



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



Dartmouth
GEISEL SCHOOL OF MEDICINE

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: Towards 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

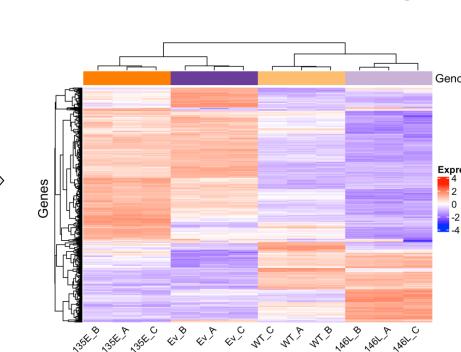


Can generate millions, 100s of millions, or billions of reads in FASTQ format (depending on instrument)

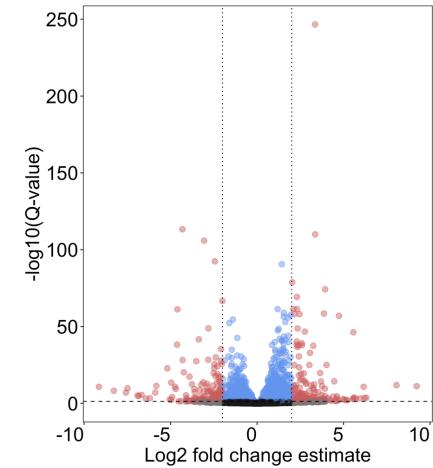
?

Biological insights

Hierarchical clustering



Differential expression



Common perception...

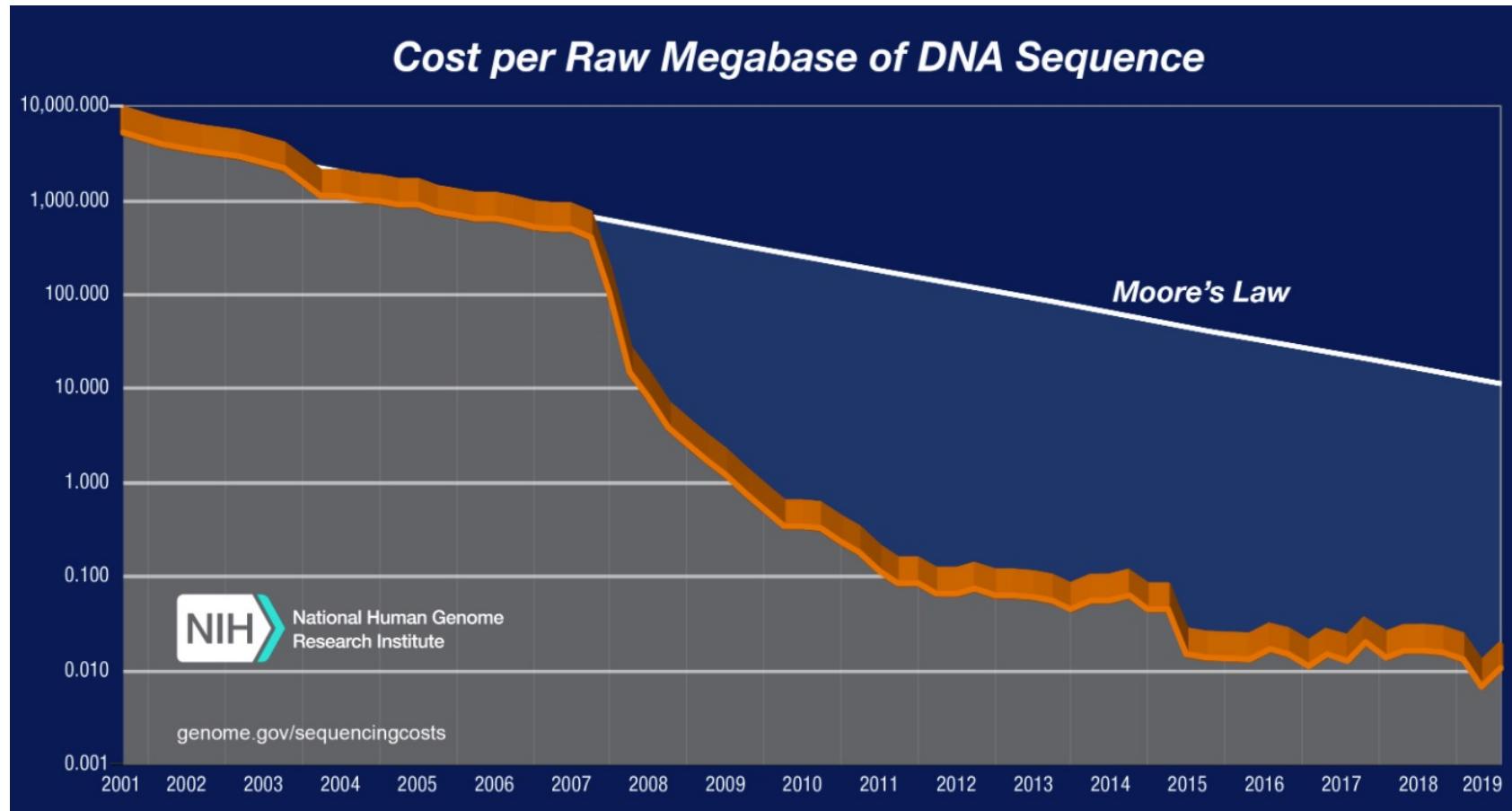


Reality..

The figure shows a Mac OS X desktop with several open windows:

- Terminal:** Multiple windows displaying command-line logs from a pipeline named "atac-seq-pipeline". The logs include details about sequencing runs, fastq files, and pipeline steps like "encode-atac-seq-pipeline".
- Genome Browser:** A window titled "IGV" showing a genomic track for "Human (hg38)" on chromosome 7. It displays genomic coordinates (chr7:140,753,191–140,753,451), gene tracks for BRAF, and coverage plots for "a74bab3c-39b..bam Coverage".
- File Manager:** A "Locations" sidebar showing paths like "/Users/omw/" and "Remote Disc". A preview pane shows files such as "hg38_annotations.bed.gz", "hg38.blacklist.bed", and "SRR04263456_1.fastq.gz".
- System Tray:** Shows battery level (35%), signal strength, and system status.

Data is getting cheaper & bigger



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

Complex analytics are playing a larger role



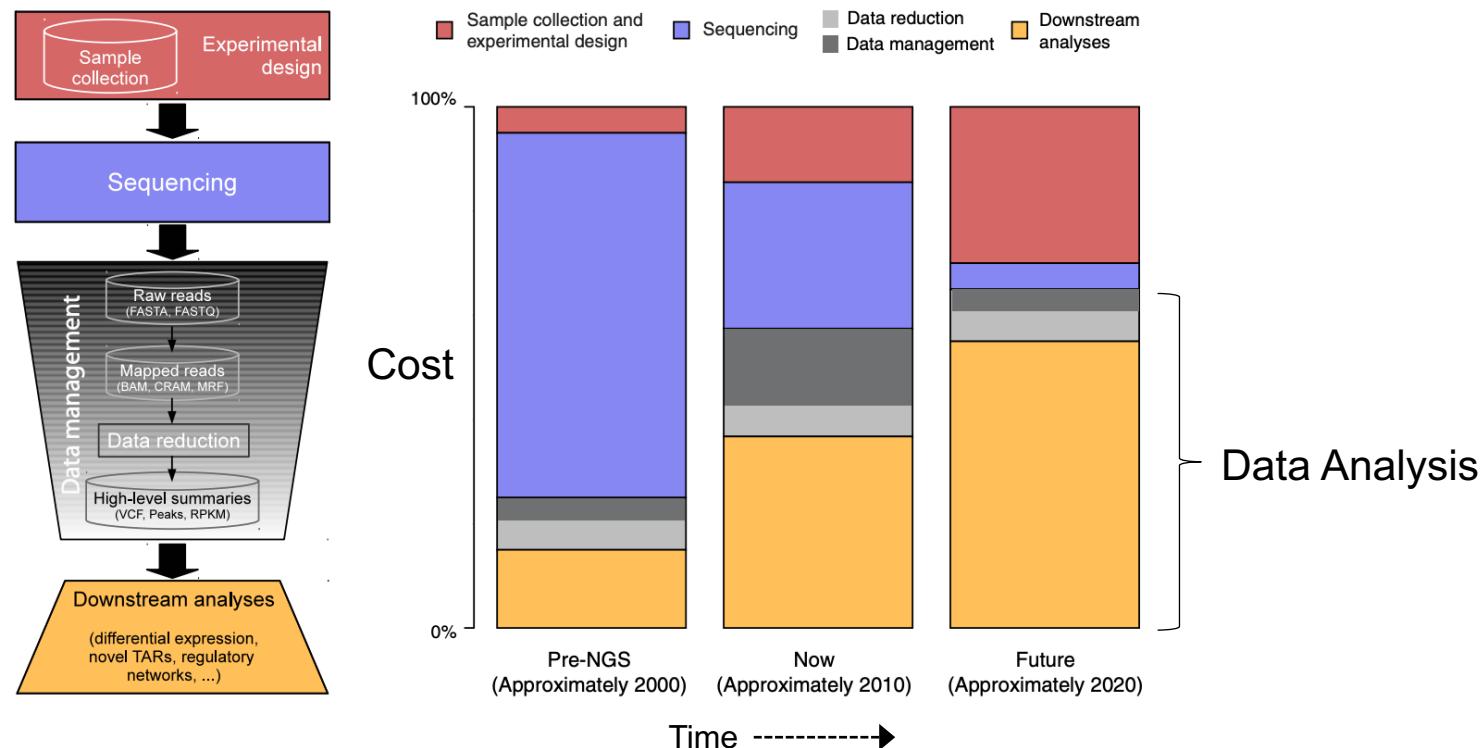
Sboner et al. *Genome Biology* 2011, 12:125
<http://genomebiology.com/2011/12/8/125>



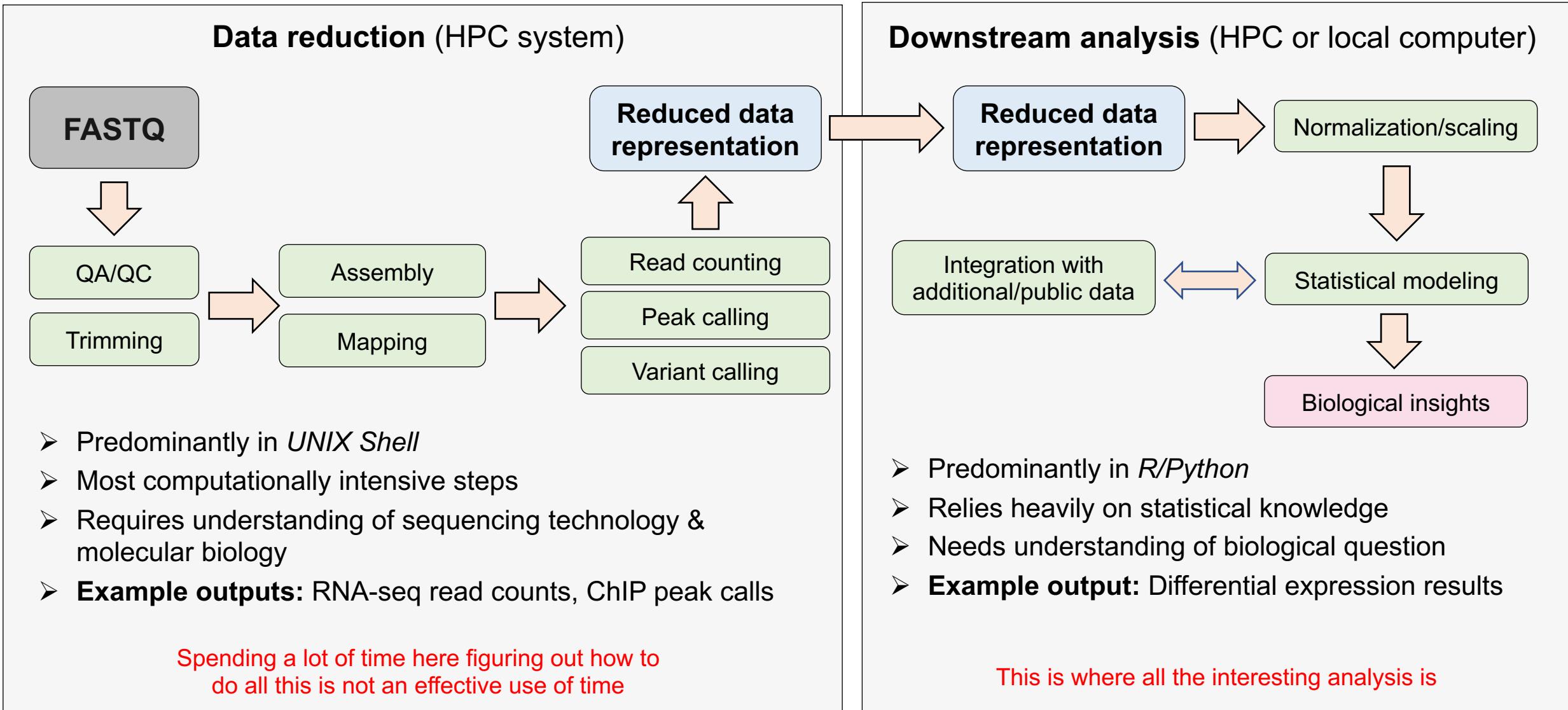
OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner^{1,2}, Xinmeng Jasmine Mu¹, Dov Greenbaum^{1,2,3,4,5}, Raymond K Auerbach¹ and Mark B Gerstein*^{1,2,6}



Two major components of genomic data analysis



How is data reduction performed?



ssh connection to discovery

```
d41294d@discovery7:/dartfs-hpc/rc/lab/G/GMBSR_bioinfo/Labs/Cramer/filtered-analysis/trim/fastq
Last login: Fri Dec 11 12:06:57 on ttys001
The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) [OwenW@vpn-two-factor-general-230-140-214 ~] $ sh ssh_to_dartfs_discovery.sh
d41294d@discovery7.dartmouth.edu's password:
Last login: Fri Dec 11 12:07:05 2020 from vpn-two-factor-general-230-140-214.dartmouth.edu

[Scheduled maintenance]

Questions? Email Research.Computing@dartmouth.edu
===== Research Computing Notices =====
[12/21/20-12/22/20] [Cluster Maintenance] the Discovery cluster will be down starting Monday
12/21/20 through Tuesday 12/22/20 5PM EST. Jobs cannot be submitted or running during this time.

Run 'notice' for more details of any announcements above.
[d41294d@discovery7 ~] $ ls
atac-seq-pipeline encode-atac-seq-pipeline.yml igv imputation public_html raw_counts.txt
bin fpkm.txt IGV_Linux_2.8.12_WithJava.zip ncbi R scripts
[d41294d@discovery7 ~] $ cd bioin
[d41294d@discovery7 GMBSR_bioinfo] $ cd Labs/Cramer/filtered-analysis/
[d41294d@discovery7 filtered-analysis] $ ls
alignment cramer-chip-rlmo-10-1 diffbind fastqscreen idr macs2 qc subsample trim
[d41294d@discovery7 filtered-analysis] $ cd trim
[d41294d@discovery7 trim] $ ls
fastq fastqc reports scripts tagged-filter-samples.txt
[d41294d@discovery7 trim] $ cd fastq
[d41294d@discovery7 fastq] $
[d41294d@discovery7 fastq] $ ls
RlMA_CASPO_1.trim.fastq.gz RlMA_Cont_1.trim.fastq.gz WT_CASPO_1.trim.fastq.gz WT_Cont_1.trim.fastq.gz
RlMA_CASPO_2.trim.fastq.gz RlMA_Cont_2.trim.fastq.gz WT_CASPO_2.trim.fastq.gz WT_Cont_2.trim.fastq.gz
RlMA_CASPO_3.trim.fastq.gz RlMA_Cont_3.trim.fastq.gz WT_CASPO_3.trim.fastq.gz WT_Cont_3.trim.fastq.gz
[d41294d@discovery7 fastq] $ zcat RlMA_CASPO_1.trim.fastq.gz | head -n 12
@NB501031:462:HFVVAFX2:1:1101:9919:1033 1:N:0:AGCGATAG+GCCCTCTAT#F0ST:Human:Mouse:Rat:AfumigatusA1163:Drosophila:Worm:Yea
st:Arabidopsis:Ecoli:rRNA:MT:PhiX:Lambda:Vectors:Adapters:0001000000000000
TACGAGTATTTCTGGCTGGATATAGTCGGATCACCTAAATCTGTGTATTCCATCCACATCTGTGTACTTCTGGCTCGTTCTCGGGCGTCTATCGAAGTCTTGTGCG
AGGACGCTCTCCCCCGAATCTTAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
EEEEAE/E/AAAAEEA<EEAA/AE<EEAA
@NB501031:462:HFVVAFX2:1:1101:7406:1037 1:N:0:AGCGATAG+GCCTCTGT#FQST:0001000000000000
TAGAATAGAAGCTAACTGTTGAGTATTAGAGCTGCGTGGGTTAGGCCATAGGAATTCAAATAAACGCTAGCTAGGTTAGGTTAAAGATAGGAAATGTAGTAC
CACCTGTAAAATAATTAGGTATATA
+
Line 40, Column 62
```

Text editor for shell scripting

```
run-bowtie2.sh
echo "Sample name is $base"
11 # set up file names for those files to be created
12 align_out=${base}_unsorted.sam
13 align_bam=${base}_unsorted.bam
14 align_sorted=${base}_sorted.bam
15 align_filtered=${base}_sorted.filtered.bam
16 align_filtered_nodups=${base}_sorted.filtered.nodups.bam
17
18 # directory with bowtie genome index
19 genome=FungiDB-46_AfumigatusA1163
20
21 # print which file we are working on
22 echo "Processing file $fq"
23
24 # Run bowtie2
25 bowtie2 -p 10 -q --local \
26 -x ../../genomic_references/fungi/bowtie2-index/$genome \
27 -U subsample/$fq \
28 -S alignment/$align_out 2> alignment/${base}_bowtie2.log
29
30 # Create BAM from SAM
31 samtools view -h -S -b -@ 6 -o alignment/$align_bam alignment/$align_out
32
33 # Sort BAM file by genomic coordinates
34 sambamba sort -t 6 -o alignment/$align_sorted alignment/$align_bam
35
36 # Filter out unmapped and multimapping
37 sambamba view -h -t 6 -f bam -F "[XS] == null and not unmapped" \
38 alignment/$align_sorted > alignment/$align_filtered
39
40 # is finding duplicates in this way appropriate for SE data?
41 picard MarkDuplicates \
42 I=alignment/$align_filtered O=alignment/$align_filtered_nodups \
43 M=alignment/${align_filtered}.dedup_metrics.txt \
44 VALIDATION_STRINGENCY=LENIENT \
45 REMOVE_DUPLICATES=true \
46 ASSUME_SORTED=true
47
Line 40, Column 62
Tab Size: 4 Bourne Again Shell (bash)
```

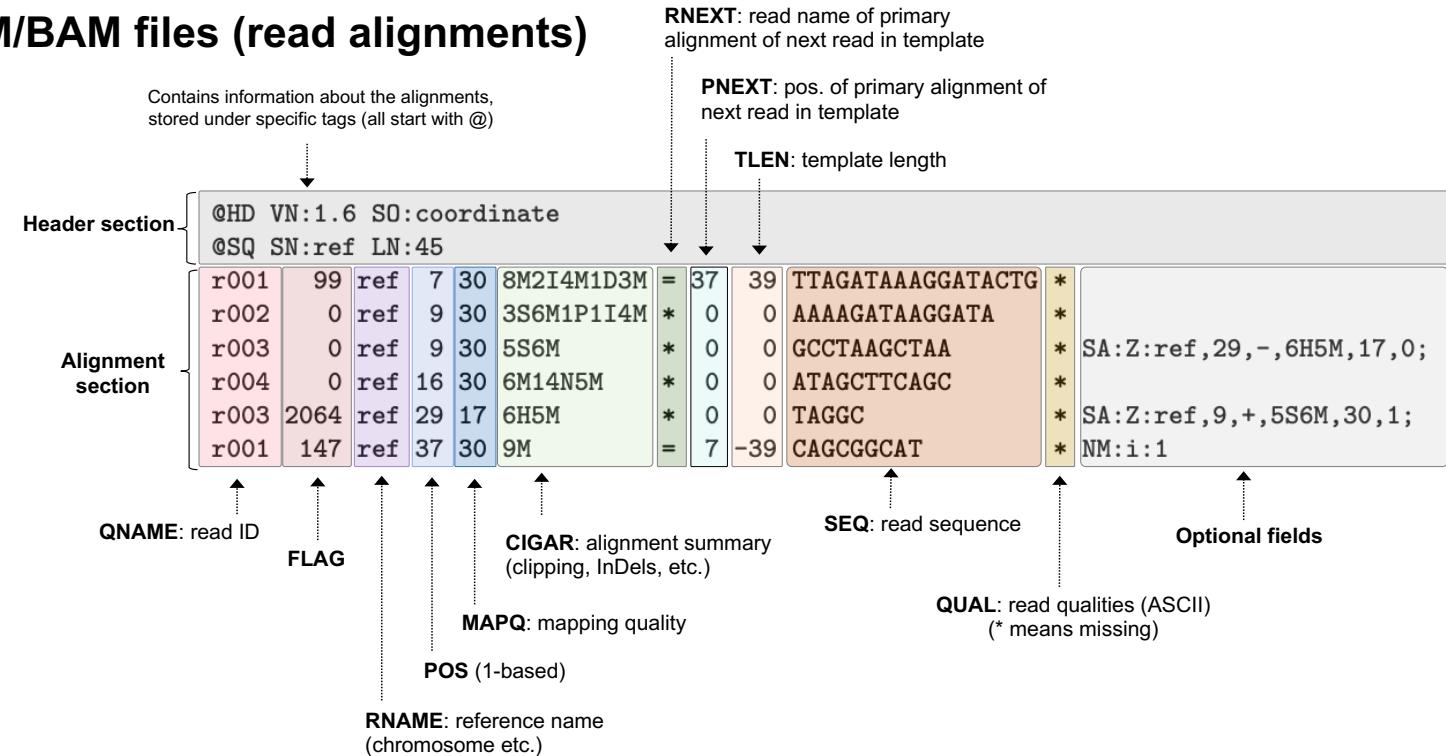
NGS data reduction involves many distinct file formats

FASTQ files (raw sequencing reads)

```
@SRR1039508.1 HWI-ST177:290:C0TECACXX:1:1101:1225:2130 length=63
CATTGCTGATACCAANNNNNNNNCATTCTCAAGGTCTCCTCCCTACGGAATTACA
+
HJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJHHHFFFFFD
```

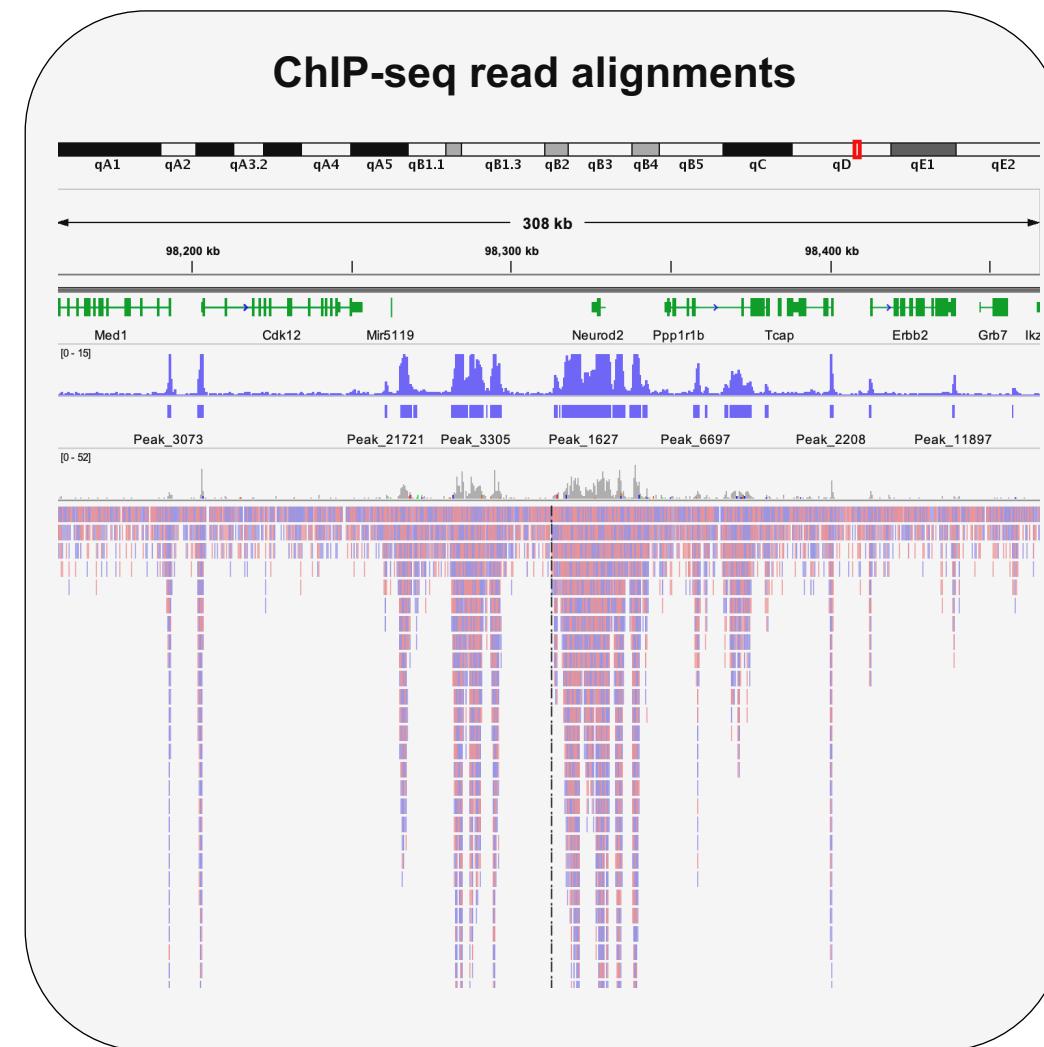
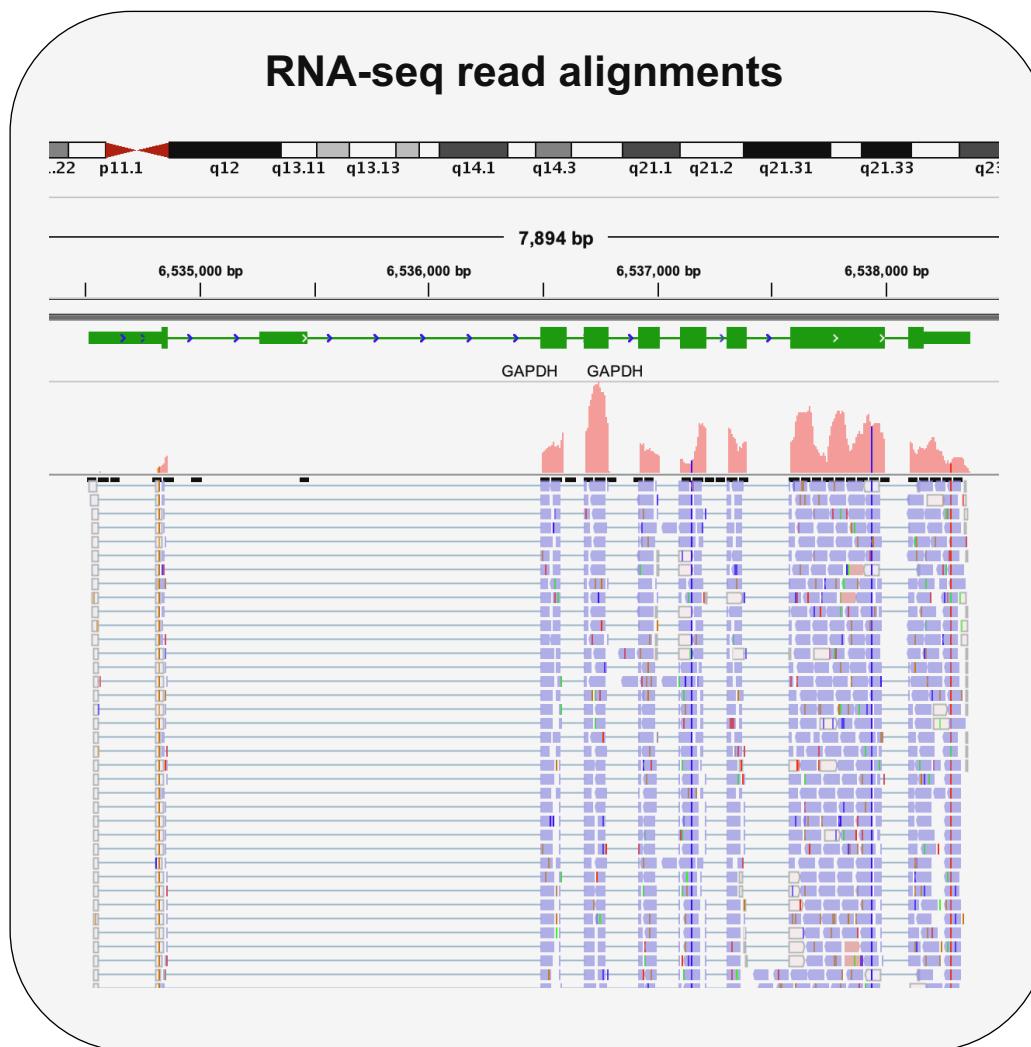
Line 1: Header line (read name)
Line 2: Base calls from sequencer
Line 3: Usually just a +
Line 4: Base quality scores

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

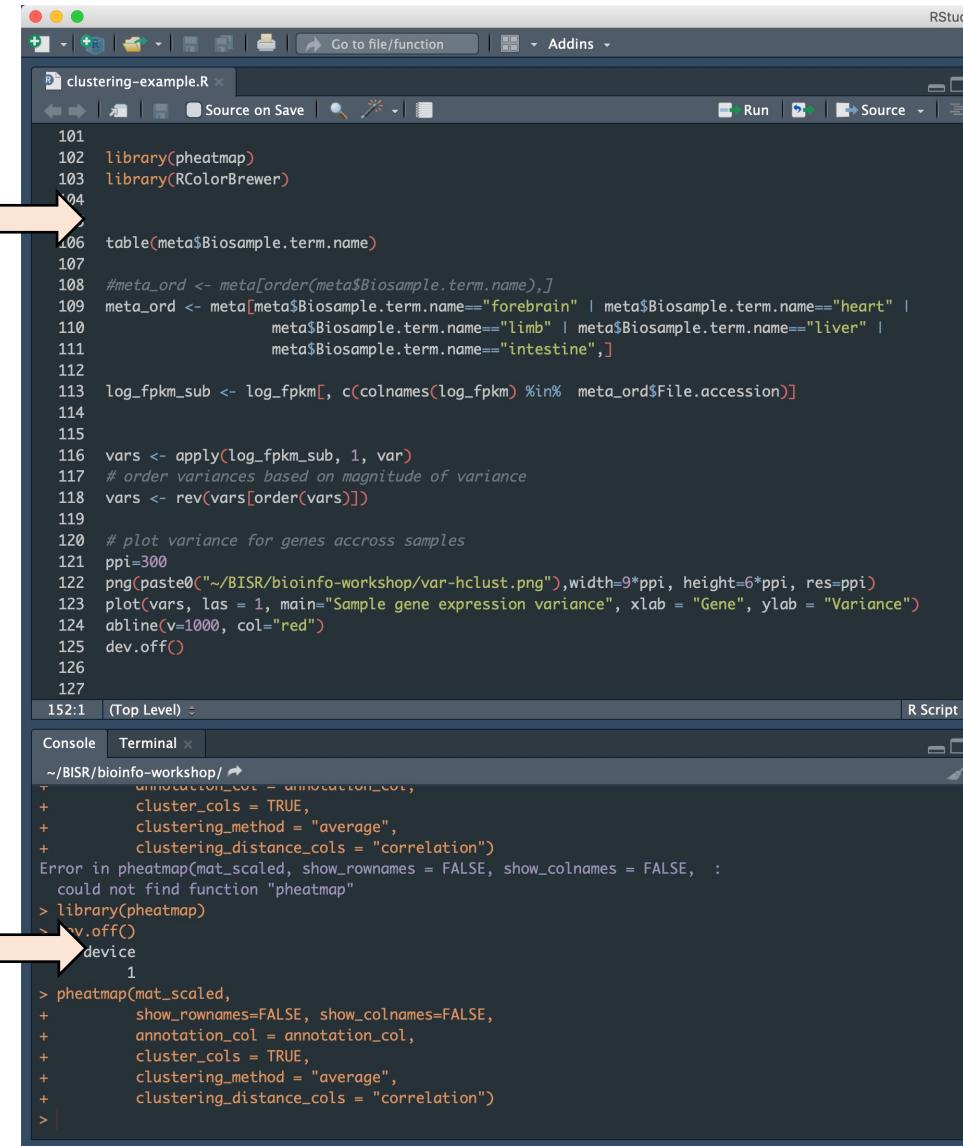


Analysis pipelines, quality control metrics, and possible hypothesis inherently differ

How is downstream analysis performed?

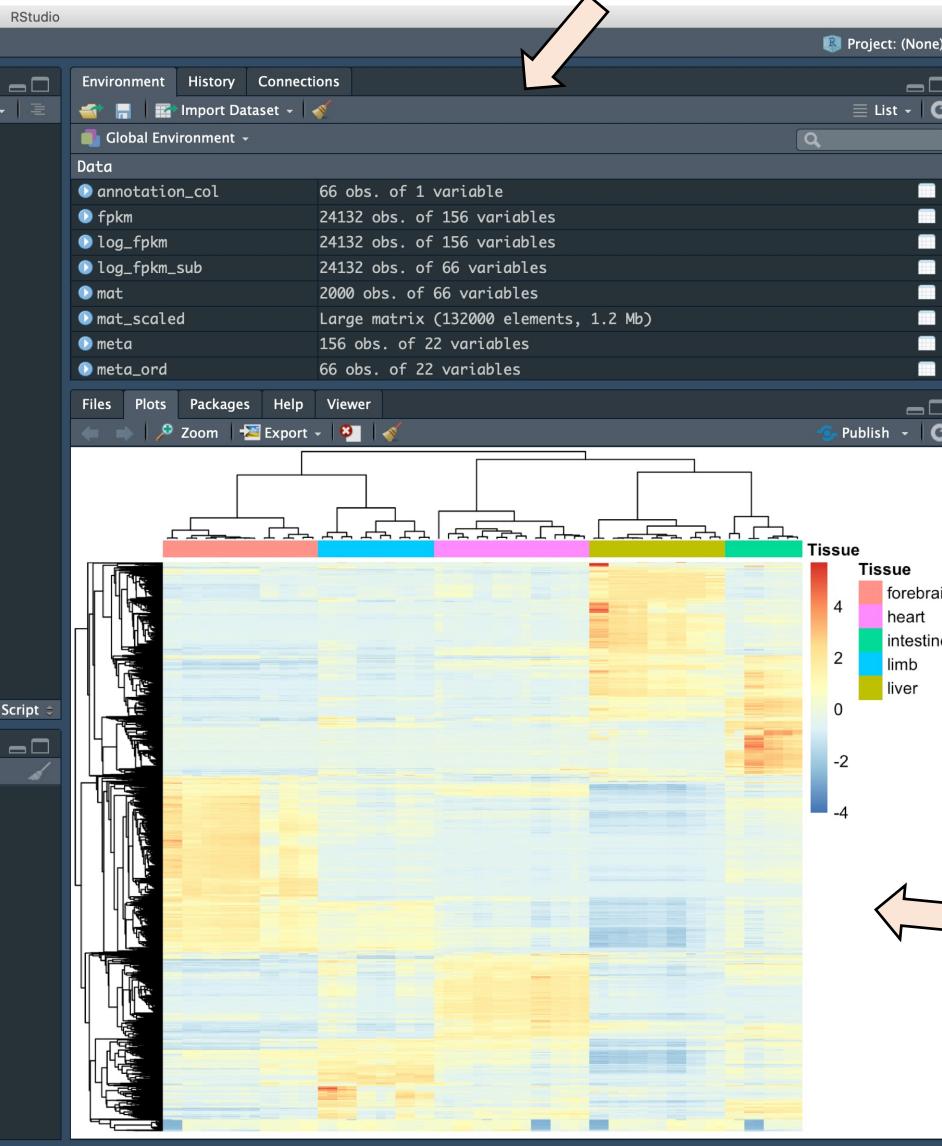
- R/python or other software that allow statistical programming

Scripting window



```
101 library(pheatmap)
102 library(RColorBrewer)
103
104 table(meta$Biosample.term.name)
105
106 #meta_ord <- meta[order(meta$Biosample.term.name),]
107 meta_ord <- meta[meta$Biosample.term.name=="forebrain" | meta$Biosample.term.name=="heart" |
108                 meta$Biosample.term.name=="limb" | meta$Biosample.term.name=="liver" |
109                 meta$Biosample.term.name=="intestine",]
110
111 log_fpkm_sub <- log_fpkm[, c(colnames(log_fpkm) %in% meta_ord$file.accession)]
112
113 vars <- apply(log_fpkm_sub, 1, var)
114 # order variances based on magnitude of variance
115 vars <- rev(vars[order(vars)])
116
117 # plot variance for genes across samples
118 ppi=300
119 png(paste0("~/BISR/bioinfo-workshop/var-hclust.png"),width=9*ppi, height=6*ppi, res=ppi)
120 plot(vars, las = 1, main="Sample gene expression variance", xlab = "Gene", ylab = "Variance")
121 abline(v=1000, col="red")
122 dev.off()
123
124
125
126
127
```

Console

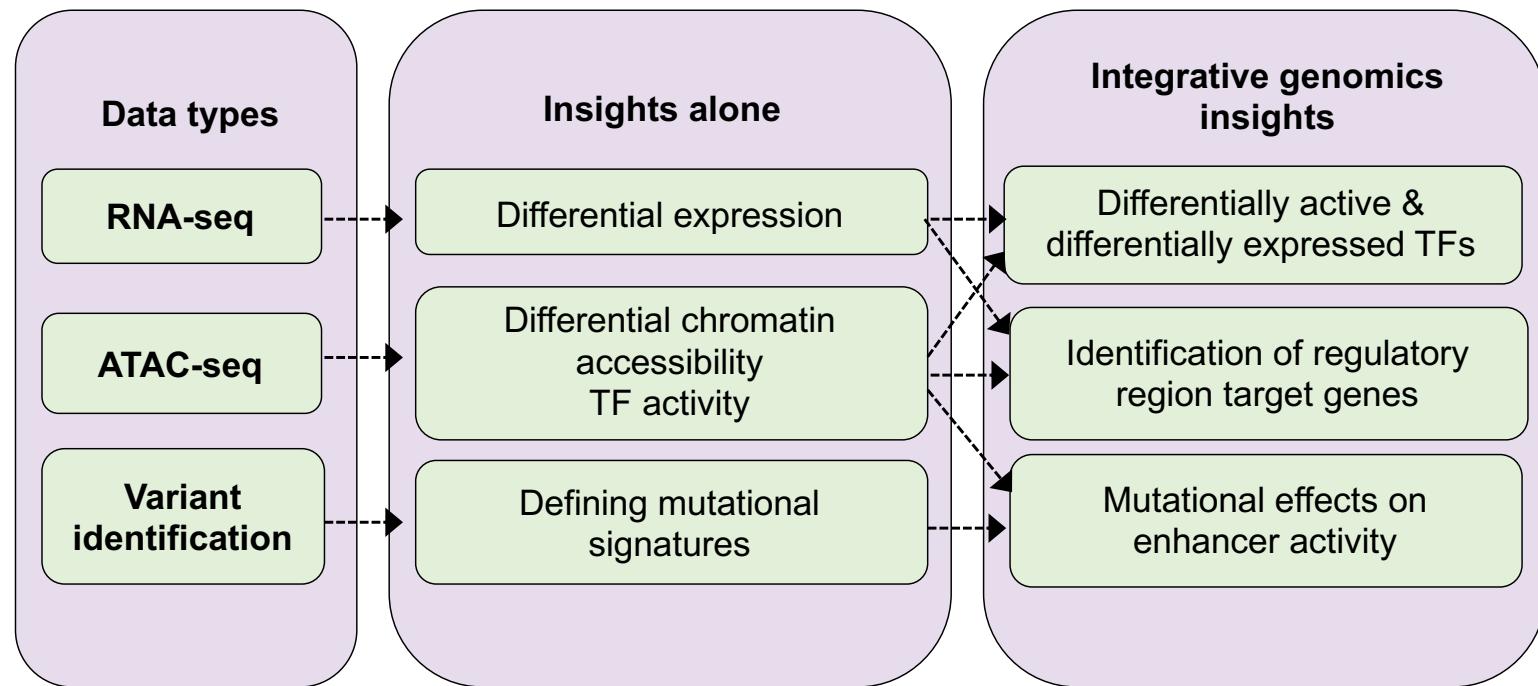


Environment

Plotting window

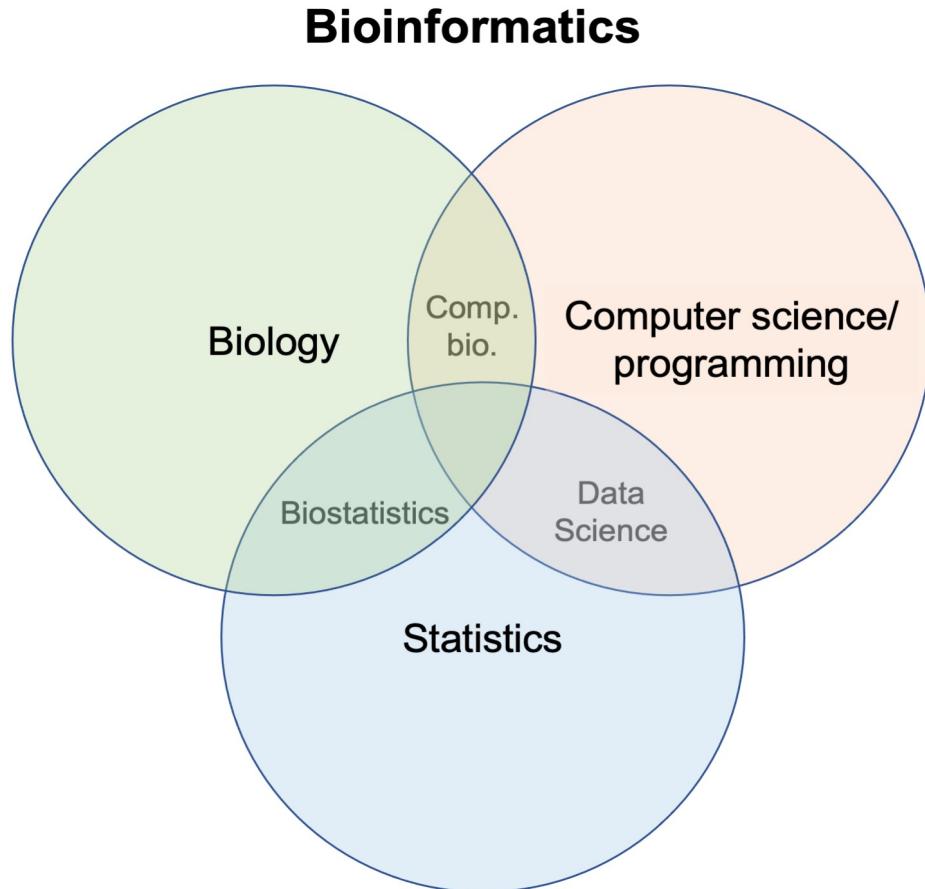
Advanced downstream analysis: Integrative genomics

- 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

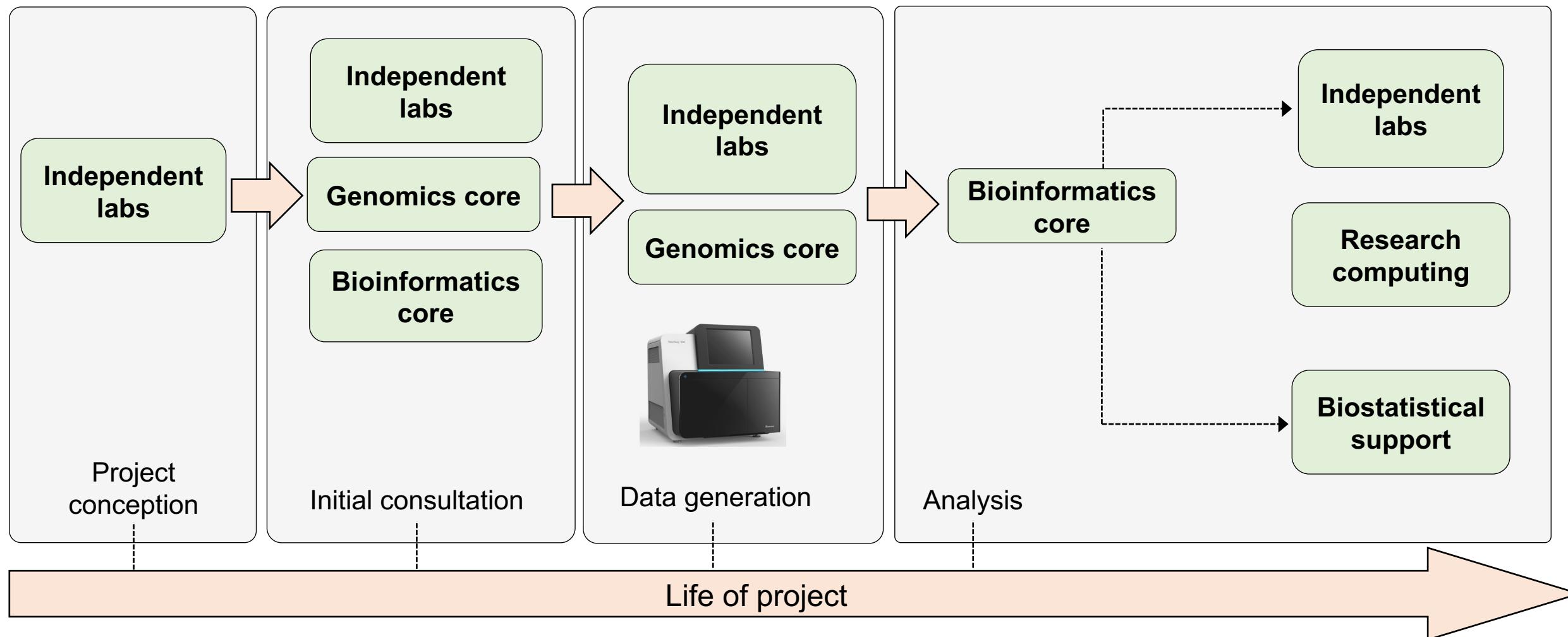


- Bioinformatics requires an interdisciplinary skill set
- Familiarity with multiple experimental & analytical frameworks
- Emphasis of skill set depends on types of analysis you intend to perform
- Working knowledge of statistics to understand and apply methodology

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?