Reproducible and Interpretable Data Management:
Looking Out For Future You (And Other People Too)

Sarah Odell
University of California Davis
January 13, 2018

Reproducibility Is A Spectrum

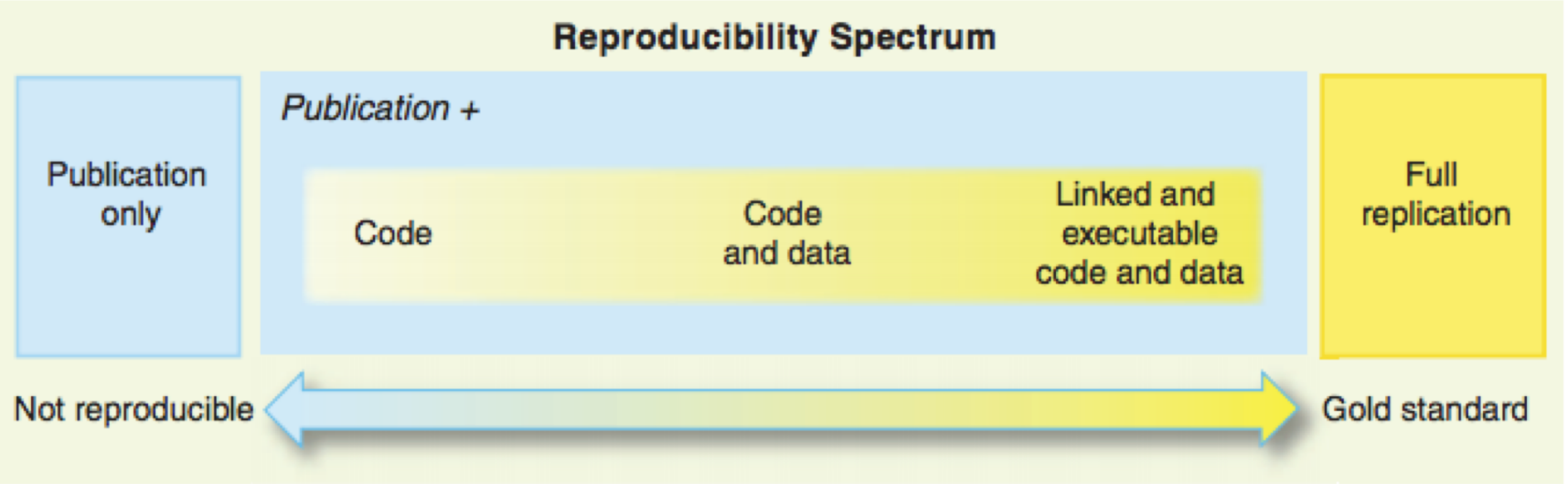
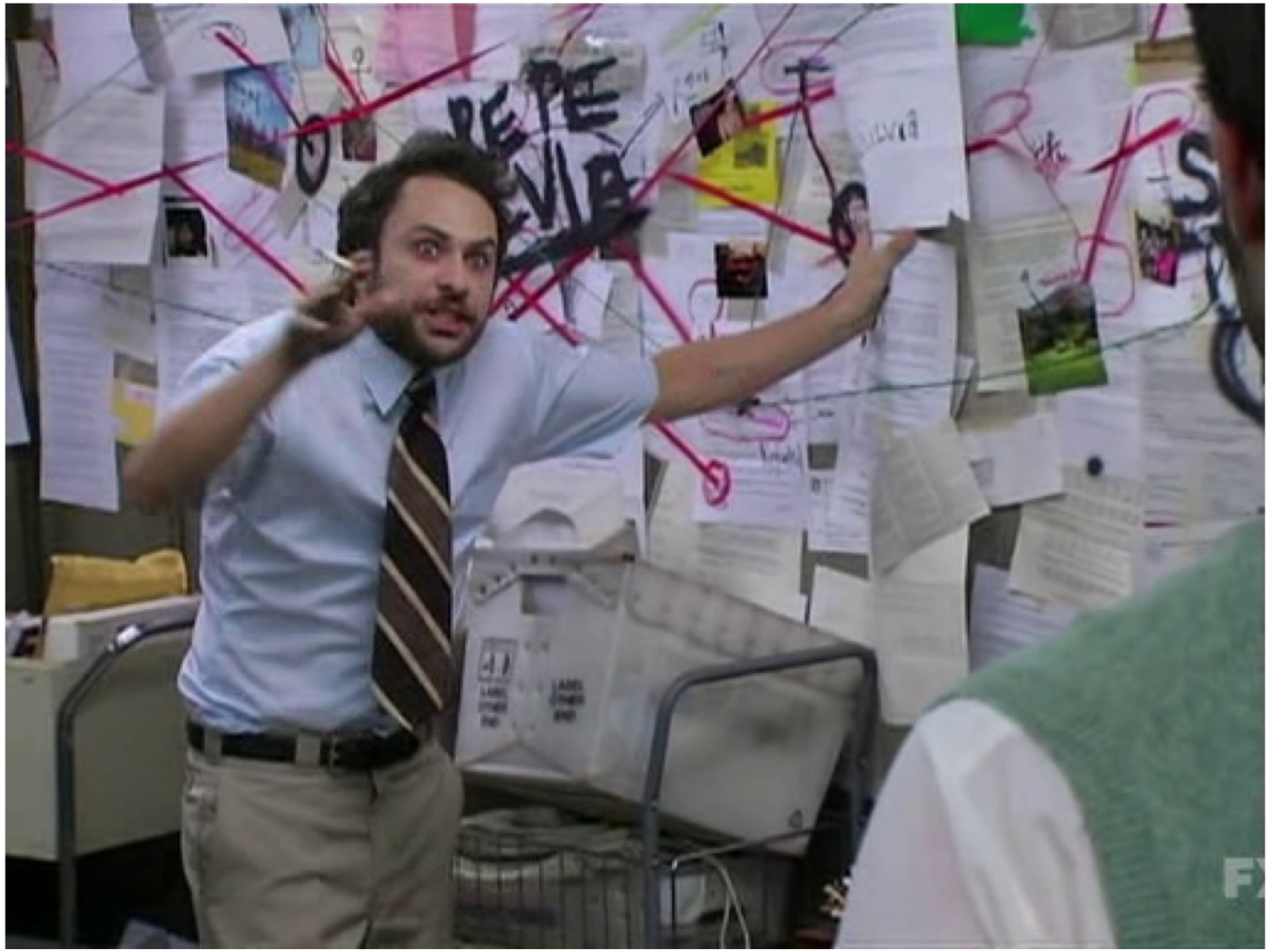


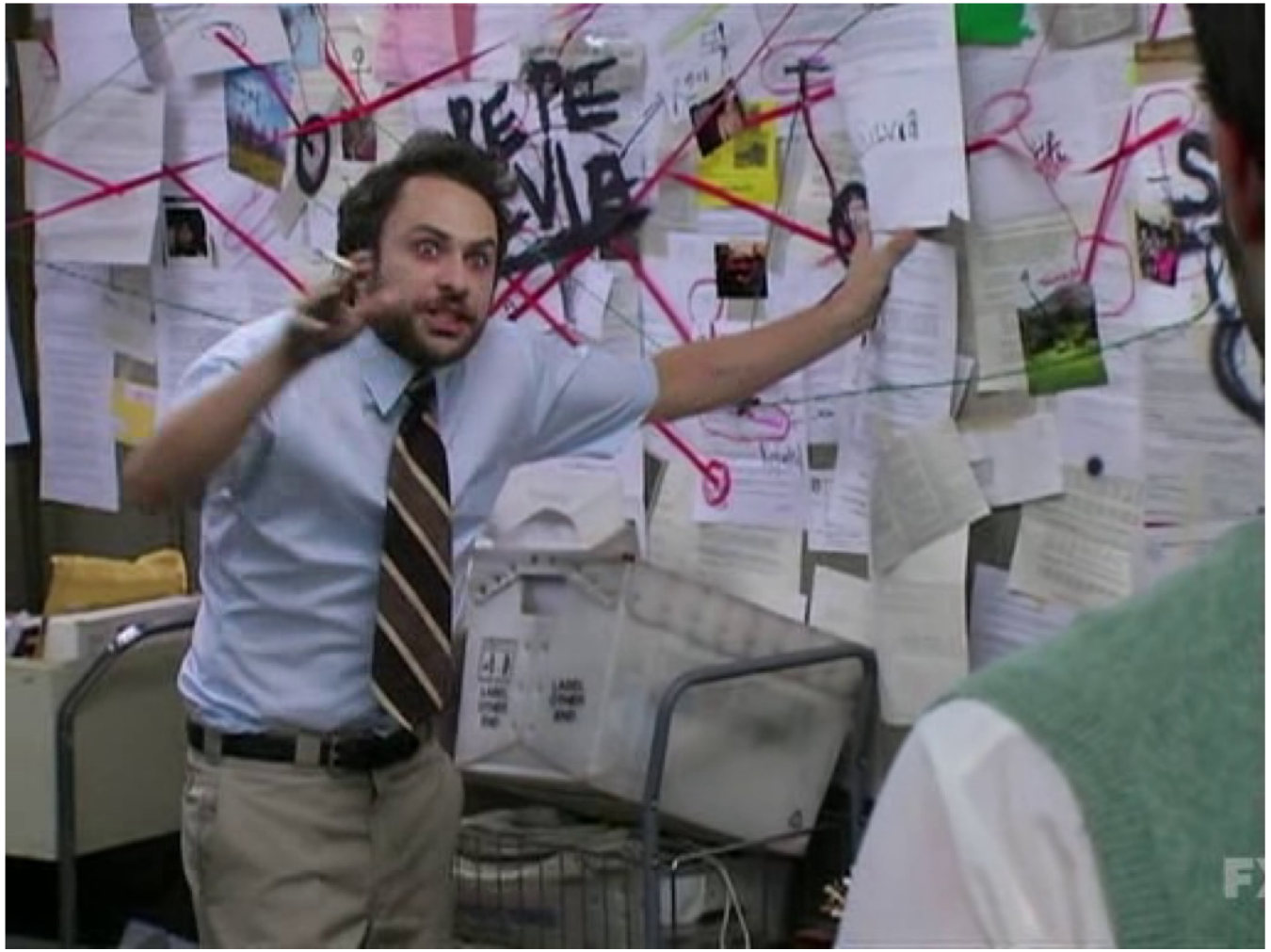
Fig. 1. The spectrum of reproducibility.

Peng, 2011

Trying to explain how you did your analysis to a collaborator in six months



Trying to explain how you did your analysis to **yourself** in six months

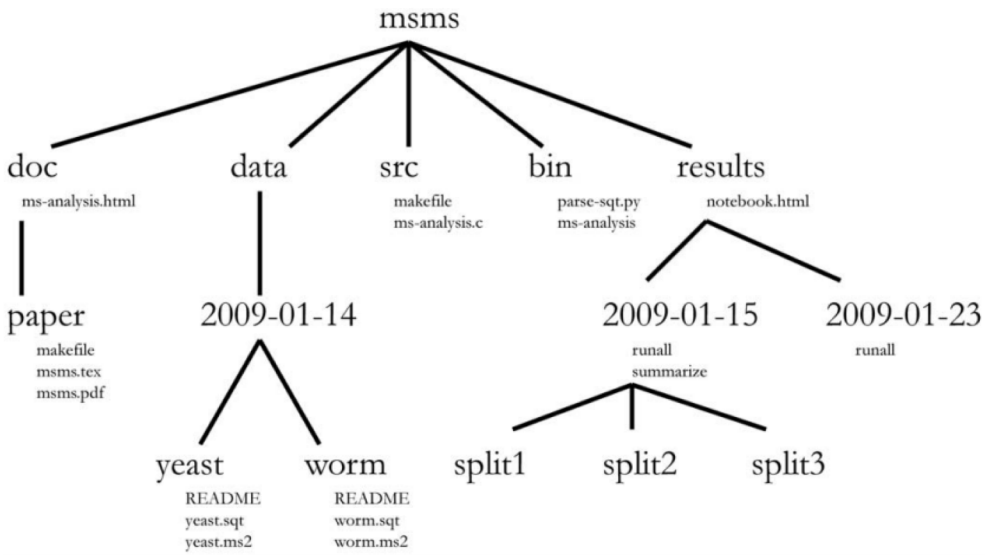


Naming Things

Naming Things

- Descriptive data files and script names
- The final is not the final – use dates

Naming Things



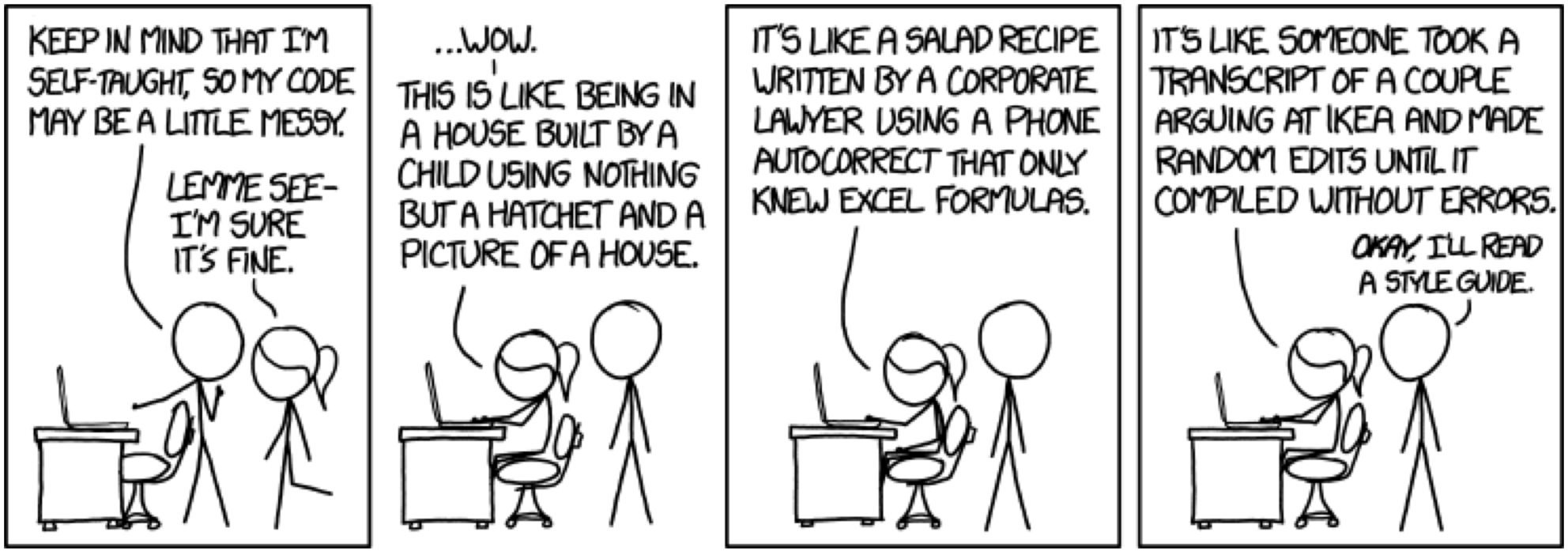
Noble, 2009

- Descriptive data files and script names
- The final is not the final – use dates
- Keep related files in the same directory
- Make a README for that directory

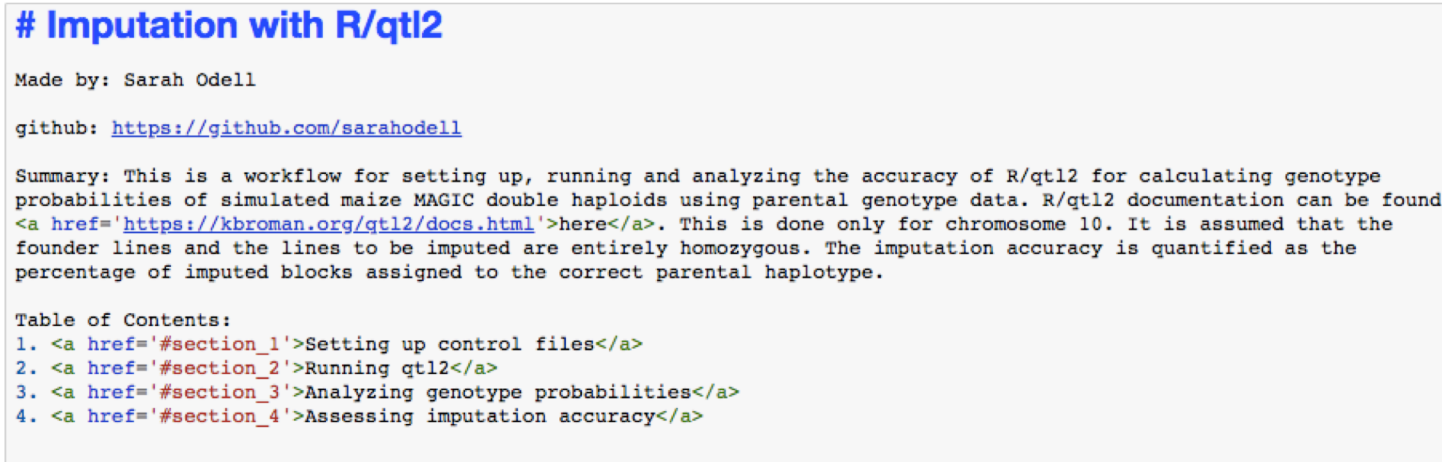
Version Control

Image Source: http://ipengineer.net/2015/04/git-for-network-engineers/

Making Code for Humans



Making Code for Humans



Imputation with R/qtl2

Made by: Sarah Odell

github: https://github.com/sarahodell

Summary: This is a workflow for setting up, running and analyzing the accuracy of R/qtl2 for calculating genotype probabilities of simulated maize MAGIC double haploids using parental genotype data. R/qtl2 documentation can be found here. This is done only for chromosome 10. It is assumed that the founder lines and the lines to be imputed are entirely homozygous. The imputation accuracy is quantified as the percentage of imputed blocks assigned to the correct parental haplotype.

Table of Contents:

1. Setting up control files
2. Running qtl2
3. Analyzing genotype probabilities
4. Assessing imputation accuracy

- Can show Markdown, Python, R and command line
- Can be converted into html, pdf, etc.

Making Code for Humans

```
# Imputation with R/qtl2

Made by: Sarah Odell

github: https://github.com/sarahodell

Summary: This is a workflow for setting up, running and analyzing the accuracy of R/qtl2 for calculating genotype
probabilities of simulated maize MAGIC double haploids using parental genotype data. R/qtl2 documentation can be found
<a href='https://kbroman.org/qtl2/docs.html'>here</a>. This is done only for chromosome 10. It is assumed that the
founder lines and the lines to be imputed are entirely homozygous. The imputation accuracy is quantified as the
percentage of imputed blocks assigned to the correct parental haplotype.

Table of Contents:
1. <a href='#section_1'>Setting up control files</a>
2. <a href='#section_2'>Running qtl2</a>
3. <a href='#section_3'>Analyzing genotype probabilities</a>
4. <a href='#section_4'>Assessing imputation accuracy</a>
```

Imputation with R/qtl2

Made by: Sarah Odell

github: https://github.com/sarahodell

Summary: This is a workflow for setting up, running and analyzing the accuracy of R/qtl2 for calculating genotype probabilities of simulated maize MAGIC double haploids using parental genotype data. R/qtl2 documentation can be found here. This is done only for chromosome 10. It is assumed that the founder lines and the lines to be imputed are entirely homozygous. The imputation accuracy is quantified as the percentage of imputed blocks assigned to the correct parental haplotype.

Table of Contents:

1. Setting up control files
2. Running qtl2
3. Analyzing genotype probabilities
4. Assessing imputation accuracy

- Can show Markdown, Python, R and command line
- Can be converted into html, pdf, etc.

Data and Software Versions

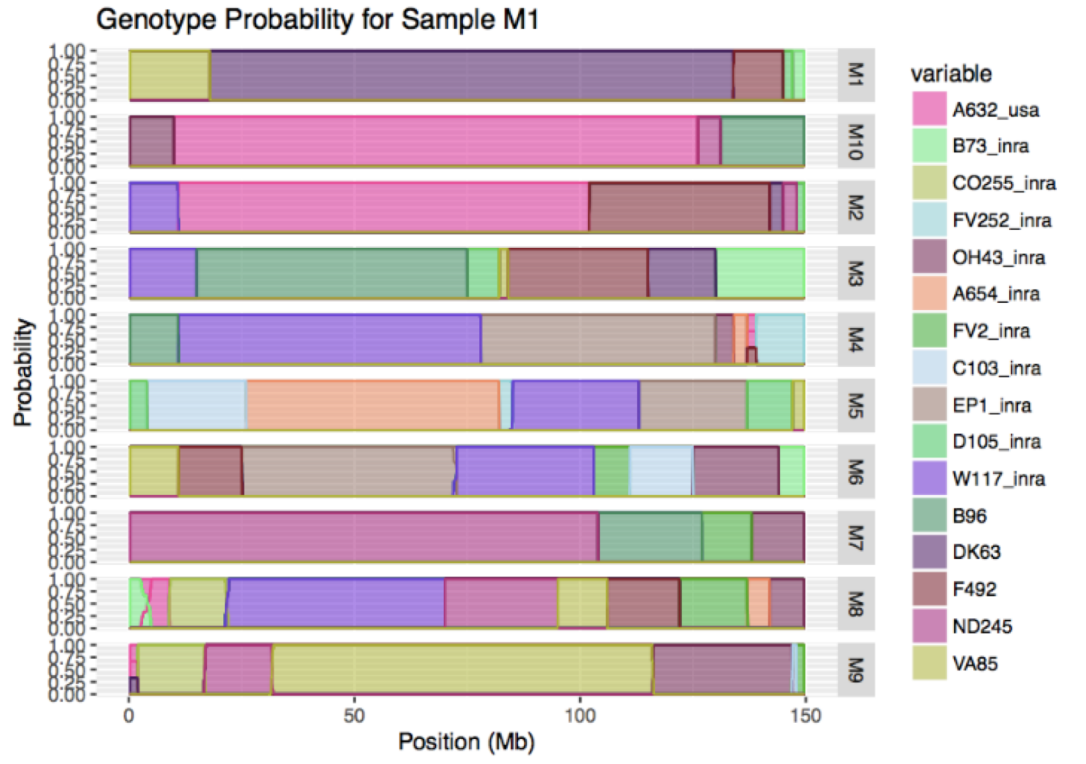
In [2]:
```
print(format(Sys.time(), "Last updated: %m/%d/%Y"))
print(sprintf("Created using: %s",R.Version()$version.string))
```

```
[1] "Last updated: 01/09/2019"
[1] "Created using: R version 3.4.2 (2017-09-28)"
```

Recreating Figures

- Provide code to recreate figures

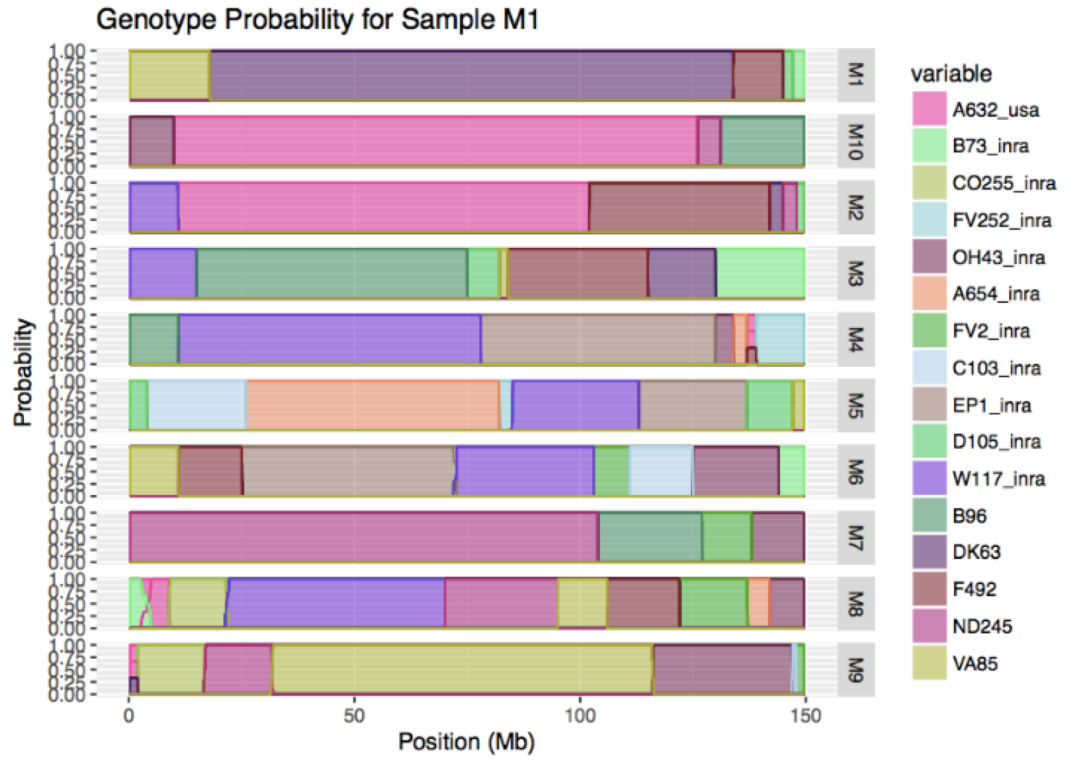
```
In [ ]: ggplot(data=all_pr,aes(x=pos,y=value,color=variable)) + facet_grid(sample ~ .) +
scale_color_manual(values=hex_colors) + scale_fill_manual(values=hex_colors)+
geom_area(aes(fill=variable),alpha=5/10) + geom_line() +
ggtitle("Genotype Probability for Sample M1") + xlab("Position (Mb)") + ylab("Probability") +
guides(color=FALSE)
```



Recreating Figures

- Provide code to recreate figures
- Imbed an image of the original for comparison

```
In [ ]: ggplot(data=all_pr,aes(x=pos,y=value,color=variable)) + facet_grid(sample ~ .) +
scale_color_manual(values=hex_colors) + scale_fill_manual(values=hex_colors)+
geom_area(aes(fill=variable),alpha=5/10) + geom_line() +
ggtitle("Genotype Probability for Sample M1") + xlab("Position (Mb)") + ylab("Probability") +
guides(color=FALSE)
```



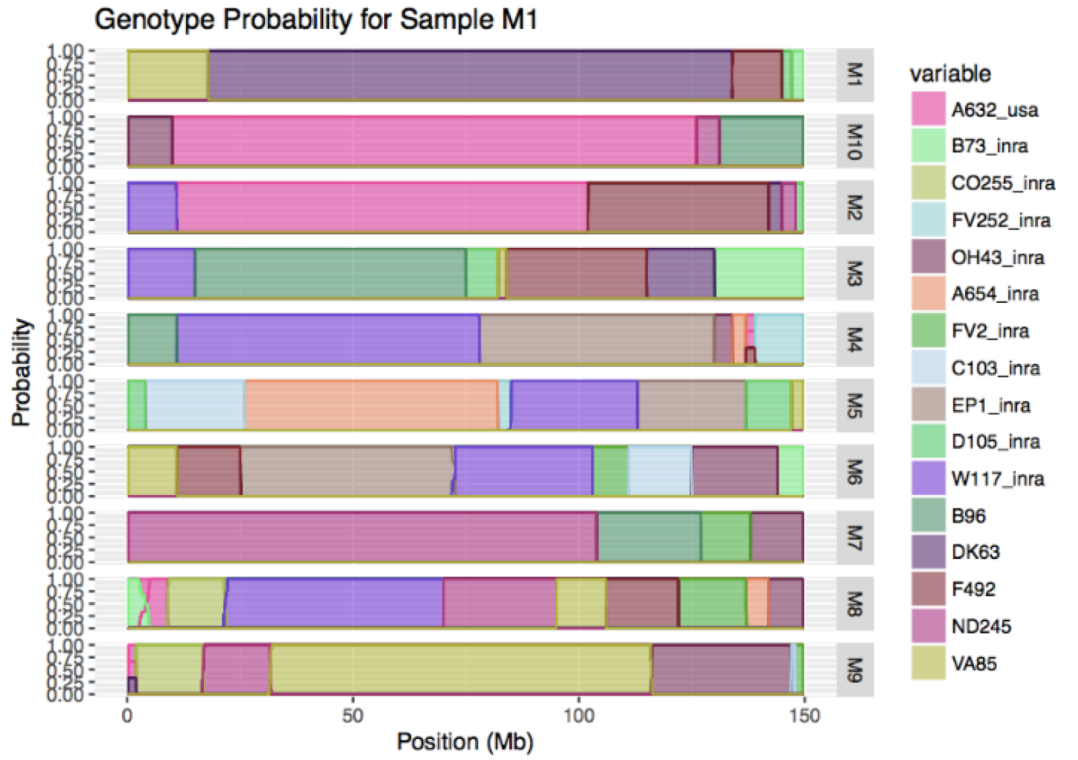

```
And here is a plot we created originally:

Original
![image](../images/qtl2_image.png)
```

In [ ]:

And here is a plot we created originally:

Original



Paths and Variables

In [5]:
```
# Set wkdir to whatever your working directory is
wkdir='~/Documents/PBGG/MAGIC/impute'
setwd(wkdir)

data=fread(sprintf('%s/DH10_Chr10_IBD_Regions.txt',wkdir),data.table=F)
head(data)
```

chrom	start	end	A632_usa	B73_inra	CO255_inra	FV252_inra	OH43_inra	A654_inra	FV2_inra	···
10	92823	94323	1	2	2	3	4	5	6	···
10	94323	123371	1	2	2	3	4	5	6	···
10	123371	125607	1	2	2	3	4	3	5	···
10	125607	545649	1	2	2	3	4	5	6	···
10	545649	598470	1	2	2	3	4	3	5	···
10	598470	601011	1	2	2	3	4	5	6	···

Noble, 2009

Paths and Variables

In [5]:
```
# Set wkdir to whatever your working directory is
wkdir='~/Documents/PBGG/MAGIC/impute'
setwd(wkdir)

data=fread(sprintf('%s/DH10_Chr10_IBD_Regions.txt',wkdir),data.table=F)
head(data)
```

chrom	start	end	A632_usa	B73_inra	CO255_inra	FV252_inra	OH43_inra	A654_inra	FV2_inra	···
10	92823	94323	1	2	2	3	4	5	6	···
10	94323	123371	1	2	2	3	4	5	6	···
10	123371	125607	1	2	2	3	4	3	5	···
10	125607	545649	1	2	2	3	4	5	6	···
10	545649	598470	1	2	2	3	4	3	5	···
10	598470	601011	1	2	2	3	4	5	6	···

Noble, 2009

Argument Descriptions for Reusable Scripts

```
In [13]: print(system(sprintf('python foundergeno.py -h'),intern=T))
```

```
 [1] "usage: foundergeno.py [-h] infile outfile founders"
 [2] ""
 [3] "Program description: Takes a vcf file and converts it to a csv format with"
 [4] "founder as rows and marker genotypes as columns with nucleotide information"
 [5] "encoded as A for reference allele and B for alternate alleles. This csv file"
 [6] "is formatted for use with R/qtl2. It requires that bcftools be installed."
 [7] ""
 [8] "positional arguments:"
 [9] "  infile      The input vcf file"
[10] "  outfile     The output csv filename"
[11] "  founders    File with a list of founders in the order they are listed in the"
[12] "              vcf file"
[13] ""
[14] "optional arguments:"
[15] "  -h, --help  show this help message and exit"
```

Building Workflows

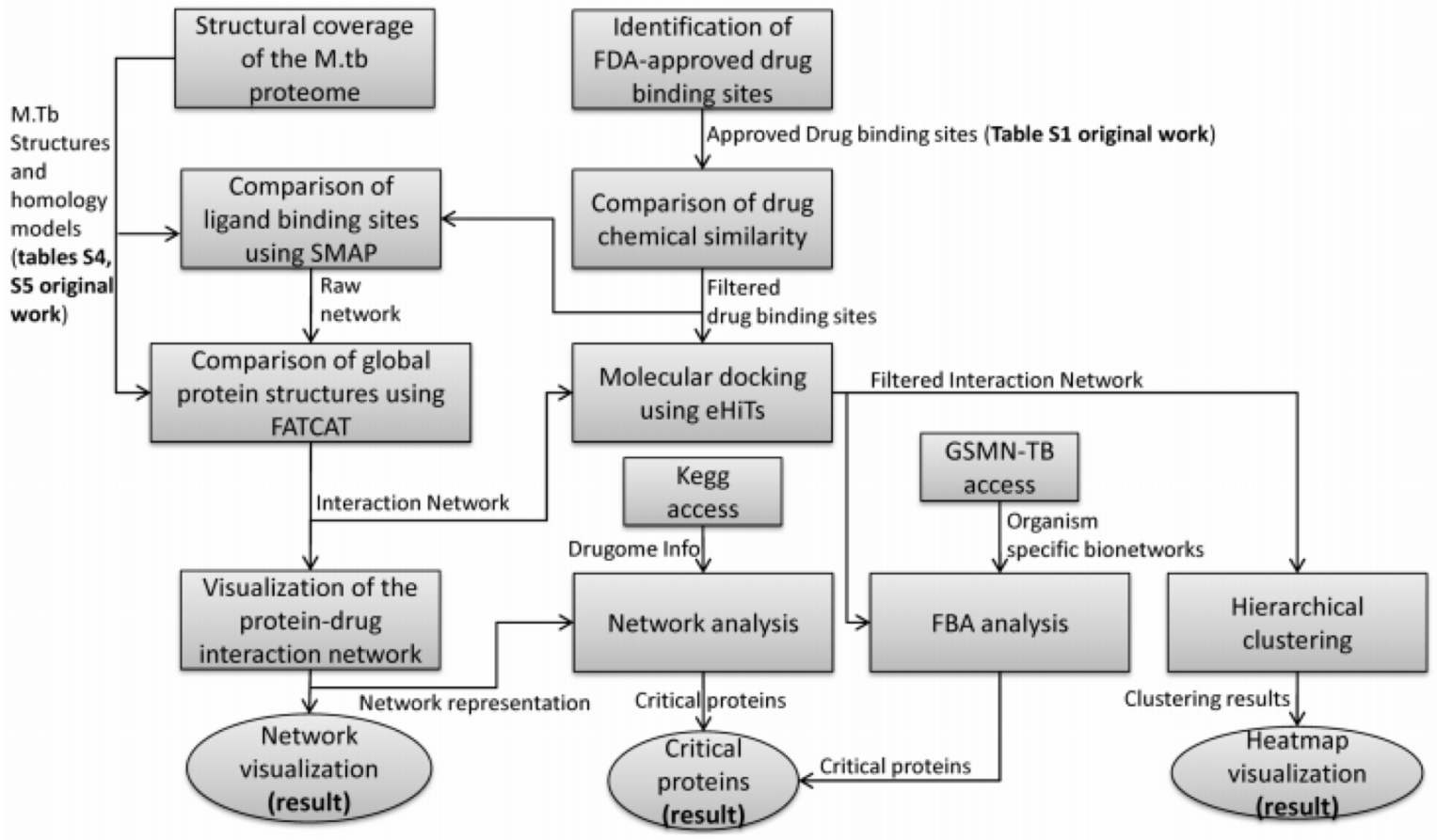


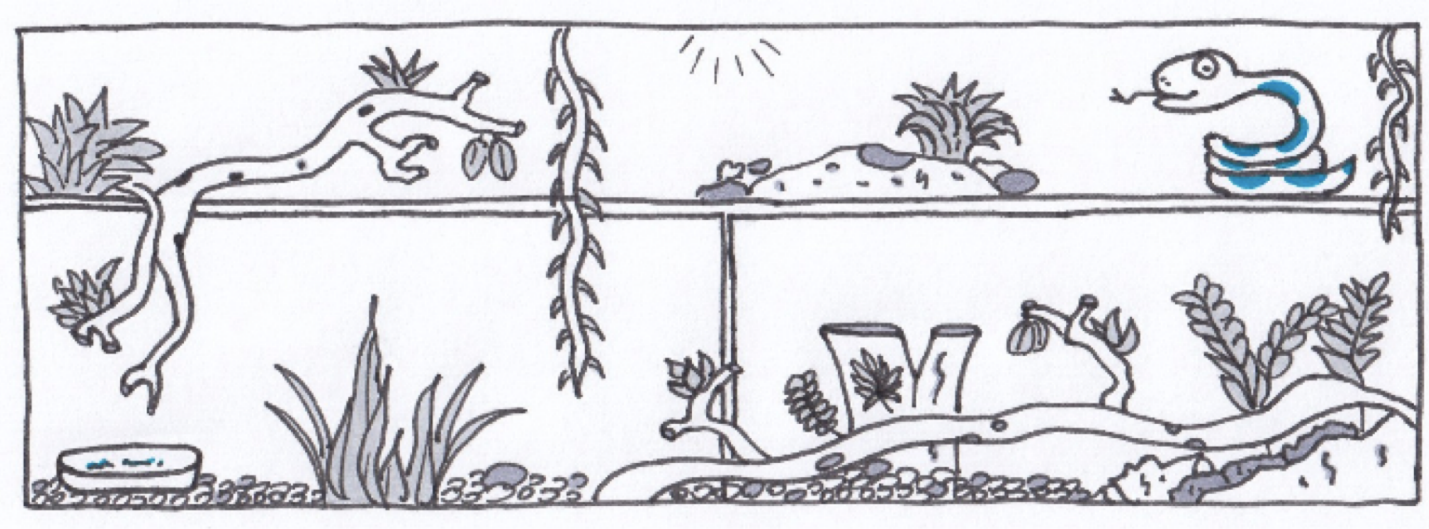
Figure 1. A high-level dataflow diagram of the TB drugome method.
doi:10.1371/journal.pone.0080278.g001

Garifjo et al., 2013

Python Tools for Workflow Building

Conda

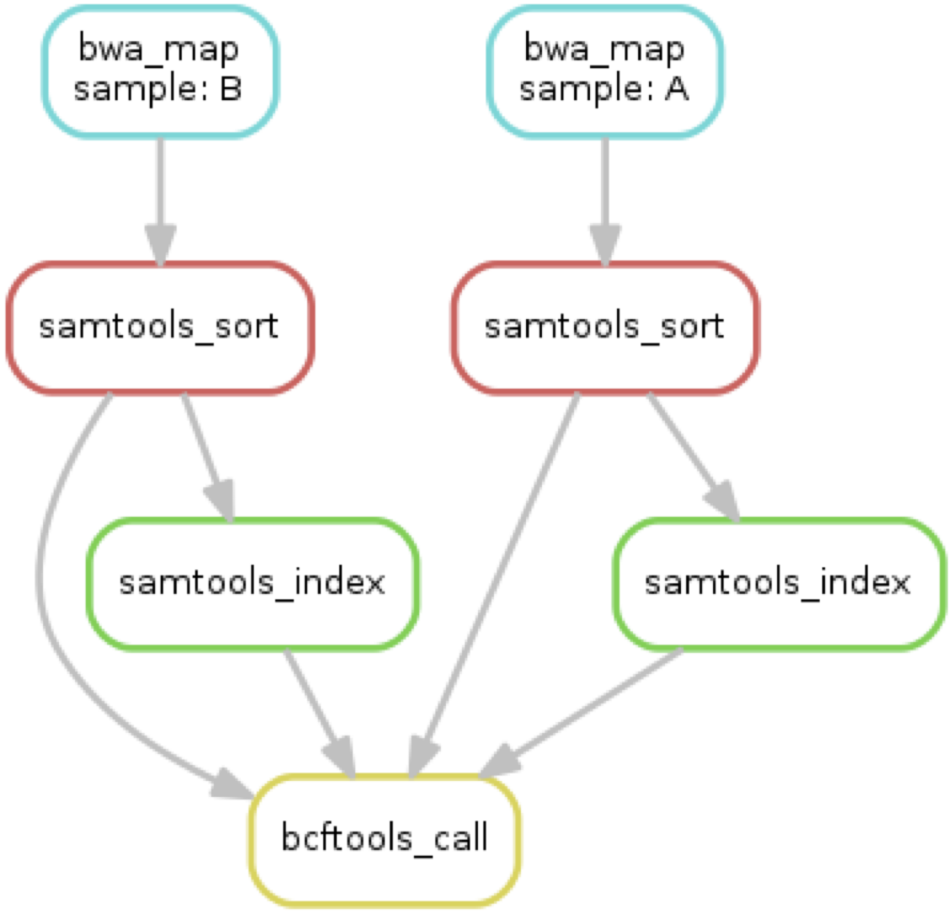
- Allows you to create **environments** with required software tools and libraries
- Can specific specific versions of software to prevent updates from breaking your code



This is how a perfect Python environment looks like ;)

Image Source: https://medium.freecodecamp.org/why-you-need-python-environments-and-how-to-manage-them-with-conda-85f155f4353c

Python Tools for Workflow Building



Snakemake –

- Workflow management system
- Useful for automating pipelines
- Can specify Conda environments
- You can create workflow diagrams of your pipeline!

Key Points

Resources

Markdown: https://www.markdowntutorial.com/

Jupyter Notebooks: https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/

Intro to Git: https://nceas.github.io/oss-lessons/version-control/1-git-basics.html

Snakemake Tutorial: https://snakemake.readthedocs.io

Conda Environments: https://conda.io/docs/user-guide/tasks/manage-environments.html

References

Broman & Woo, 2018

Noble, 2009

Peng, 2011

Stodden et al., 2016

Garijo et al., 2013