Introduction to Bioinformatics

Owen M. Wilkins, PhD

Co-director, Data Analytics Core

Center for Quantitative Biology, Geisel School of Medicine at Dartmouth

Email: DataAnalyticsCore@groups.dartmouth.edu

Website: (https://sites.dartmouth.edu/cqb/projects-and-cores/data-analytics-core/

December 2022

What is Bioinformatics?

Bioinformatics *is an interdisciplinary field that develops and applies computational methods to analyze large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or discover new biology. The computational methods used include analytical methods, mathematical modelling and simulation.*

Source: T*owards Data Science (Medium)*
https://towardsdatascience.com/what-is-bioinformatics-703170763999

Draws from numerous fields:
- Molecular biology
- Programming
- Computer science
- Statistics

Spans various disciplines
- **Genomics**
- Proteomics
- Structural Bioinformatics

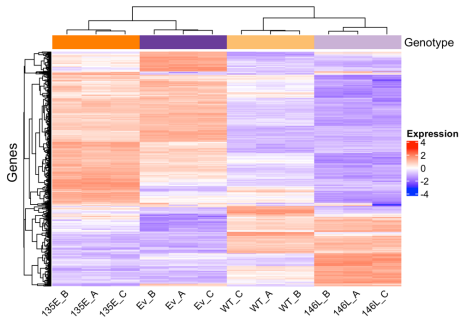
Sequencer
DNA/RNA

Can generate millions, 100s of millions, or billions of reads in FASTQ format
(depending on instrument)

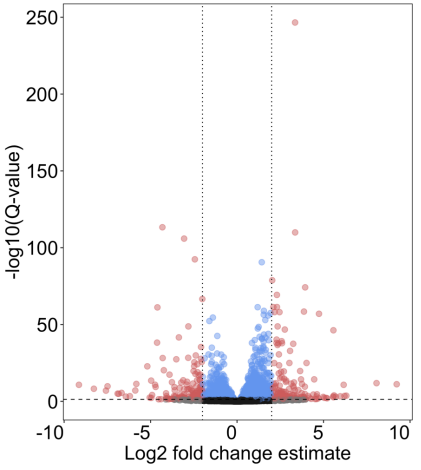
?

Biological insights

Hierarchical clustering

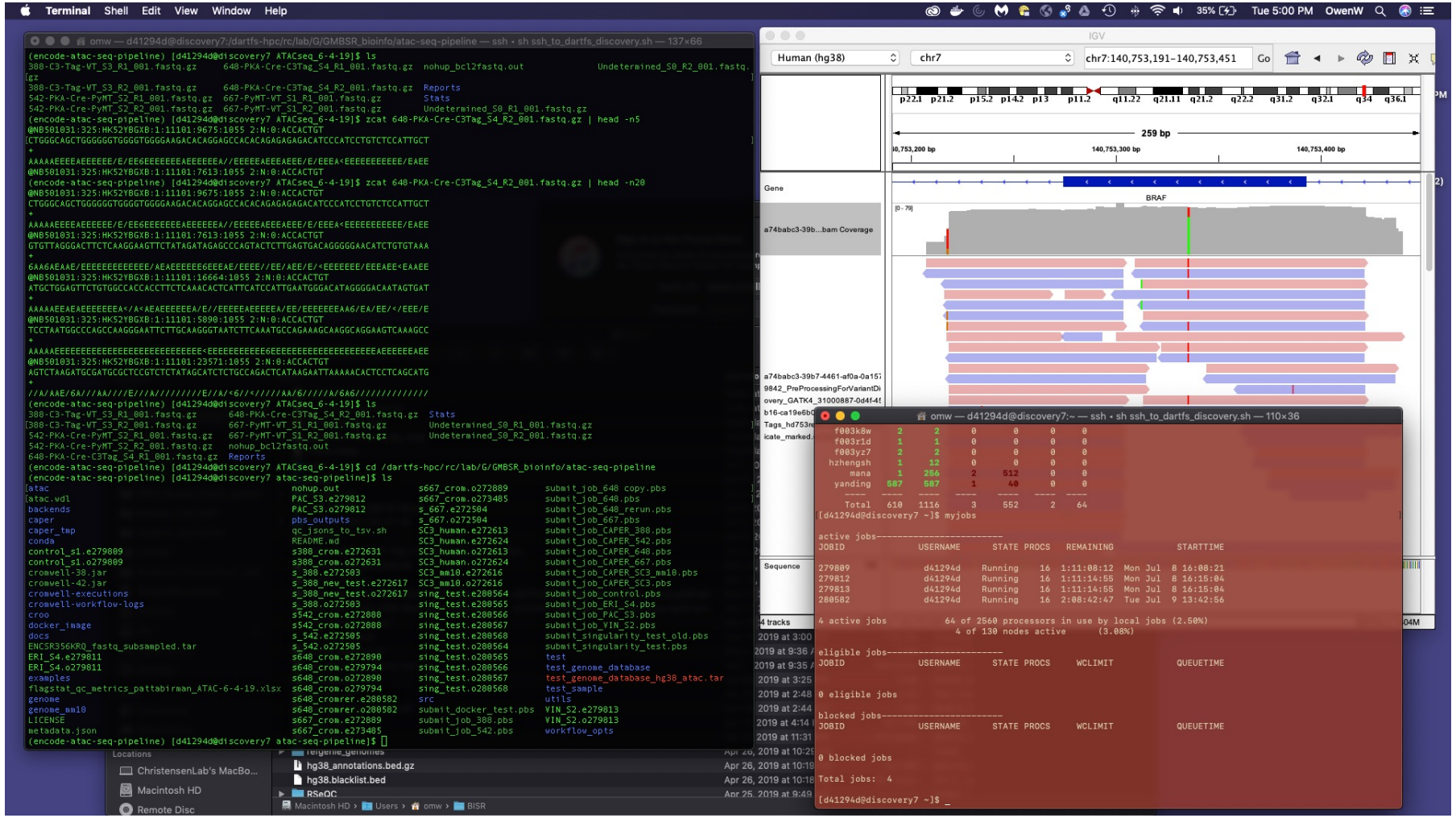


Differential expression

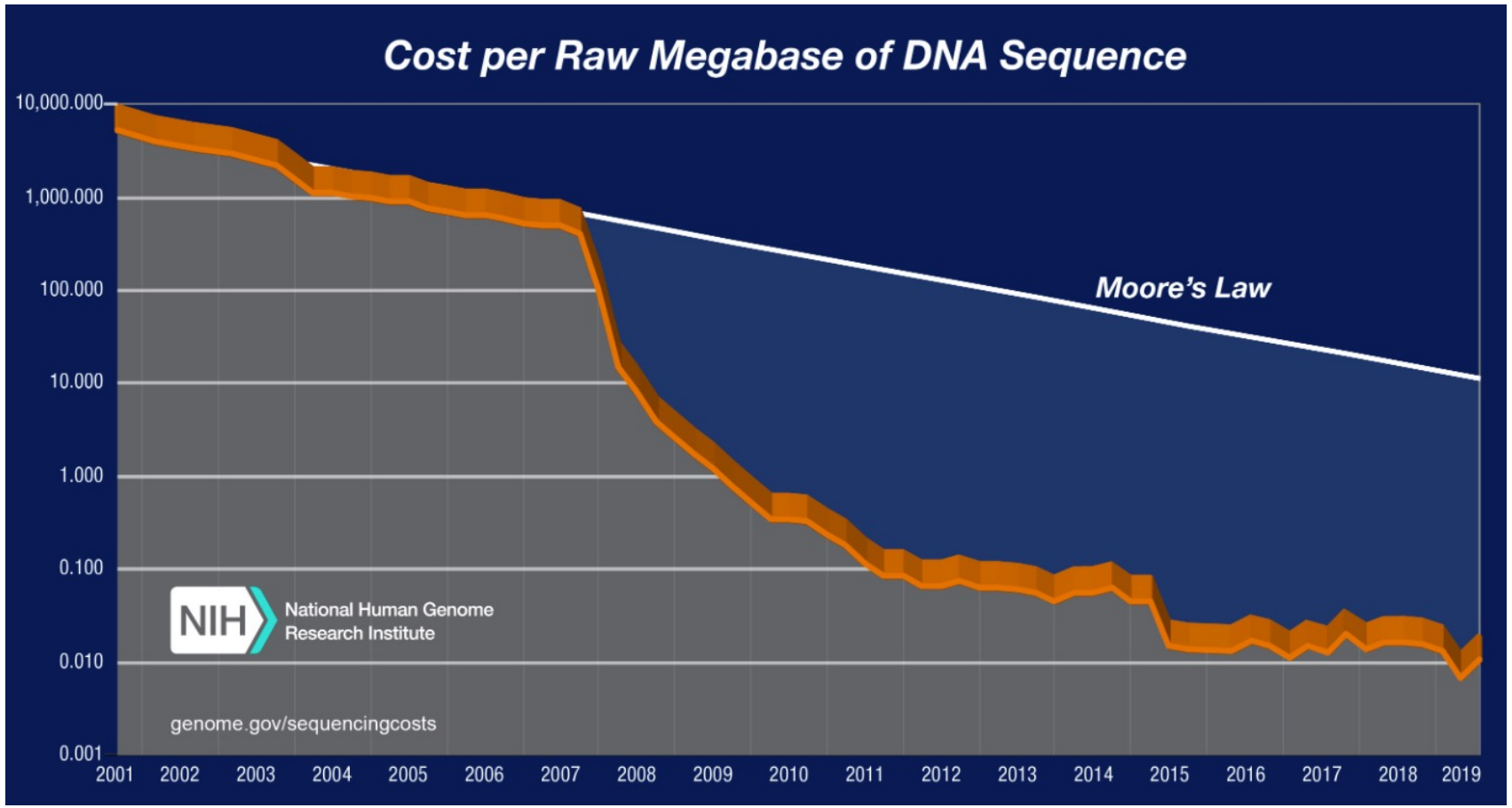


Common perception…

Reality..


Terminal Shell Edit View Window Help
IGV
Human (hg38)
chr7
chr7:140,753,191-140,753,451
Go
259 bp
BRAF
Gene
Sequence
omw — d41294d@discovery7:~ — ssh • sh ssh_to_dartfs_discovery.sh — 110×36
myjobs
active jobs
eligible jobs
blocked jobs
Total jobs: 4

Data is getting cheaper & bigger



https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Complex analytics are playing a larger role

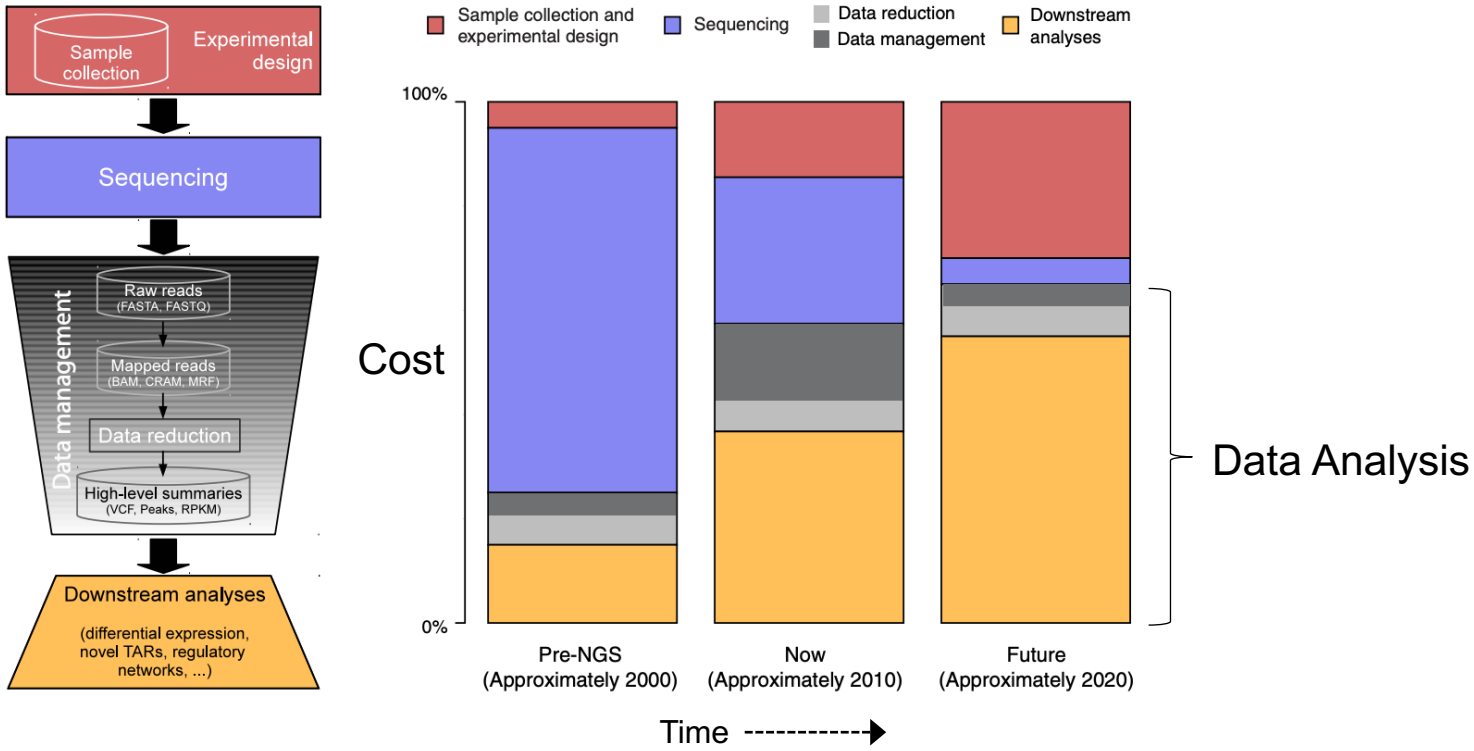
Sboner *et al. Genome Biology* 2011, **12**:125
http://genomebiology.com/2011/12/8/125

Genome Biology

OPINION

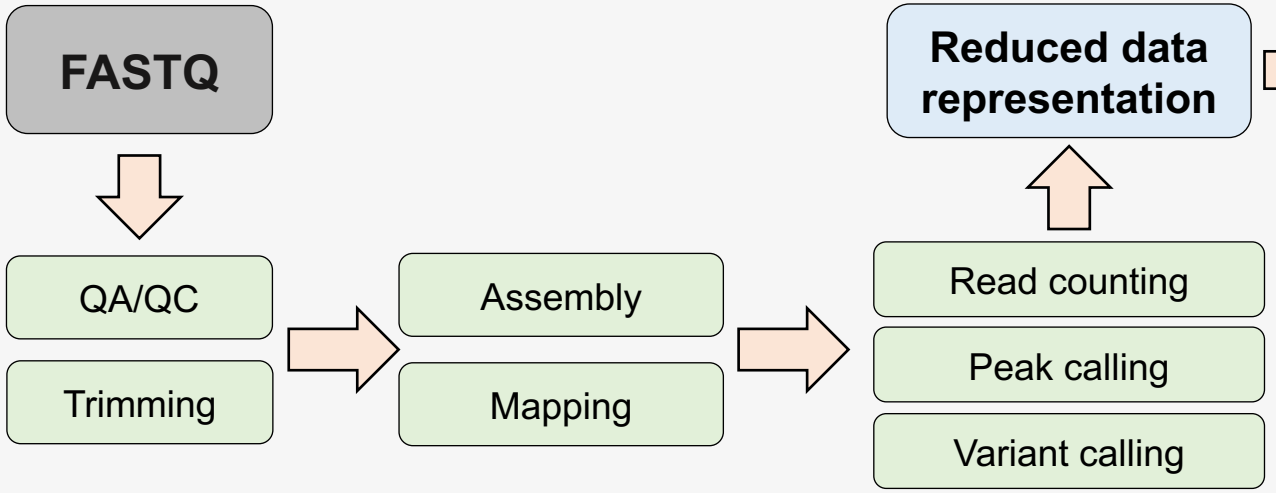
The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], Xinmeng Jasmine Mu[1], Dov Greenbaum[1,2,3,4,5], Raymond K Auerbach[1] and Mark B Gerstein*[1,2,6]



Two major components of genomic data analysis

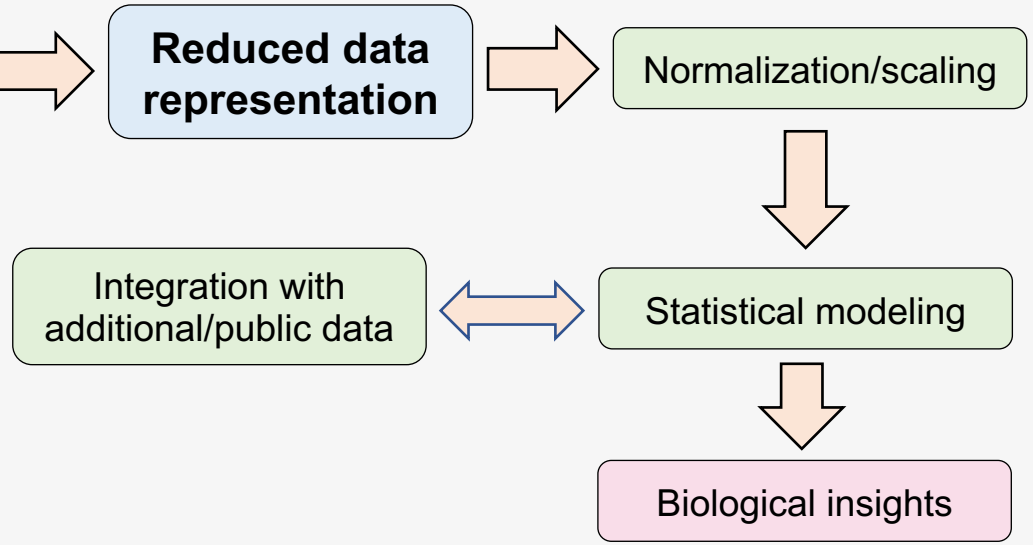
Data reduction (HPC system)



- Predominantly in *UNIX Shell*
- Most computationally intensive steps
- Requires understanding of sequencing technology & molecular biology
- **Example outputs:** RNA-seq read counts, ChIP peak calls

Spending a lot of time here figuring out how to do all this is not an effective use of time

Downstream analysis (HPC or local computer)



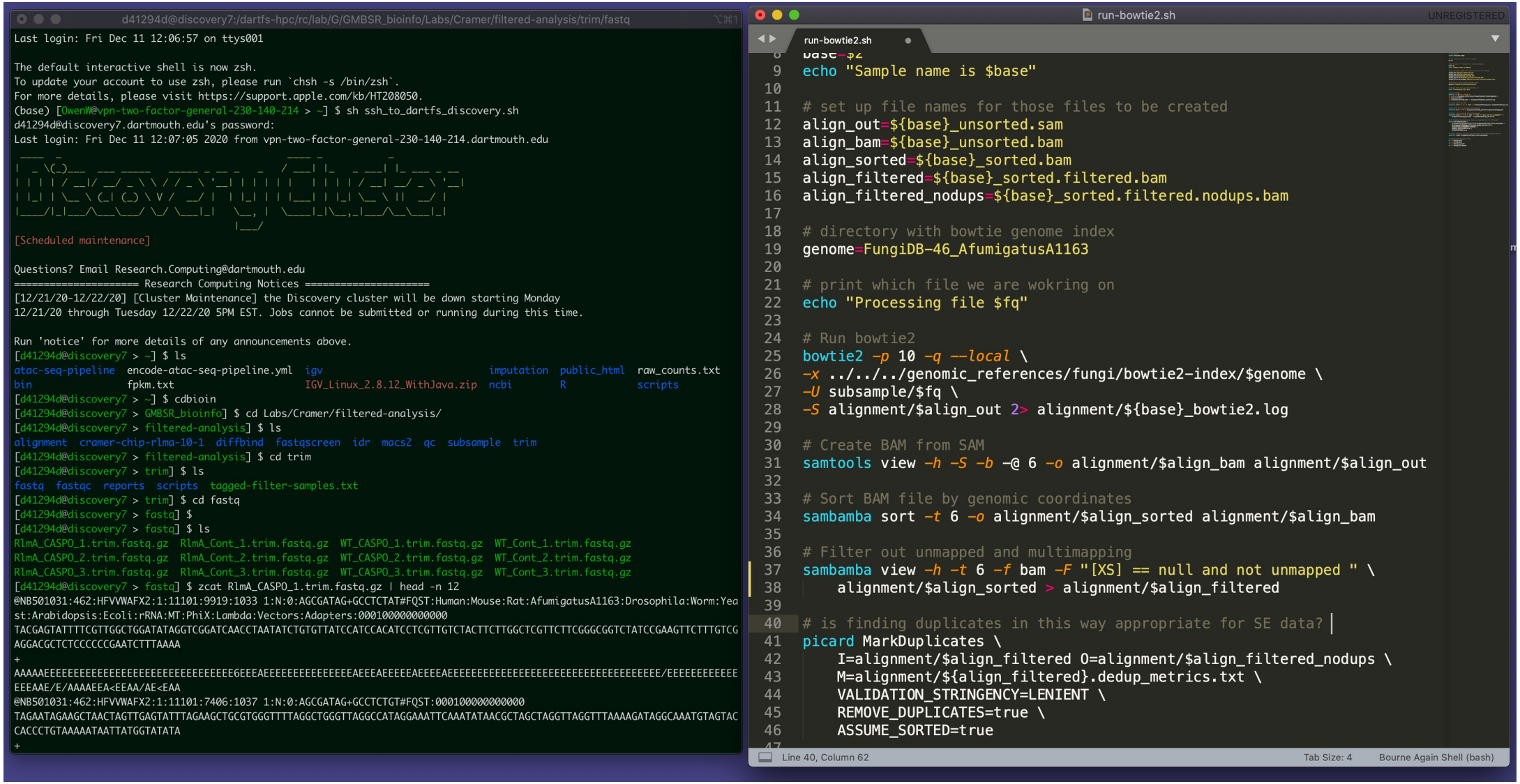
- Predominantly in *R/Python*
- Relies heavily on statistical knowledge
- Needs understanding of biological question
- **Example output:** Differential expression results

This is where all the interesting analysis is

How is data reduction performed?

ssh connection to discovery

Text editor for shell scripting



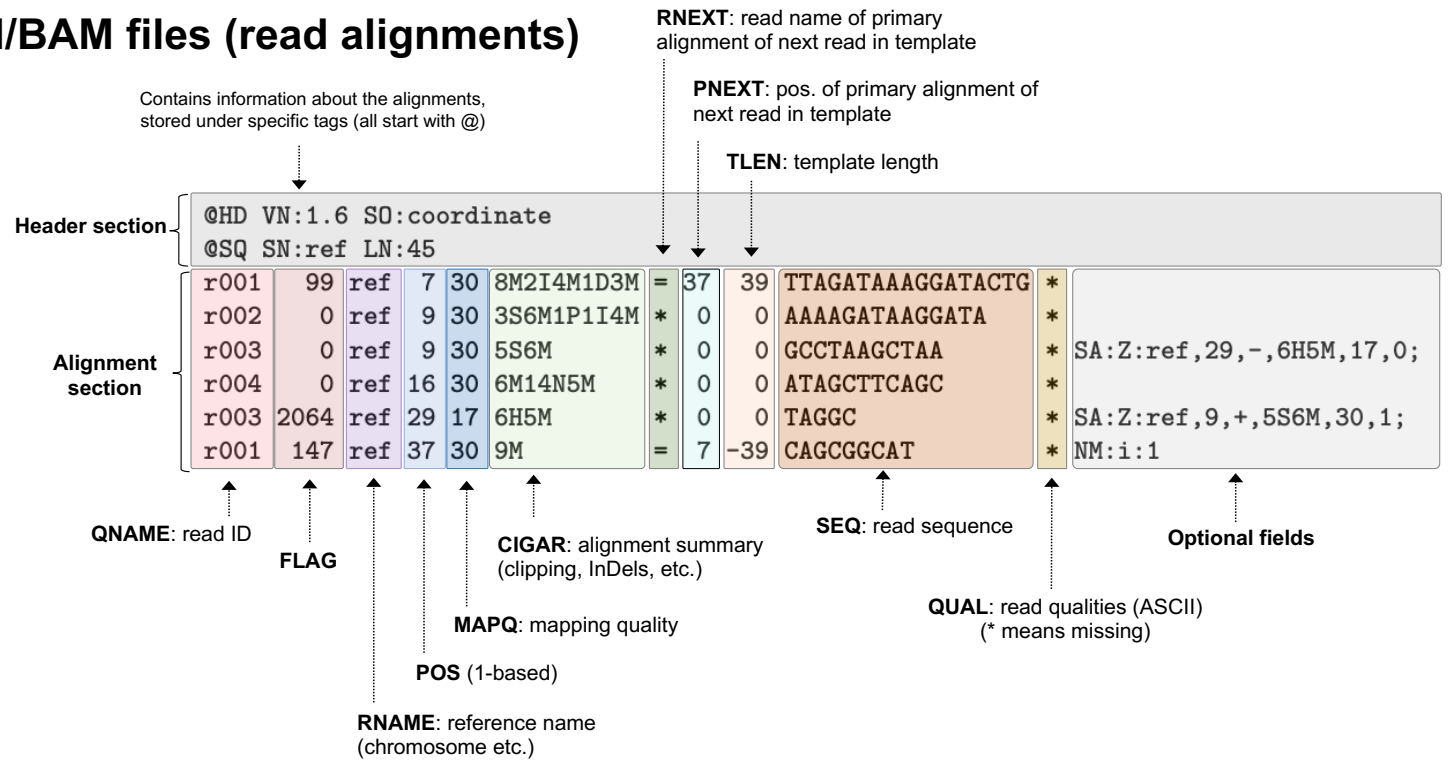
NGS data reduction involves many distinct file formats

FASTQ files (raw sequencing reads)

```
@SRR1039508.1 HWI-ST177:290:C0TECACXX:1:1101:1225:2130 length=63
CATTGCTGATACCAANNNNNNNNGCATTCCTCAAGGTCTTCCTCCTTCCCTTACGGAATTACA
+
HJJJJJJJJJJJJJJ########00?GHIJJJJJJJIJJJJJJJJJJJJJJJJJJHHHFFFFFD
```

- **Line 1:** Header line (read name)
- **Line 2:** Base calls from sequencer
- **Line 3:** Usually just a +
- **Line 4**: Base quality scores

SAM/BAM files (read alignments)

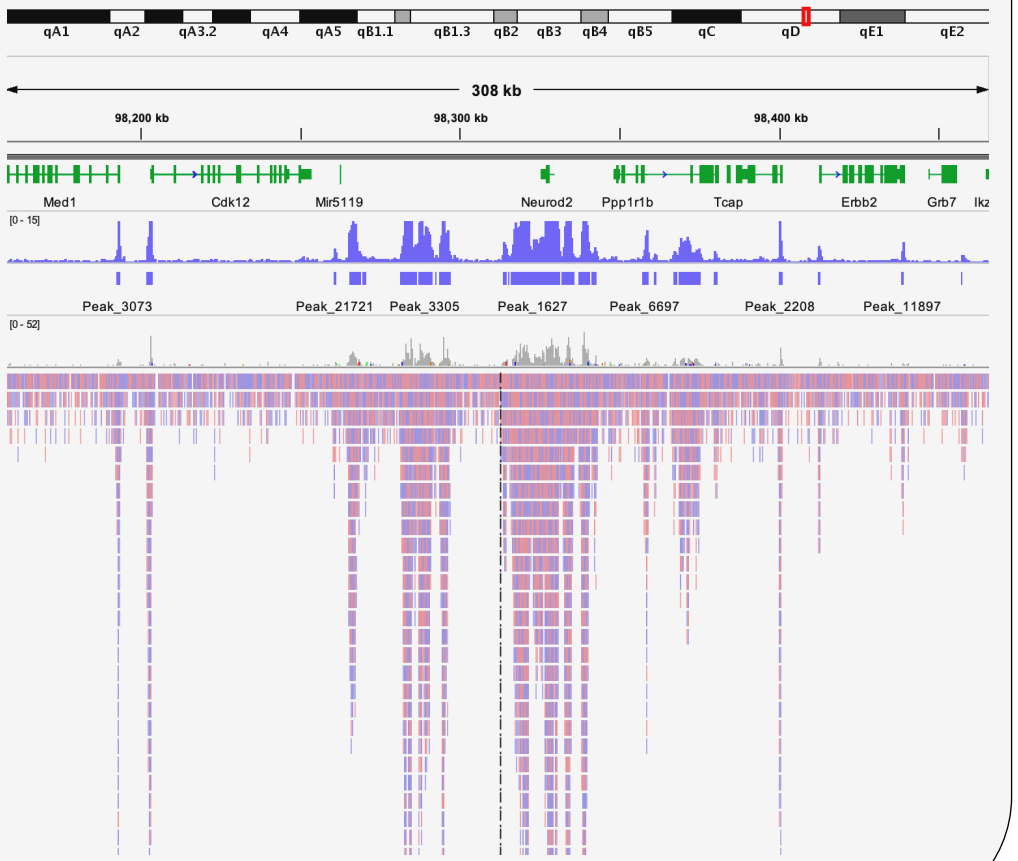


- **Many other file formats exist for storing specific types of data:**
  - FASTA
  - GTF/GFF
  - VCF
  - MAF

NGS data look very different depending on workflow

RNA-seq read alignments

ChIP-seq read alignments



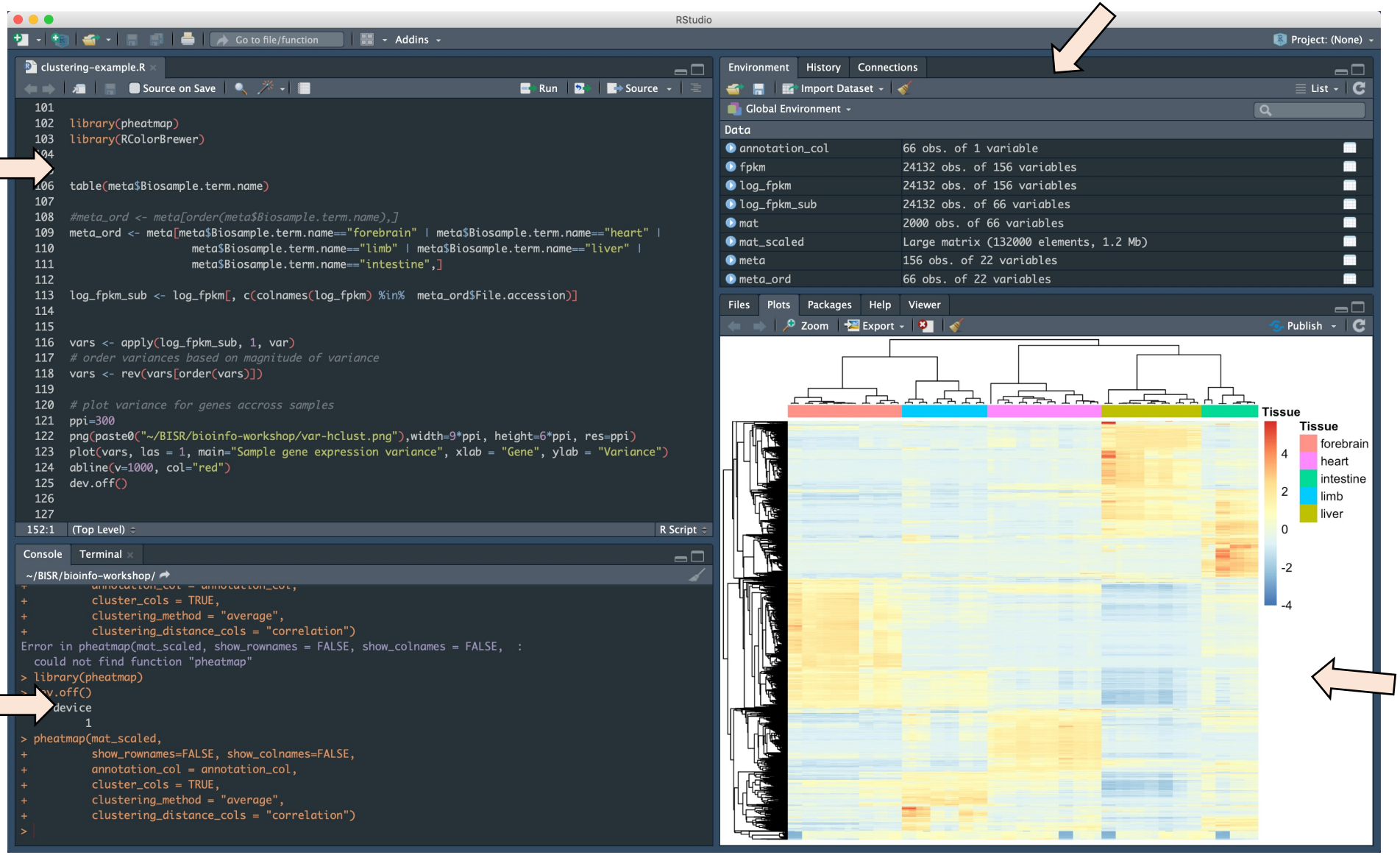
Analysis pipelines, quality control metrics, and possible hypothesis inherently differ

How is downstream analysis performed?

- R/python or other software that allow statistical programming

Environment

Scripting window

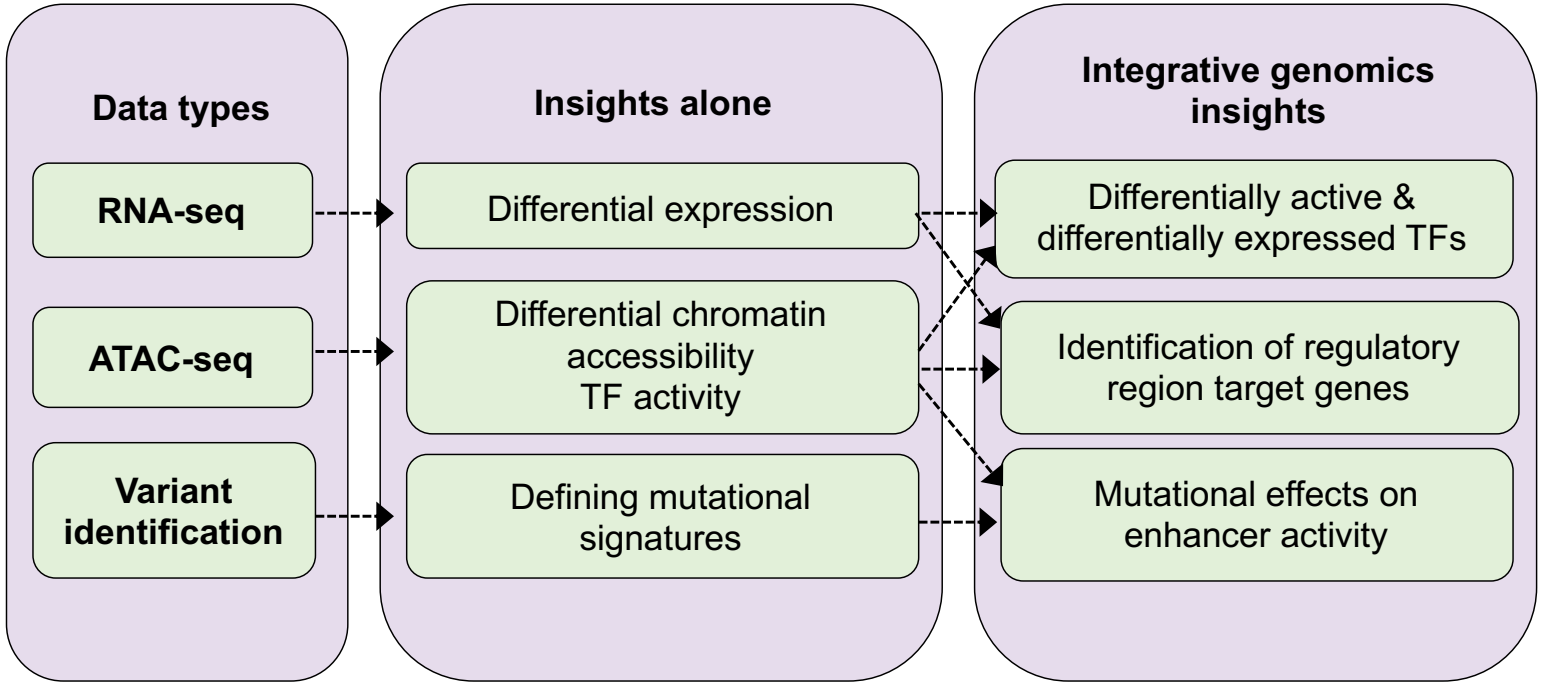


Console

Plotting window

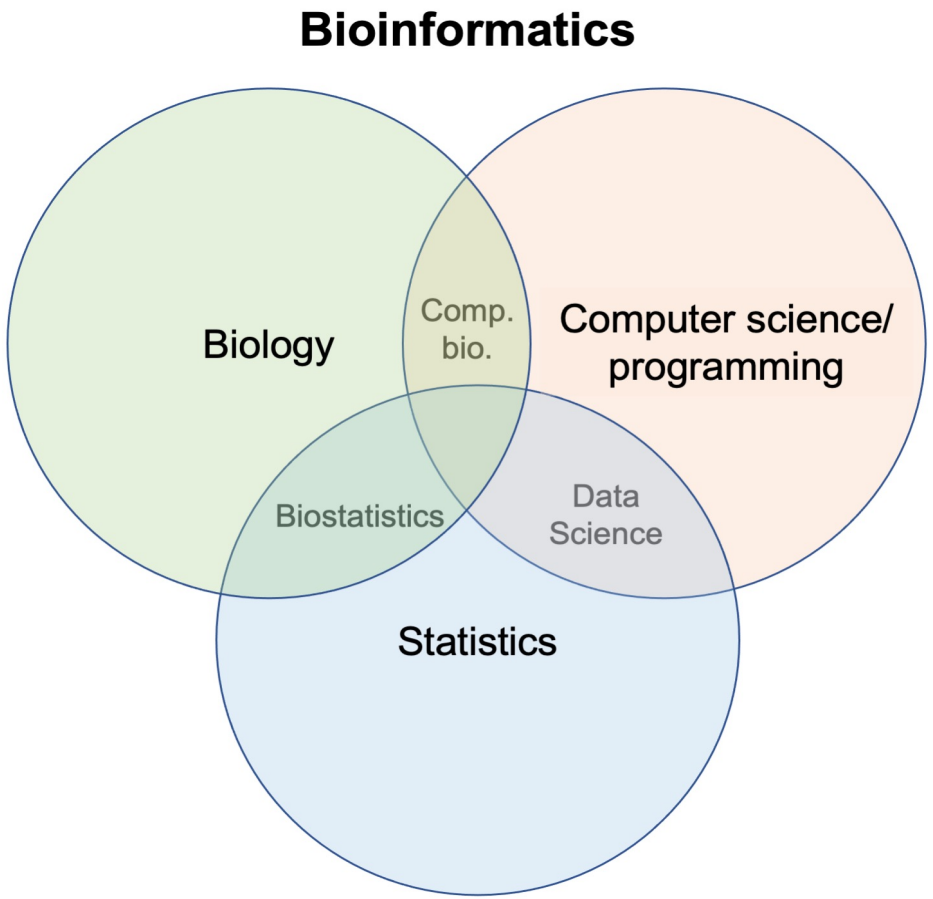
Advanced downstream analysis:
Integrative genomics

- *Leveraging data integration across more than one 'omics platform, to reveal insights not possible with each data type alone*
- *Requires some creativity*



- You may not have generated each dataset in-house (e.g. combine in house & public data)

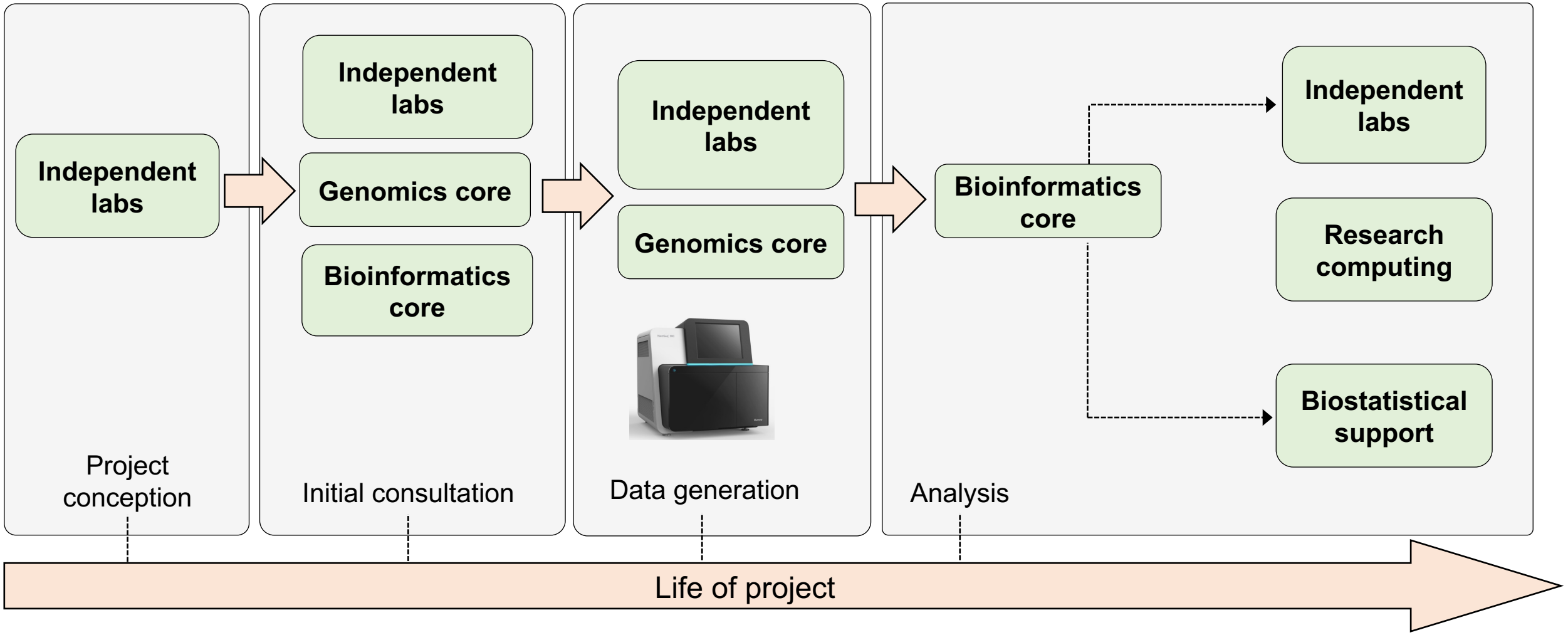
Skills sets for bioinformatics



- Bioinformatics requires an interdisciplinary skill set
- Familiarity with multiple experimental & analytical frameworks
- Emphasis of skill set depends on types of analysis you intend to perform
- Working knowledge of statistics to understand and apply methodology

When to start thinking about analysis?

- Ideally, before data generation..



Summary

- **Bioinformatics**: the use of complex computational & statistical approaches to analyze large and diverse collections of biological data
- **NGS bioinformatics involves two major steps: data reduction & downstream analysis**
- **Data reduction** involves producing a reduced representation of the dataset from which biological insights can be extracted
- **Downstream analysis** involves fitting statistical models to the data to generate insights
- **Bioinformatics** requires an interdisciplinary skill set

Questions?