

Epigenomics Websites and Bioinformatics

Introduction

- Associate Bioinformatics Scientist at Coriell Institute for Medical Research
- Help other people with their data analysis
- Independent data analysis
- All materials for my lecture will be online at



Where is the data?

Databases

Reference Information

- Variants
 - 1000 Genomes <https://www.internationalgenome.org/>
 - ClinVar <https://www.ncbi.nlm.nih.gov/clinvar/>
 - Catalogue of Somatic Mutations in Cancer (COSMIC) <https://cancer.sanger.ac.uk/cosmic>
- Epigenetics
 - Encyclopedia of DNA Elements (ENCODE) <https://www.encodeproject.org/>
 - Roadmap Epigenomics Project <http://www.roadmapepigenomics.org/>
- Single Cell (everything is under active development)
 - Human Cell Atlas <https://www.humancellatlas.org/>
 - Mouse Cell Atlas <http://bis.zju.edu.cn/MCA/>
 - Tabula Muris <https://tabula-muris.ds.czbiohub.org/>

Publicly Available Data

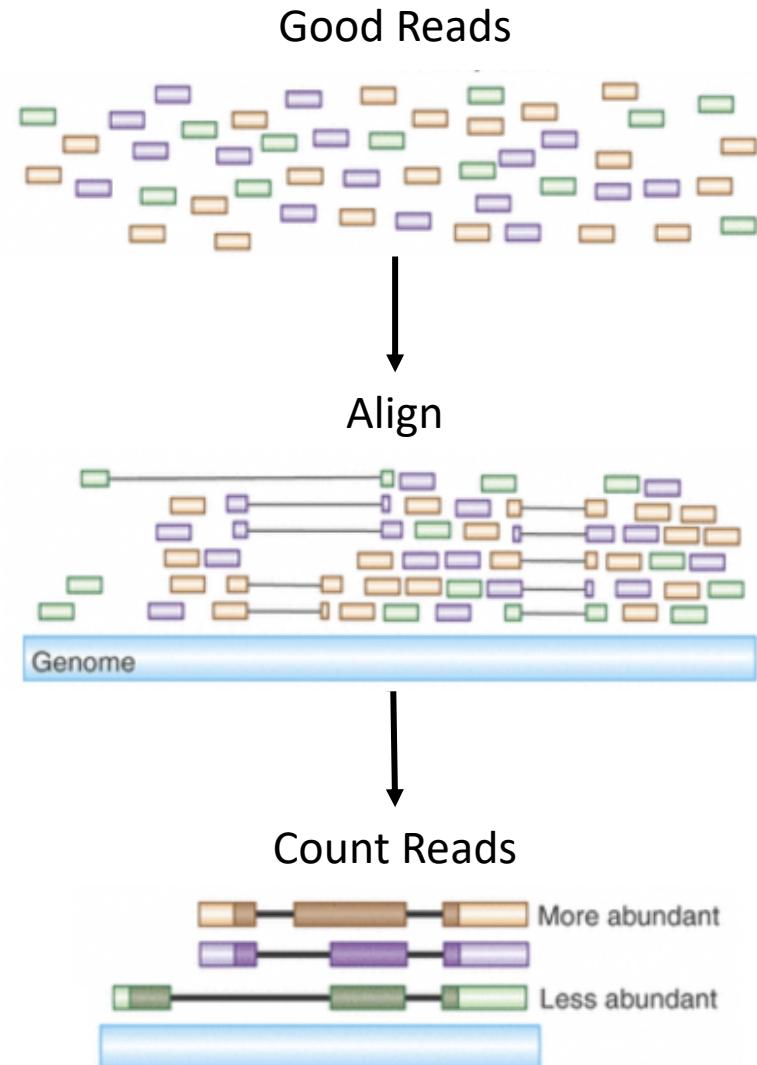
- dbGaP <https://www.ncbi.nlm.nih.gov/gap/>
- The Cancer Genome Atlas (TCGA) <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>
- DeepBlue <https://deepblue.mpi-inf.mpg.de/>
- NCBI Data
 - Gene Expression Omnibus (GEO) for processed data <https://www.ncbi.nlm.nih.gov/geo/>
 - Short Read Archive (SRA) for raw data <https://www.ncbi.nlm.nih.gov/sra>

DATABASE DEMO

General Genomics Analysis

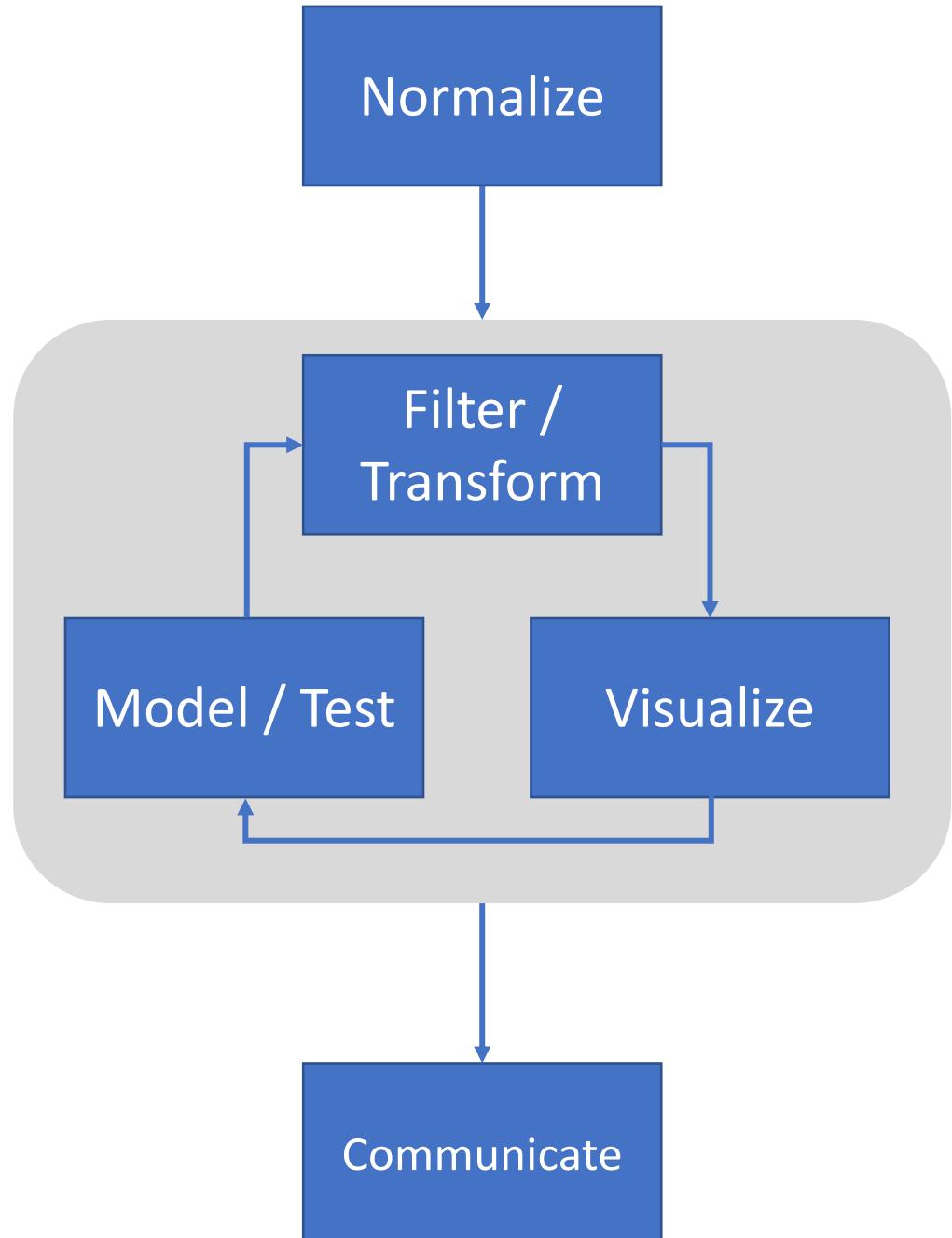
Process Data

1. Quality Control
2. Alignment
3. Count Features of Interest



Analyze Data

1. Normalize Data
2. Explore Data
 1. Filter / Transform
 2. Visualize
 3. Model / Test
3. Communicate
 - Poster
 - Paper
 - Thesis
 - Blog post
 - New Project



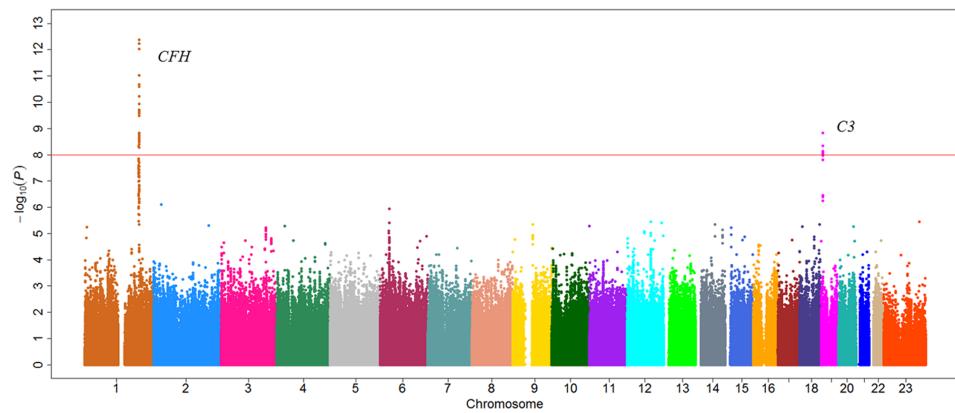
DO A BIT OF ANALYSIS

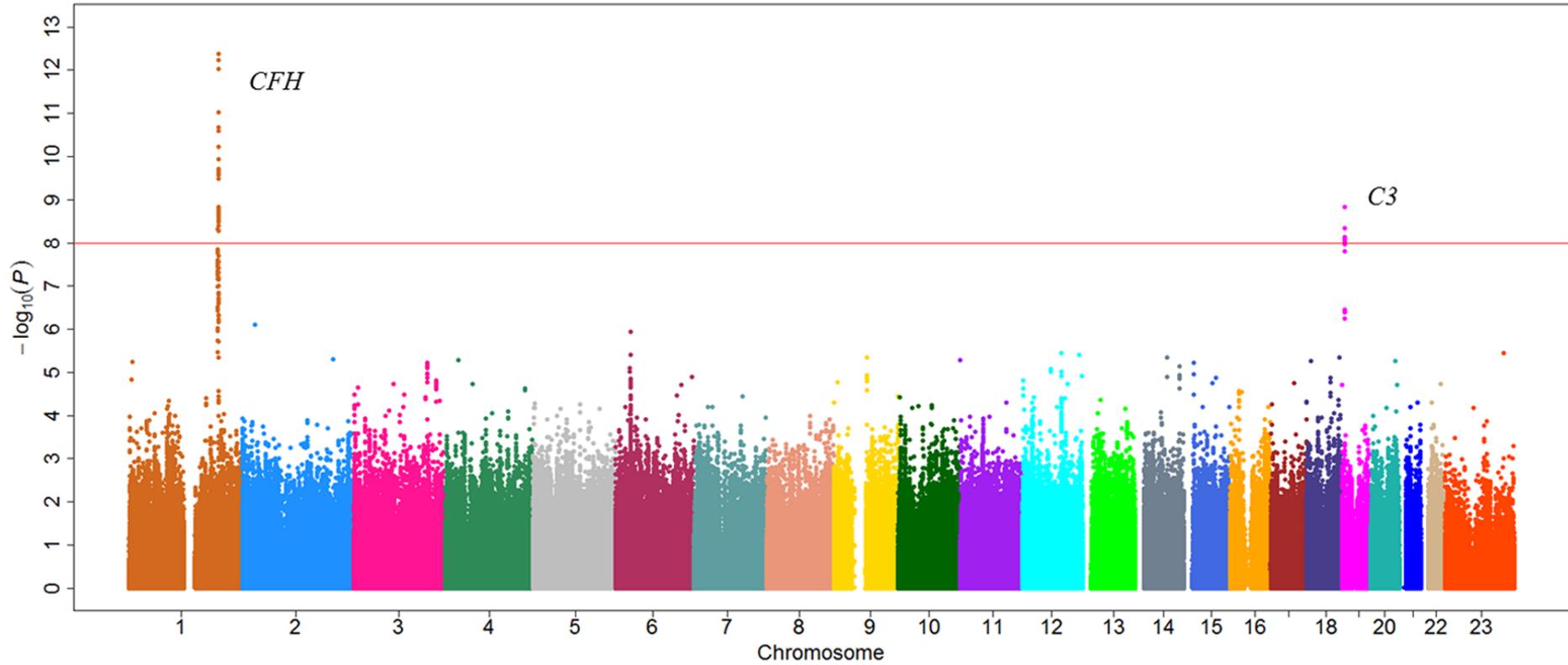
<https://infinityloop.shin-yapps.io/TCC-GUI/>

What Are the Variations
When Looking at
Different Features?

DNA-seq

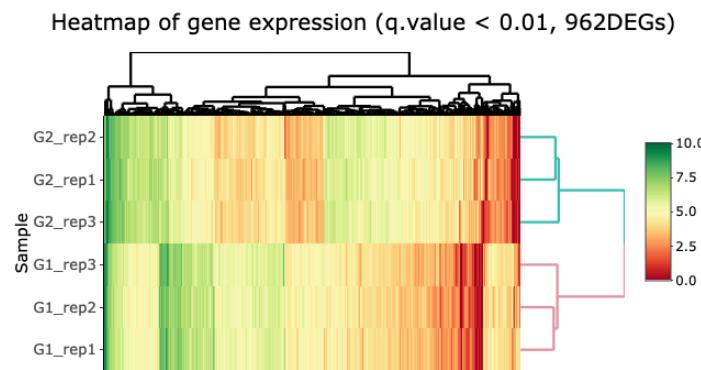
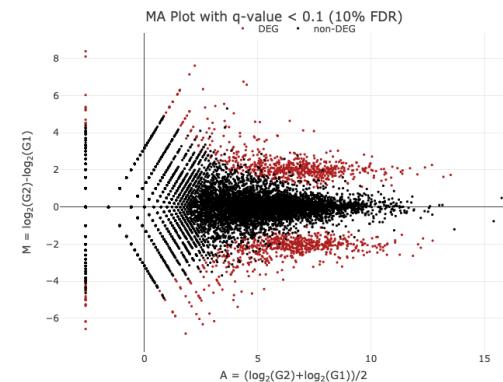
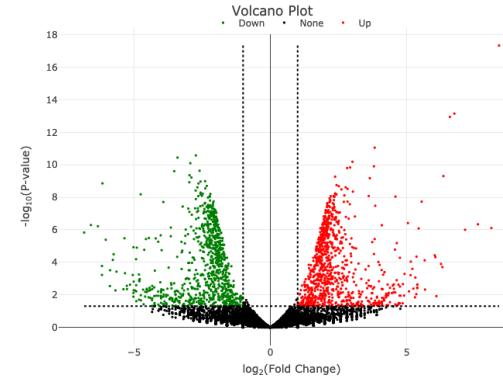
- Analyzes DNA sequence
- Detect
 - Small nucleotide variants
 - Insertions and deletions
 - Copy number alterations
- What variants are present in treatment that aren't in my control?
- Tests
 - Differential analysis
 - Disease vs non-diseases
 - Cancer vs normal in the same individual
 - Genome wide association study
 - Parent-child trios
 - Population genetics
- Visualize
 - Manhattan plot





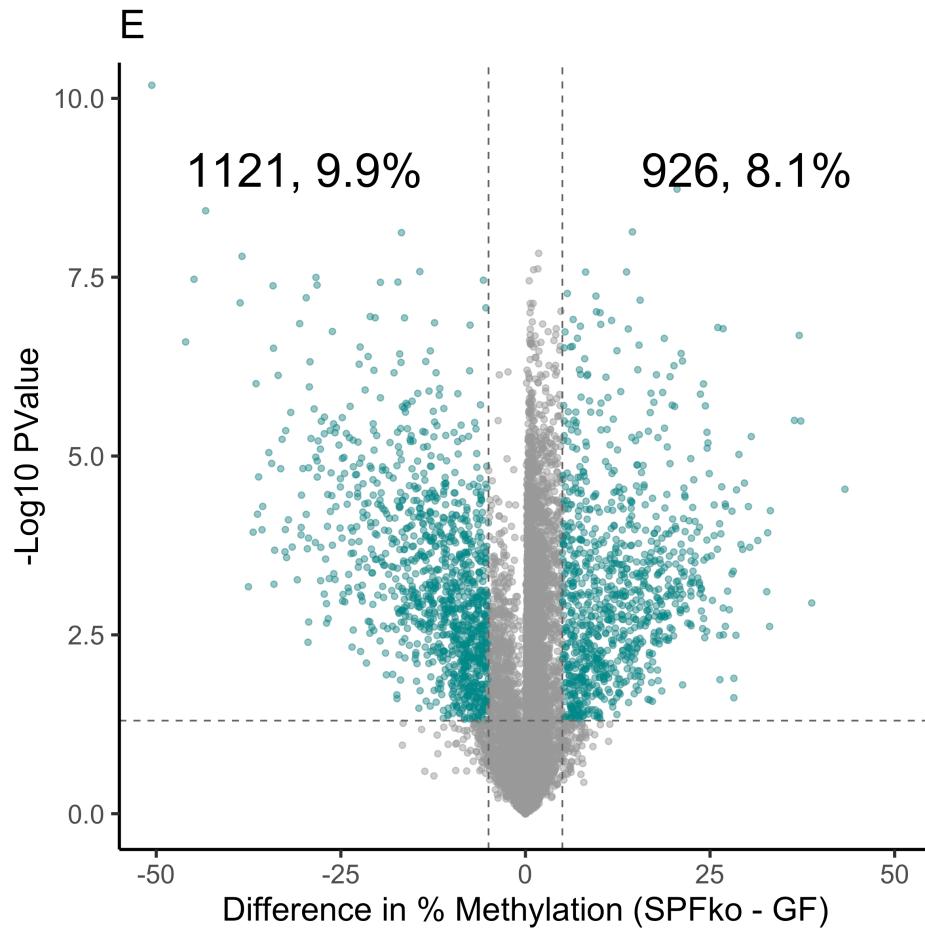
RNA-seq

- Analyzes gene expression levels
- Differential Expression: What is the relative change in expression between my conditions?
- Visualize
 - Volcano plot
 - MA plot
 - heatmap



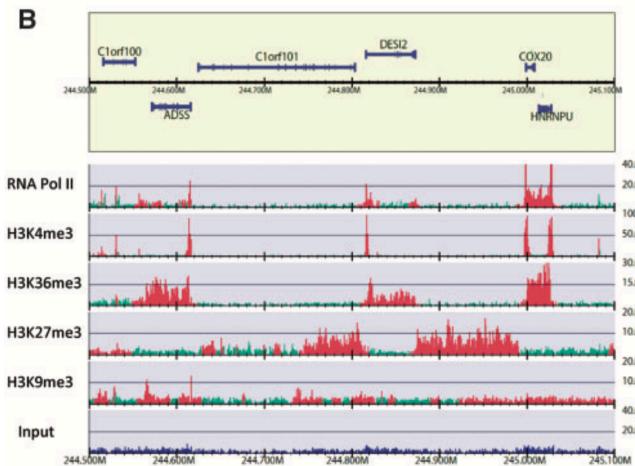
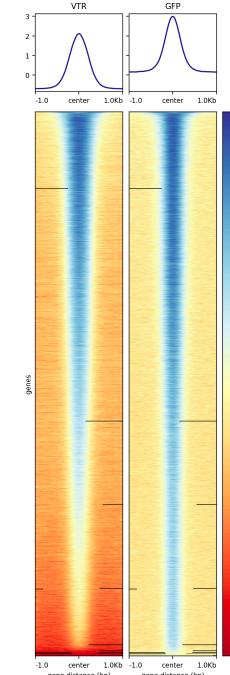
BS-seq

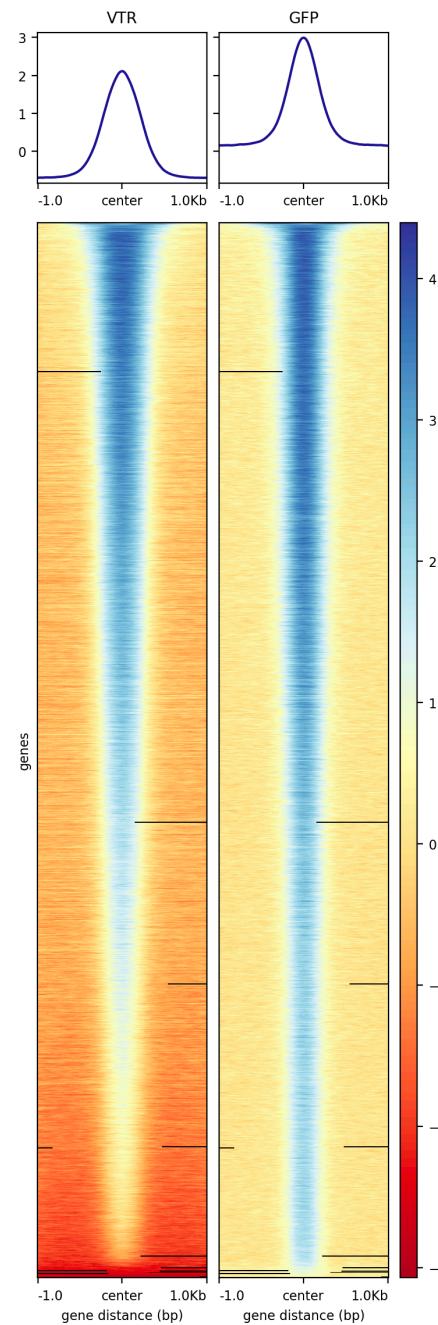
- Analyzes gene methylation levels
- Assays
 - Reduced Representation Bisulfite Sequencing (RRBS)
 - Whole Genome Bisulfite Sequencing (WGBS)
- Differential Methylation: Which CpG sites changed methylation between my conditions?
- Visualize
 - Volcano plot



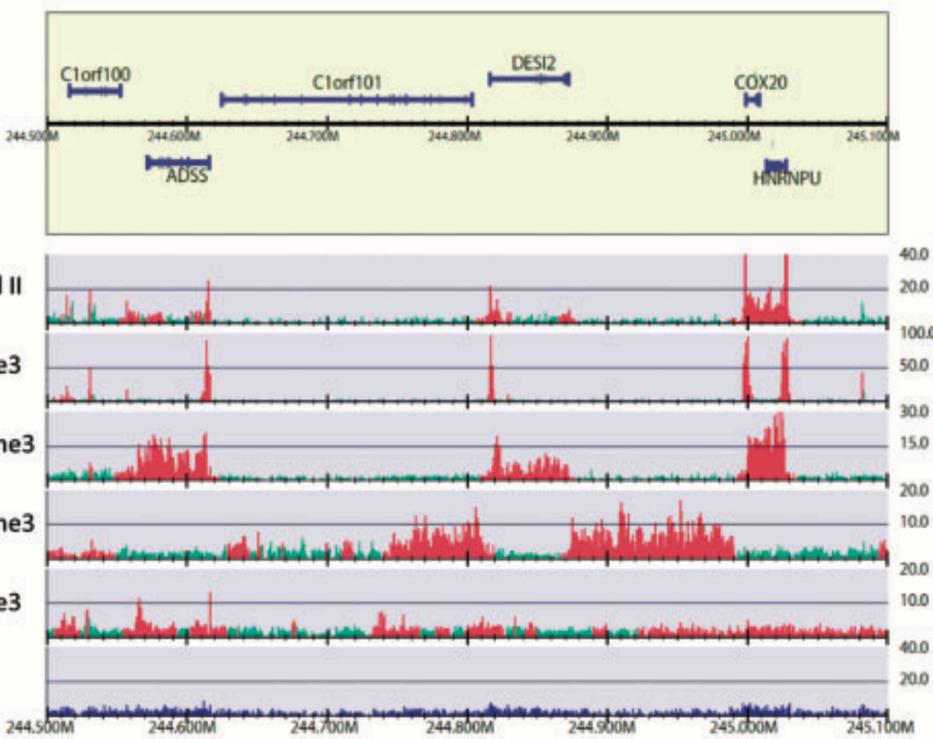
ChIP-seq

- Analyzes features associated with DNA by targeting with an antibody against the feature of interest
 - DNA-associated proteins
 - Histone modifications
- Instead of counting reads, call peaks
- Where is my feature of interest? Does the location change or the signal change with the experimental condition?
- Visualize
 - Example pileup
 - Traces
 - Heatmaps





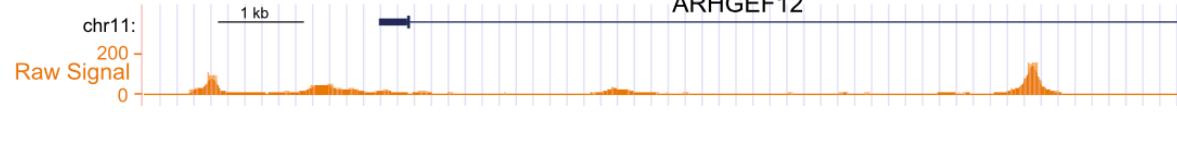
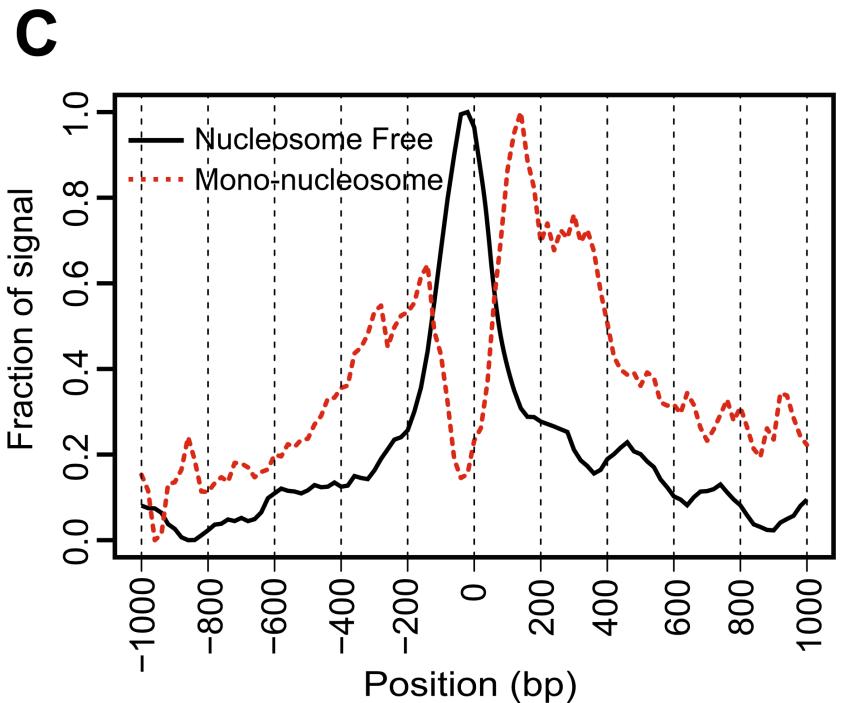
B



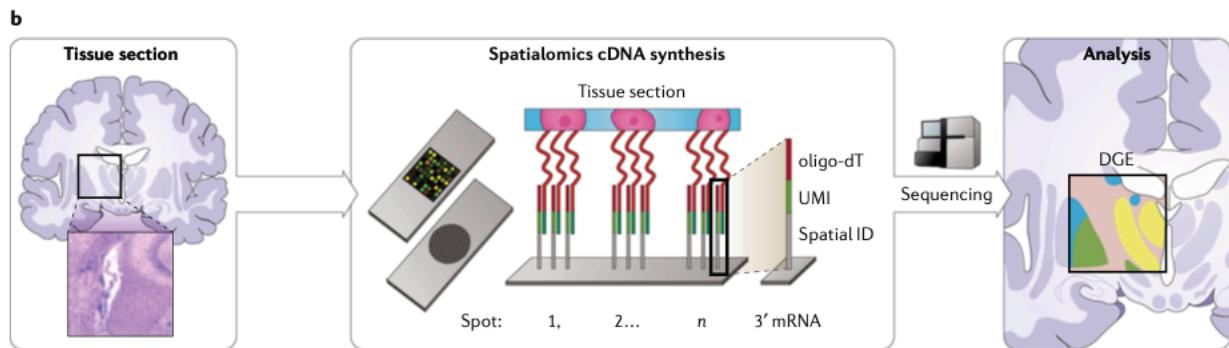
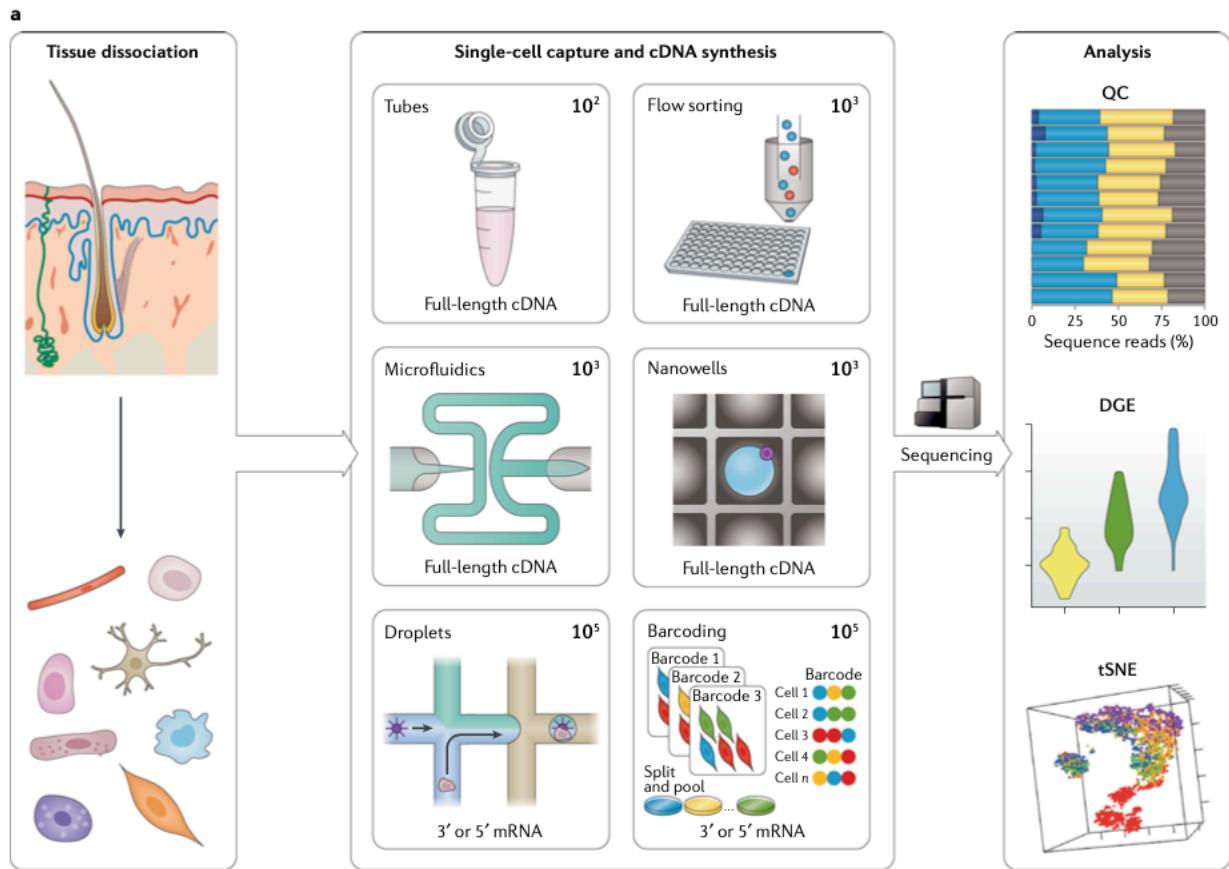
Images: KK (unpublished), Nakato 2017, More Info: Nakato 2017

ATAC-seq

- Assay for chromatin accessibility
 - Get signal in open chromatin
 - Most chromatin is closed, get well-defined signal in promoters and enhancers
- Instead of counting reads, call peaks
- Does the location change or the signal change with the experimental condition?
- Visualize
 - Example pileup
 - Traces
 - Heatmaps



Single Cell Assays



Images: Stark 2019, More Info: Stark 2019, Lahneman 2020

How to can I get started in bioinformatics on my own? (A collection of free online resources)

1. Learn a scripting language

- R <https://r4ds.had.co.nz/>
- Python <https://jakevdp.github.io/PythonDataScienceHandbook/>

2. Take a work through a genomics, biostatistics, and/or bioinformatics course

- Biomedical Data Science <http://genomicsclass.github.io/book/>
- Modern Statistics for Modern Biology
<http://web.stanford.edu/class/bios221/book/index.html>
- Bioinformatics Data Skills <https://vincebuffalo.com/book/> (book not free sorry!)

3. Go to a MeetUp

- RLadies Philly <https://www.meetup.com/r-ladies-philly/>
- Philadelphia Python Users Group <https://vincebuffalo.com/book/>

References

Literature:

1. **DNA-seq:** Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Communications Biology* 2, (2019).
2. **RNA-seq:** Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* 20, 631–656 (2019). <https://doi.org/10.1038/s41576-019-0150-2>
3. **BS-seq:** Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* 21, (2020).
4. **ChIP-seq:** Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in Bioinformatics* bbw023 (2016) doi:10.1093/bib/bbw023.
5. **ATAC-seq:** Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* 21, (2020).
6. **Single Cell:** Lähnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biology* 21, (2020).

Images:

1. Mapping reads to a reference or de novo assembly. Reprinted by permission on EMBL-EBI (<https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/read-mapping-or>) from Macmillan Publishers Ltd: *Nature Biotechnology*. Haas BJ and Zody MC. Advancing **RNA-seq** analysis. 28:421-423, copyright 2010 ([10](#)).
2. Yang, Xiaobo; Sun, Jielin; Gao, Yong; Tan, Aihua; Zhang, Haiying; Hu, Yanling; et al. (2015): Manhattan plot of genome-wide association analyses for C3 level.. *PLoS Genetics*. Figure. <https://doi.org/10.1371/journal.pgen.1002916.g001>