# Statistical and Computational Analysis of the Human Genome

BCHM 5631 Course Syllabus, Spring 2020

**Instructor:** John Rinn john.rinn@colorado.edu **Office:** JSCBB Room B417
**Classes:** MWF 1-2:50 pm, JSCBB Room B231
This lab course covers fundamental statistical and computational approaches to large scale data. Students will learn the unix command line to: access public human genome data, learn what statistics apply to which types of data and apply them to study specific regions of the human genome involved in development and disease. This lab course will cover fundamental aspects of Virtual computing, Container analysis pipelines (e.g. NextFlow, GitHub) in an intuitive and practical learning framework. Same as BCHM 4631. Add Consent: Department Consent Required.

## JANUARY 13: Introduction and overview of course

*Lecture 1: Introduction and overview*
Read Chapters 1-3 for Monday Jan 22
Bioinformatics Data Skills by Vince Buffalo

## JANUARY 15: Set up tools and practice

examples in Chapters 1-3
First we will install the following tools, and after that work through some command line exercises.

## JANUARY 17: ENCODE data reproducibility and Example datasets

*Lecture 2: Data Reproducibility in Science / Intro to Transcriptional regulation*
Overview of Encode / ChIP and Transposons as missing regulatory regions
What is a promoter and transcription factor?

### Encode Data portal

Go through each category and get familiar — we will specifically be looking at:
DNA binding / TF-CHIPseq / K562 / paired-ended
Exploring metadata in unix select samples, click columns add control, click on table and then download .tsv
Use ls, head, tail, cat, awk / grep to explore this metadata table.

## JANUARY 22: Start organizing metadata for data

Retrieval
Sorting hat / intro to bash scripts
Taking notes in Markdown
Organize input files for our ChIP-seq pipeline

- In addition to the sample-level metadata you retrieved last time, download this table of file-level metadata from ENCODE
- As a group, choose one transcription factors that you would like to analyze.
- We're going to subset and organize our metadata file to include just those files that you would like to download and the columns that will be useful to you using awk and grep.
- We'll also make a file which contains the URLs to retrieve the fastq files from Encode.
- Read chapters 4-5

## Statistical and Computational Analysis of the Human Genome

BCHM 5631 Course Syllabus

### JANUARY 24: Let’s go get data!

*Lecture 3: Where does data live in Biology, how do we get it, and did we get the right file?*

We will each go retrieve a ENCODE data file from our sample sheet.

SFTP, SSH, SCP

wget -i file.txt

md5sum

### JANUARY 27: git & GitHub

*Lecture 4: git & gitting GitHub*

Class Exercises:

- Create a git repository and commit some changes
- Create one GitHub repository per group and commit your sample sheet script

Beyond the samplesheet

We're going to create a file that matches the ChIP samples to their control samples. The format of this file is specified by the pipeline that we will be running.

THESE ARE THE REQUIRED COLUMNS FOR THE DESIGN FILE

group,replicate,fastq_1,fastq_2,antibody,control (**** fastq1 and fastq2 URLS ****)

Make a design file by Friday January 31 for your TF

- Hint: this maybe easiest in excel. Look up file accession numberfor YTF. Then look for "paired with" you will see a new File accession number -- that needs to be in your control column.
- If your "paired with" identifier is not in the sample sheet (Jan 22 lecture notes) -- then go to encode portal and find it :)
- Advanced exercise : Script this in bash (going to need a few greps & joins :)

### JANUARY 29: IT lecture on Fiji

Tour of Fiji data center

Meet at Space Sciences

Please take notes on the key rules and regulations — to do and not to do’s!

### JANUARY 31: Connect to Fiji

- Layout of class directories -- where will you be doing work?
- Get a local git repo -- set up ssh key for fiji-GitHub
- Moving files to and from fiji

**HOME : Data you really want to keep and back up not intermediate analyses**

/Users/<identikey>

**Scratch: THe wild west no limits (within reason) here is where we will start doing analysis and set up git etc.**

scratch/Users/<identikey>

# Statistical and Computational Analysis of the Human Genome

BCHM 5631 Course Syllabus

**Folder to submit final files/analysis -- more later**
/Shares/rinn_class/students/<identikey>
**The precooked class**
/Shares/rinn_class/data
Design File presentations
Rsync

## FEBRUARY 3: Set up Fiji to get ready to run nextflow

- SCREEN (screen -list / ctr-d + a/ screen -r)
- Get fastq's for your TF
- SLURM review (interactive & batch jobs)
- md5sum -c
- Go over class design file

exchange design files to have a total of 3 TFs (e.g., collaborate with another group)
`cp` design files.
What happened? How can we solve this?
*Discuss and catch up on what we have learned about unix and commands, etc.*

## FEBRUARY 5: NextFlow / nf-core chipseq

*Lecture 5: Flowing with NEXTFlow*

- Nextflow paper: Nextflow enables reproducible computational workflows
- Nextflow
- NF-CORE

Read basic documentation and install nextflow in your path!

## FEBRUARY: NextFlow / nf-core chipseq

Set up design and sample files — folder structures —
Run.sh
GOAL -- Set up your project directory run for 3 TFs

- design.csv
- nextflow.config
- run.sh
- blacklist
- fastq directory w/ fastqs downloaded
- checked by John or Michael
- run pipeline

sbatch run.sh
queue -u X000
scancel jobid

**Familiarize yourself and take notes on file types**
https://www.encodeproject.org/help/file-formats/
Read next flow documentation and nextflow.out
Homework: google the programs used in nextflow.out – Fastqc, TimaGalore, BWAMem, SortBAM, MergeBAM, BigWig, MACSCallPeak, Peak QC

Each group presents three questions they would like to address based on the TE-DNA, TE-RBP, E-CLIP study designs. Each person 3 questions.

Presentation outline:

- Introduce yourself and your research
- Present the question that you would like to pursue with the class dataset
- Discuss how you'd like to use lessons from this class in your own research

## FEBRUARY 26: R Part II

*Lecture 9: Intro to R -- part II*

- Continuation of R data types
- Introduction to ggplot2 and tidyverse
- Exercise -- plotting gene profiles
- Git from R

Good R tutorial: https://www.youtube.com/watch?v=fDRa82lxzaU

## FEBRUARY 28: R Part III: R for Genomics (GRanges /rtracklayer)

*Lecture 10: R for Genomics -- part I*

Install the following packages in fiji-viz/RStudio

```
install.packages("BiocManager")
BiocManager::install("GenomicRanges")
BiocManager::install("rtracklayer")
```

- Review your solutions to the for loop/plotting exercise
- Introduce GRanges and findOverlaps
- Read in peak files, repeatMasker files, and find overlaps

## MARCH 2: Class on own -- peak plotting exercise

Exercise: Make some plots to characterize the overlap of ChIP-seq peaks with TEs.

Can be as simple as plotting the number of overlaps of one particular TF with a class of TEs - OR - since you have data for all the TFs, you can plot each protein's peaks and where they fall in relation to the center of the repeat -- i.e. a metaplot heatmap or profile plot.

If you get stuck, ask your group members for help and if you're still stuck, ask in the general slack channel. We will go over your plots and code on Wednesday

## March 4: Review TE / TF metaplots exercise

<u>3 Groups of 5</u>

Group 1 (mRNA): Kristen, Shelby, Ben, Soroya, Arpan
Group 2 (lncRNA): Savannah, Michael, Tao, Dan, Graycen
Group 3 (TE): Alison, Devon, Tom, Kevin, Guilia

Granges Gencode
Granges consensus.peak.file
Intersect Granges
Go over TE intersection plots and problems

## MARCH 6:

- Fix RMarkdown with Jon
- Introduction to RMarkdown and functions
- Git structure -- how teams will be committing to class repository /scratch/Users/<identikey>
- Discussion: Clustering

**Exercise 1: Practicing with Git**
Each person contributes commmits to the README.md in each group.
Submit a pull request to the master branch.

**Exercise 2: Creating a gold-standard peak set**
Write a function that will require peaks to be present in all replicatesper TF. Then iterate over all TFs to create peak sets (Granges objects) that consist of peaks present in all replicates. Write these peaks to one bed file per TF. Copy these peak files to your class directory /Shares/rinn_class/students/<identikey> .
We will be reviewing these files on Monday.

Bonus: Write the function such that the number or percent of replicates required is adjustable.

Considerations: Do you want to merge the peak regions? What is the minimum overlap required? How do the results change when this parameter is varied? How many peaks do we lose by doing this approach?

- **Going remote as of Friday March 13**
- Browsing / spot checking consensus peaks in UCSC (session
- example "consensus_peaks" in UCSC class session list --
- Randomly sampled peaks to check out)

**UCSC Resources:**

- Peak files for each replicate
- /Shares/rinn_class/data/ucsc_peaks
- Consensus peak files /Shares/rinn_class/data/k562_chip/analysis/00_consensus_peaks/ucsc_peak_tracks
- BigWig file link bigWigs

Profile plot RMarkdown output
**Class Exercise:**
Look through the profile plots and remake the plots for your favorite TF(s) or all of them.
Find two TFs that have different profile plots.
Find examples of their consensus peaks with bigwig replicates.
Present interesting aspects about these TFs from literature (NCBI Gene).
Prepare a presentation per group for Friday.
Slack a ppt or keynote to the general channel before class on Friday.

# Statistical and Computational Analysis of the Human Genome

BCHM 5631 Course Syllabus

## March 16: Remote class orientation (precooked .Rmd)

- **Welcome to Zoooooom !**
- Break out rooms for groups
- Slack and zoom / trello
- Presentations

## March 18: Intersect annotation features in GRanges for mRNA, lncRNA and TE

Intersect_excercise.Rmd

Do intersects in class for your "biotype"

Class exercise (presentations Friday March 20): Find some interesting examples for your group (5 TFs).
Is there a trend with number of peaks and number of overlaps? How could we "shuffle" to understand if this is significant or happens by chance?

Which ones bind your biotype more than others? What is the most unique DNA binding protein for your group?

## MARCH 30: Functions, Features and Fun and git organization for analyses

[Paper to read on mRNA and lncRNA promoter properties]
(https://www.dropbox.com/s/ux3e7xzl9lsflxz/Mele_et_al.pdf? dl=0)
[Second paper to read on promoter properties]
(https://www.dropbox.com/s/m4832lsedpt826f/Genome%20Res.-2019-Mattioli-gr.242222.118.pdf?dl=0)

## APRIL 1: No class : APRIL-FOOLs <- Present interesting promoters that have many DNA binding protein events.

Clustering

## APRIL 3: Findings from clustering & paper figure presentations

- Present a figure and associated analysis/findings from each paper (Mele et al. & Mattioli et al.)
- Present findings from your clustering exercise:
  - What groupings make sense?
  - Are there different clustering groupings when you compare all promoters vs your subset?

## APRIL 6: Expression comparisons -- recapitulate

Mele et al finding that more TFs higher expression.

- Class exercise: are there promoters with lots of TFs that are not expressed? At what point would we say there are a lot of TFs bound :) ? Hint: histogram of cooccurrence matrix.

## APRIL 8:

- Walk through results (ghosts)
- Other questions to analyze? Distribute analyses.
- Prepare questions for Michael Snyder

## APRIL 10: *Michael Snyder guest lecture/interview*

### APRIL 13: Permutation test class exercise I

[Intuitive Statistics Lecture]
(https://www.dropbox.com/s/95iq9veg5e7qp1y/Permuation_false_discovery. dl=0)
Groups will work on `permutation_test_class.Rmd`

### APRIL 15: Permutation test class exercise II

### APRIL 17: Design manuscript outline

### APRIL 20: Work through making figures – clean code and figures in .Rmd

### APRIL 22: Work through making figures -- cleancode and figures in .Rmd

### APRIL 24: Figure from each group due in .Rmd

### APRIL 27: Finalize Figures and git

### APRIL 29: Sweep up the workshop!