Getting Started in the Flaherty Lab

Tete Zhang

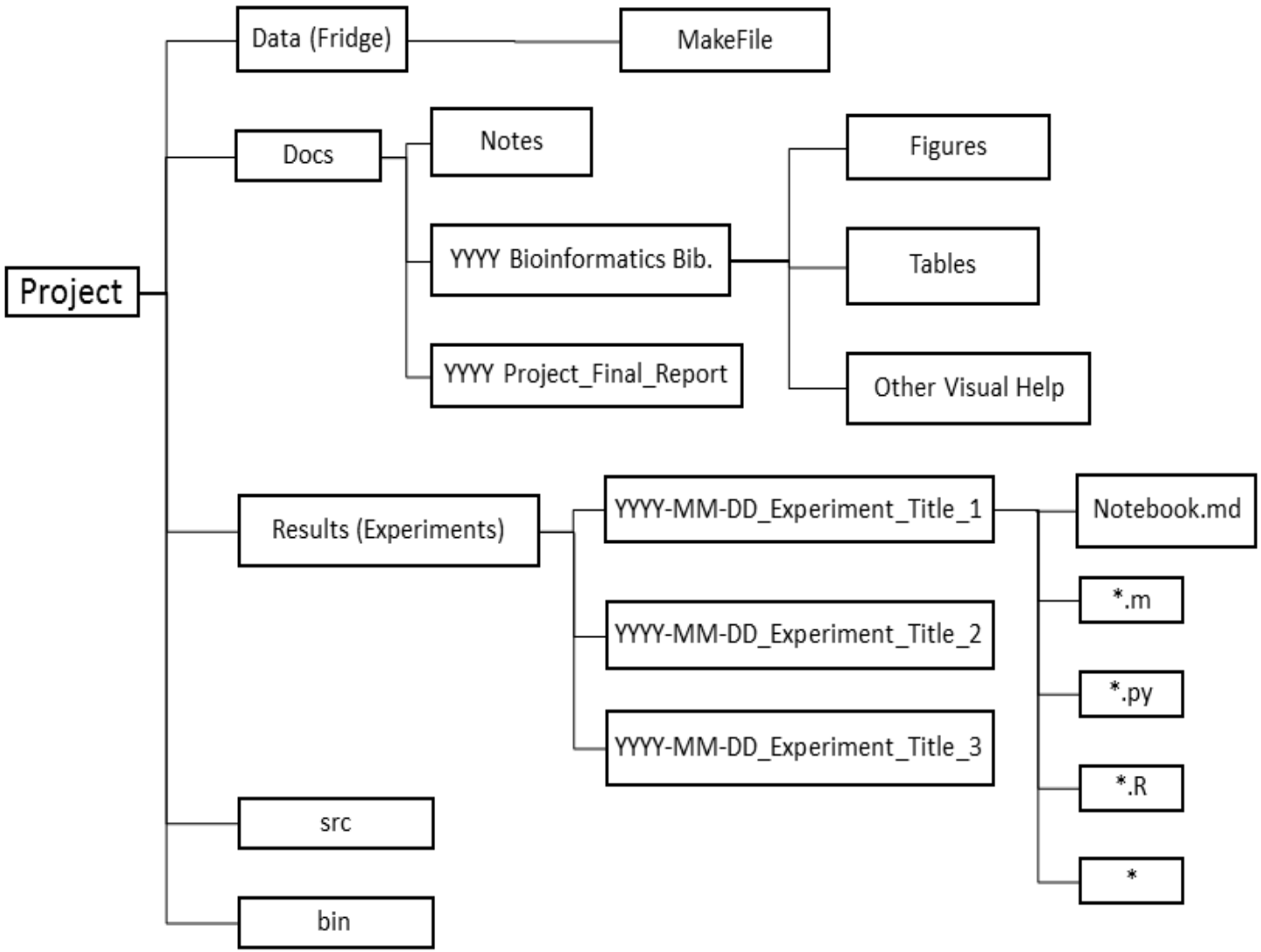
Department of Bioinformatics and Computational Biology
Worcester Polytechnic Institute
Worcester, MA 01609

tzhang3@wpi.edu

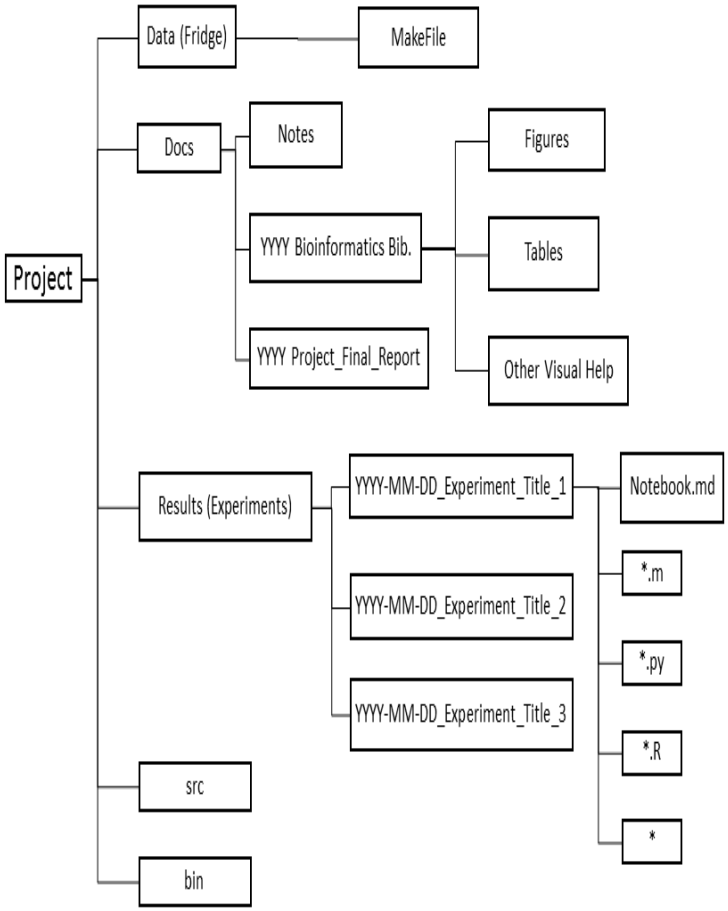
May 29, 2014

Data Management

Data Management



Data Management - Data Fridge



Data Fridge

- Makefile
- target:dependent
  [tab] rules
-

```
yourfile.dc : yourfile.pileup
    pileup2dc yourfile.pileup > yourfile.dc
yourfile.pileup : yourfile.bam
    samtools ipileup yourfile.bam -o yourfile.pileup
yourfile.bam : yourfile.sam
    samtools view -b -S -o yourfile.bam
```

Data Management - Docs

- Notebook
- Figures and Tables
- Project Final Report

Data Management - src (source)

Completed code for your data analysis program.

DataManagement - bin

Executable files, such as:

- Resources helpful to the experiment
- Compiled programs - execute with one command

Version Control

- Why Version Control?
- What are the options for Version Control?
- Why GIT?

Version Control - GIT

- GIT GUI or Command Line
- clone, fetch, pull
- stage, commit, push
- Conflict Management

Flaherty Lab Environment

Flaherty Lab Environment

- Redwood Server
- Amazon Machine Image
- Starcluster? MPI? What else?

Flaherty Lab Environment - Redwood Server

- Flaherty Lab Linux Server for Computation
- 64 core + 256GB RAM
- 9TB high speed drive + 1TB solid state drive via NFS

Flaherty Lab Environment - Amazon Machine Image (AMI)

- When you need more than 64 cores for calculation..
- numpy, scipy, h5py, pytables, matplotlib, pyramid, scikit-learn, pandas, statsmodels, networkx, theano, gdal, pysal, and shapely
- on-demand price: 0.145 dollar per worker per hour

  spot instance price: 0.018 dollar per worker per hour

Thank you!