R for data analysis and visualization

Data Carpentry workshop, Edinburgh
12th June 2018

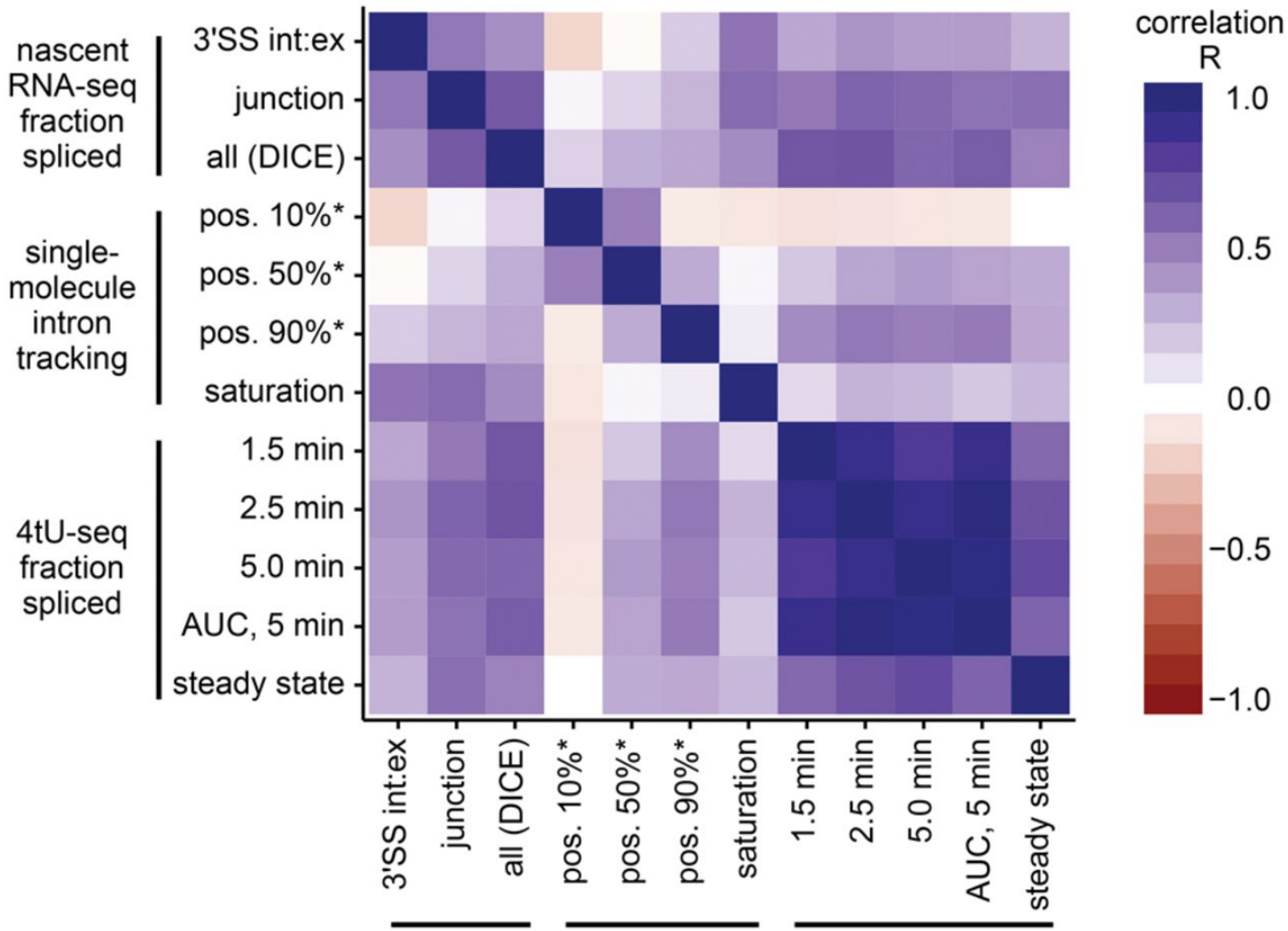
Edward Wallace
Edward.Wallace@ed.ac.uk
Institute for Cell Biology & SynthSys,
University of Edinburgh

Lesson website here:
http://www.datacarpentry.org/R-ecology-lesson/

Who am I?

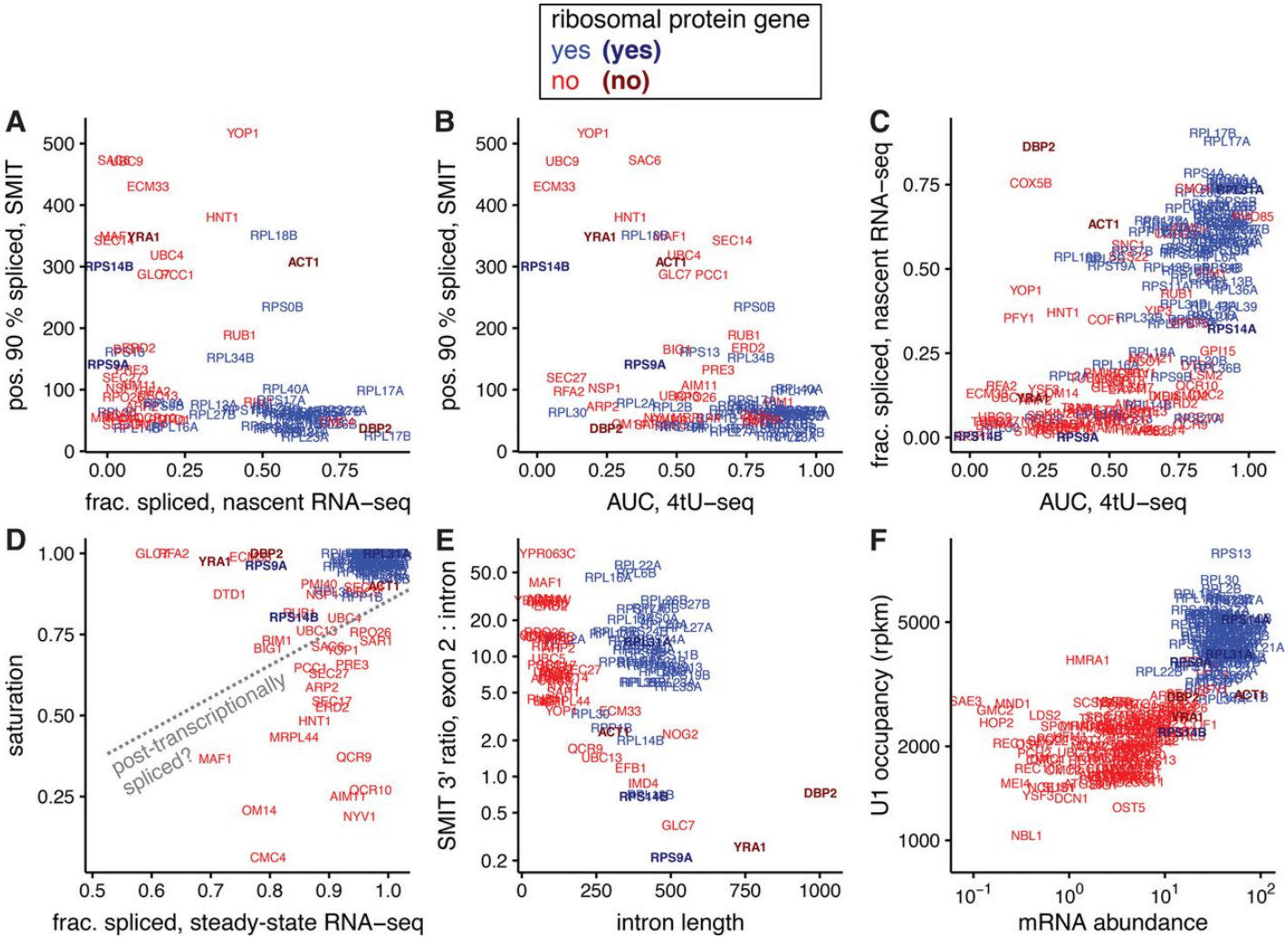
- Mathematics BA, Cambridge
- PhD in mathematics, Chicago - *MATLAB*
- Postdoc in protein synthesis, Chicago – *R, python*
- PI Studying RNA processing in fungi, Edinburgh
- Biological data scientist? Quantitative Biologist? Systems Biologist? RNA Biologist? Mycologist?
- I work with many kinds of biological data
  - sequences, RNA-seq, qPCR, proteomics, …

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



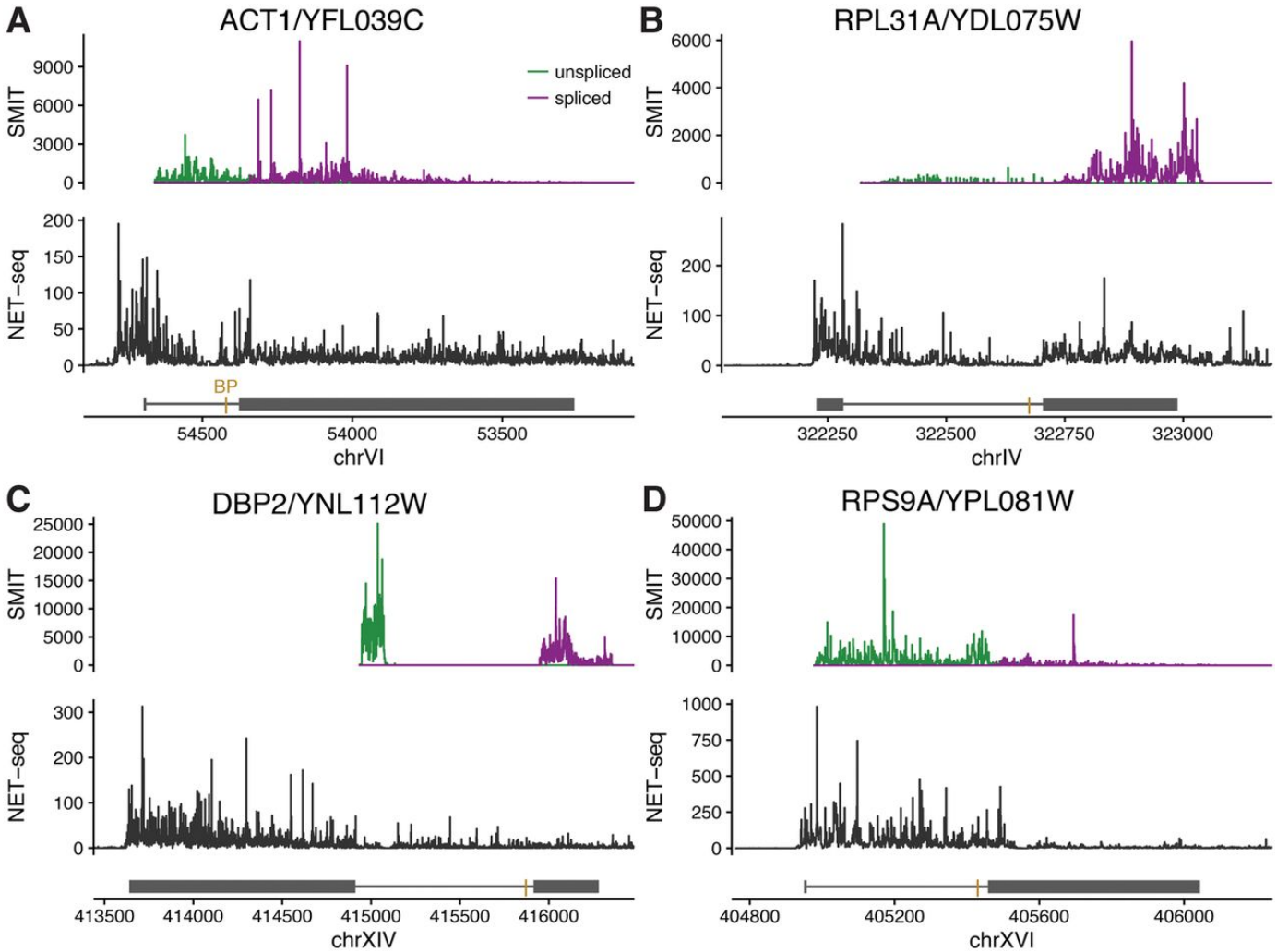
Edward W.J. Wallace, and Jean D. Beggs RNA 2017;23:601-610

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

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



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

Why R? I wanted to use data from this paper to plan an experiment:

MOLECULAR AND CELLULAR BIOLOGY, June 2004, p. 5534–5547 Vol. 24, No. 12
0270-7306/04/$08.00+0 DOI: 10.1128/MCB.24.12.5534–5547.2004

Genome-Wide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors

Jörg Grigull, Sanie Mnaimneh, Jeffrey Pootoolal, Mark D. Robinson, and Timothy R. Hughes*

Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5G 1L6, Canada

Received 2 December 2003/Returned for modification 6 January 2004/Accepted 9 March 2004

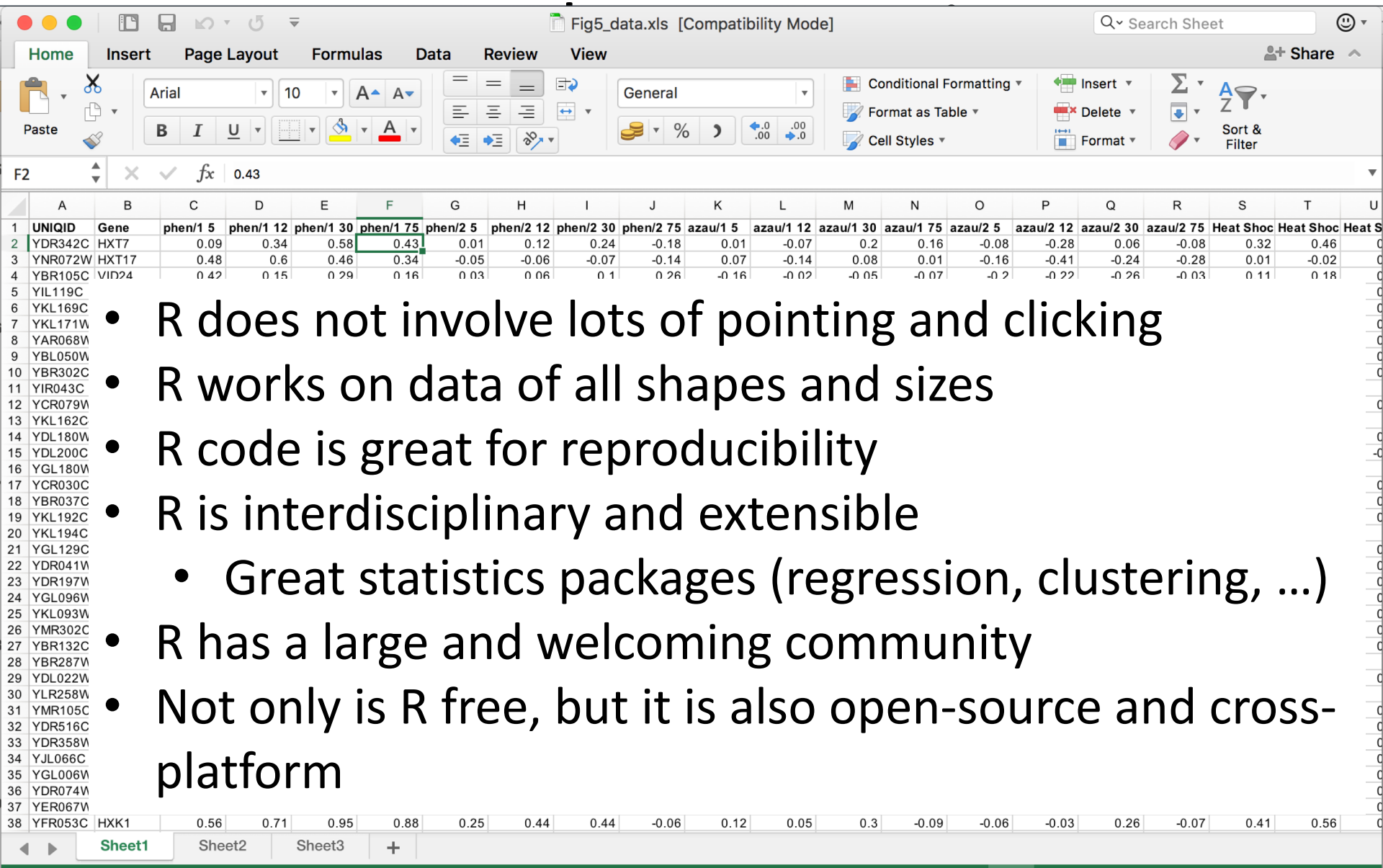
Why R? I wanted to use data from this

Fig5_data.xls [Compatibility Mode]

F2 | 0.43

UNIQID	Gene	phen/1 5	phen/1 12	phen/1 30	phen/1 75	phen/2 5	phen/2 12	phen/2 30	phen/2 75	azau/1 5	azau/1 12	azau/1 30	azau/1 75	azau/2 5	azau/2 12	azau/2 30	azau/2 75	Heat Shoc	Heat Shoc	Heat S
YDR342C	HXT7	0.09	0.34	0.58	0.43	0.01	0.12	0.24	-0.18	0.01	-0.07	0.2	0.16	-0.08	-0.28	0.06	-0.08	0.32	0.46	0
YNR072W	HXT17	0.48	0.6	0.46	0.34	-0.05	-0.06	-0.07	-0.14	0.07	-0.14	0.08	0.01	-0.16	-0.41	-0.24	-0.28	0.01	-0.02	0
YBR105C	VID24	0.42	0.15	0.29	0.16	0.03	0.06	0.1	0.26	-0.16	-0.02	-0.05	-0.07	-0.2	-0.22	-0.26	-0.03	0.11	0.18	0
YIL119C	RPI1	0.1	0.21	0.35	0.35	-0.04	-0.06	0.02	-0.1	-0.07	-0.06	0.06	0.04	-0.06	-0.12	0.11	-0.04	0.16	0.12	0
YKL169C	YKL169C	0.42	0.06	0.51	0.22	0.09	0.02	0.1	-0.13	0.07	-0.07	0.09	-0.02	0.07	0.15	0.25	-0.01	0.02	0.06	0
YKL171W	YKL171W	0.34	0.12	0.49	0.49	0.01	0.1	0.09	0	-0.13	-0.15	0.05	-0.07	-0.02	-0.05	-0.02	-0.08	0.1	0.11	0
YAR068W	YAR068W	0.06	0.2	0.34	0.56	0.07	0.22	0.39	0.48	-0.02	0.09	0.2	0.28	-0.04	0.01	0.08	0.17	0.02	0.08	0
YBL050W	SEC17	0	0.24	0.34	0.56	0.21	0.27	0.3	0.29	-0.01	0.06	0.3	0.28	0.03	-0.02	0.19	0.21	0.01	0.07	0
YBR302C	COS2	0.28	0.24	0.6	0.64	0.09	0.23	0.32	0.38	-0.11	0	0.05	0.11	0	-0.01	0.09	0.13	0.04	0.13	0
YIR043C	YIR043C	0.25	0.2	0.49	0.37	0.05	0.18	0.21	0.19	0.09	-0.02	0.05	0.04	0.14	0.05	0.1	0.15	0.04	0.11	
YCR079W	YCR079W	0.08	0.36	0.43	0.38	0.09	0.17	0.21	0.26	-0.03	0.09	0.09	0.1	-0.1	-0.18	0.03	0.24	0.1	0.15	0
YKL162C-/	YKL162C-/	0.02	0.02	0.93	0.76	0.04	0.09	0.2	0.34	0.04	0.05	0.29	0.23	0.04	0.11	0.38	0.52	0	0	
YDL180W	YDL180W	0.06	0.54	0.38	0.24	0.05	0.11	0.23	0.21	0.17	0.08	0.32	0.22	0.02	-0.01	0.33	0.41	0.03	0.09	0
YDL200C	MGT1	-0.08	0.27	0.47	0.46	0.08	0.22	0.24	0.22	0.05	0.14	0.23	0.16	-0.02	0.16	0.27	0.17	-0.05	-0.04	-0
YGL180W	APG1	0.31	0.61	0.74	0.63	0.33	0.42	0.24	-0.15	0.06	0.22	0.31	0.41	-0.01	0.1	0.31	0.52	0.08	0.09	
YCR030C	YCR030C	0.33	0.19	0.31	0.35	0.09	0.11	0.14	0.12	0.14	0.13	0.21	0.01	-0.06	0.07	0.14	0.16	0.07	0.04	0
YBR037C	SCO1	0.08	0.26	0.37	0.33	0.24	0.35	0.44	0.16	0.2	0.08	0.17	0.13	0.12	0.09	0.17	0.24	0.06	0.15	0
YKL192C	ACP1	-0.05	0.16	0.23	0.36	0.04	0.16	0.34	0.12	0.16	0.07	0.08	0.16	0.07	0.04	0.08	0.2	0.01	0.09	0
YKL194C	MST1	-0.1	0.1	0.32	0.21	0.06	0.17	0.31	0.08	0.13	0.08	0.13	0.1	0.08	0.01	0.22	0.18	0.16	0.08	
YGL129C	YGL129C	-0.24	0.25	0.26	0.41	0.1	0.21	0.21	0.04	-0.04	0.11	0.19	0.09	0.08	0.02	0.17	0.11	0.03	0.06	0
YDR041W	YDR041W	-0.07	0.13	0.19	0.16	0.1	0.2	0.27	0.23	0.09	0.06	0.08	0.08	0.06	0.06	0.12	0.16	0.01	0.08	0
YDR197W	CBS2	-0.05	0.08	0.33	0.33	0	0.04	0.13	0.27	0.13	0.08	0.05	0.19	0.06	-0.01	0.06	0.19	0.08	0.12	0
YGL096W	YGL096W	0.31	0.35	0.67	0.58	0.2	0.17	0.41	0.27	0.4	0.1	0.07	0.1	0.06	-0.05	0.08	0.18	0.17	0.12	0
YKL093W	MBR1	0.14	0.23	0.4	0.32	0.06	0.11	0.34	0.03	0.08	0.12	0.23	0.16	0.01	0.09	0.09	0.14	0.09	0.15	0
YMR302C	RNA12	-0.12	0.24	0.33	0.42	0.05	0.05	0.14	0.03	0.1	0.08	0.11	0.06	0.08	0.02	0.01	0.06	0.05	0.05	0
YBR132C	AGP2	0.45	0.38	0.48	0.52	0.15	0.08	0.08	-0.13	0.04	0.15	0.15	0.1	-0.06	0.55	0.14	0.13	0.19	0.23	0
YBR287W	YBR287W	0.47	0.25	0.45	0.4	0.06	0.05	0.04	0	-0.03	0.05	-0.04	-0.02	-0.05	0.05	-0.01	0.18	0.08	0.14	
YDL022W	GPD1	0.37	0.46	0.61	0.25	0.12	0.26	0.07	-0.21	0.04	0.17	0.15	0.1	0.05	0.15	0.15	0.13	0.32	0.39	0
YLR258W	GSY2	0.23	0.38	0.49	0.39	0.19	0.3	0.21	-0.08	0.22	0.17	0.2	0.17	0.05	-0.01	0.15	0.04	0.34	0.48	
YMR105C	PGM2	0.72	0.85	1.25	1.07	0.42	0.64	0.68	0.04	0.16	0.22	0.22	0.35	0.11	-0.07	0.22	0.33	0.52	0.71	0
YDR516C	YDR516C	0.45	0.61	0.63	0.64	0.25	0.35	0.27	-0.11	0.17	0.15	0.23	-0.04	0.14	0.06	0.21	-0.13	0.32	0.42	0
YDR358W	GGA1	0.3	0.63	0.51	0.58	0.14	0.22	0.21	0.07	0	0.28	0.22	0.11	0.2	0.13	0.21	0.22	0.22	0.22	0
YJL066C	YJL066C	0.01	0.33	0.49	0.41	0.1	0.17	0.27	0.18	0.02	0.21	0.2	0.23	0.11	0.19	0.22	0.29	0.14	0.24	0
YGL006W	PMC1	0.09	0.29	0.41	0.47	0.05	0.13	0.15	0.06	0.12	0.13	0.22	0.07	0.06	0.02	0.11	0.04	0.13	0.13	0
YDR074W	TPS2	0.66	0.66	0.64	0.36	0.23	0.3	0.16	-0.36	0.1	0.11	0.1	0.17	-0.19	-0.13	0.03	-0.01	0.41	0.41	0
YER067W	YER067W	0.63	0.87	1.1	0.77	0.53	0.86	0.72	-0.03	0.05	-0.02	0.18	-0.17	-0.15	-0.31	0.17	-0.14	0.47	0.49	0
YFR053C	HXK1	0.56	0.71	0.95	0.88	0.25	0.44	0.44	-0.06	0.12	0.05	0.3	-0.09	-0.06	-0.03	0.26	-0.07	0.41	0.56	0

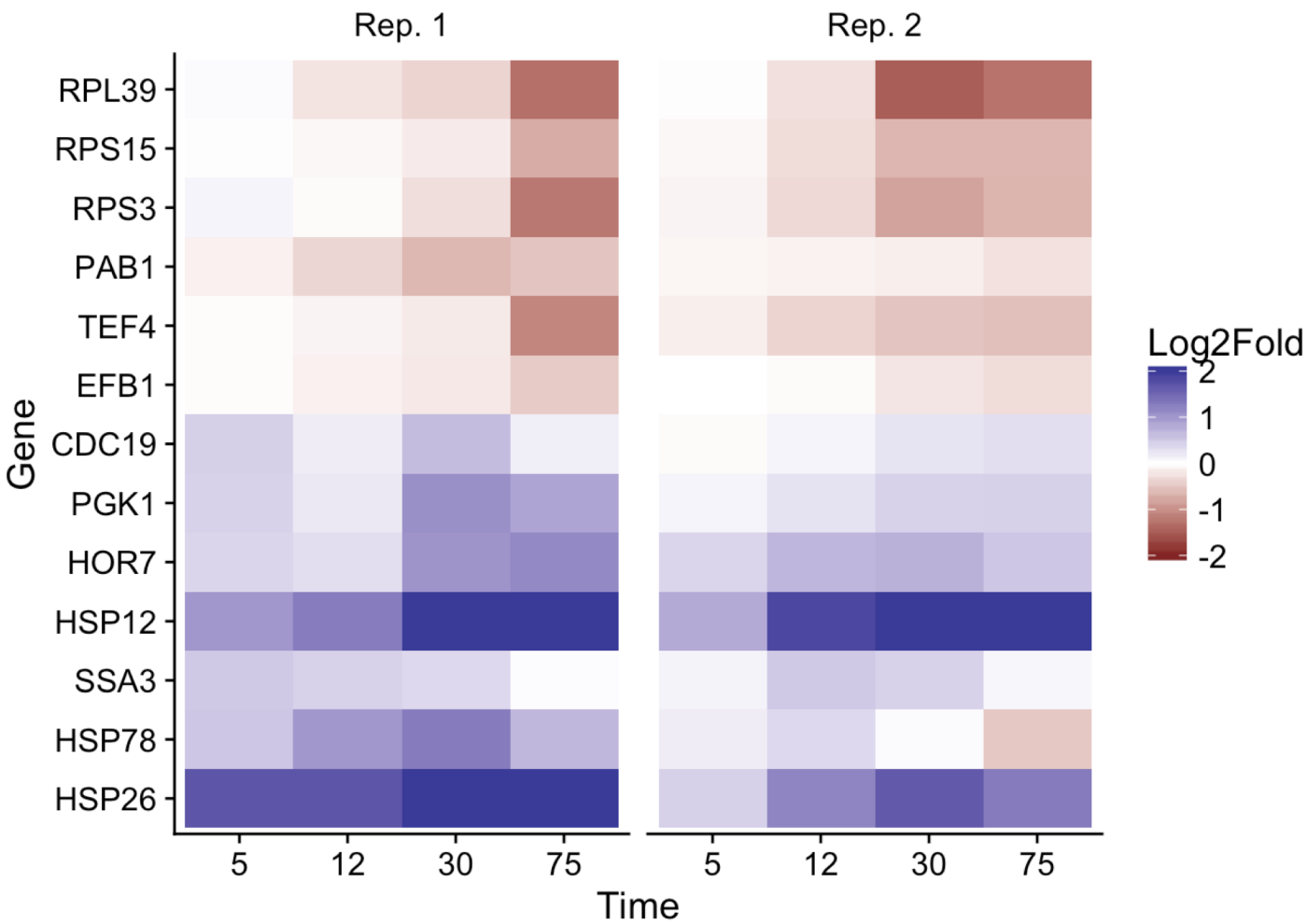
Why R? I wanted to use data from this



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	UNIQID	Gene	phen/1 5	phen/1 12	phen/1 30	phen/1 75	phen/2 5	phen/2 12	phen/2 30	phen/2 75	azau/1 5	azau/1 12	azau/1 30	azau/1 75	azau/2 5	azau/2 12	azau/2 30	azau/2 75	Heat Shoc	Heat Shoc	Heat S
2	YDR342C	HXT7	0.09	0.34	0.58	0.43	0.01	0.12	0.24	-0.18	0.01	-0.07	0.2	0.16	-0.08	-0.28	0.06	-0.08	0.32	0.46	0
3	YNR072W	HXT17	0.48	0.6	0.46	0.34	-0.05	-0.06	-0.07	-0.14	0.07	-0.14	0.08	0.01	-0.16	-0.41	-0.24	-0.28	0.01	-0.02	0
4	YBR105C	VID24	0.42	0.15	0.29	0.16	0.03	0.06	0.1	0.26	-0.16	-0.02	-0.05	-0.07	-0.2	-0.22	-0.26	-0.03	0.11	0.18	0
5	YIL119C																				0
6	YKL169C																				0
7	YKL171W																				0
8	YAR068W																				0
9	YBL050W																				0
10	YBR302C																				0
11	YIR043C																				0
12	YCR079W																				0
13	YKL162C																				0
14	YDL180W																				0
15	YDL200C																				-0
16	YGL180W																				0
17	YCR030C																				0
18	YBR037C																				0
19	YKL192C																				0
20	YKL194C																				0
21	YGL129C																				0
22	YDR041W																				0
23	YDR197W																				0
24	YGL096W																				0
25	YKL093W																				0
26	YMR302C																				0
27	YBR132C																				0
28	YBR287W																				0
29	YDL022W																				0
30	YLR258W																				0
31	YMR105C																				0
32	YDR516C																				0
33	YDR358W																				0
34	YJL066C																				0
35	YGL006W																				0
36	YDR074W																				0
37	YER067W																				0
38	YFR053C	HXK1	0.56	0.71	0.95	0.88	0.25	0.44	0.44	-0.06	0.12	0.05	0.3	-0.09	-0.06	-0.03	0.26	-0.07	0.41	0.56	0

- R does not involve lots of pointing and clicking
- R works on data of all shapes and sizes
- R code is great for reproducibility
- R is interdisciplinary and extensible
  - Great statistics packages (regression, clustering, …)
- R has a large and welcoming community
- Not only is R free, but it is also open-source and cross-platform

R produces high-quality graphics that help me understand data **quickly**



There are many places to get help:

- 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 group: http://edinbr.org/

Key ideas for this workshop

- Organize your workspace
- Basic elements of R (objects, functions, …)
- Starting with data
- Manipulating data frames
- Visualizing data - *tomorrow*
- Live coding, stop me if you have questions.

Lesson website here:
http://www.datacarpentry.org/R-ecology-lesson/