Introduction to data manipulation with R

Natural History Museum Image Processing Course
11th January 2018
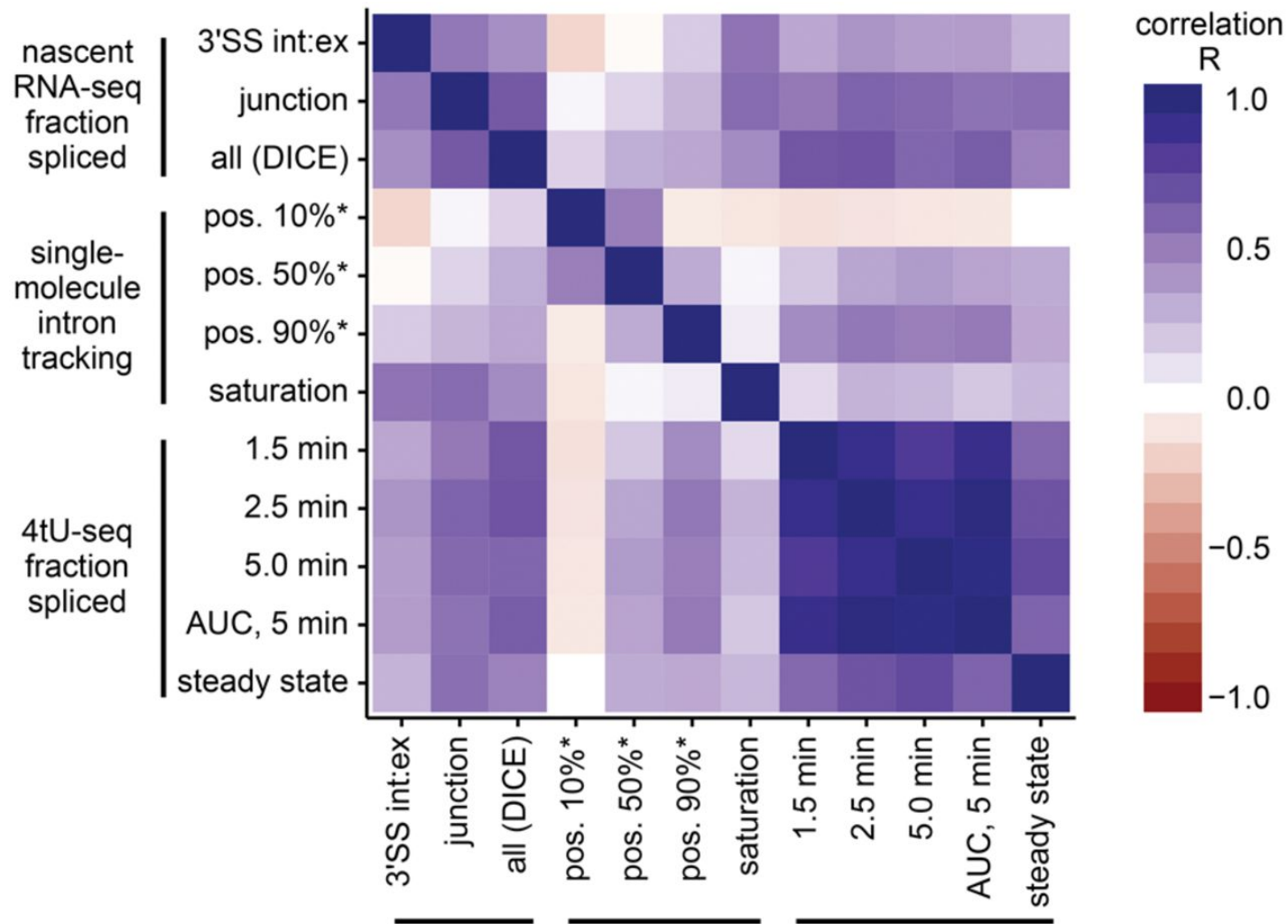
Edward Wallace
ewjwallace@gmail.com
Institute for Cell Biology & SynthSys,
University of Edinburgh

Download data from here:
https://github.com/ewallace/R-lesson-pores-data

Who am I?

- PhD in mathematics, Chicago
- Postdoc in protein synthesis, Chicago
- Studying RNA processing in fungi, Edinburgh
- Biological data scientist? Quantitative Biologist? Systems Biologist? Mycologist?
- I work with many kinds of biological data
  - trained as data carpentry instructor
  - examples from my recent paper next

# Estimates of cotranscriptional splicing, or splicing speed, mostly agree.

**Edward W.J. Wallace, and Jean D. Beggs RNA
2017;23:601-610**

RNA

Intron-containing ribosomal protein transcripts (blue) tend to be spliced faster and more cotranscriptionally, compared to nonribosomal transcripts (red).

Edward W.J. Wallace, and Jean D. Beggs RNA
2017;23:601-610

RNA
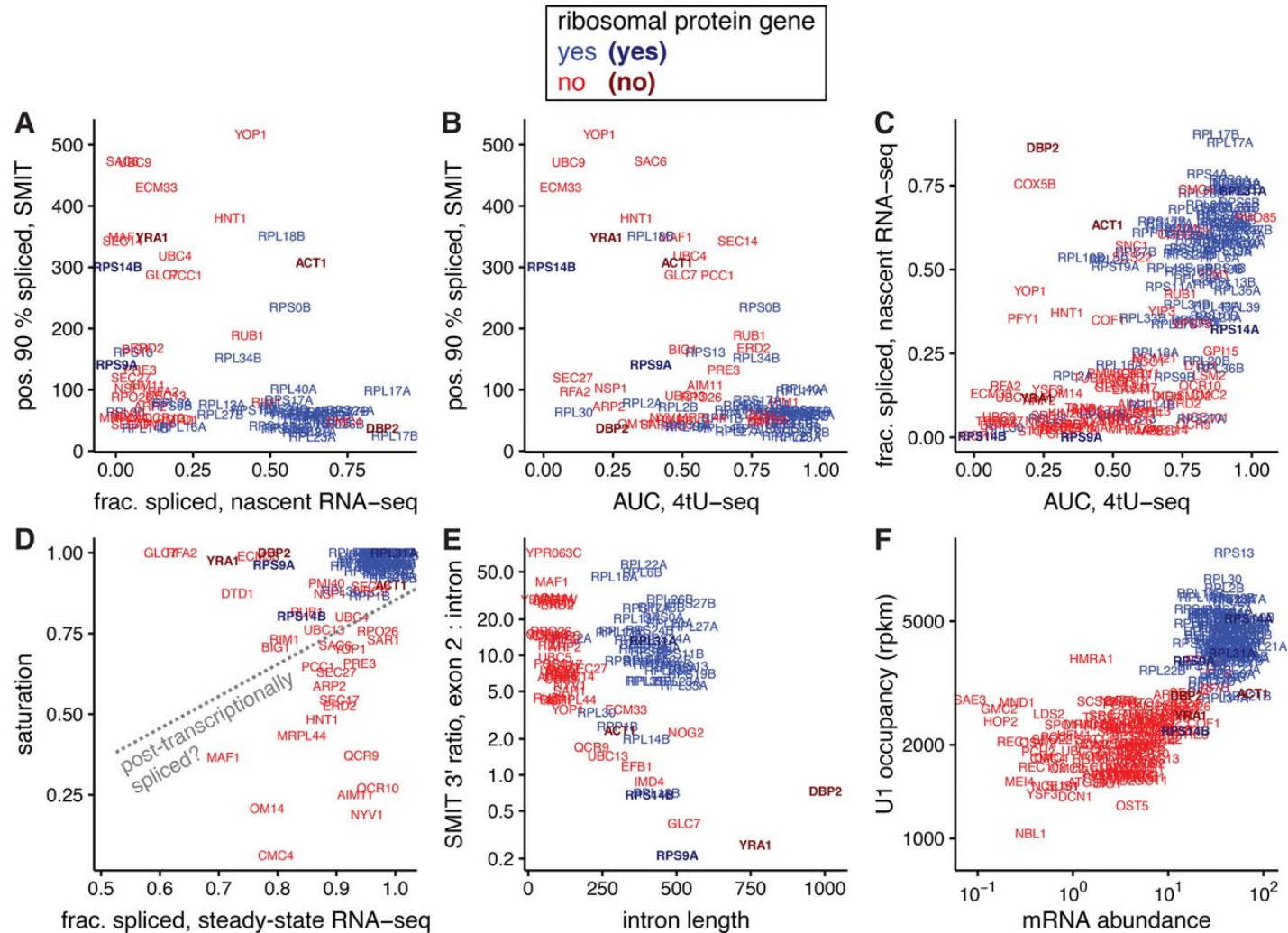
Comparison of SMIT and NET-seq profiles along individual genes, plotted in genomic co-ordinates.

RNA
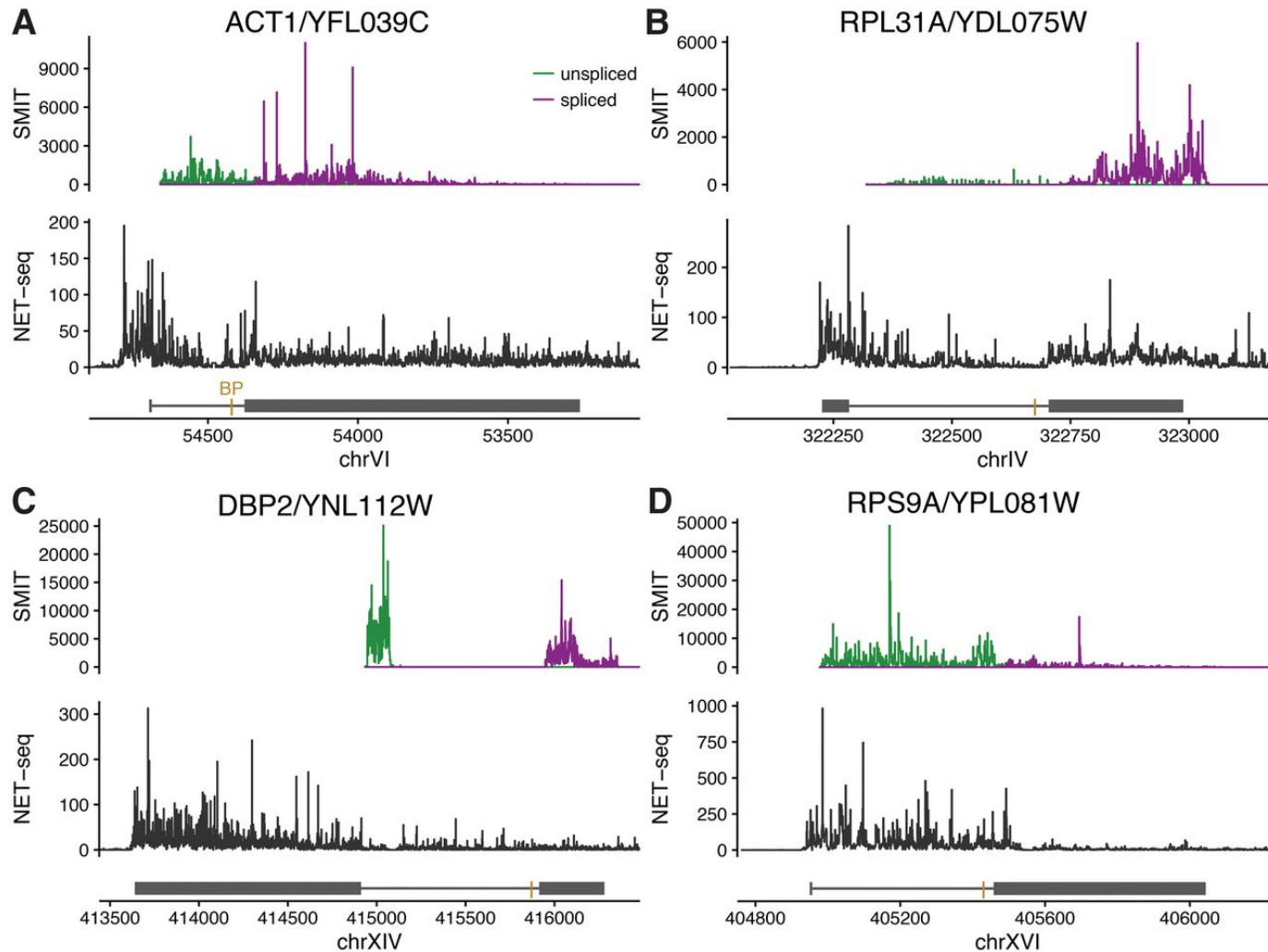
Why I use R to interpret and understand my data

- This has many steps:
  - import
  - tidy
  - transform
  - visualize
  - model data
- R does all of these well, and is
  - reproducible
  - publication-quality
  - free
  - extensible
- But there is a learning curve!

Helpful free resources for R

- Data Carpentry: http://www.datacarpentry.org/R-ecology-lesson/

- R for Data Science, Garrett Grolemund and Hadley Wickham: http://r4ds.had.co.nz/

- Fundamentals of Data Visualization, Claus Wilke: http://serialmentor.com/dataviz/

- Stack Overflow: https://stackoverflow.com/questions/tagged/r

- Your local R meetup: https://www.meetup.com/LondonR/

Key ideas for today

- Use tidyverse packages to
  - import with readr
  - transform with dplyr
  - visualize with ggplot2
- Data structure matters!
- Live coding, stop me if you have questions.

Download data from here:
https://github.com/ewallace/R-lesson-pores-data