta, 1/\theta) \).
  - Mean: \( \theta \cdot 1/\theta = 1 \), Variance: \( \theta \cdot 1/\theta^2 = 1/\theta \).
- Assume that \( Y|E \sim Poisson(\mu E) \)
- Then \( Y \) has a negative binomial distribution with mean \( \mu \) and variance \( \mu + \mu^2/\theta = \mu(1 + \mu/\theta) \)
  - Variance of \( Y \) increases quadratically with the mean rather than linearly.

\(^9\alpha = 1/\theta \) in the DEXSeq paper

\(^{10}\text{Source: 140.654 2012 slides by Roger Peng} \)
Main log-linear model

\[
\log \mu_{ijl} = \beta_i^G + \beta^E_{il} + \beta_i^C + \beta_{i\rho_j l}^{EC}
\]  

- $\beta_i^G$: baseline expression strength of gene $i$
- $\beta^E_{il}$: log of the expected fraction of the reads mapped to gene $i$ that overlap counting bin $l$
- $\beta_i^C$: log of the fold change in overall expression of gene $i$ under condition $\rho_j$
- $\rho_j$: experimental condition of sample $j$
- $\beta_{i\rho_j l}^{EC}$: effect condition $\rho_j$ has on the fraction of reads falling into bin $l$
Variability: gene expression + exon usage

- Var. in gene expression: when the total number of transcripts for a gene $i$ differs from the expected value under $\rho_j$
- Var. in exon usage: using different exons or counting bins

$$\log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ij}^S + \beta_{i\rho_j l}^{EC}$$

(3)

Change $\beta_{i\rho_j}^C$ by $\beta_{ij}^S$. Absorbs var. in gene expression.
Dispersion estimates

[Source: Anders, Reyes, Huber; Detecting differential usage of exons from RNA-seq data, 2012, Genome Research]
Analysis of Deviance

- Deviance \( D(\hat{\beta}) = 2\ell^* - 2\ell(\hat{\beta}; y) \) where \( \ell^* \) is the saturated likelihood
- Two spaces for \( \beta \): small \( S \) (nested) and large \( L \) with \( H_0 : \beta \in S \) and \( H_a : \beta \in L - S \).
- Likelihood ratio
  \[
  LR = \frac{\mathcal{L}(\hat{\beta}_S; y)}{\mathcal{L}(\hat{\beta}_L; y)}
  \]
- Under \( H_0 \), \(-2 \log LR \sim \chi^2_{|L| - |S|}\)
- Note \( D(\hat{\beta}_S) - D(\hat{\beta}_L) = -2[\ell(\hat{\beta}_S; y) - \ell(\hat{\beta}_L; y)] = -2 \log LR \)

---

Source: 140.654 2012 slides by Roger Peng
Testing for DEX: ANODEV

Fit two models

\[ \log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ij}^S \]  \hspace{0.5cm} (4)

\[ \log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ij}^S + \beta_{i\rho_jl}^{EC} \delta_{ll'} \]  \hspace{0.5cm} (5)

where

\[ \delta_{ll'} = \begin{cases} 1 & \text{if } l = l' \\ 0 & \text{otherwise} \end{cases} \]

Then test using analysis of deviance (ANODEV)
Control FDR by adjusting p-values using Benjamini-Hochberg’s method.
Finding DEX: knockdown of *pasilla* on *Drosophila melanogaster* example

Source: http://www-huber.embl.de/pub/DEXSeq/analysis/brooksetal/
Detection power depends on mean $^{14}$

$^{14}$Source: reproduced with code from http://genome.cshlp.org/content/suppl/2012/08/20/gr.133744.111.DC1/Supp_II.html
Without considering biological variation \(^{15}\)

\(^{15}\)Source: http://www-huber.embl.de/pub/DEXSeq/analysis/brooksetal/
Interesting comparison

- Mock comparison: check for DEX between replicates from a control condition
- Used an FDR of 10%
- DEXSeq: 8 genes (159 in the real control vs treatment comparison)
- Cuffdiff v 1.3.0: 639 genes (37 in real comp.)

This trend continues with other data sets.
Thanks!

- Main source: Anders, Reyes, Huber; Detecting differential usage of exons from RNA-seq data, 2012, Genome Research
- PMID: 22722343.