Engineering annotation-agnostic tools for differential expression analysis

Alyssa Frazee
Johns Hopkins University
@acfrazee
Why annotation-agnostic?
Why annotation-agnostic?

(1) isoform-level analysis
Scientifically important

cell differentiation (Trapnell 2010)

organism development (Graveley 2010)

cancer (Govindan 2012)
but read-counting is challenging
Why annotation-agnostic?

(2) allows for new discoveries
The hero scientist who defeats cancer will likely never exist.

No exalted individual, no victory celebration, no Marie Curie or Jonas Salk, who in 1955, after he created the first polio vaccine, was asked, So what’s next? Cancer? — as if a doctor finished with one disease could simply shift his attention to another, like a chef turning from the soup to the entrée.

Cancer doesn’t work that way. It’s not just one disease; it’s hundreds, potentially thousands. And not all cancers are caused by just one agent — a virus or bacterium that can be flushed and crushed. Cancer is an intricate and potentially lethal collaboration of genes gone awry, of growth inhibitors gone missing, of hormones and epigenomes changing and rogue cells breaking free. It works as one great armed force, attacking by the equivalent of air and land and sea and stealth, and we think we’re going to take it out with what? A lab-coated sniper?

“This disease is much more complex than we have been treating it,” says MIT’s Phillip Sharp. “And the complexity is stunning.”
One solution: assembly
Approach 1: avoid assembly altogether
idea: scan genome base-by-base, highlight segments showing differential expression signal
DER Finder

chr22: 17684448–17684670

read coverage

DE signal

genomic position
The DER Finder tool is used to analyze gene expression changes between normal and tumor samples. The graph shows a comparison of read coverage and DE signal across different genomic positions. The x-axis represents the genomic position, while the y-axis shows the log2(expression count + 1) for both normal and tumor samples. The DE signal is highlighted in yellow, indicating significant expression differences. The chr22 region from 17684448 to 17684670 is expanded to show more detail. The t-statistic graph is used to infer the significance of these expression changes.
find signal at each nucleotide

expression

covariate of interest

confounders

\[ g(Y_{ij}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^{K} \gamma_k(l_j)W_{ik} + \varepsilon_{ij} \]
find signal at each nucleotide

$$g(Y_{ij}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^{K} \gamma_k(l_j)W_{ik} + \varepsilon_{ij}$$

samples indexed by $i$
locations indexed by $l_i$
confounders indexed by $k$
segment genome into groups of nucleotides with similar signal

hidden states (unknown truth)

DE → DE → DE → not DE → not DE

t₁ → t₂ → t₃ → t₄ → t₅

emissions (observed)
segment genome into groups of nucleotides with similar signal

Hidden Markov Model
linear models

HMM

chr22: 17684448–17684670

log2(count+1)

t statistic

exons

states

genomic position

normal
tumor

HMM (candidate DERs)
linear models

HMM

permutation tests for statistical significance

log2(count+1)

chr22: 17684448–17684670

t statistic
match to annotation if desired: CECR1, “may play a role in regulating cell proliferation”
engineering challenges

• creating and handling nucleotide-by-sample matrix
• efficient linear model fitting (solution: lmFit)
• efficient segmentation with HMM
• efficient p-value calculations

Initial solution: https://github.com/alyssafrazee/derfinder
Approach 2: improve the current software infrastructure for analysis of transcriptome assemblies
Ballgown

transcriptome assembly pipelines

Align Reads (e.g. TopHat)

Assemble Transcripts (e.g. Cufflinks)

Estimate Expression (Cufflinks via Tablemaker, RSEM)

paired-end RNA-seq reads

Ballgown as connecting framework

R/Bioconductor DE analysis

Differential Expression Tests (default Ballgown models, limma, EdgeR, DESeq, …)

biorXiv preprint: http://biorxiv.org/content/early/2014/03/30/003665, Bioconductor package “ballgown”
S4 class for transcript assemblies

Ballgown object

- Expr
  - Exon
  - Intron
  - Transcript
  - Matrices
  - GRanges

- Structure
  - Exon
  - Intron
  - Transcript

- Indexes
  - pData
  - Bamfiles

- Data frames
  - e2t
  - i2t
  - t2g
easy exploration and DE analysis

plotting functions
easy exploration and DE analysis

```r
stat_results = stattest(my_assembly, feature='transcript',
                    meas='FPKM', covariate='group')

head(stat_results)
##   feature id      pval      qval
## 1 transcript 10 0.0138158 0.1052123
## 2 transcript 25 0.2677362 0.7911498
## 3 transcript 35 0.0108507 0.0895183
## 4 transcript 41 0.4710802 0.9025375
## 5 transcript 45 0.0840295 0.4893481
## 6 transcript 67 0.2731739 0.7911498
```

statistical tests
(drop-in replacement for Cuffdiff)
easy exploration and DE analysis

easily connects to existing DE packages

statistical tests
easy exploration and DE analysis

Assembled and Annotated Transcripts

annotation functions

get corresponding gene names, match assembled and annotated transcripts, plot assembly alongside annotation
highly flexible
freely available!

**Bioconductor (devel):**

```r
source("http://bioconductor.org/biocLite.R")
biocLite("ballgown")
```

**GitHub:**

https://github.com/alyssafrazee/ballgown

Cufflinks users will also need **Tablemaker:**

https://github.com/alyssafrazee/tablemaker
Thanks

Collaborators: Jeff Leek (advisor), Steven Salzberg, Ben Langmead, Andrew Jaffe, Rafa Irizarry, Sarven Sabunciyan, Kasper Hansen, Geo Pertea, Leonardo Collado Torres

Contact: alyssafrazee.com, alyssa.frazee@jhu.edu, @acfrazee (Twitter)
differential expression model

for each transcript, compare the fits of the following models using an F-test. Null hypothesis is that the fits of model (a) and model (b) are equally good; alternative is that (a) fits better.

\[
\text{(a)} \quad \text{expression}_i = \alpha + \beta_0 \text{group}_i + \sum_{p=1}^{P} \gamma_p \text{confounder}_{ip} + \text{noise}_{ip}
\]

\[
\text{(b)} \quad \text{expression}_i = \alpha^* + \sum_{p=1}^{P} \gamma_p^* \text{confounder}_{ip} + \text{noise}_{ip}^*
\]