Differential Gene Expression with RNA-Seq

GrasPods Coding Workshop

Rod Docking

2017-07-26

Experimental Medicine Program, Department of Medicine, University of British Columbia

Outline

- Introduction to differential gene expression analysis (10m)
- Hands-on runthrough of example analysis (40m)
- Next steps and discussion (10m)

R Preamble

If you want to run the code during today's workshop:

- Check that you can open the main `.Rmd` document
- Check that necessary libraries are installed - if not, start installing them now:

```
install.packages(c('tidyverse', 'knitr',
                   'RColorBrewer', 'stringr'))
source("https://bioconductor.org/biocLite.R")
biocLite(c("edgeR", "limma",
           "Mus.musculus", "Glimma", "gplots"))
```

Learning Goals

What we'll cover today:

- An overview of what kinds of questions DE analysis can be used to tackle
- The main problems that can occur in DE analysis, and how to address them
- A worked example of DE analysis

What we *won't* cover today:

- Quantification of RNA-Seq count data
- Deep dives into the statistical models underlying DE analysis
- Nitty-gritty R coding details
- Everything else (isoforms, exons, etc.)

Differential Expression Analysis Overview

- Underyling rationale/assumptions:
    - The pool of RNA molecules present in a cell is reflective of that cell's functional state
    - By quantifying the abundances of different RNA molecules, and *comparing* those abundances between different conditions, we can make inferences and test hypotheses about gene regulation and cell function
- Example questions:
    - When Gene A is knocked out, which other genes change their expression levels?
    - When different mutations occur in Gene B, do the same genes always change in their expression levels?
    - What's the difference in expression between, say, different types of breast progenitor cell?

Workthrough Background

- Adapted from Law, Alhamdoosh, Su, Smyth, & Ritchie (2016), based on data from Sheridan et al. (2015)
- Experimental design:
    - Sorted cell populations (basal, luminal progenitor, and mature luminal) from female mice
    - 3x replicates per condition, sequenced using Illumina 2x100bp reads
- Research questions:
    - Which genes are differentially expressed between these different cell populations?
    - Do those genes correspond to particular pathways?
- **So Why Not Just Do 20,000 *t* tests?**

Why Not To Do 20,000 t-tests (1/2)

- What if we sequenced a different number of reads for each sample?
- What if lowly-expressed genes behave very differently from highly-expressed genes?
- What if some genes aren't expressed at all?
- What if short genes behave differently from long genes?
- What if a few highly-expressed genes 'suck up' most of the reads in particular samples?
- What if there are batch effects due to differing lab conditions?

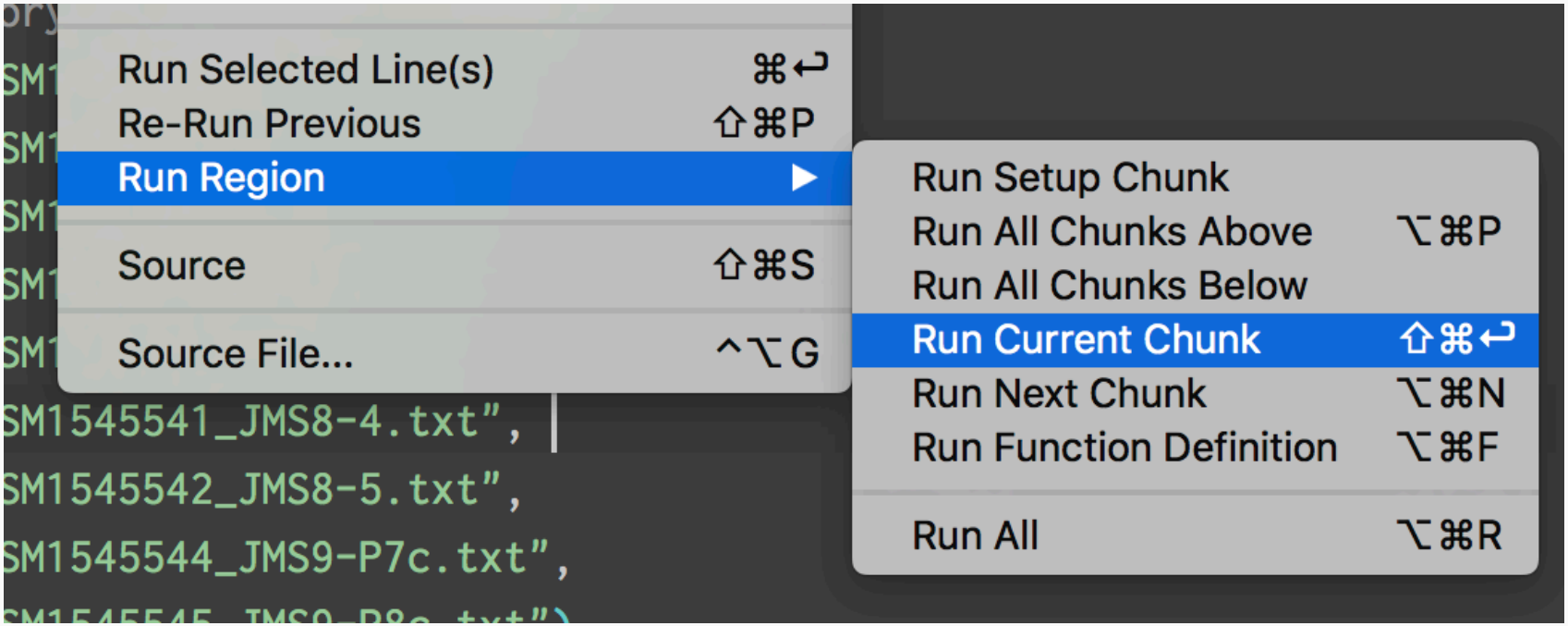
Why Not To Do 20,000 *t*-tests (2/2)

- What if counts *aren't* normally distributed at all? If they aren't, which model should we use? Poisson? Negative binomial?
- What if the variance in counts is different for long vs. short genes? Or high- vs. low-expression genes?
- How we adjust our *p* values to account for the thousands of tests we're doing?
- How do we decide what effect size (i.e. fold change) is biologically meaningful?

Disclaimers

- There are a *lot* of different ways to do this - the material presented today uses the popular `limma` package
- The R code here uses some fairly advanced data structures and is somewhat terse - I'll try to explain as we go. But - there *are* a lot gentler ways to learn R.

Workthrough



Where to Go From Here

- Gentler introductions to learning R:
    - `http://r4ds.had.co.nz/`
    - `http://stat545.com/`
    - `https://stat540-ubc.github.io/`
- A survey of best practices for RNA-seq data analysis: Conesa et al. (2016)

Designing Differential Expression Experiments

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Questions?

- Raise a GitHub issue
- Email: `rdocking` at `bcgsc.ca`
- Twitter: `@rdocking`

References

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. doi:10.1186/s13059-016-0881-8

Law, C. W., Alhamdoosh, M., Su, S., Smyth, G. K., & Ritchie, M. E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, glimma and edgeR. *F1000Research*, *5*, 1408. doi:10.12688/f1000research.9005.2

Sheridan, J. M., Ritchie, M. E., Best, S. A., Jiang, K., Beck, T. J., Vaillant, F., ... Visvader, J. E. (2015). A pooled shRNA screen for regulators of primary mammary stem and progenitor cells identifies roles for asap1 and prox1. *BMC Cancer*, *15*, 221. doi:10.1186/s12885-015-1187-z