Bioinformatics Tools
High Performance Computing

BMI701 Introduction of Biomedical Informatics
Lab Session 8

Wei-Hung Weng

November 4, 2016

HMS DBMI — MGH LCS

Bioinformatics

- Gene expression
  - Microarray (RMA, LIMMA), Ontology and pathway (GSEA, KEGG), Batch removal (COMBAT, SVA), NGS DNA-seq (FastQC, BWA, STAR), RNA-seq (DESeq, Kalisto, Sleuth)
- Transcription factor regulation / Epigenetics
  - ChIP-seq / DNase-seq (MACS, BETA), methylation, HiC
- Genome-wide association
  - GWAS (Plink, GCTA), SNP array
- Whole genome sequencing (GATK, Mutect)
- Resources
  - DAVID, GEO, UCSC, CBioPortal

Methods / Algorithms

- Programming
  - R, BioConductor and python
- Statistical tests
  - median polish, distribution, qq-plot, t-test, χ^2 test, K-S test, FDR
- Modeling
  - Generalized linear models, logistic regression, LASSO, k-means and hierarchical clustering, PCA, SVM, Burrows-Wheeler alignment, EM, Gibbs sampling, HMM, dynamics programming, survival analysis

Bioinformatics tools

- Google is your good friend
- Bioconductor
- `source("https://bioconductor.org/biocLite.R")`
- `biocLite("PACKAGE_NAME")`
- `library(PACKAGE_NAME)`

High Performance Computing

- Impossible to do all computing on your laptop
- Harvard FAS Odyssey
- HMS Orchestra
- Partners ERISOne
- How is the submission looks like?

Workflow

1. Request for the account
2. Prepare tools
   - FileZilla
   - Putty for Windows, Terminal for Linux/Mac
3. Connect to the server
   - `ssh USERNAME@SERVER_ADDRESS`
   - `ssh -Y jimmy@orchestra.med.harvard.edu` (With X11)
   - `ssh jimmy@erisone.partners.org`
4. Install whatever you need
   - Odyssey and Orchestra have most of the common use tools, e.g. R, python, their packages, ...
   - Create virtual environment to install your tools

Workflow

5. (Optional) Go to interactive node
6. Upload/Edit your script
7. Create a submission file (`.lsf` for ERISOne, `.slurm` for Odyssey, ...)
8. How is the submission looks like?
    - `bsub -q normal -n 4 -R 'rusage[mem=2000]' < job.lsf`
    - `job.lsf` is a normal exec file
9. Use `bjobs` to check the status
10. Use `bkill` to kill the job

- Demonstration

Some Bioinformatics Learning Resources

- edx Bioinformatics
- Coursera Bioinformatics
- Harvard STAT115/BIO512: Computational biology and bioinformatics

Some Advanced NLP Materials

- NLTK book (very useful!)
- Coursera Jurafusky
- Coursera Radev
- Coursera NLP provided by Michael Collins is also good, but it's gone now (you can try if you can find it)
- Harvard CS287: Natural Language Processing (Rush)
- MIT 6.864: Advanced Natural Language Processing (Barzilay)

Some Visualization Resources

- Harvard CS171: Visualization
- Harvard BMI706? (Gehlenborg)

Some Machine Learning Courses

- Coursera Andrew Ng
- Caltech LFD
- Harvard CS109: Data Science (2015 archive)
- Harvard CS181: Machine Learning (Rush, Parkes)
- MIT 6.036: Introduction of Machine Learning (Jaakkola)
- Stanford CS231n: Convolutional Neural Networks for Visual Recognition (Li)

Take Home Message

- Brief overview of bioinformatics / computational biology (Courtesy by Dr. XS Liu)
- Tools and algorithms
- How to use cluster
- Some recommendations
- Contact
  - Github repository
  - ckbjimmy@gmail.com
  - Linkedin: Wei-Hung Weng