Reproducible quantitative proteomics analyses using Jupyter notebooks and R

Phillip Wilmarth

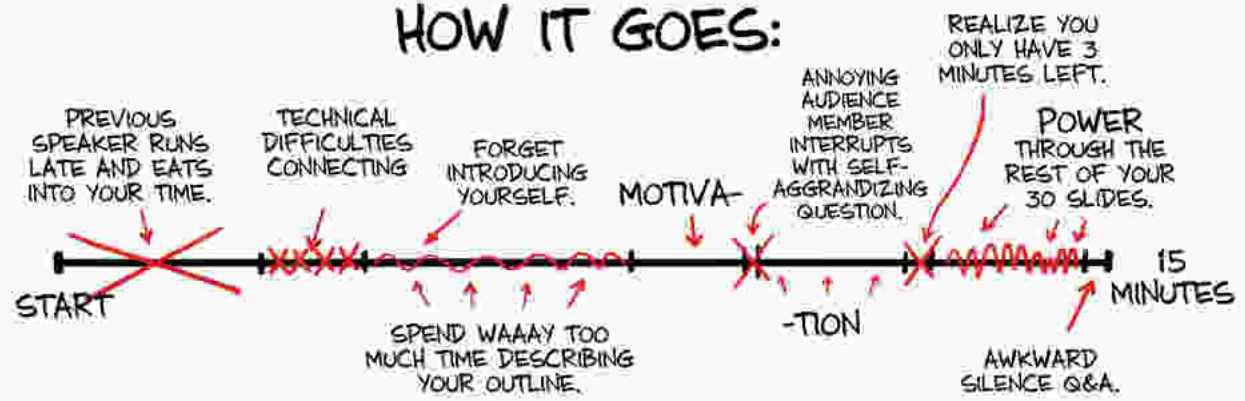
Proteomics Core, OHSU

2018 Cascadia Proteomics Symposium

YOUR CONFERENCE PRESENTATION

HOW YOU PLANNED IT:

HOW IT GOES:



JORGE CHAM © 2013

WWW.PHDCOMICS.COM

Somers J. “The scientific paper is obsolete.” The Atlantic. 2018 Apr 5.

...omic analysis of elongated skulls
...tensive female-biased immigration
...arly Medieval Bavaria

Krishna R. Veeramah[a], Andreas Rott[b,1], Melanie Groß[c,1], Lucy van Dorp[d], Saioa López[e], Karola Kirsanow[c], Christian Sell[c], Jens Blöcher[c], Daniel Wegmann[f,g], Vivian Link[f,g], Zuzana Hofmanová[f,g], Joris Peters[b,h], Bernd Trautmann[b], Anja Gairhos[i], Jochen Haberstroh[j], Bernd Päffgen[k], Garrett Hellenthal[d], Brigitte Haas-Gebhard[i], Michaela Harbeck[b,2,3], and Joachim Burger[c,2,3]

[a]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245; [b]State Collection for Anthropology and Palaeoanatomy, Bavarian Natural History Collections, 80333 Munich, Germany; [c]Palaeogenetics Group, Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, 55099 Mainz, Germany; [d]UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, WC1E 6BT London, United Kingdom; [e]Cancer Institute, University College London, WC1E 6DD London, United Kingdom; [f]Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland; [g]Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland; [h]ArchaeoBioCenter and Institute for Palaeoanatomy, Domestication Research and the History of Veterinary Medicine, Ludwig Maximilian University, 80539 Munich, Germany; [i]Bavarian State Archaeological Collection, 80538 Munich, Germany; [j]Bavarian State Department of Monuments and Sites, 80539 Munich, Germany; and [k]Institute of Prehistoric and Protohistoric Archaeology, Ludwig Maximilian University, 80799 Munich, Germany

Edited by Eske Willerslev, University of Copenhagen, Copenhagen, Denmark, and approved January 30, 2018 (received for review November 21, 2017)

Modern European genetic structure demonstrates strong correlations with geography, while genetic analysis of prehistoric humans has indicated at least two major waves of immigration from outside the continent during periods of cultural change. However, population-level genome data that could shed light on the demographic processes occurring during the intervening periods have been absent. Therefore, we generated genomic data from 41 individuals dating mostly to the late 5th/early 6th century AD from present-day Bavaria in southern Germany, including 11 whole genomes (mean depth 5.56×). In addition we developed a capture array to sequence neutral regions spanning a total of 5 Mb and 486 functional polymorphic sites to high depth (mean 72×) in all individuals. Our data indicate that while men generally had

to form in the 5th century AD, and that it emanated from a combination of the romanized local population of the border province of the former Roman Empire and immigrants from north of the Danube (2). While the Baiuvarii are less well known than some other contemporary groups, an interesting archaeological feature in Bavaria from this period is the presence of skeletons with artificially deformed or elongated skulls (Fig. 1*A*).

Artificial cranial deformation (ACD), which is only possible during early childhood, is a deliberate and permanent shaping of the head performed with great effort. In some societies reshaping the human skull has been seen as an ideal of beauty, while it

Significance

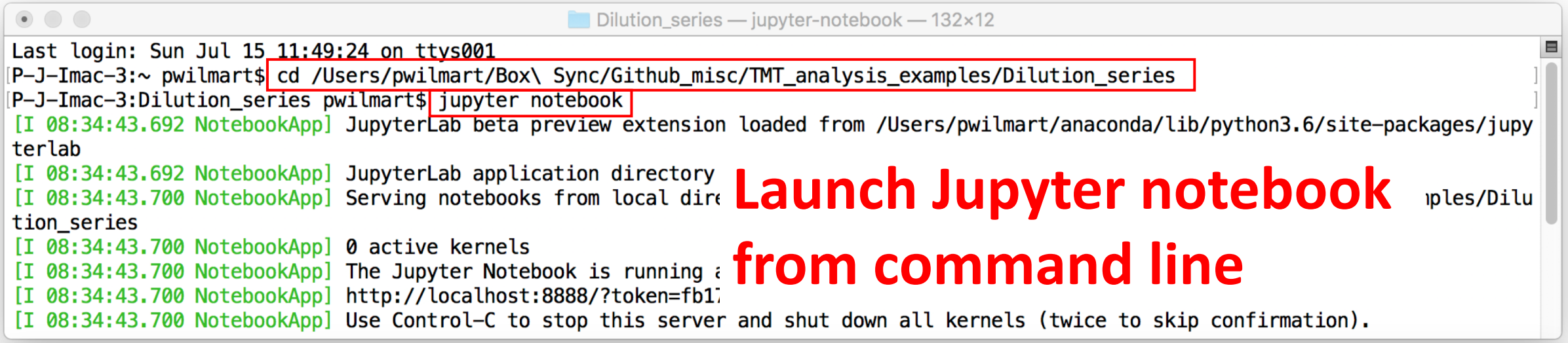
POPULATION BIOLOGY

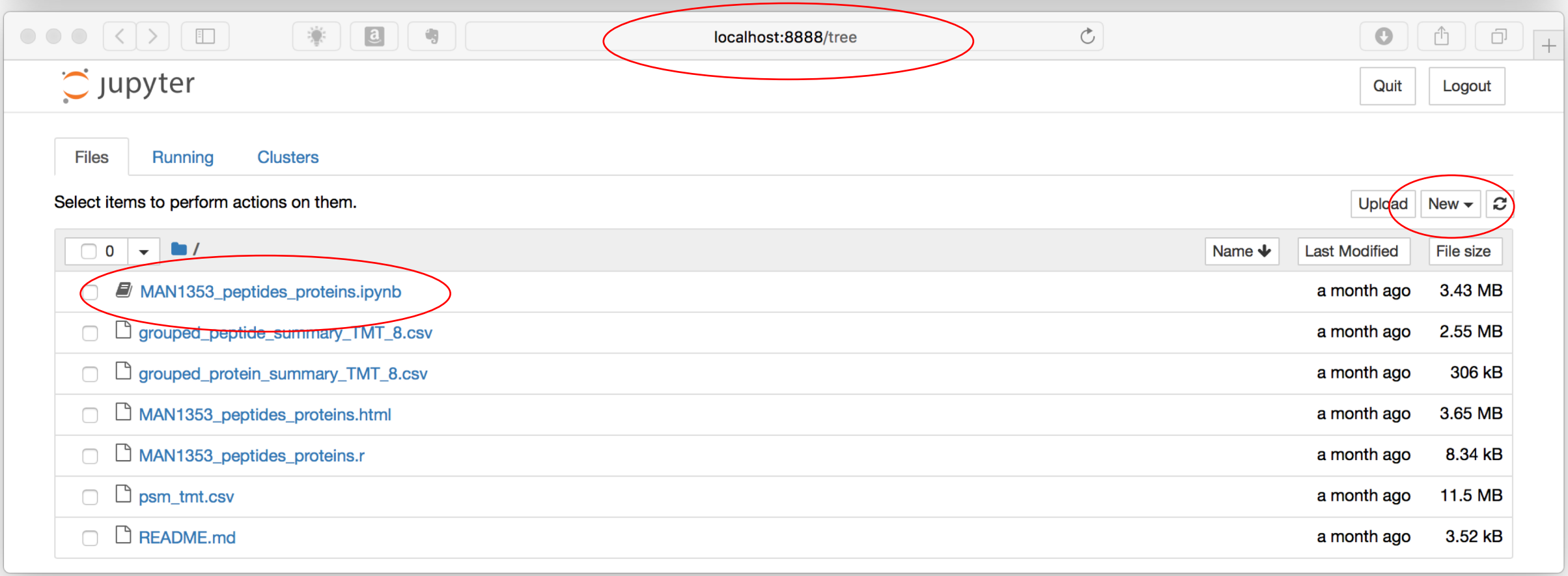
Atlantic article says notebooks can replace papers – so what are notebooks?

- Notebooks are web server applications
- You work on notebooks with web browsers
- Notebooks format text, run code, and show output
- Jupyter notebooks or R Markdown (RStudio) + knitr

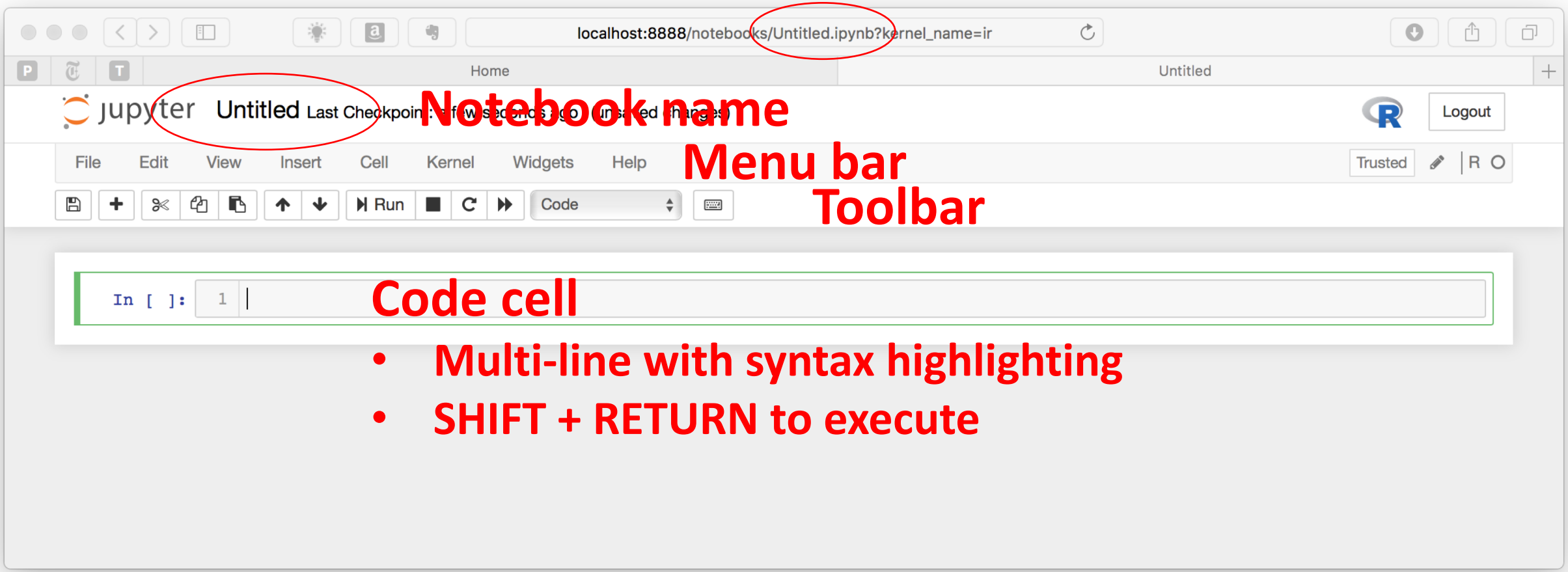
Installing Jupyter notebooks

- Part of scientific python distributions
  - www.anaconda.com
- Extensive documentation online
  - www.jupyter.org/documentation
- Supports Python by default
- R support needs separate installation
  - https//irkernel.github.io/installation/

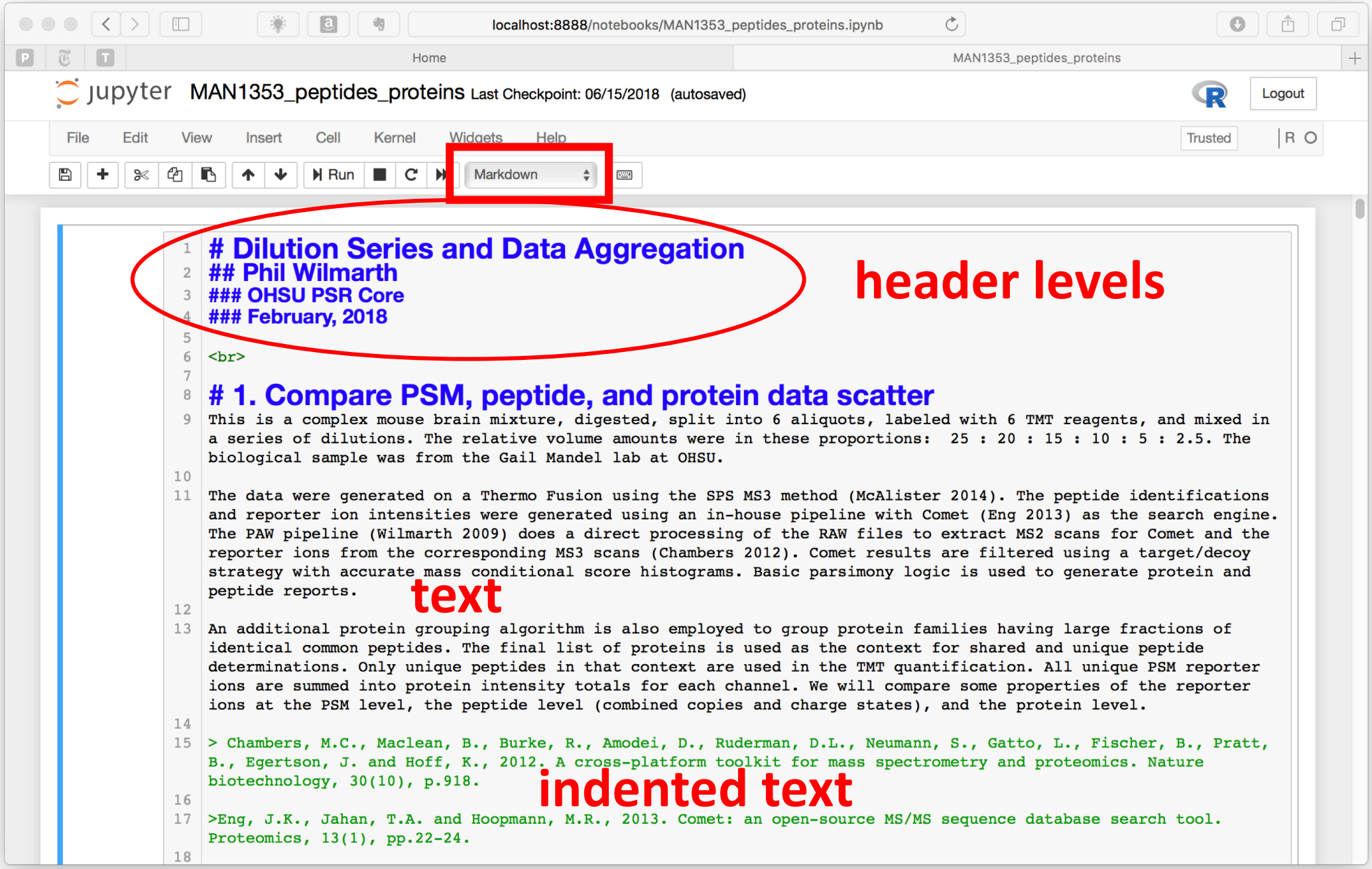

Dilution_series — jupyter-notebook — 132×12
Last login: Sun Jul 15 11:49:24 on ttys001
P-J-Imac-3:~ pwilmart$ cd /Users/pwilmart/Box\ Sync/Github_misc/TMT_analysis_examples/Dilution_series
P-J-Imac-3:Dilution_series pwilmart$ jupyter notebook
[I 08:34:43.692 NotebookApp] JupyterLab beta preview extension loaded from /Users/pwilmart/anaconda/lib/python3.6/site-packages/jupyterlab
[I 08:34:43.692 NotebookApp] JupyterLab application directory
[I 08:34:43.700 NotebookApp] Serving notebooks from local dire
tion_series
[I 08:34:43.700 NotebookApp] 0 active kernels
[I 08:34:43.700 NotebookApp] The Jupyter Notebook is running a
[I 08:34:43.700 NotebookApp] http://localhost:8888/?token=fb17
[I 08:34:43.700 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
Launch Jupyter notebook
from command line


localhost:8888/tree
jupyter
Quit
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
0
/
Name
Last Modified
File size
MAN1353_peptides_proteins.ipynb
a month ago
3.43 MB
grouped_peptide_summary_TMT_8.csv
a month ago
2.55 MB
grouped_protein_summary_TMT_8.csv
a month ago
306 kB
MAN1353_peptides_proteins.html
a month ago
3.65 MB
MAN1353_peptides_proteins.r
a month ago
8.34 kB
psm_tmt.csv
a month ago
11.5 MB
README.md
a month ago
3.52 kB

Notebook anatomy



https://jupyter-notebook.readthedocs.io/en/stable/notebook.html



https://medium.com/ibm-data-science-experience/markdown-for-jupyter-notebooks-cheatsheet-386c05aeebed

localhost:8888/notebooks/MAN1353_peptides_proteins.ipynb
Home
MAN1353_peptides_proteins
jupyter MAN1353_peptides_proteins Last Checkpoint: 06/15/2018 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
R
Run
Markdown

Dilution Series and Data Aggregation

Phil Wilmarth

OHSU PSR Core

February, 2018

1. Compare PSM, peptide, and protein data scatter

This is a complex mouse brain mixture, digested, split into 6 aliquots, labeled with 6 TMT reagents, and mixed in a series of dilutions. The relative volume amounts were in these proportions: 25 : 20 : 15 : 10 : 5 : 2.5. The biological sample was from the Gail Mandel lab at OHSU.

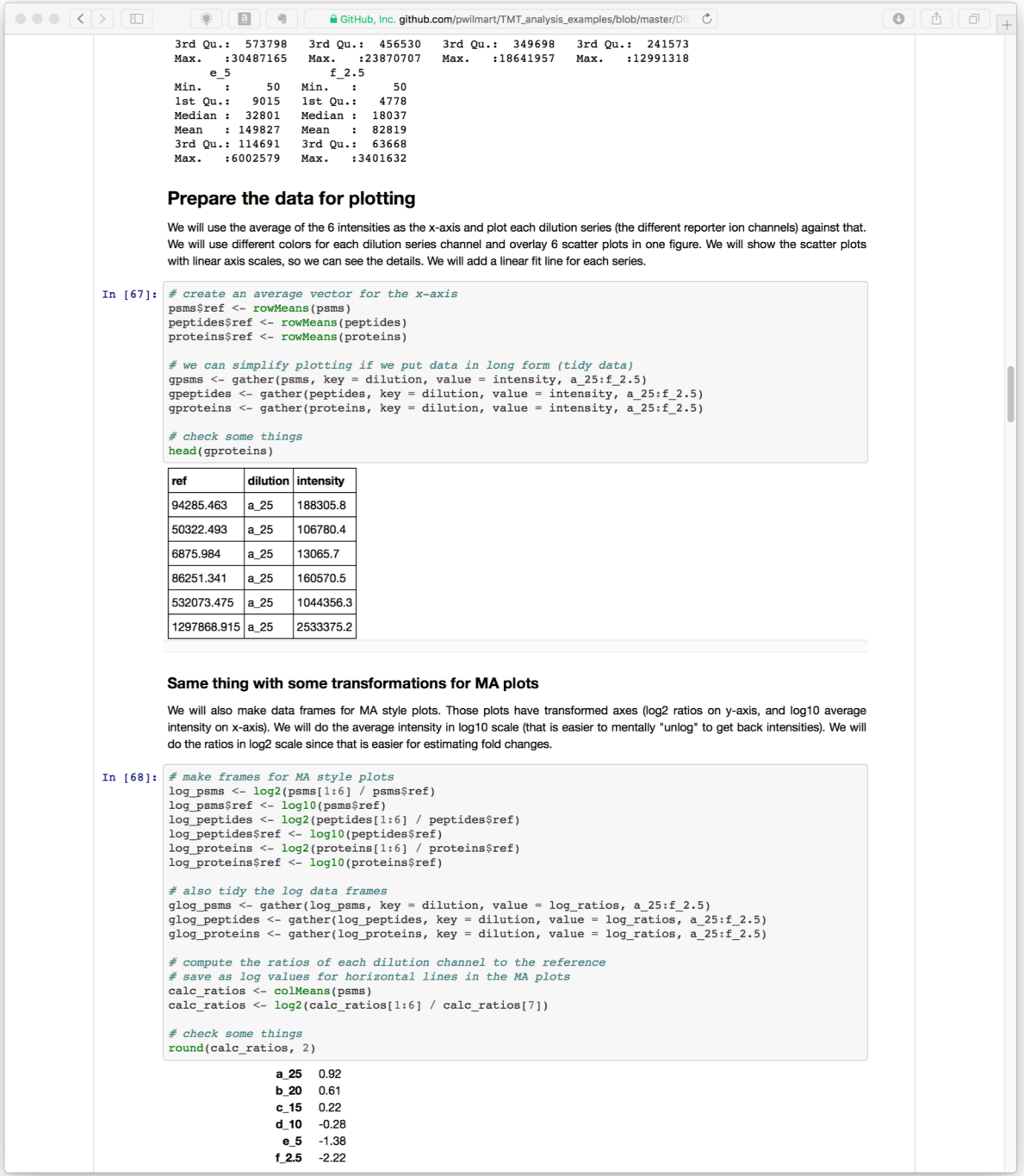
The data were generated on a Thermo Fusion using the SPS MS3 method (McAlister 2014). The peptide identifications and reporter ion intensities were generated using an in-house pipeline with Comet (Eng 2013) as the search engine. The PAW pipeline (Wilmarth 2009) does a direct processing of the RAW files to extract MS2 scans for Comet and the reporter ions from the corresponding MS3 scans (Chambers 2012). Comet results are filtered using a target/decoy strategy with accurate mass conditional score histograms. Basic parsimony logic is used to generate protein and peptide reports.

An additional protein grouping algorithm is also employed to group protein families having large fractions of identical common peptides. The final list of proteins is used as the context for shared and unique peptide determinations. Only unique peptides in that context are used in the TMT quantification. All unique PSM reporter ions are summed into protein intensity totals for each channel. We will compare some properties of the reporter ions at the PSM level, the peptide level (combined copies and charge states), and the protein level.

> Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J. and Hoff, K., 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nature biotechnology, 30(10), p.918.

> Eng, J.K., Jahan, T.A. and Hoopmann, M.R., 2013. Comet: an open-source MS/MS sequence database search tool. Proteomics, 13(1), pp.22-24.

Notebooks are a linear collection of cells



Pictures are worth a thousand words

- R and ggplot2 are very powerful for data visualization
- TMT labeling of same digest in 6 dilutions

  25 : 20 : 15 : 10 : 5 : 2.5

- Comparisons at PSM, peptide, protein levels
- https://github.com/pwilmart/TMT_analysis_examples

Markdown

Code

Output

localhost:8888/notebooks/MAN1353_peptides_proteins.ipynb

Home | MAN1353_peptides_proteins

jupyter MAN1353_peptides_proteins Last Checkpoint: 06/15/2018 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | R

Run | Code

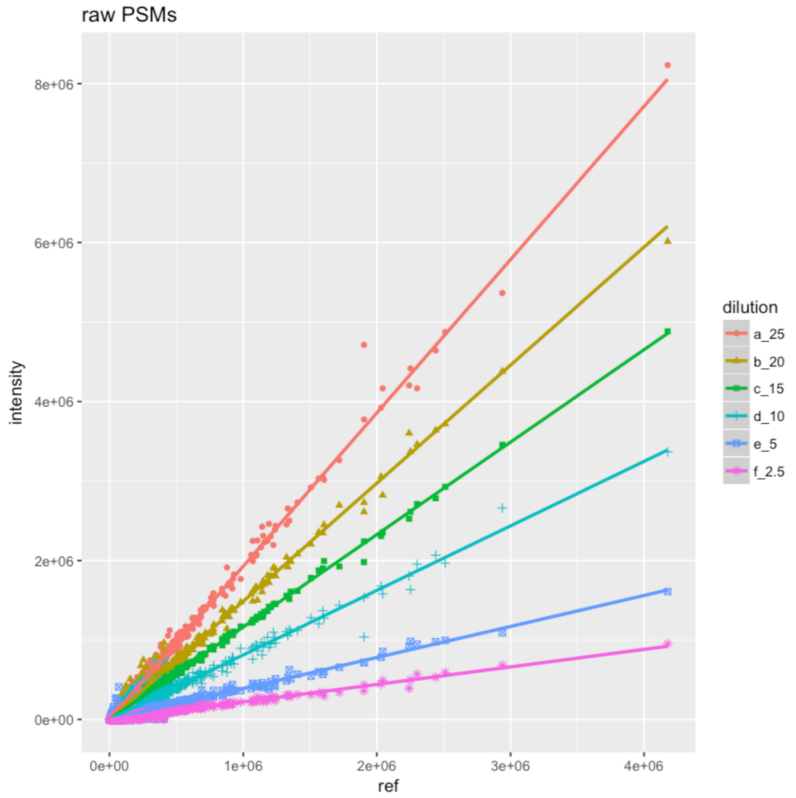
Look at the PSMs first

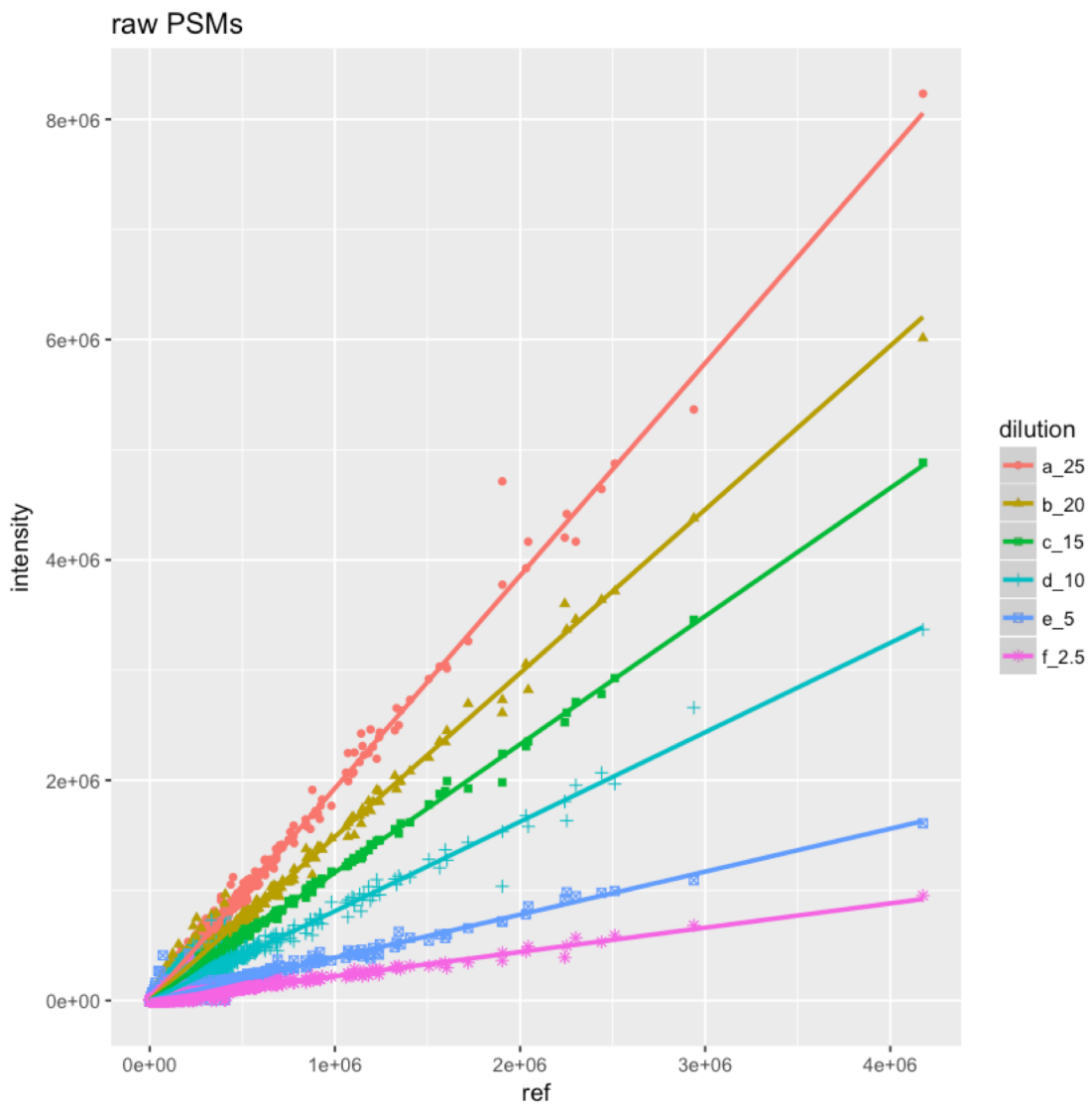
We have not done any filtering for the PSMs other than removing zeros. They are the raw peak heights. There will be duplicates of the same peptides. There will also be signals from contaminants and decoys.

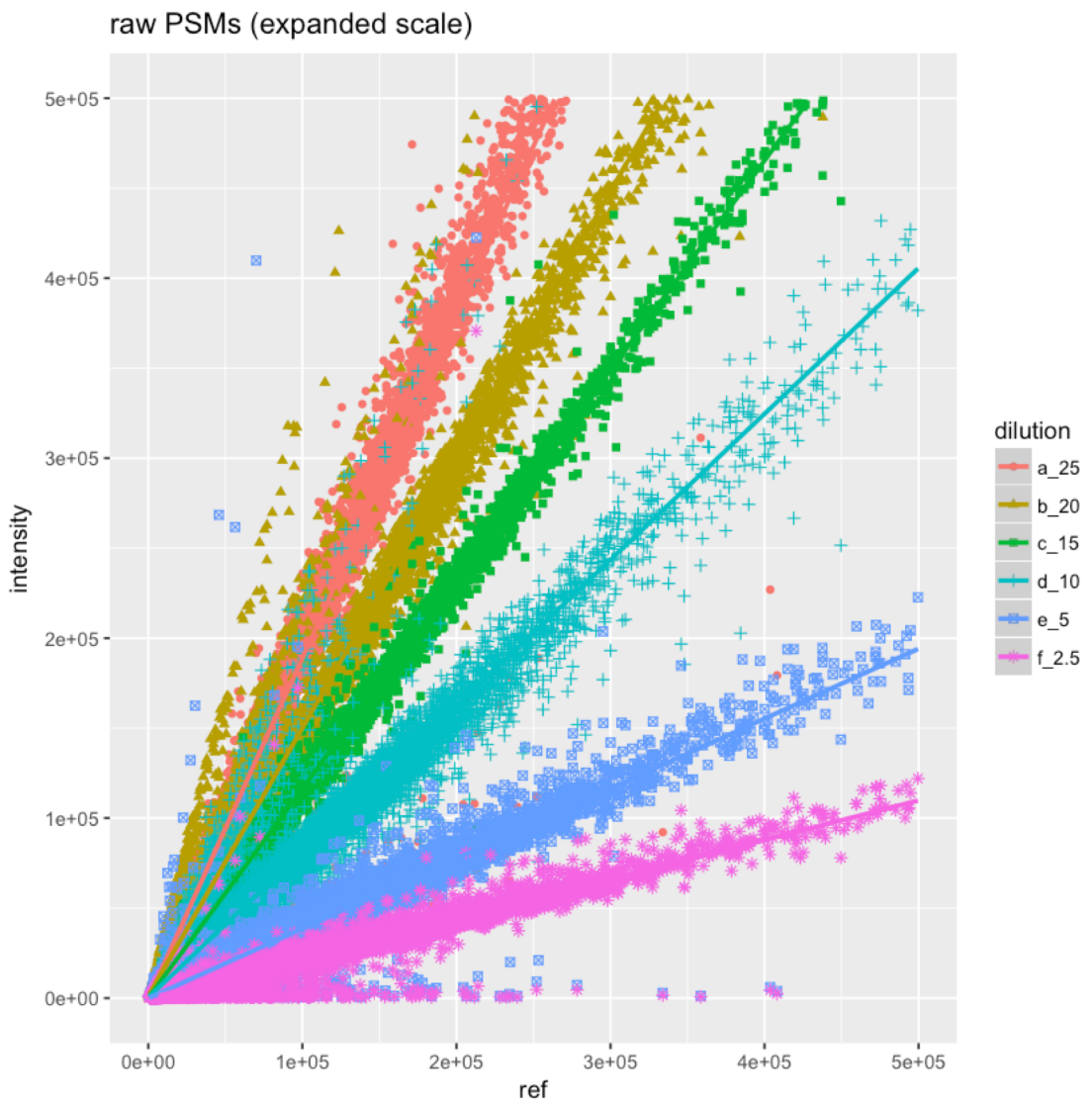
In [69]:

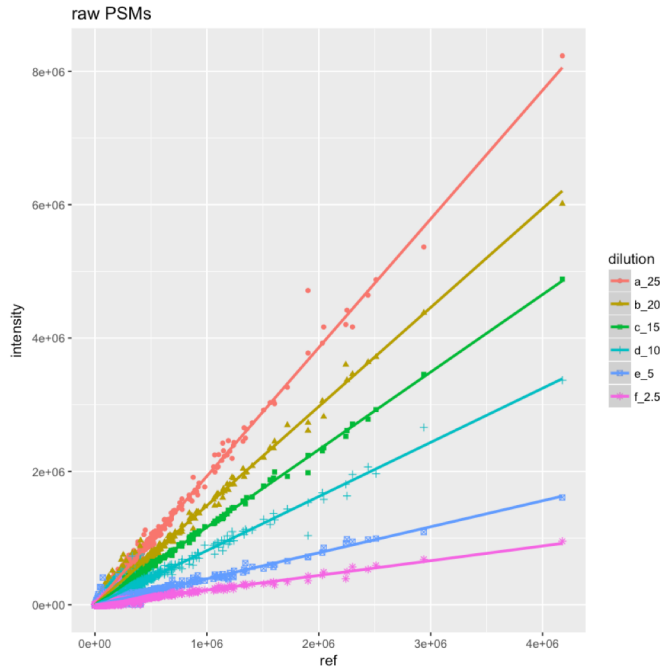
```
# full scale, linear axes
ggplot(data = gpsms, aes(x = ref, y = intensity)) +
  geom_point(aes(color = dilution, shape = dilution)) +
  geom_smooth(aes(color = dilution), method = "lm") +
  ggtitle("raw PSMs")

# expanded scale
ggplot(data = gpsms, aes(x = ref, y = intensity)) +
  geom_point(aes(color = dilution, shape = dilution)) +
  geom_smooth(aes(color = dilution), method = "lm") +
  xlim(c(0, 500000)) + ylim(c(0, 500000)) +
  ggtitle("raw PSMs (expanded scale)")
```

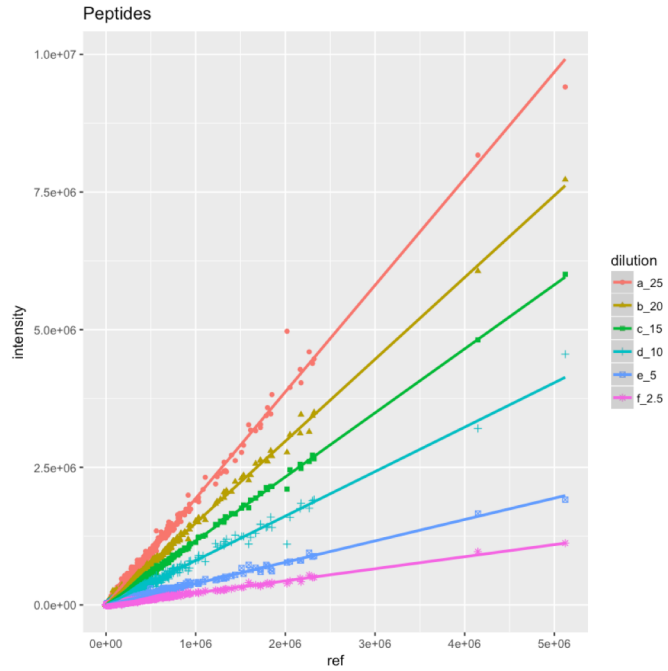
```
Warning message:
"Removed 1277 rows containing non-finite values (stat_smooth)."Warning message:
"Removed 1277 rows containing missing values (geom_point)."Warning message:
"Removed 52 rows containing missing values (geom_smooth)."
```

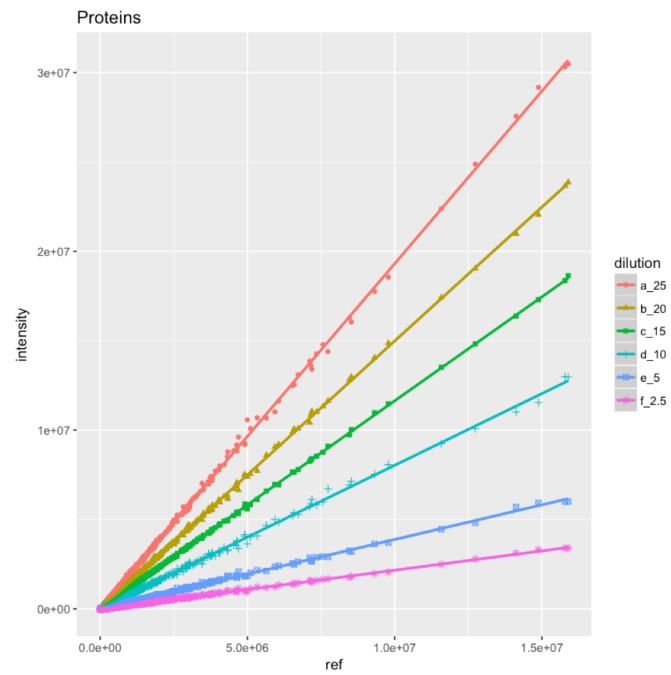



raw PSMs
intensity
ref
dilution
a_25
b_20
c_15
d_10
e_5
f_2.5
0e+00
2e+06
4e+06
6e+06
8e+06
0e+00
1e+06
2e+06
3e+06
4e+06

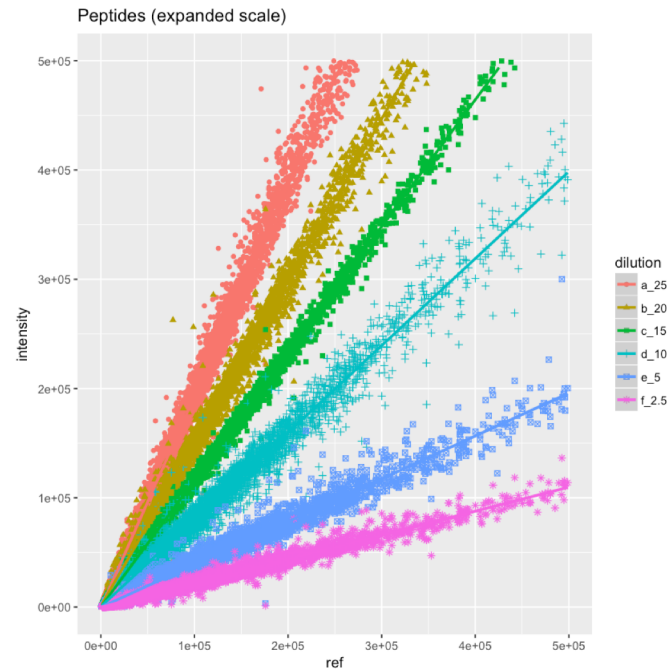

raw PSMs (expanded scale)
intensity
ref
dilution
a_25
b_20
c_15
d_10
e_5
f_2.5
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05

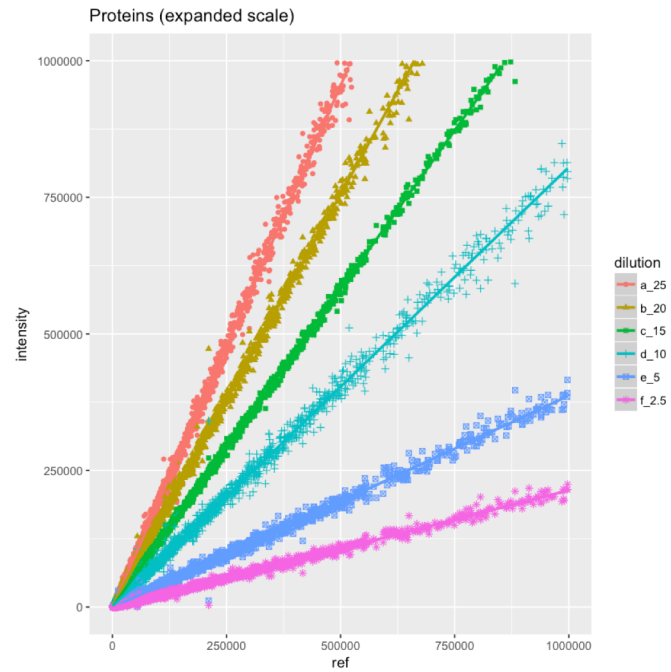

raw PSMs
intensity
ref
dilution
a_25
b_20
c_15
d_10
e_5
f_2.5


Peptides
intensity
ref
dilution
a_25
b_20
c_15
d_10
e_5
f_2.5

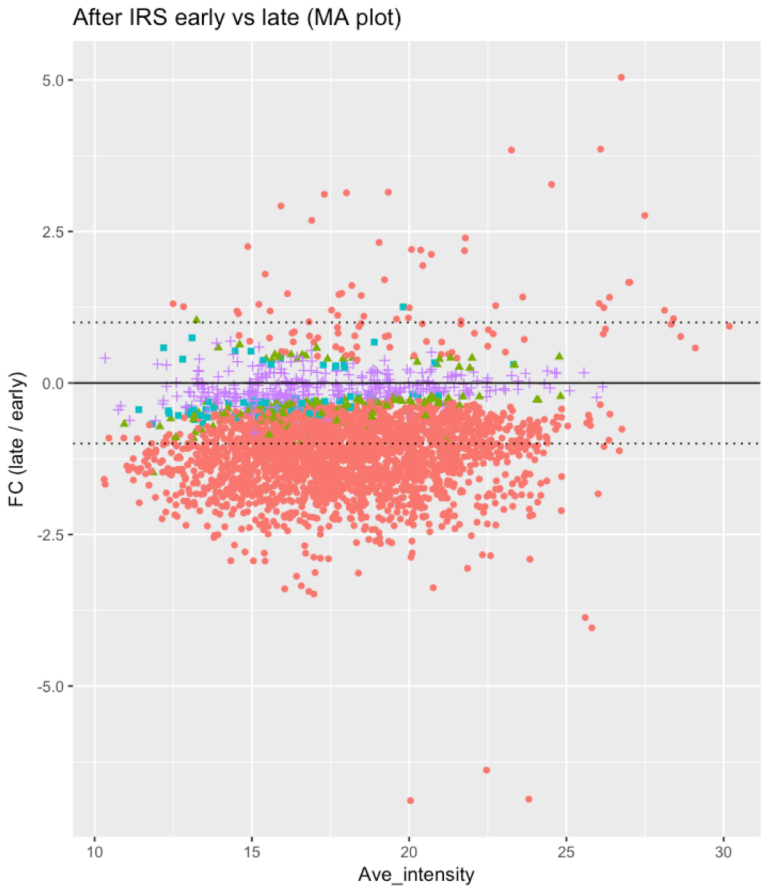

Proteins
intensity
ref
dilution
a_25
b_20
c_15
d_10
e_5
f_2.5


raw PSMs (expanded scale)
intensity
ref
dilution
a_25
b_20
c_15
d_10
e_5
f_2.5

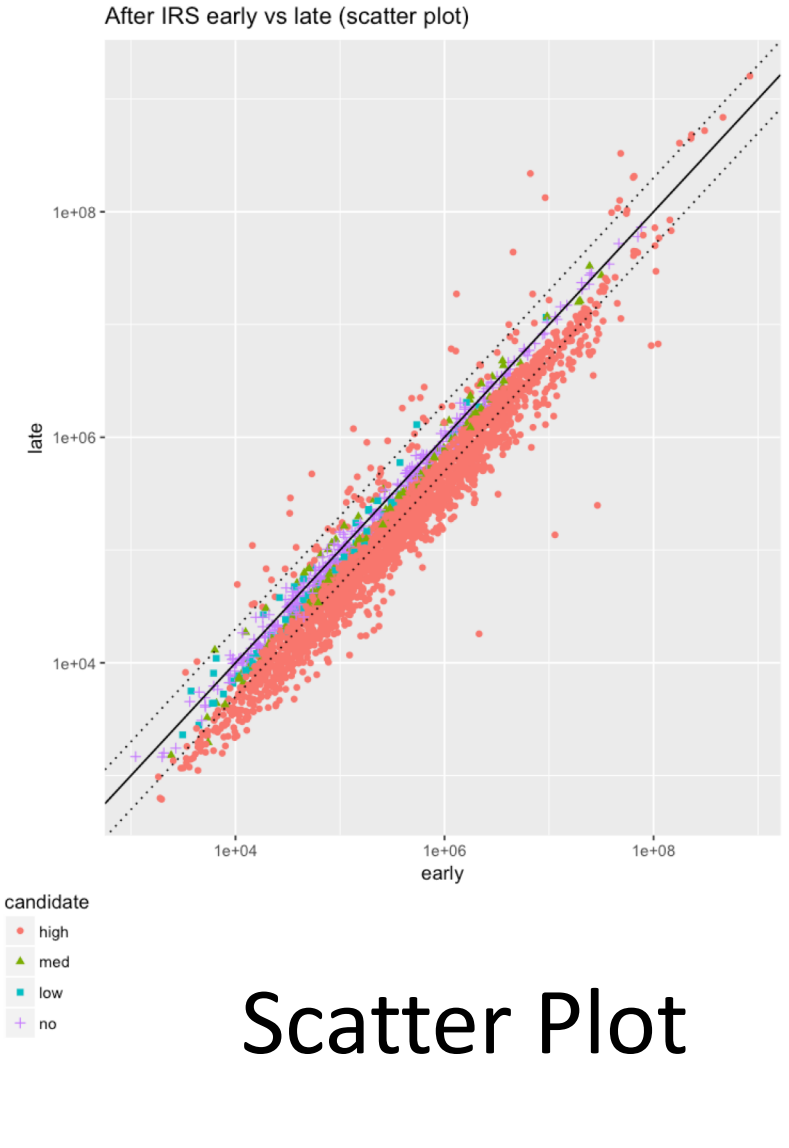

Peptides (expanded scale)
intensity
ref
dilution
a_25
b_20
c_15
d_10
e_5
f_2.5


Proteins (expanded scale)
intensity
ref
dilution
a_25
b_20
c_15
d_10
e_5
f_2.5

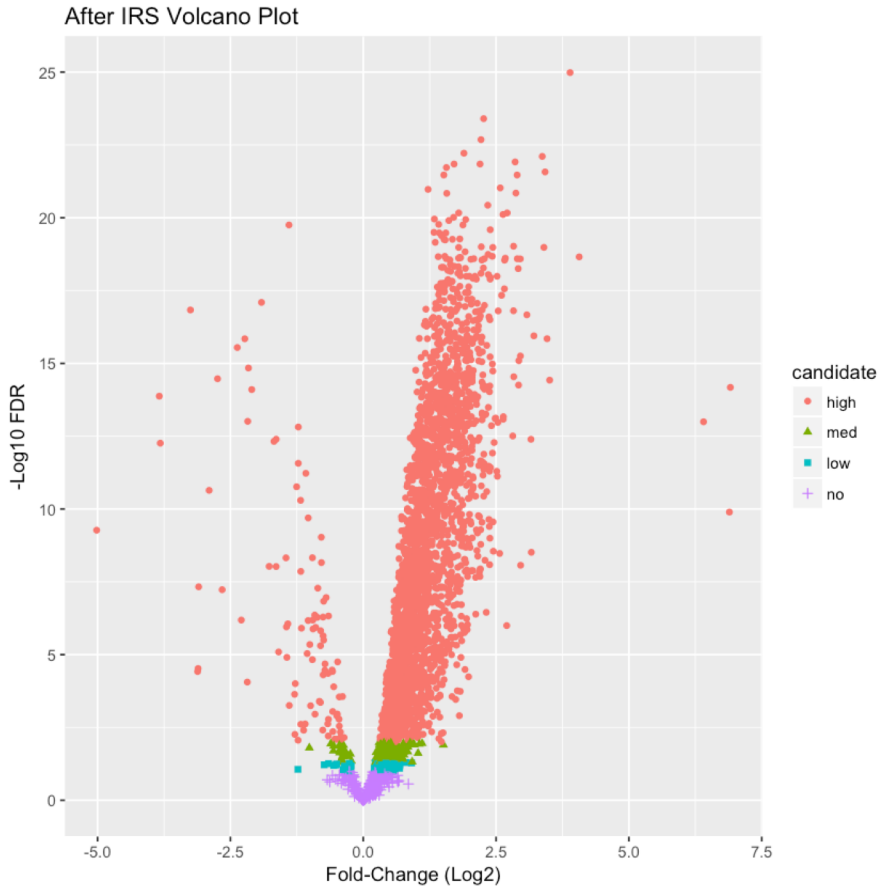
MA Plot



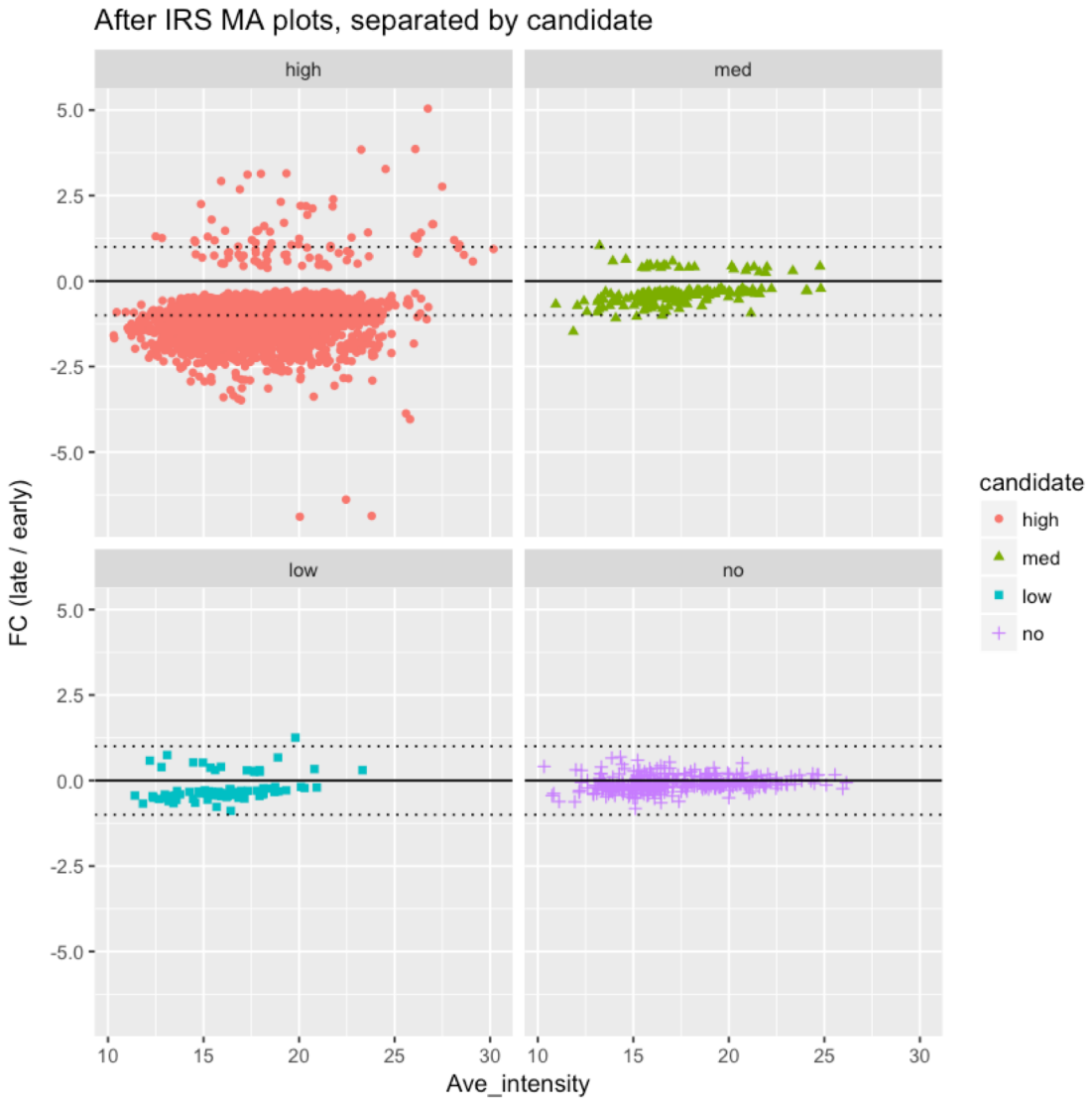
Scatter Plot

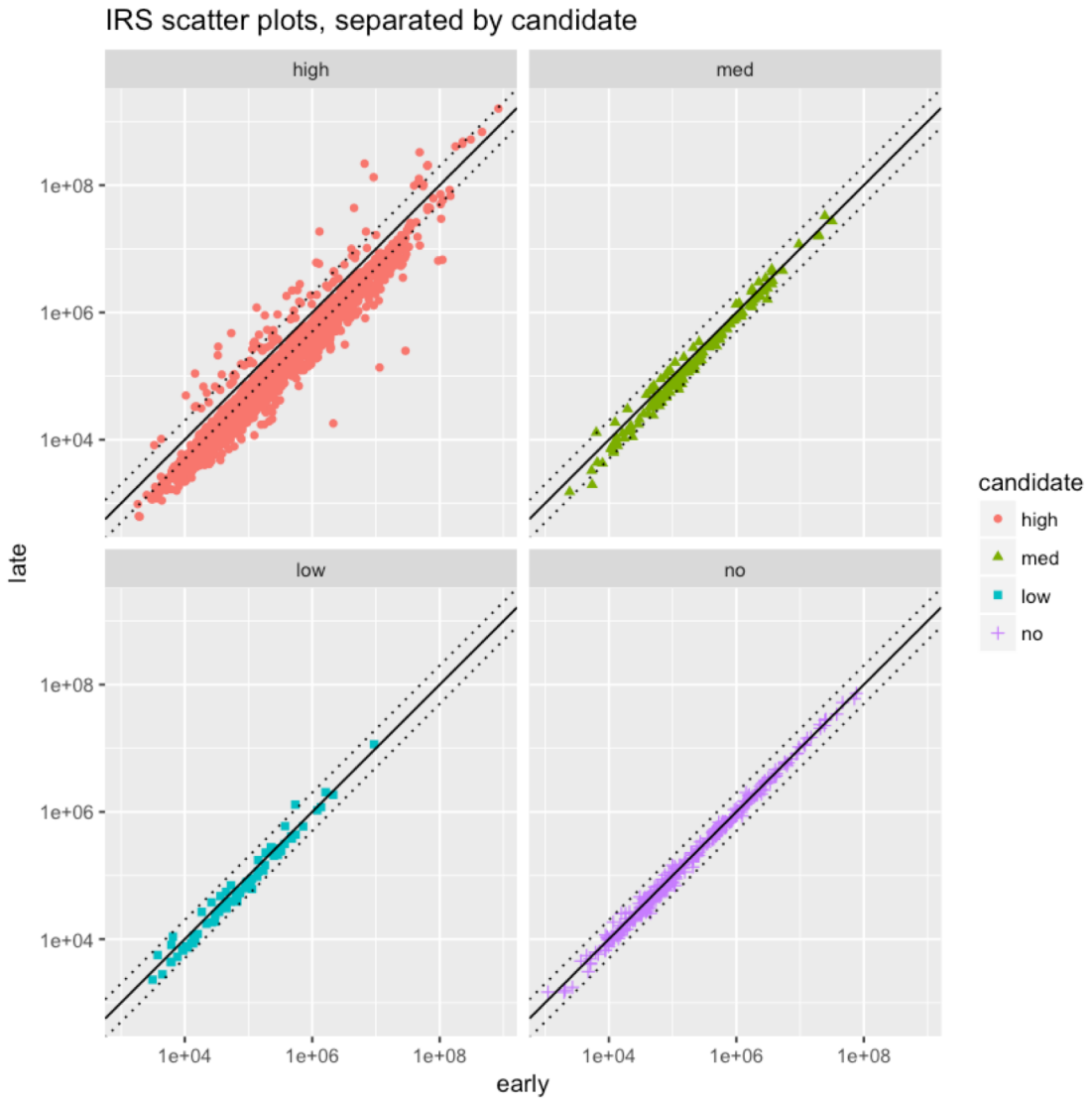


Volcano Plot



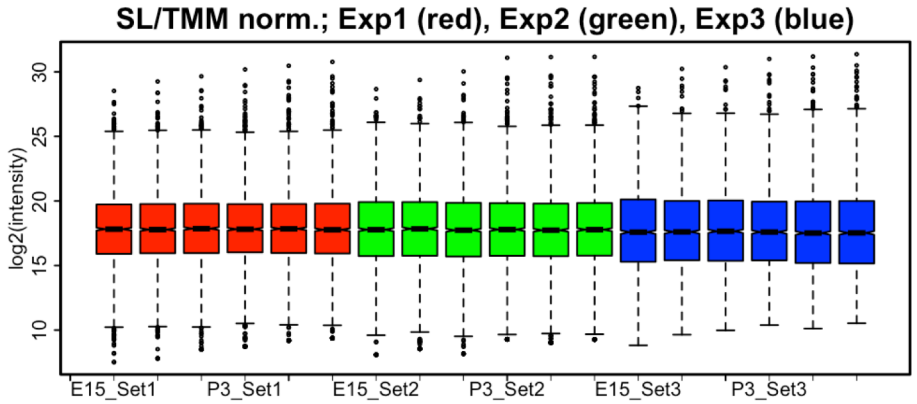
Facet plots – separate candidates by significance

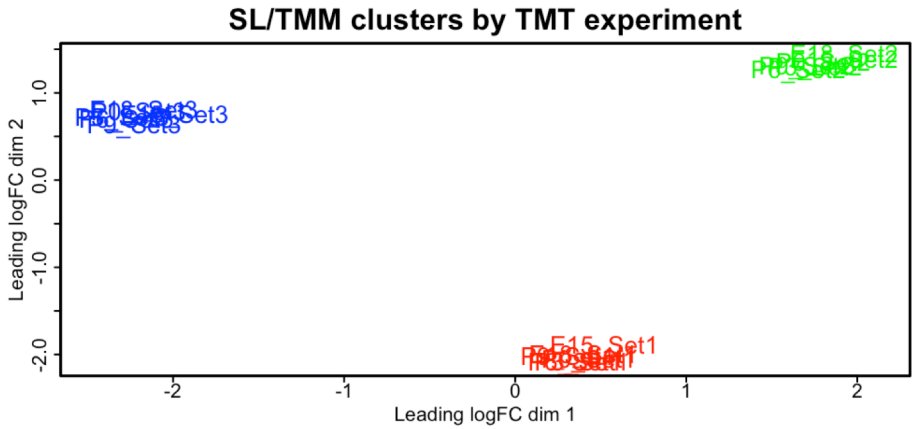


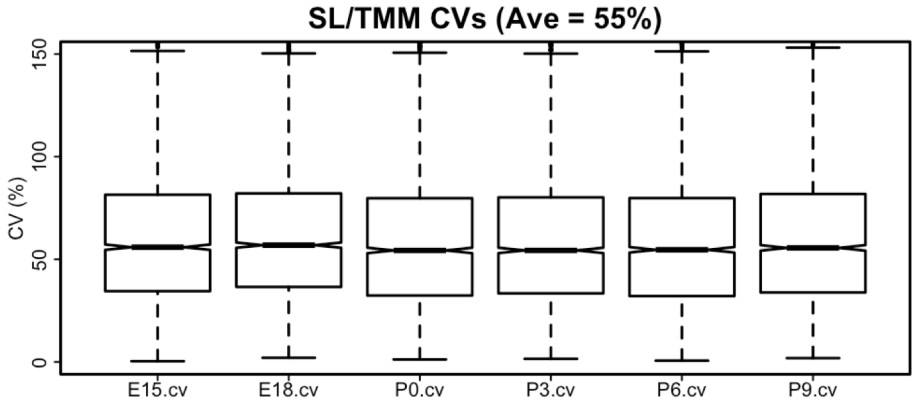


https://github.com/pwilmart/IRS_normalization

Without IRS







Mouse lens: 6 developmental times
biological replicates in separate TMT experiments

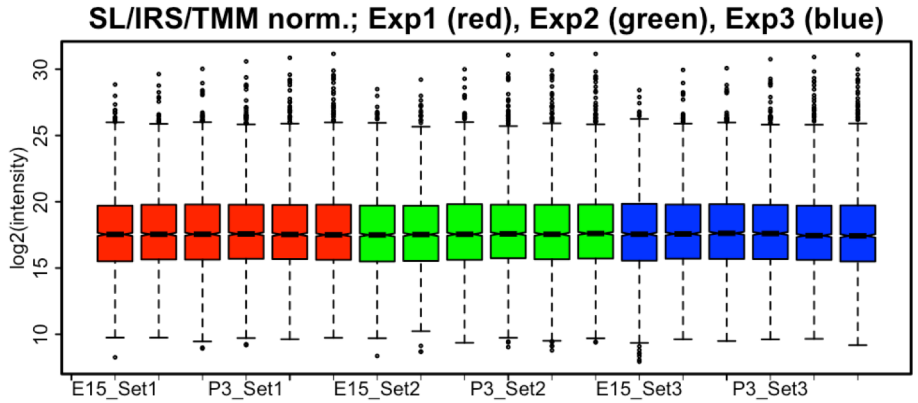
Single factor normalizations seem to work

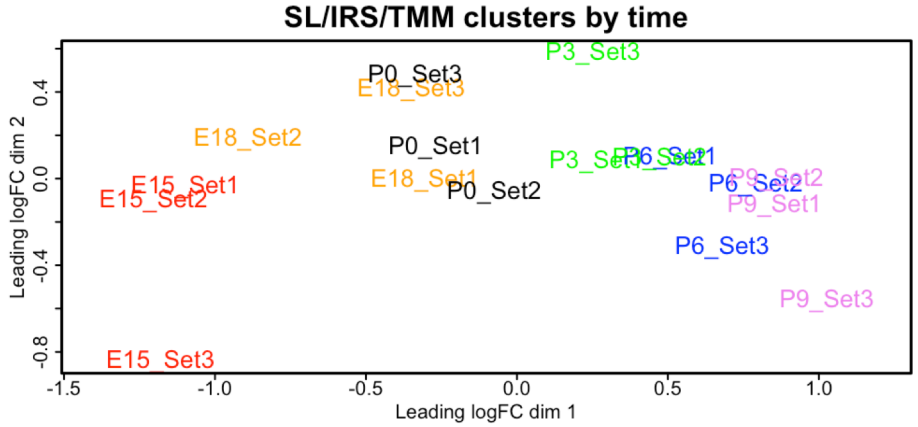
They don’t actually work

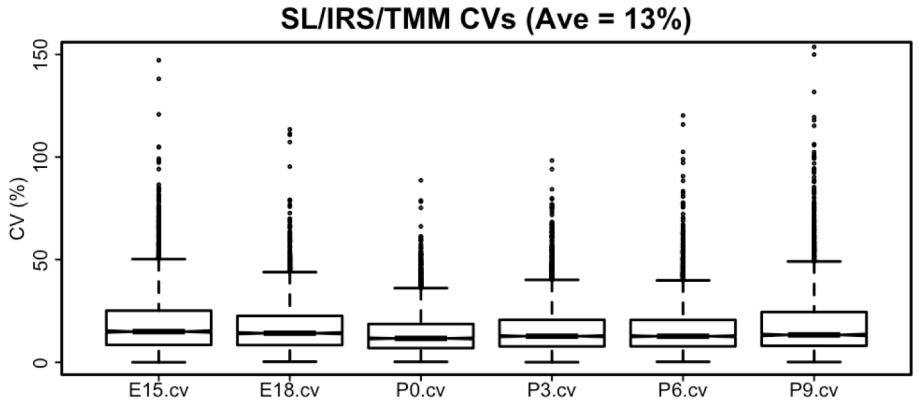
Internal Reference Scaling: corrects random MS2 sampling that affects reporter ion intensities between scans.

Plubell, et al., MCP, 2017, v16, p873-890.

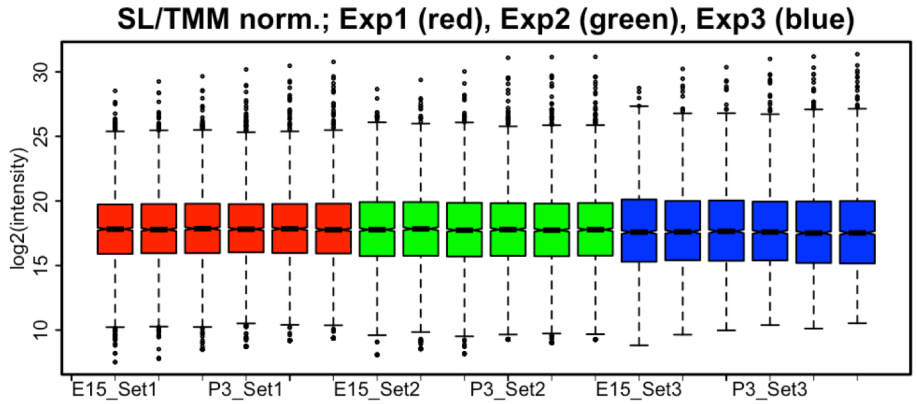
With IRS

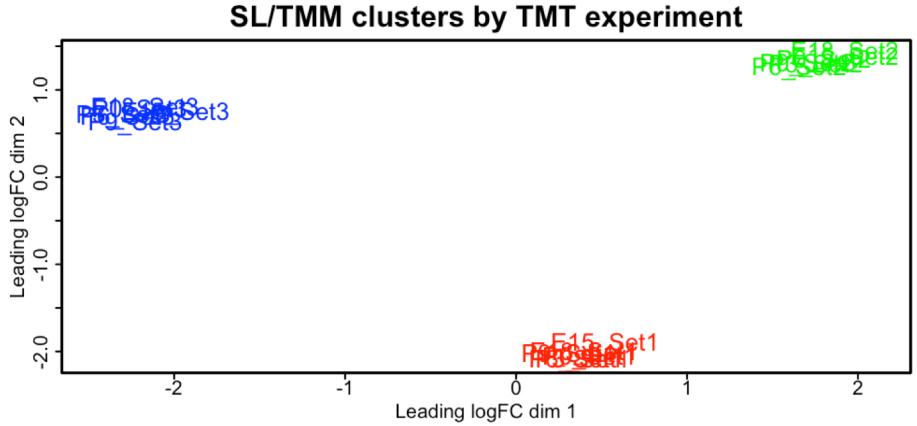


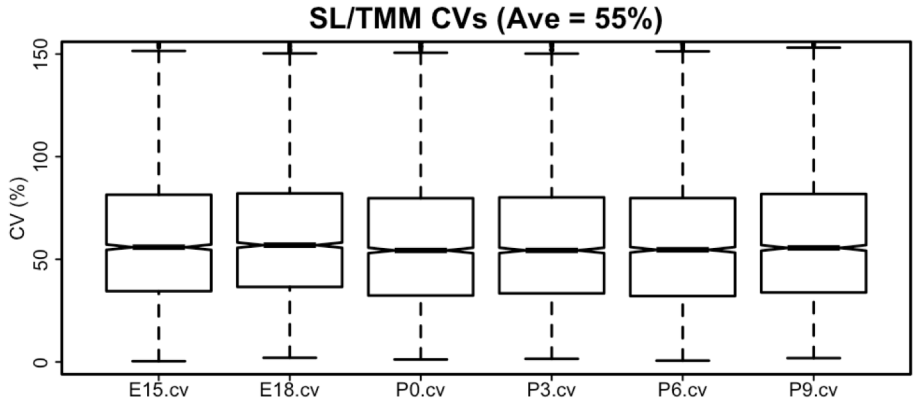




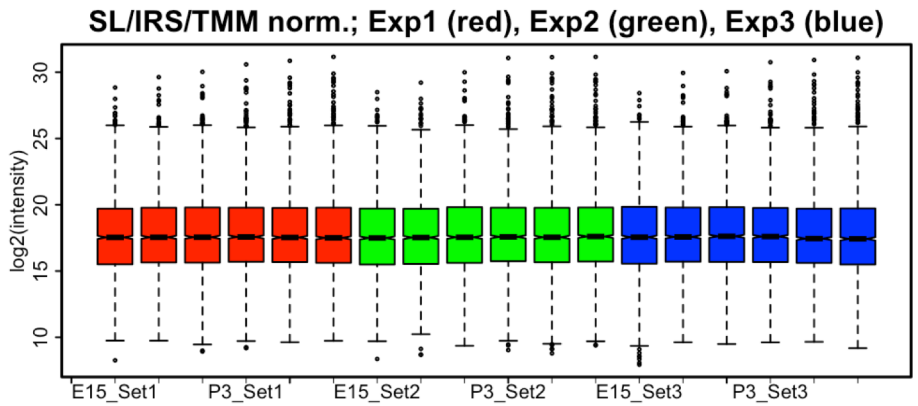
Without IRS

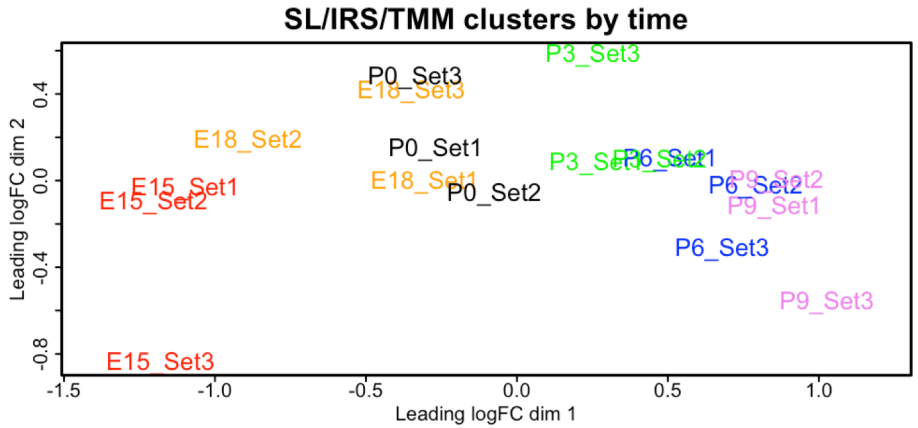


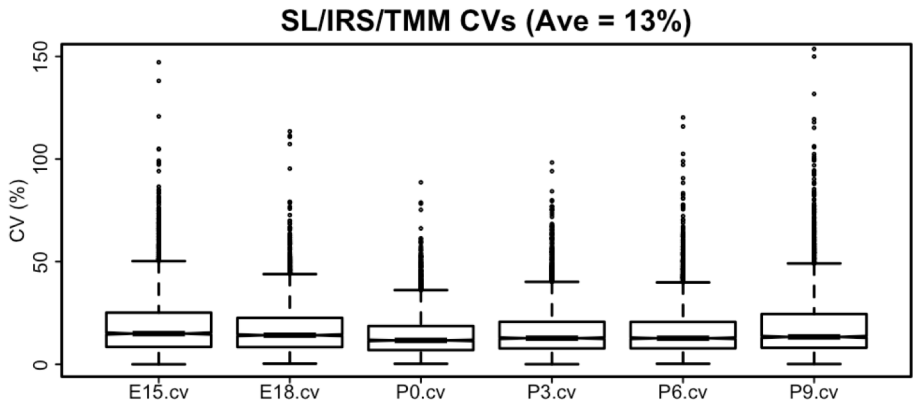




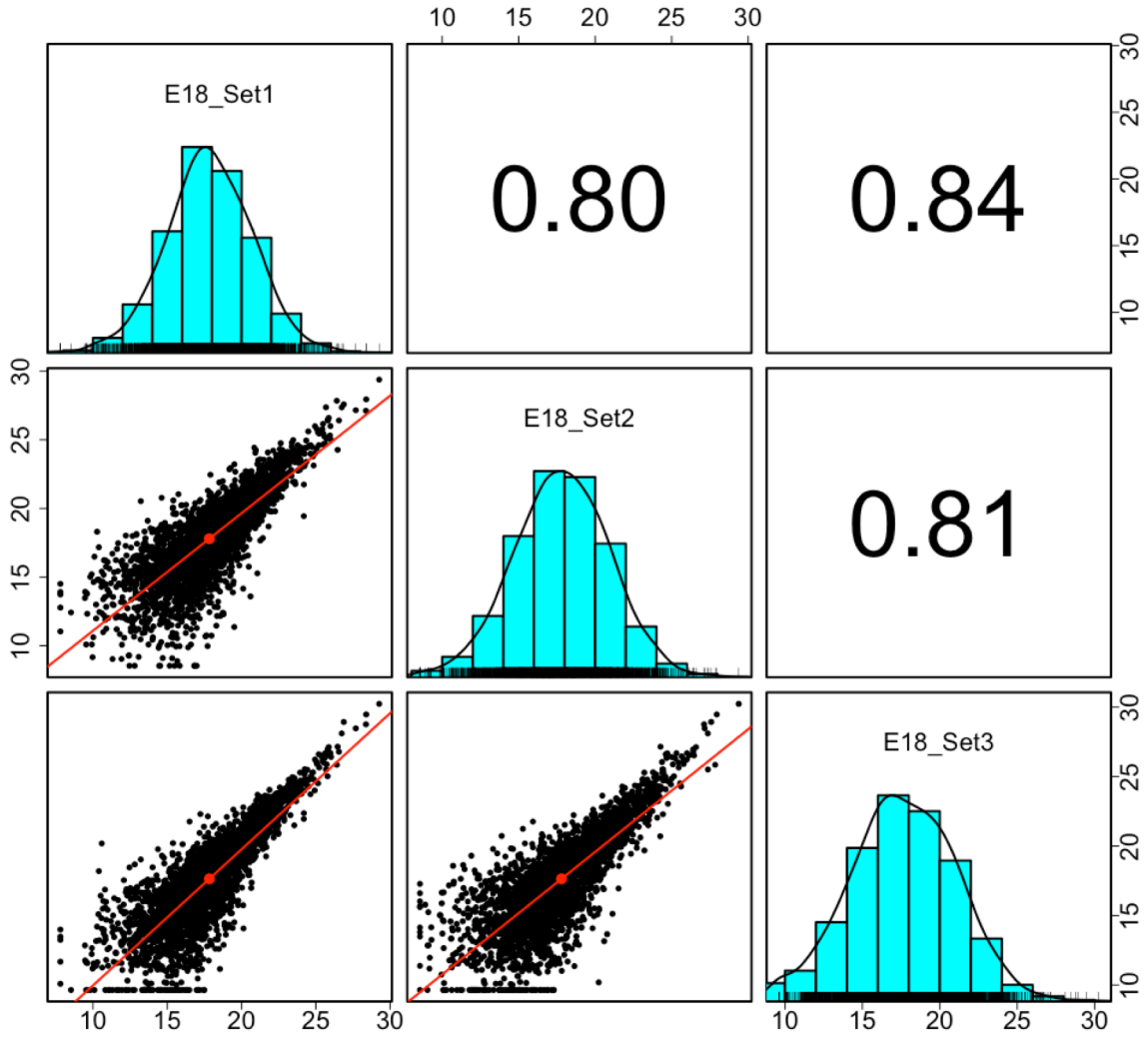
With IRS



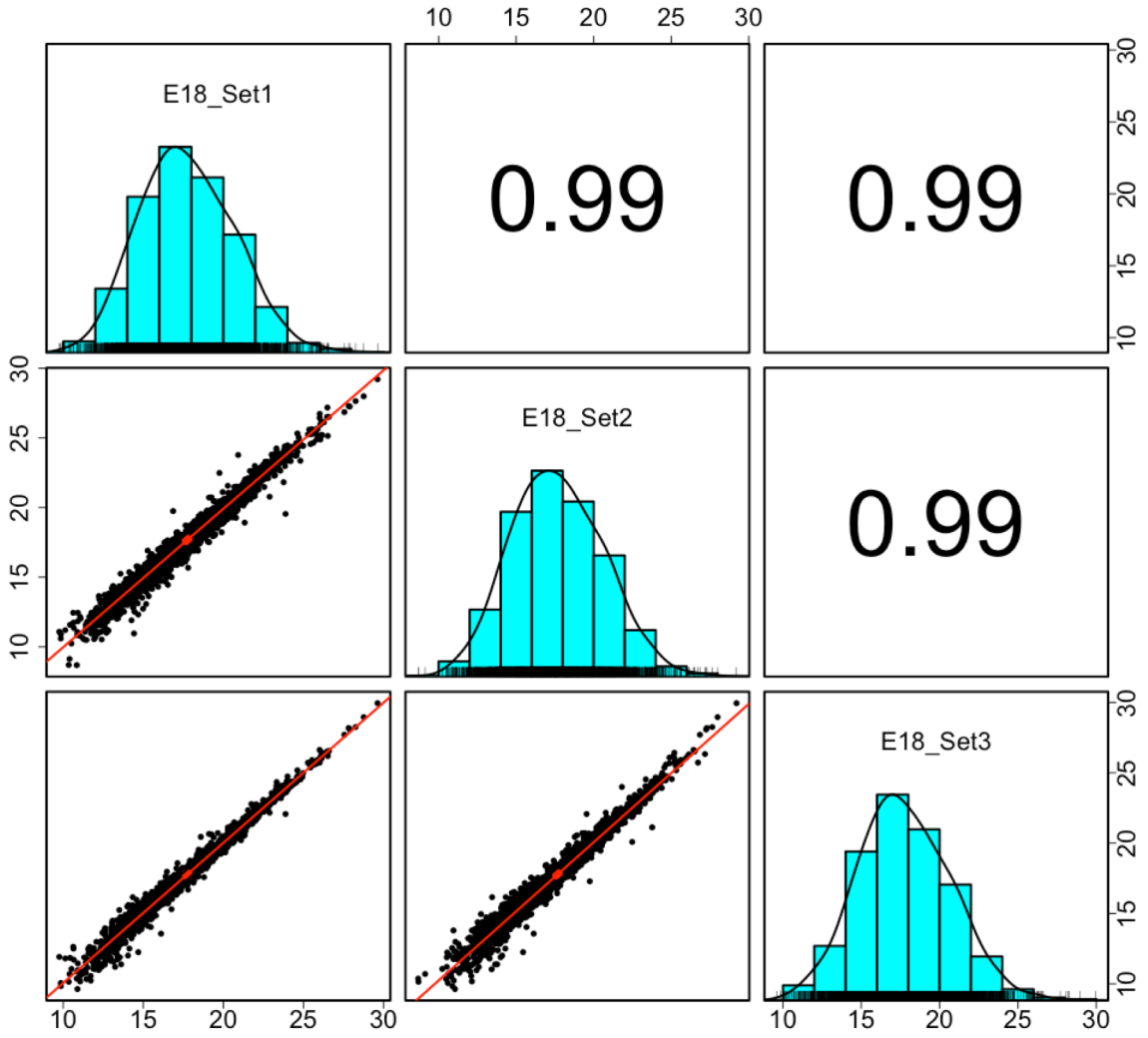




Without IRS



With IRS



https://github.com/pwilmart/ASMS2018

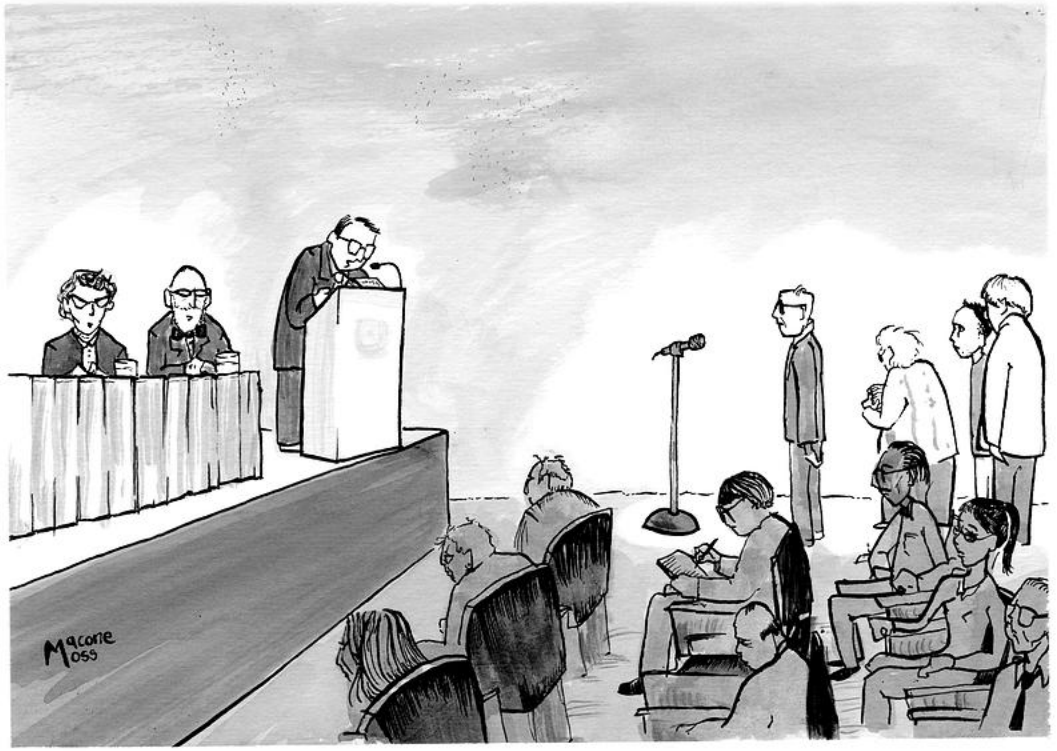
Atlantic article says notebooks can replace traditional scientific papers…

- Notebooks don’t work for all content
- Notebooks make great Supplemental Materials
- Notebooks are logical choices for:
  - QC and sanity checks
  - Normalizations, transformations, correlations
  - Statistical testing and modeling

Disclaimers

- Scientific python and Jupyter notebooks are big topics
- R is confusing and there is a lot to learn
  - Traditional (old) R
  - RStudio and modern R
  - Statistics
  - Bioconductor
- Git/Github also extensive
- “data science” field is interesting: Hadley Wickham, Wes McKinney, Jake VanderPlas (UW)

See: https://github.com/pwilmart



"We'd now like to open the floor to shorter speeches disguised as questions."

https://goo.gl/images/x2PZXP (Steve Macone, New Yorker, October 18th, 2010)