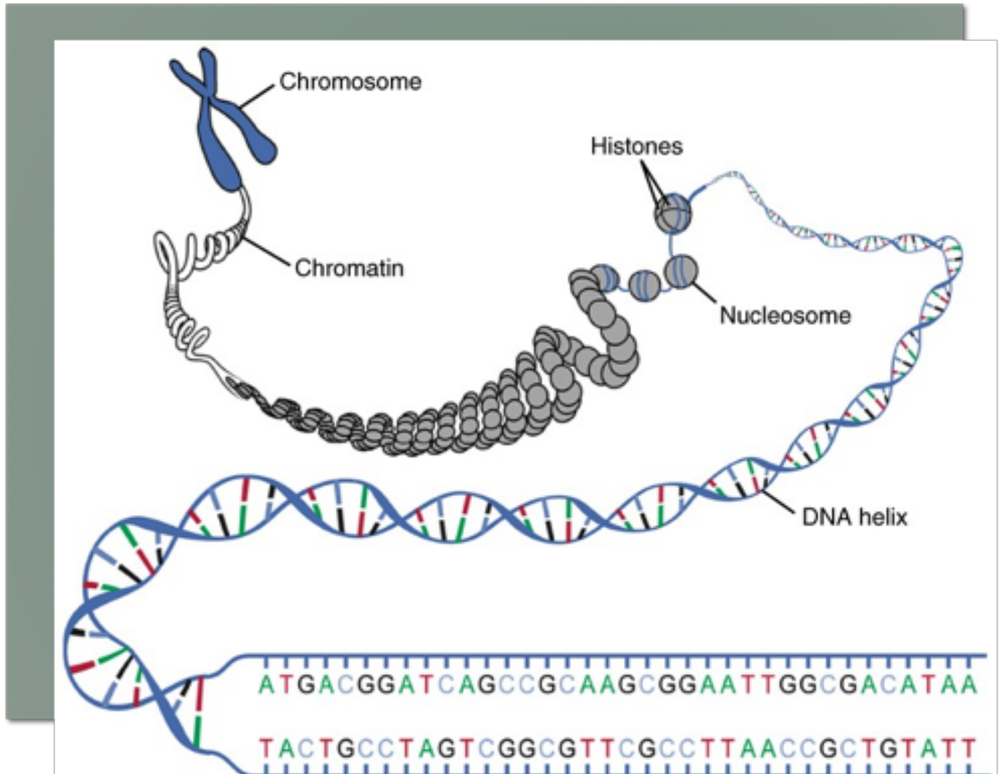
Intro to Bioinformatics

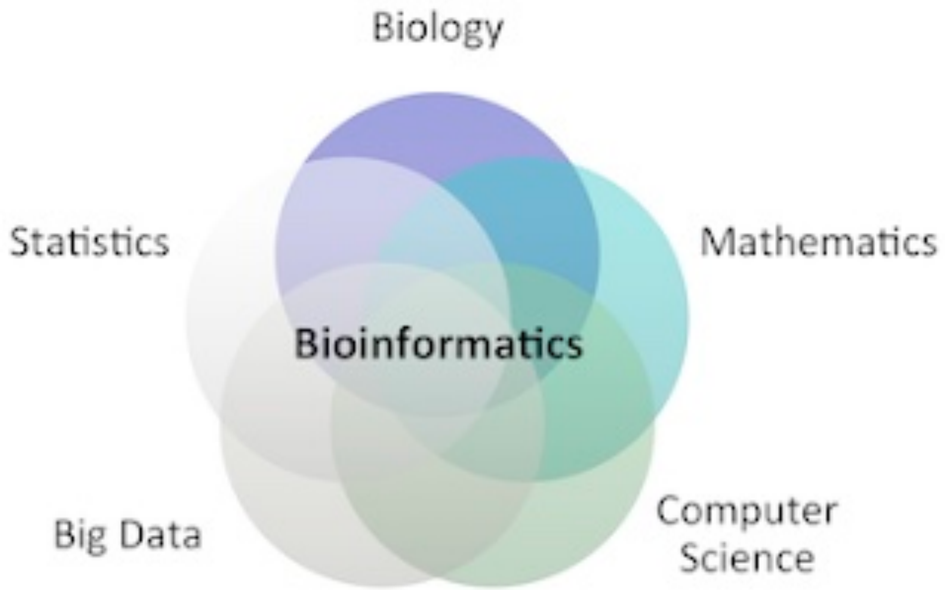
Susan Strickler



What we will cover:

1. Background
2. ‘Omics
3. Databases
4. Operating systems used in bioinformatics
5. Programming languages used in bioinformatics
6. Data Best Practices
7. Computational infrastructure
8. Careers

1. Background: What is bioinformatics?

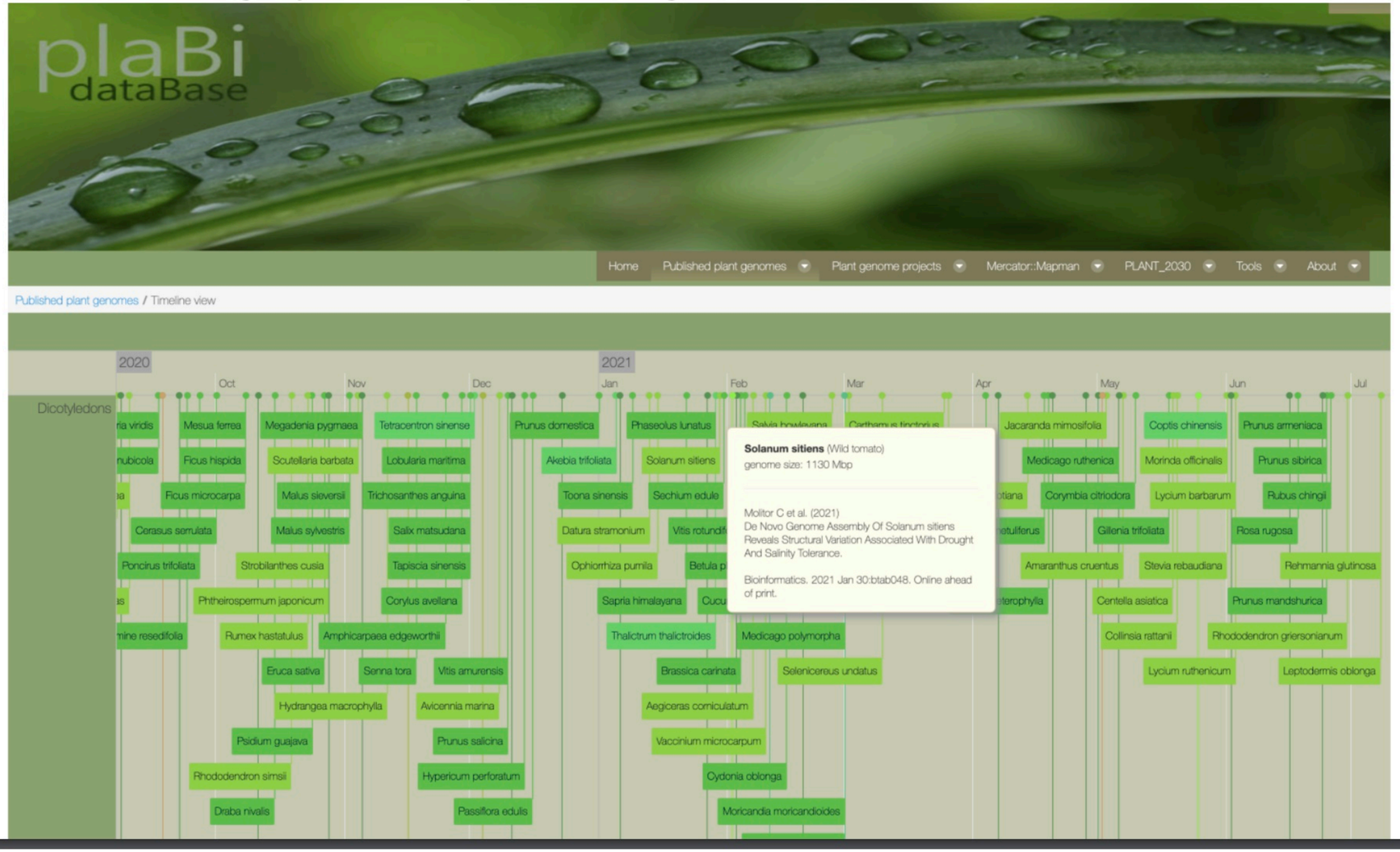


Bioinformatics can...

- Identify similar sequences
- Provide a putative function for a sequence
- Assemble sequences (genomes, transcriptomes)
- Annotate genomes
- Identify differentially expressed genes
- Build networks of genes or metabolites
- Determine phylogenetic relationships
- Mine literature for biological information
- Uncover differences between two genomes
- Calculate how a protein folds
- Find trends in large datasets
- Speed up your research
- Enable you to ask new questions

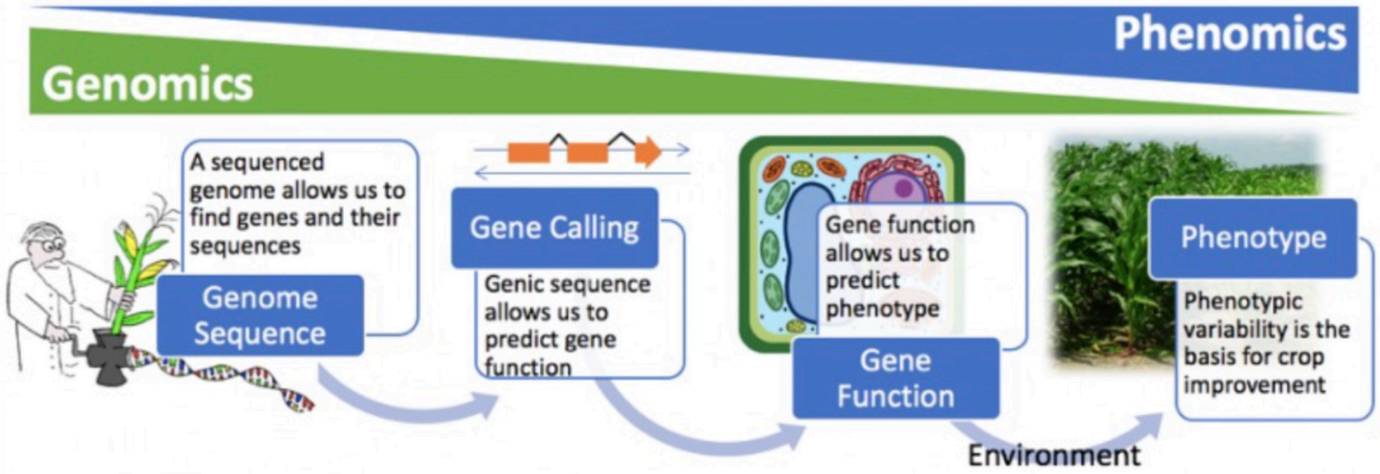
- ‘Omics: genomics, transcriptomics, proteomics, metabolomics……

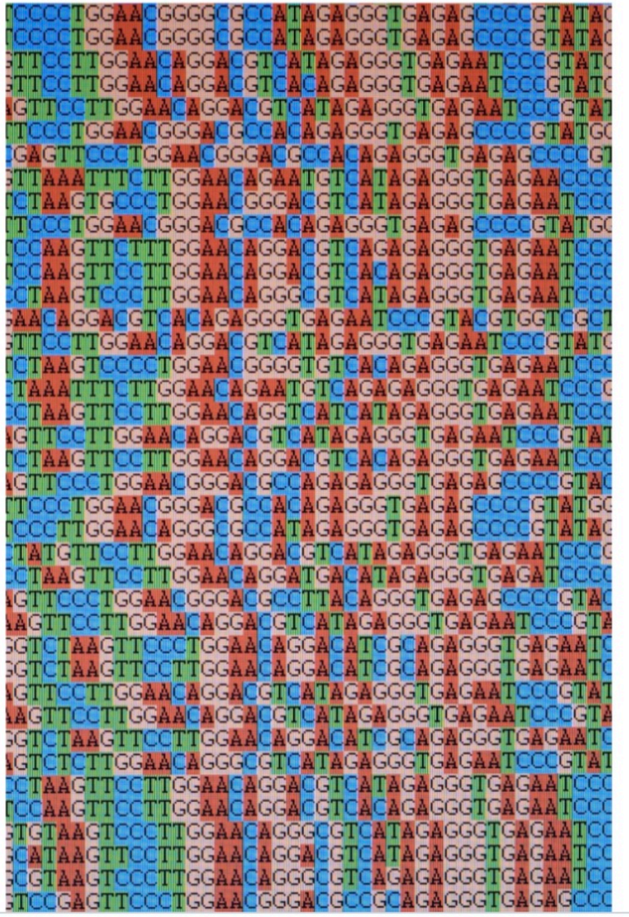
Genomics



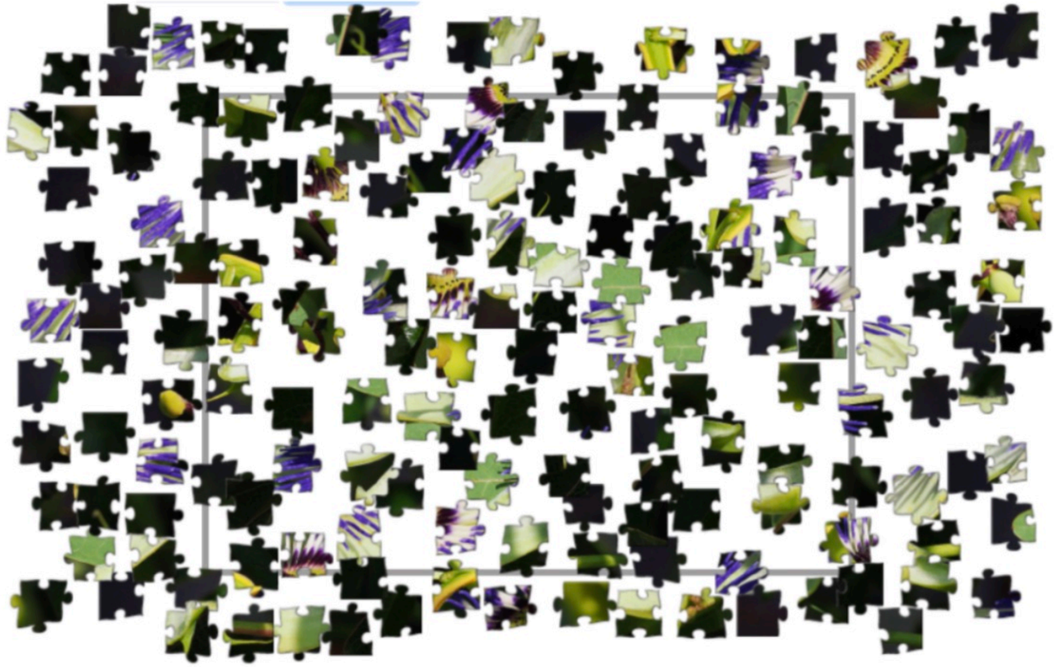
- The study of genomes

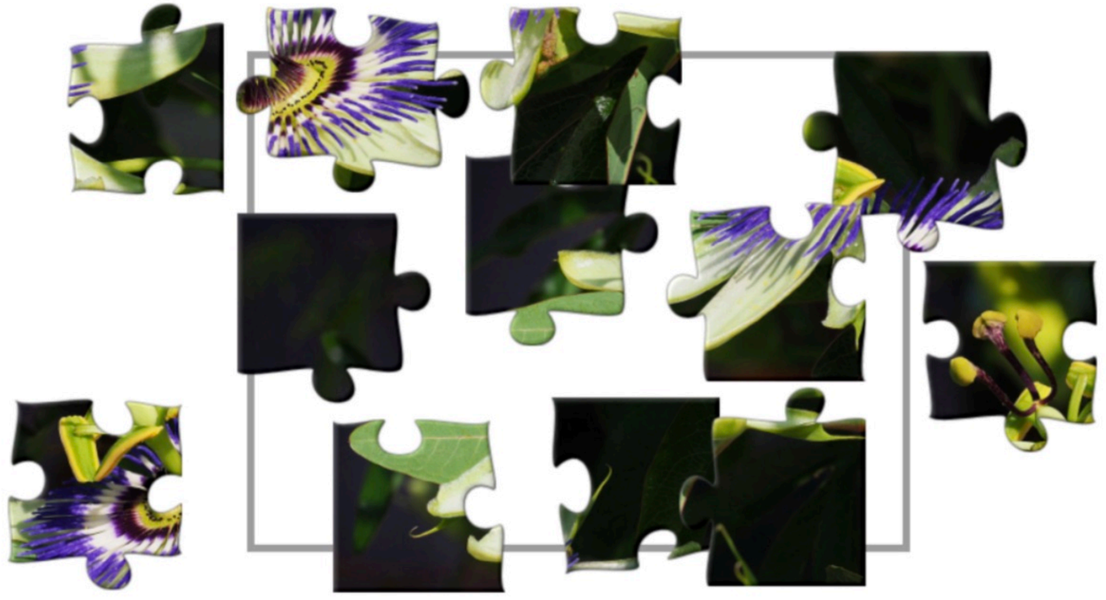
Genome as a puzzle



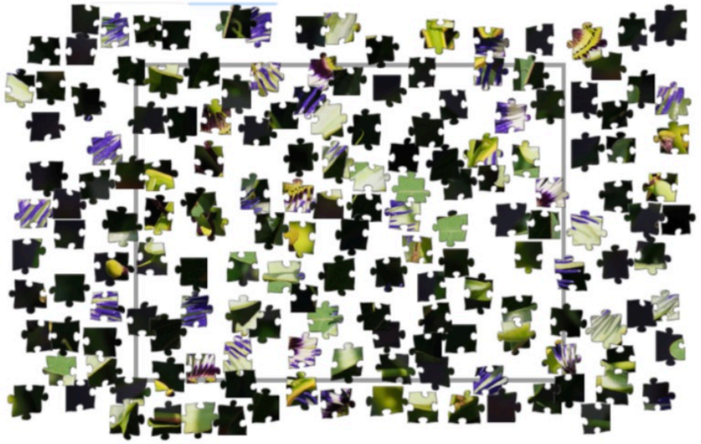


Which puzzle is easier to put together?



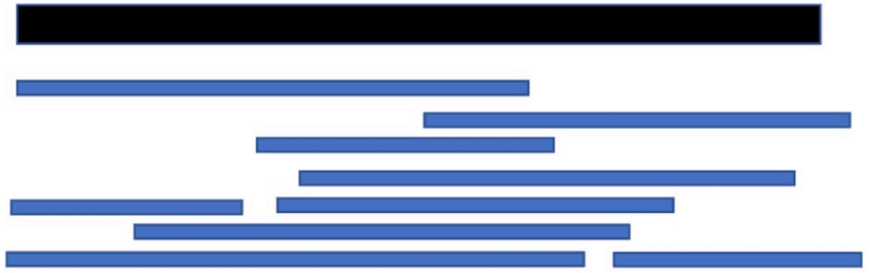


In the world of genomes



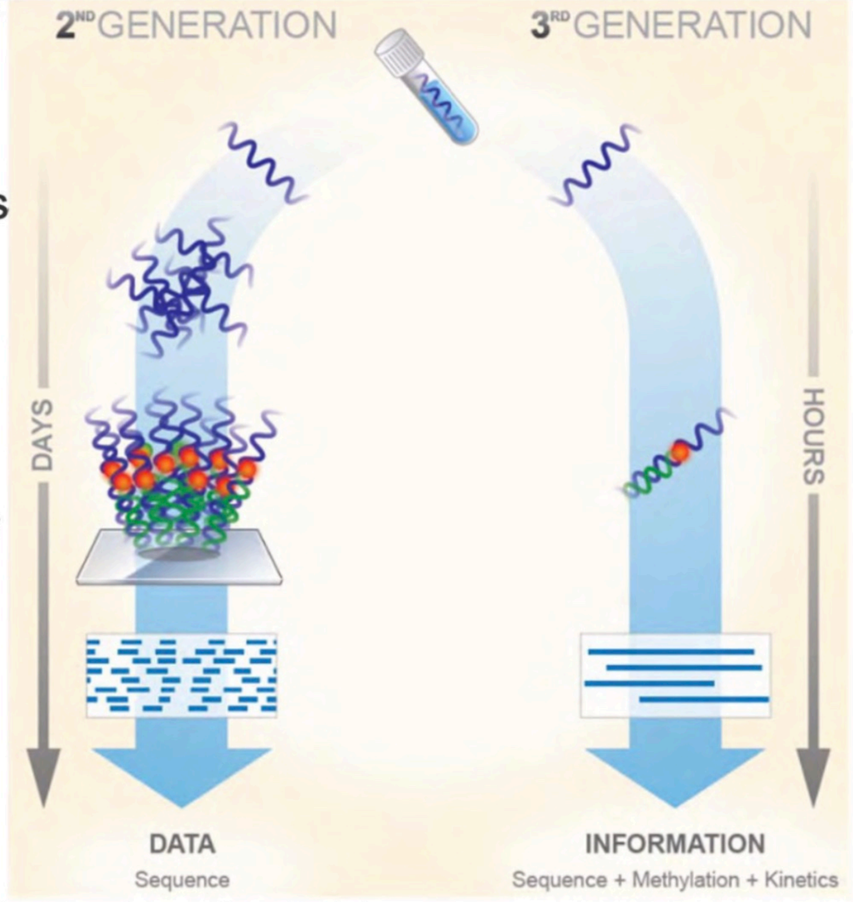
Reference Genome

Sequencing Reads



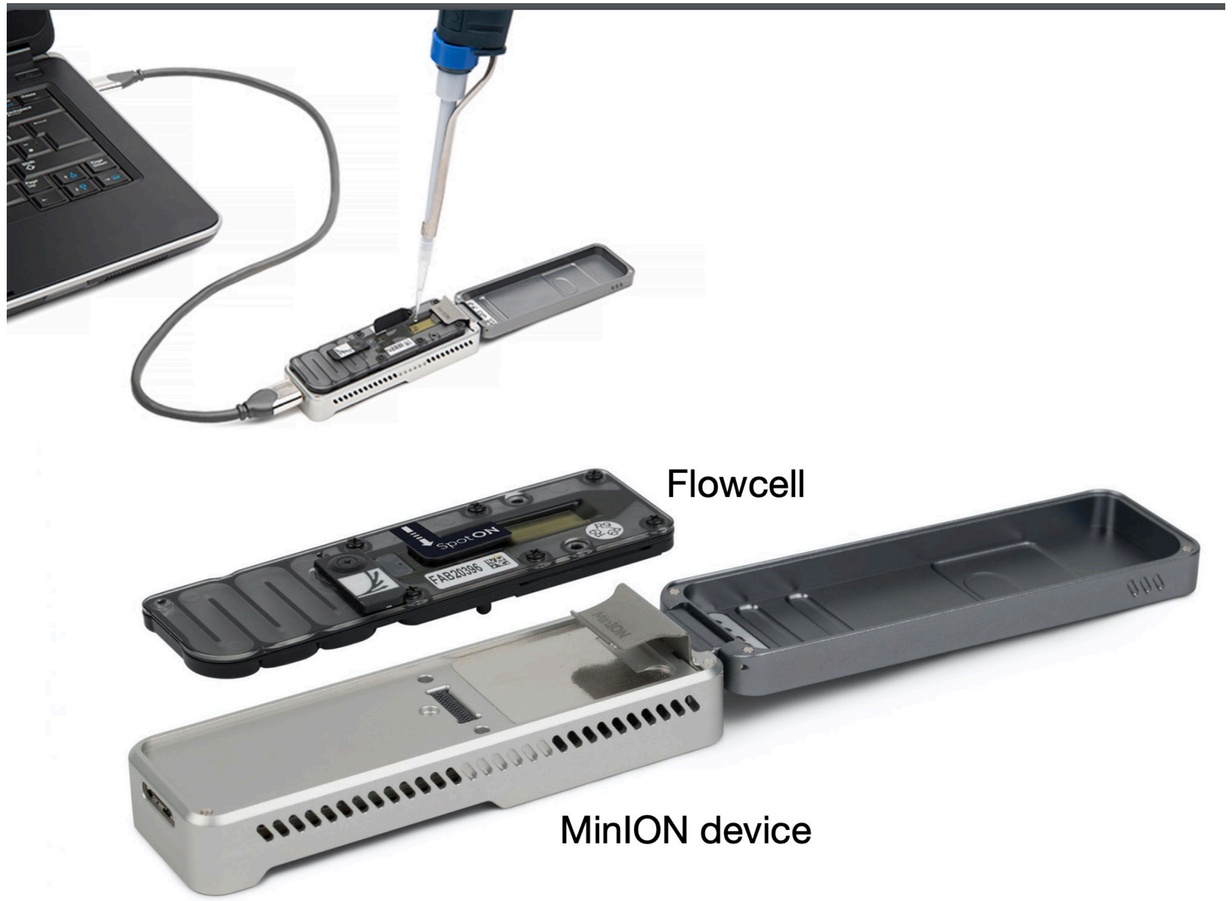
Short vs Long-reads

- Short reads
- Amplification errors and bias
- Several enzymatic steps
- Multi-molecule raw accuracy
- Errors tend to be systematic
- More coverage required



https://get.genotoul.fr/wp-content/uploads/2017/06/150519_Technology_and_Applications_PACBIO_GerritKuhn_low.pdf

- Long reads
- No required amplification
- Simple sample prep
- Single molecule raw accuracy
- Errors tend to be random (vs. systematic)
- Less coverage required


Flowcell
MinION device

de novo assembly

- Illumina only
  - High quality reads with fewer errors
- Hybrid option
  - Nanopore or PacBio + Illumina
  - Either raw or error corrected long-reads
- Long-read only
  - Raw typically works better
  - Need to polish after with Illumina data to fix errors
- Recommendation is 50x coverage short-reads and 50x long-reads
- So how much data do I need?

Short-read

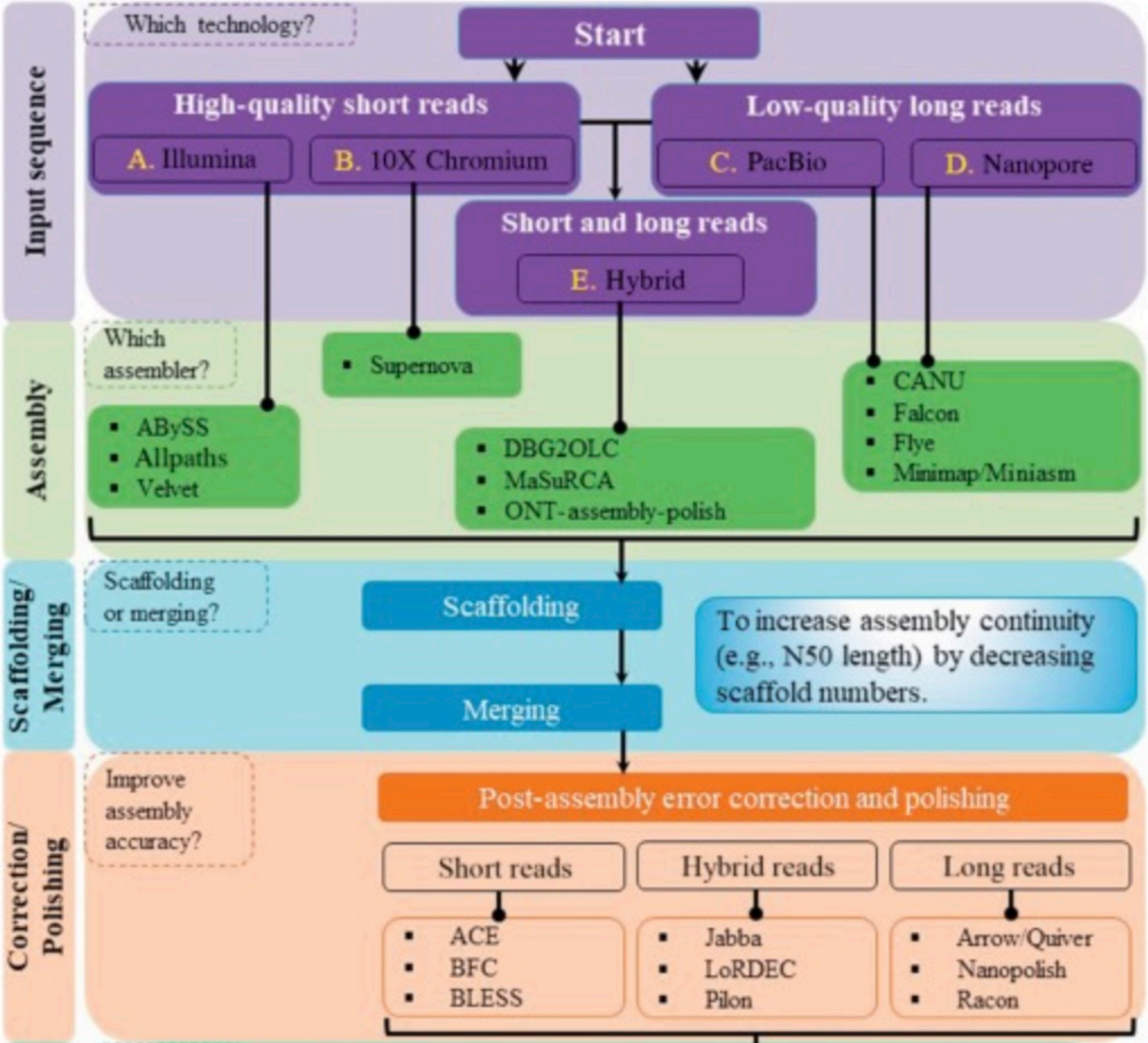
Assemblies
- Fragmented
- Small N50: 10s–100s of kbp
- Very accurate

Uses
- SNPs
- Phylogenetics
- Specific alleles

Long-read

Assemblies
- Complete
- 98+% accuracy

Uses
- Large-scale structure
- Horizontal gene transfer



Trends in Plant Science
Volume 24, Issue 8, August 2019, Pages 700-724

Feature Review

Tools and Strategies for Long-Read Sequencing and *De Novo* Assembly of Plant Genomes

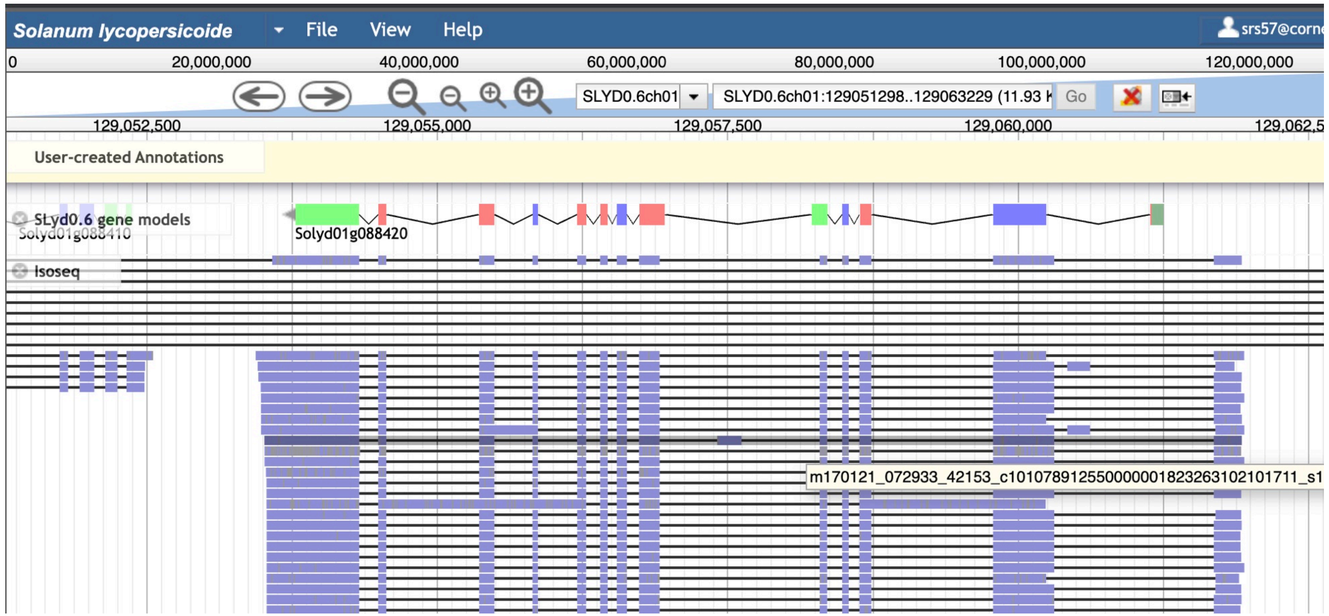
Hyungtaek Jung [1], Christopher Winefield [2], Aureliano Bombarely [3, 4], Peter Prentis [5], Peter Waterhouse [1, 6]

For more on genome assembly

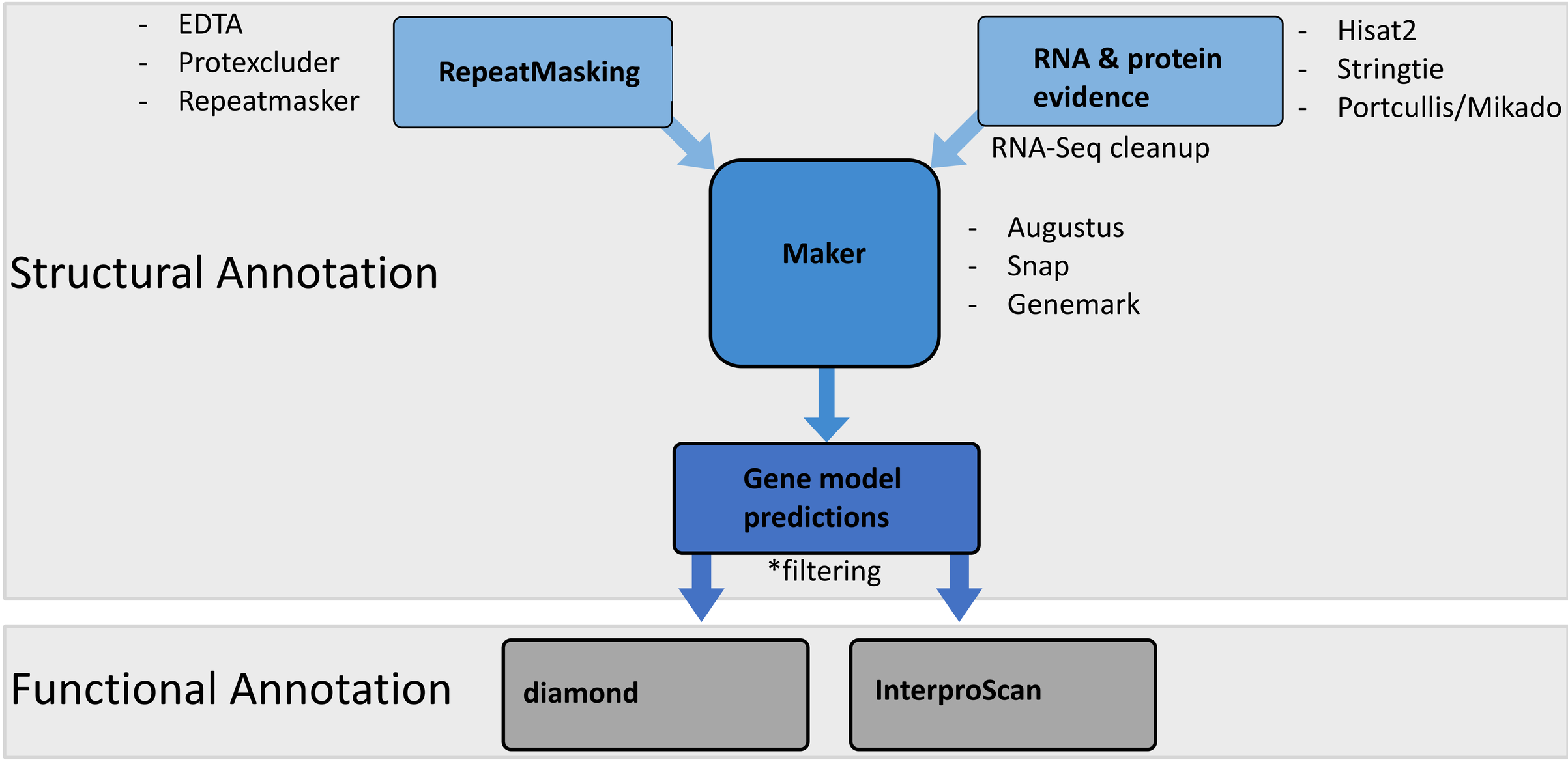
- https://github.com/bcbc-group/Botany2022NMGWorkshop/blob/main/4.GenomeAssembly/Botany2022_assembly_workshop_slides.pdf

Goals of genome annotation

- Predict, categorize, and mask repetitive elements
- Determine gene structures as accurately as possible
- Predict possible functions of predicted genes
- Associate GO terms, domains, etc for downstream analyses



Annotation pipeline



More info on annotation

- https://github.com/bcbc-group/Botany2022NMGWorkshop/tree/main/5.Annotation

- Can launch on CyVerse VICE

Variant Detection

Why Call SNPs?

> Current Issue > vol. 108 no. 17 > Jesse A. Poland, 6893–6898, doi: 10.1073/pnas.1010894108

Check for updates

Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize

Jesse A. Poland[a,1], Peter J. Bradbury[a,b], Edward S. Buckler[a,b], and Rebecca J. Nelson[a,c,2]

UNIT 7.18 Next-Gen Sequencing-Based Mapping and Identification of Ethyl Methanesulfonate-Induced Mutations in *Arabidopsis thaliana*

Xue-Cheng Zhang[1], Yves Millet[2], Frederick M. Ausubel[1], Mark Borowsky[1]

Published Online: 1 OCT 2014

DOI: 10.1002/0471142727.mb0718s108

Lab Protocol Title

CURRENT PROTOCOLS in Molecular Biology

Current Protocols in Molecular Biology

ASEReadCounter

Calculate read counts per allele for allele-specific expression analysis

Category Diagnostics and Quality Control Tools

Traversal LocusWalker

PartitionBy LOCUS

SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data

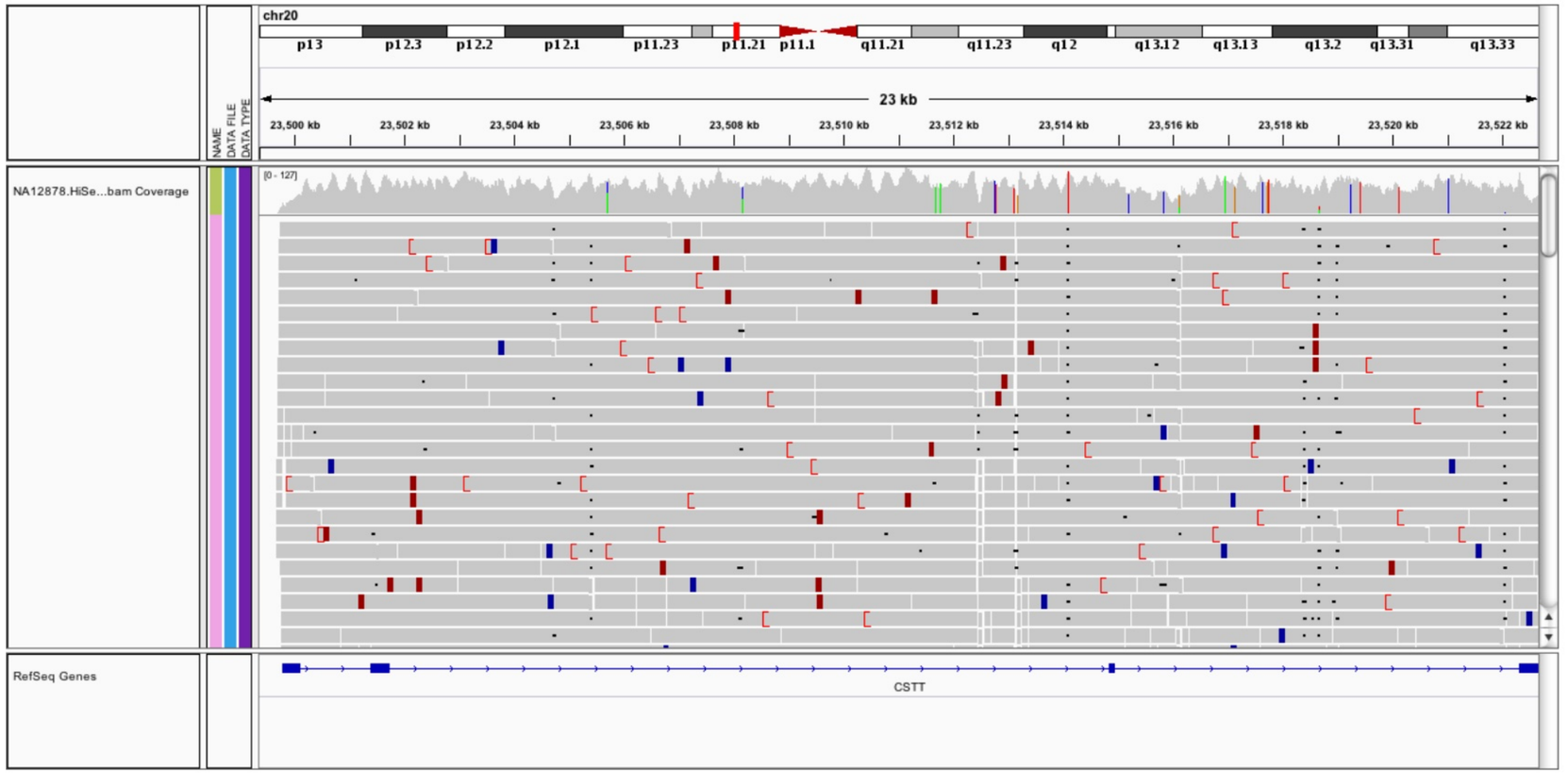
Tae-Ho Lee, Hui Guo, Xiyin Wang, Changsoo Kim and Andrew H Paterson

BMC Genomics 2014 15:162 | DOI: 10.1186/1471-2164-15-162

Received: 25 September 2013 | Accepted: 18 February 2014 | Published: 26 February 2014

BTI

Which mismatches are real mutations and which are noise/error?



https://www.broadinstitute.org/gatk/events/3391/GATKw1310-BP-0A-Intro_to_NGS.pdf

Table 1 Algorithms and short descriptions of the seven variant calling tools

From: Evaluation of variant calling tools for large plant genome re-sequencing

Variant tool	Version	Algorithm	Pipelines	Default filter	Reference
FreeBayes	v1.2.0–2	Haplotype-based	FreeBayes	[b]10,[m]1	Garrison E, et al, 2012 [29]
		Bayesian			
GATK	4.0.11.0	Haplotype-based	MarkDuplicates	[b]10,[m]20	DePristo M, et al, 2011 [27]
		significant test	BaseRecalibrator		
			HaplotypeCaller		
Platypus	0.8.1	Haplotype-based	Platypus callVariants	[b]20,[m]20	Rimmer A, et al, 2014 [30]
		significant test			
Samtools /mpileup	1.9	Site align-based	Samtools/mpileup	[b]13,[m]0	Li H, 2011 [28]
		gt likelihoods	bcftools call		
SNVer	0.5.3	Site align-based	SNVerIndividual	[b]17,[m]20	Wei Z, et al, 2011 [31]
		MAF *p*-value		[f]0.25,[r]1,[p]0.05	
VarScan	v2.3.9	Site-based	Samtools/mpileup	[b]15,[m]0	Koboldt D, et al, 2012 [33]
		allele frequency	mpileup2snp	[f]0.2,[r]2,[p]0.01	
VarDict	2018	Site-based	VarDict	[b]22.5,[m]0	Lai Z, et al, 2016 [32]
		alleles Fisher's	var2vcf_valid	[f]0.01,[r]2	

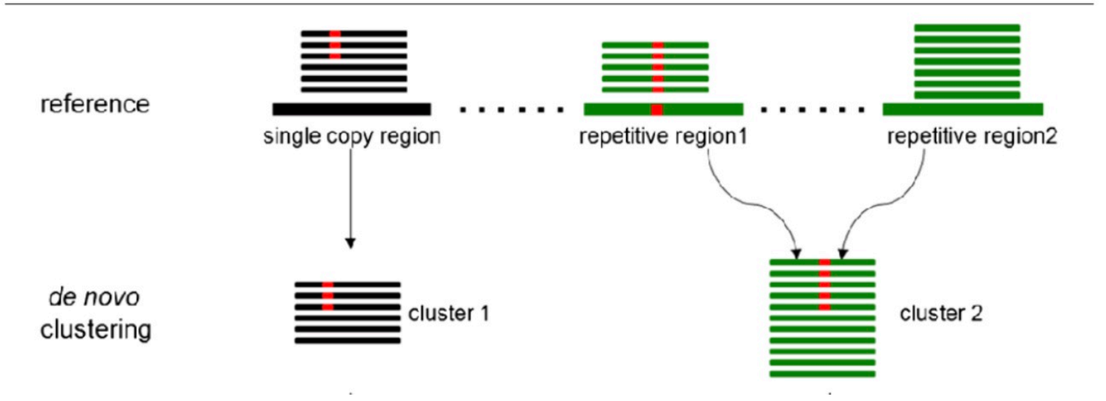
[a]Only default settings were listed. [b]BQ Base quality; [m]MQ Mapping quality; [r]VR Variant containing reads or total reads containing variants (TR); [f]VF Variant frequency; [p] P-value; [d]DP Depth coverage

Yao, Z., You, F.M., N'Diaye, A. *et al.* Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics* **21,** 360 (202
10.1186/s12859-020-03704-1

Downstream analyses using SNPs

Population genetic studies

- Many options and different requirements for SNP calling
- RAD-Seq
  - Stacks, both reference guided or *de novo*
- Hyb-Seq, RNA-Seq or Whole Genome Sequencing
  - GATK, Freebayes, mpileup
  - Must have a reference genome or at the very least something to map reads to

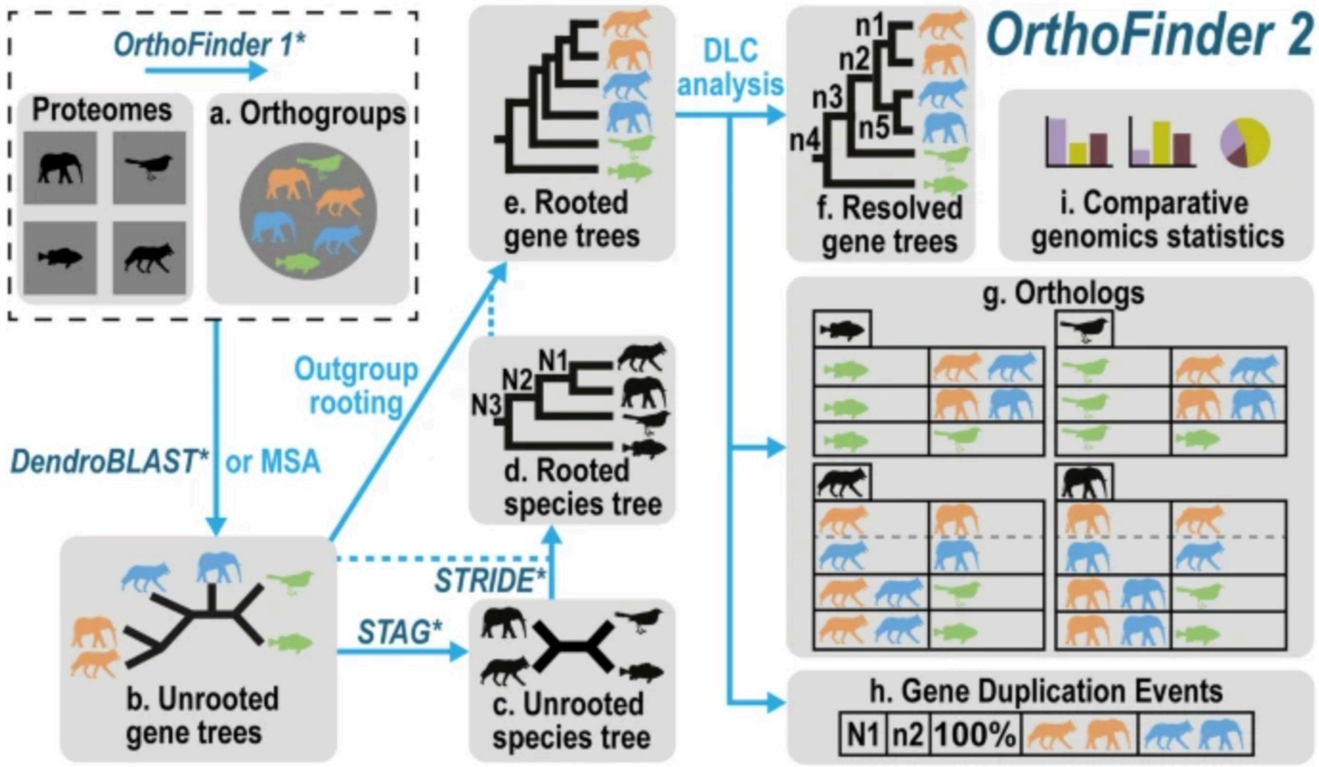


https://github.com/bcbc-group/Botany2022NMGWorkshop/tree/main/6.DownstreamAnalyses

Phylogenomics

- How do sequences/genomes relate to each other?
- Align sequences
  - ClustalW
  - Muscle
  - MAFFT
- Build phylogenetic trees
  - Parsimony (PAUP)
  - Neighbor join (QuickTree)
  - Maximum likelihood (IQ-TREE, PHYLIP)
  - Bayesian (MrBayes)

Gene Families



The OrthoFinder workflow. The method used for each step is shown by the arrow. Published algorithms are shown in italics and are followed by an asterisk. A dotted blue line connecting with a solid arrow indicates additional data that are used in order to carry out the transformation indicated by the solid arrow. MSA, multiple sequence alignment-based tree inference; DLC, duplication-loss-coalescence. (**a**) Orthogroup inference using the original OrthoFinder algorithm (an orthogroup is the set of genes descended from a single gene in the last common ancestor of all the species under consideration). (**b**) Gene tree inference. (**c**) Species tree inference. (**d**) Species tree rooting (**e**) Gene tree rooting (**f**) Hybrid overlap + DLC analysis of rooted gene trees to infer orthologs and gene duplication events. (**g**) Illustration of the ortholog results table for the genes in each input species (four main boxes). The horizontal divisions within these show the orthologs for each individual species pair. (**h**) Illustration of the gene duplication event table showing the location of the gene duplication events mapped to the species tree, the location in the gene tree, the percent retention of the duplicate genes in the sampled species, and the genes descended from the gene duplication event. (**i**) Comparative genomics statistics

Single copy orthologs

CAFE: a computational tool for the study of gene family evolution FREE

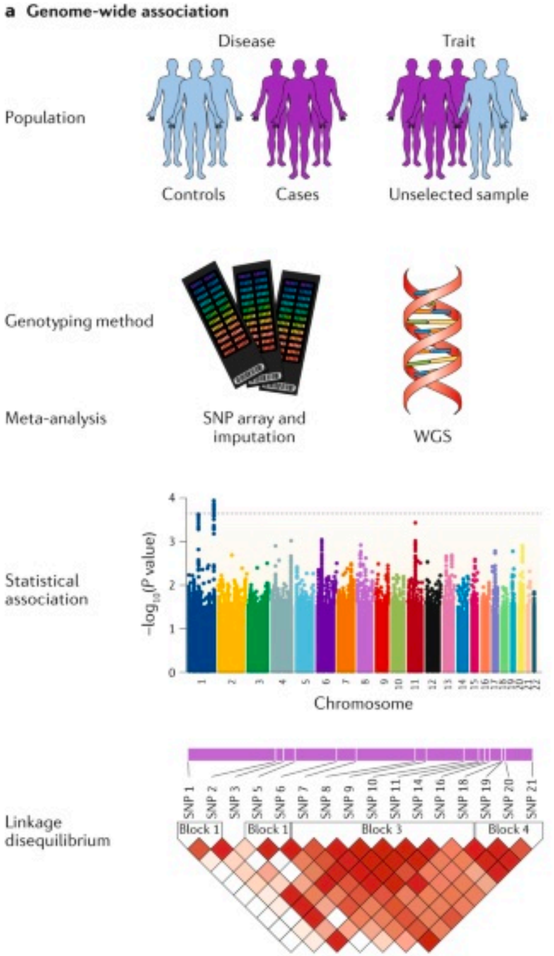
Tijl De Bie, Nello Cristianini, Jeffery P. Demuth, Matthew W. Hahn ✉ Author Notes

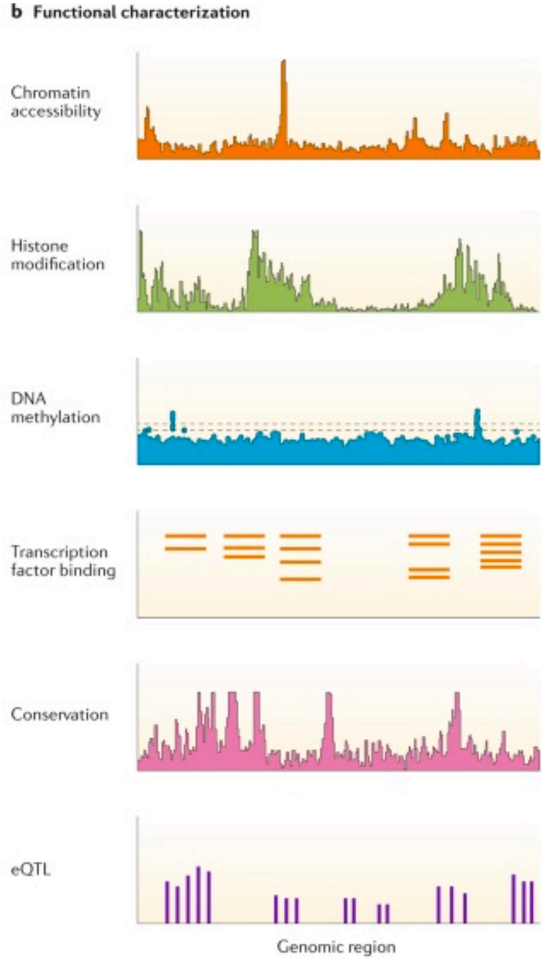
Bioinformatics, Volume 22, Issue 10, 15 May 2006, Pages 1269–1271,

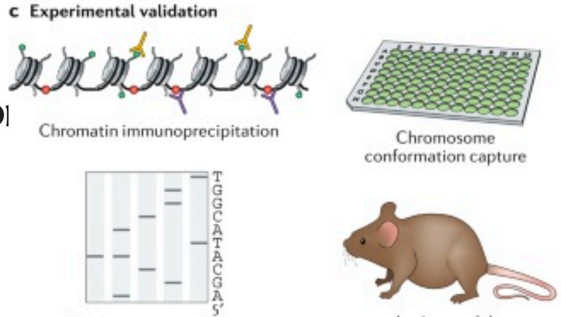
GWAS

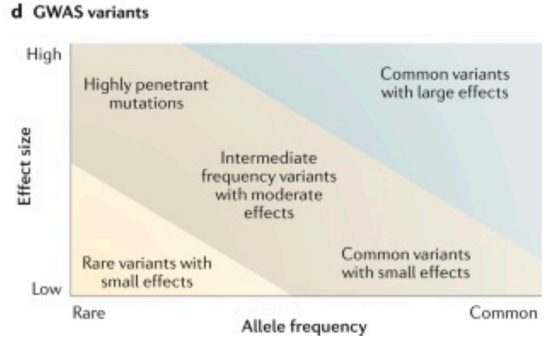
- Genome-Wide Association Study
- Correlate SNPs with phenotypic effects
- Requires genotyped and phenotyped populations
- Allows identification of genes responsible for a phenotype

High LD = haplotype blocks









Review Article | Published: 08 May 2019

Benefits and limitations of genome-wide association studies

Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré & David Meyre

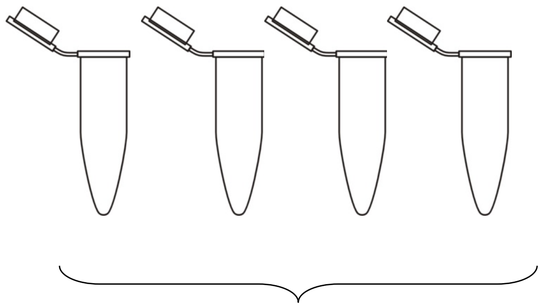
Nature Reviews Genetics **20**, 467–484 (2019) | Cite this article

https://btiplantbioinfocourse.wordpress.com/bti-bioinformatics-course-2018/qtl-gwas-using-r/

Transcriptomics

The study of transcriptomes

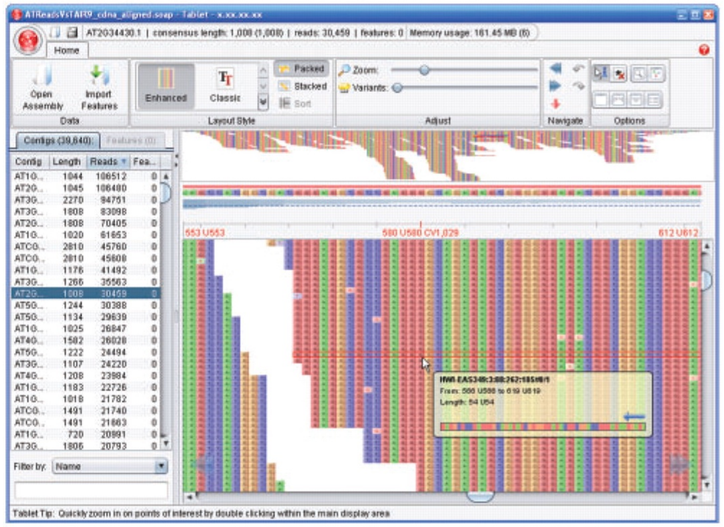
RNASeq



RNA samples

cDNA synthesis, multiplex and sample prep (bridge amplification)

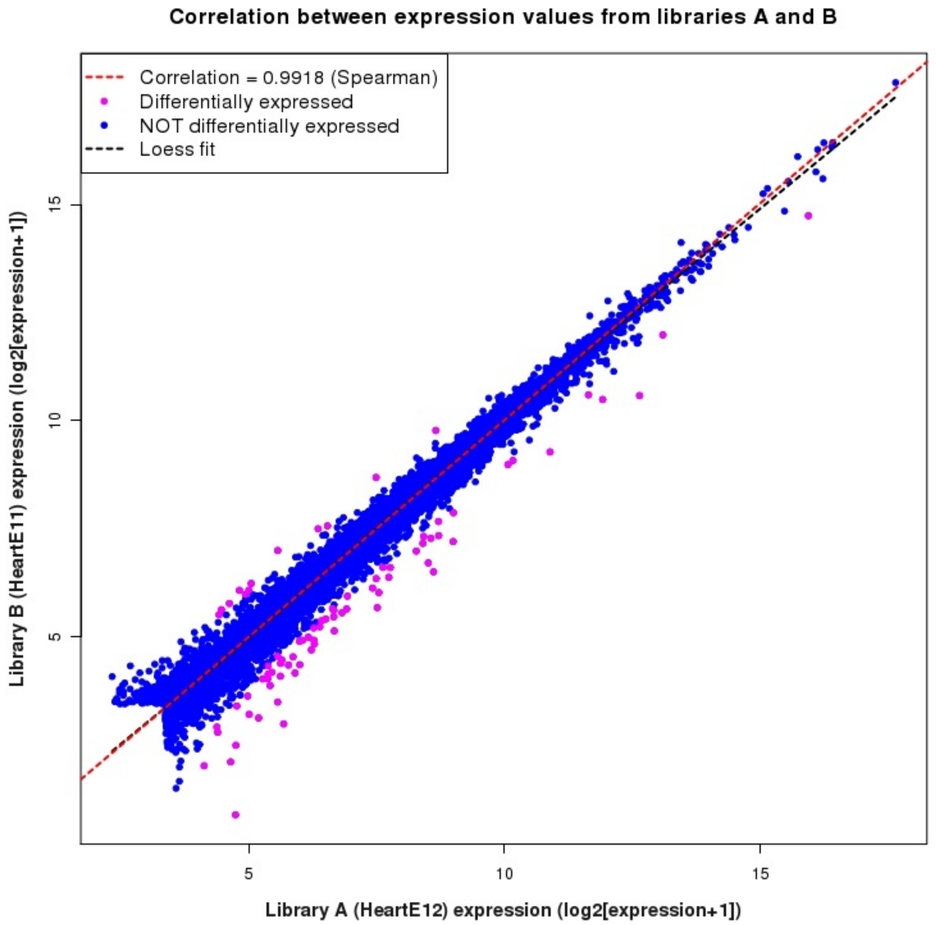
Align to reference genome/transcriptome Using bowtie2

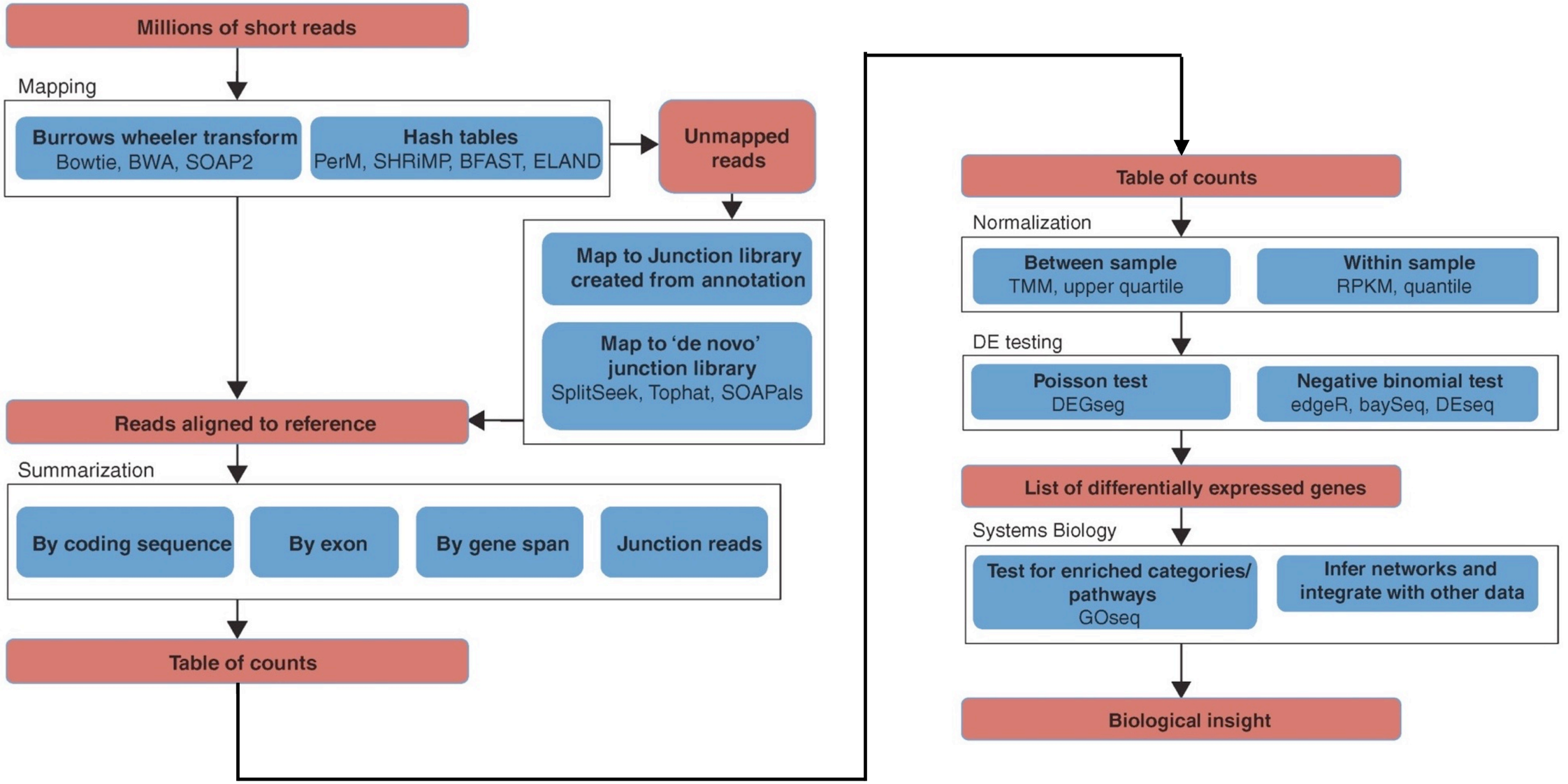


Visualize in Tablet software, determine rpkm values (number of reads per kilobase of exon per million reads)

Differential Expression

- Identify differentially expressed genes




Millions of short reads
Mapping
Burrows wheeler transform
Bowtie, BWA, SOAP2
Hash tables
PerM, SHRiMP, BFAST, ELAND
Unmapped reads
Map to Junction library created from annotation
Map to 'de novo' junction library
SplitSeek, Tophat, SOAPals
Reads aligned to reference
Summarization
By coding sequence
By exon
By gene span
Junction reads
Table of counts
Table of counts
Normalization
Between sample
TMM, upper quartile
Within sample
RPKM, quantile
DE testing
Poisson test
DEGseq
Negative binomial test
edgeR, baySeq, DEseq
List of differentially expressed genes
Systems Biology
Test for enriched categories/ pathways
GOseq
Infer networks and integrate with other data
Biological insight

Differential Expression Analysis

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Robinson MD, McCarthy DJ and Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, 26, pp. -1.

http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/
edgeRUsersGuide.pdf

DESeq

Differential expression analysis for sequence count data

Simon Anders & Wolfgang Huber

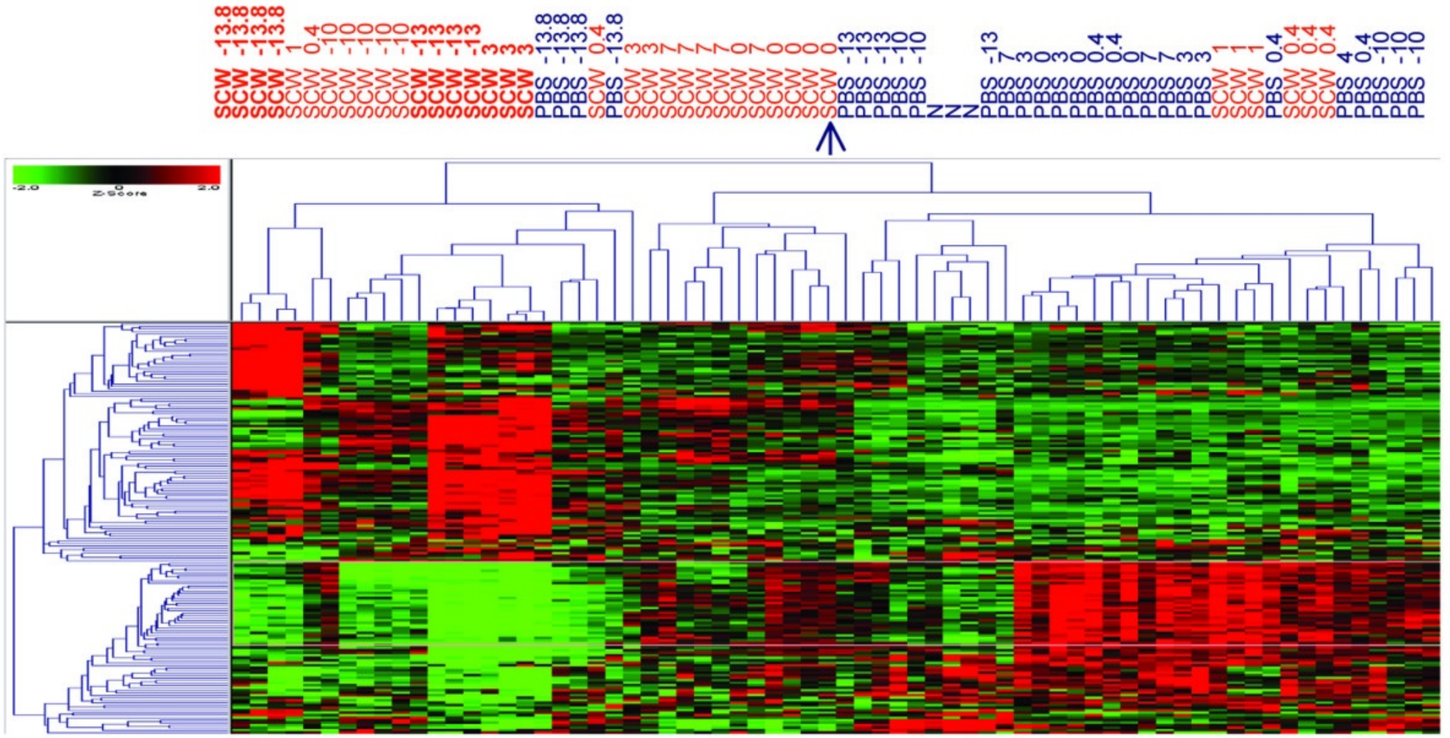
Genome Biology **11**, Article number: R106 (2010) | Cite this article

A Beginner's Guide to Analysis of RNA Sequencing Data

Clarissa M. Koch [1]*, Stephen F. Chiu [1,2]*, Mahzad Akbarpour [2], Ankit Bharat [1,2], Karen M. Ridge [1,3], Elizabeth T. Bartom [4‡], and Deborah R. Winter [5‡]

Hierarchical Clustering

- Identify genes with similar expression profiles

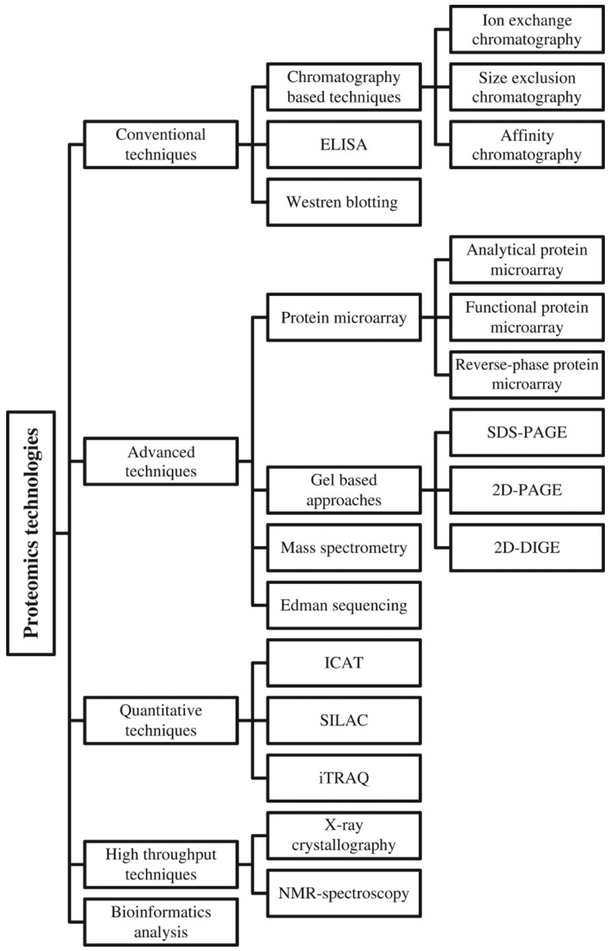


pheatmap2: Pretty and parallel heatmap
In jokergoo/pheatmap2: Pretty Plus Parallel Heatmap

Proteomics

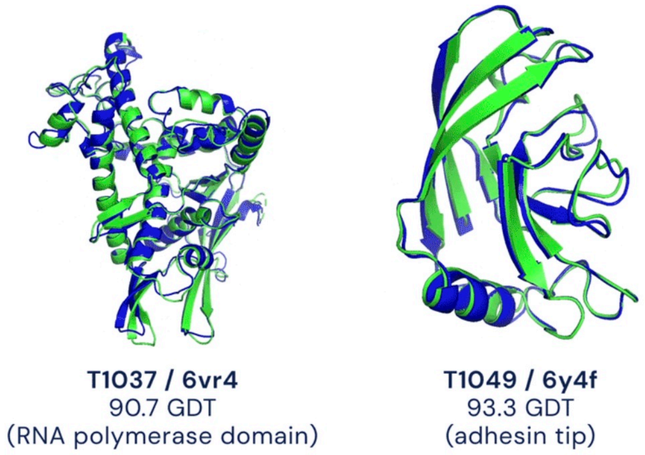
The study of proteomes

Figure 1. An overview of proteomics techniques.
Applications of proteomics techniques.
Figure 3.Schematic ...



J Chromatogr Sci, Volume 55, Issue 2, 1 February 2017, Pages 182–196, https://doi.org/10.1093/chromsci/bmw167

AlphaFold2



nature › articles › article

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, ... Demis Hassabis + Show authors

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

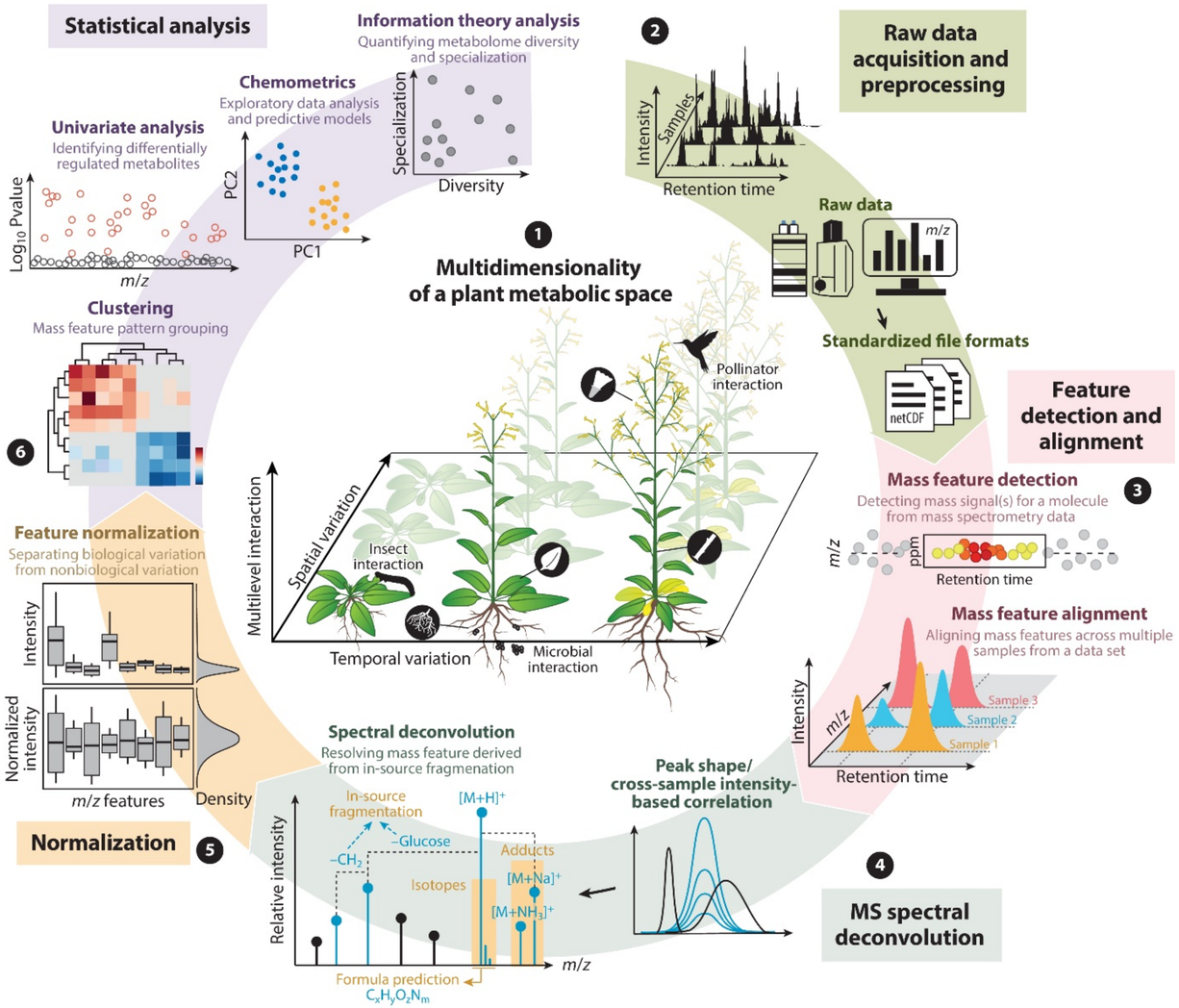
Metabolomics

- study of chemical processes involving metabolites, the small molecule substrates, intermediates, and products of cell metabolism.

- Detected by mass spectrophotometry

Next-Generation Mass Spectrometry Metabolomics Revives the Functional Analysis of Plant Metabolic Diversity

Annual Review of Plant Biology
Vol. 72:867-891 (Volume publication date June 2021)
First published as a Review in Advance on March 29, 2021
https://doi.org/10.1146/annurev-arplant-071720-114836



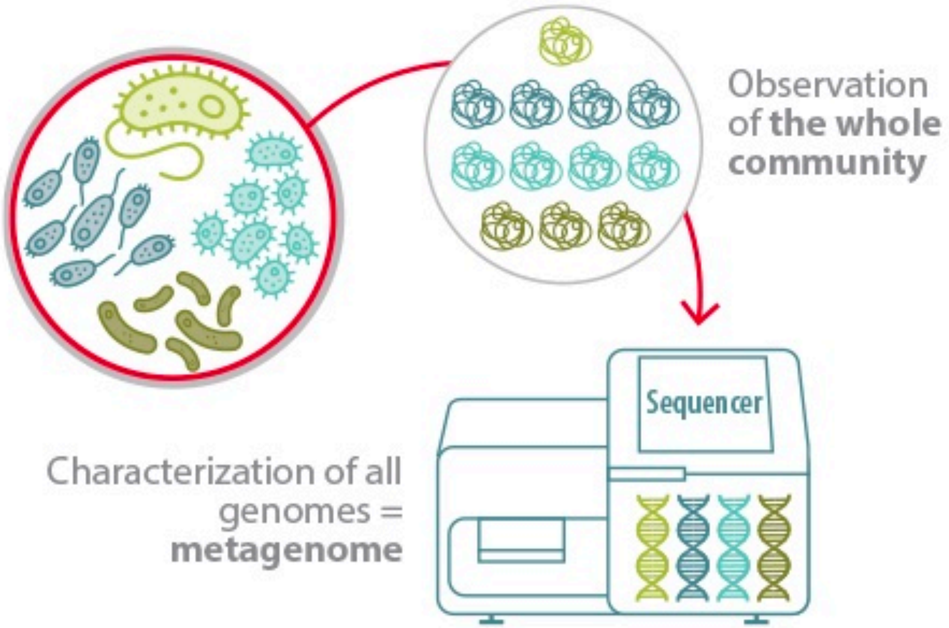
Li D, Gaquerel E. 2021
Annu. Rev. Plant Biol. 72:867–91

Annual Reviews

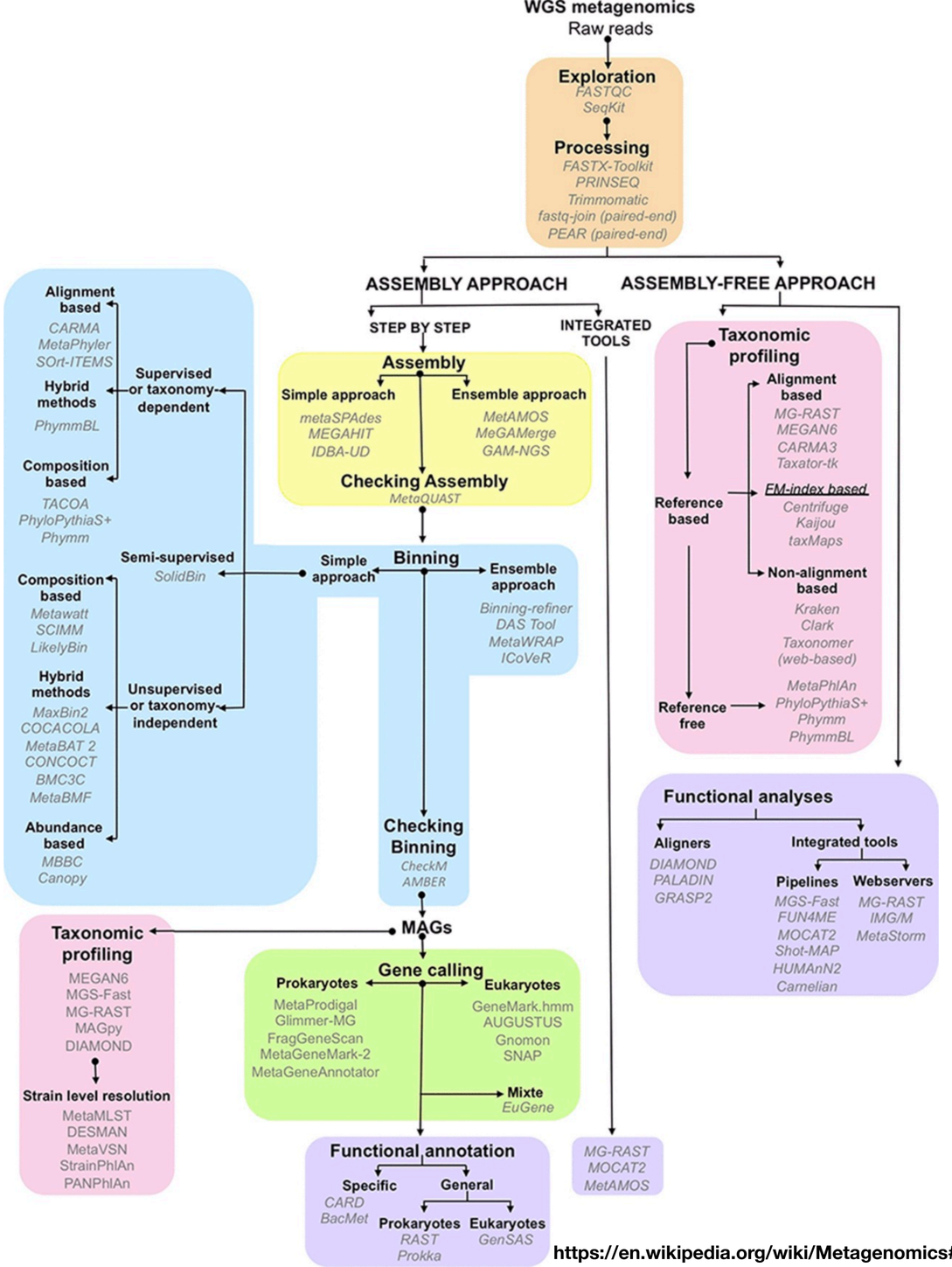
Metagenomics

the study of genetic material recovered directly from environmental samples.

What is metagenomics?



LALLEMAND



https://en.wikipedia.org/wiki/Metagenomics#/media/File:WGS_metagenomics_analysis_steps.gif

Table 1. A List of Benchmarked Classifiers and Their Various Characteristics

Type	Classifier	Custom Databases	Generates Abundance Profile	Memory Required	Time Required	Reference
DNA	Bracken	yes	yes	<1 Gb	<1 min	Lu et al., 2017
	Centrifuge	yes	yes	20 Gb	7 min	Kim et al., 2016
	CLARK	yes	yes	80 Gb	2 min	Ounit et al., 2015
	CLARK-S	yes	yes	170 Gb	40 min	Ounit and Lonardi, 2016
	Kraken	yes	yes	190 Gb	1 min	Wood and Salzberg, 2014
	Kraken2	yes	yes	36 Gb	1 min	Wood and Salzberg, 2014
	KrakenUniq	yes	yes	200 Gb	1 min	Breitwieser et al., 2018
	k-SLAM	yes	yes	130 Gb	2 h	Ainsworth et al., 2017
	MegaBLAST	yes	no	61 Gb	4 h	Morgulis et al., 2008
	metaOthello	no	no	30 Gb	1 min	Liu et al., 2018
	PathSeq	yes[a]	no	140 Gb	5 min	Walker et al.,
	prophyle	yes	no	40 Gb	40 min	Břinda et al., 2017
	taxMaps	yes	yes	65 Gb	25 min	Corvelo et al., 2018
Protein	DIAMOND	yes	no	110 Gb (varies)	10 min	Buchfink et al., 2015
	Kaiju	yes	yes	25 Gb	1 min	Menzel et al., 2016
	MMseqs2	yes	no	85 Gb (varies)	9 h	Steinegger and Söding, 2017
Markers	MetaPhlAn2	no	yes	2 Gb	1 min	Truong et al., 2015
	mOTUs2	no	yes	2 Gb	1 min	Milanese et al., 2019

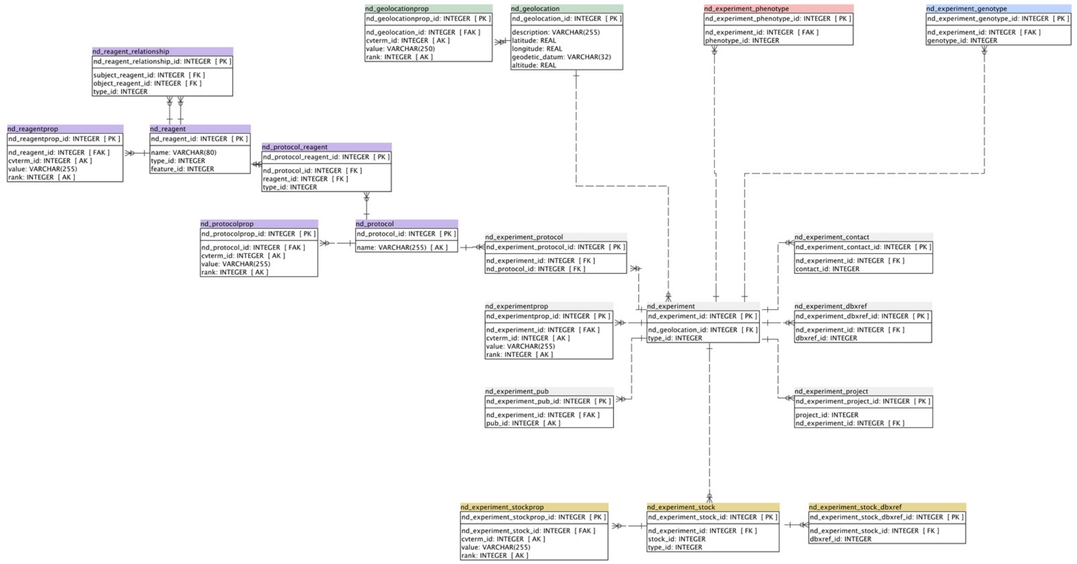
Benchmarking Metagenomics Tools for Taxonomic Classification

Simon H. Ye [1, 2], Katherine J. Siddle [2, 3], Daniel J. Park [2], Pardis C. Sabeti [2, 3, 4, 5]

3. Databases

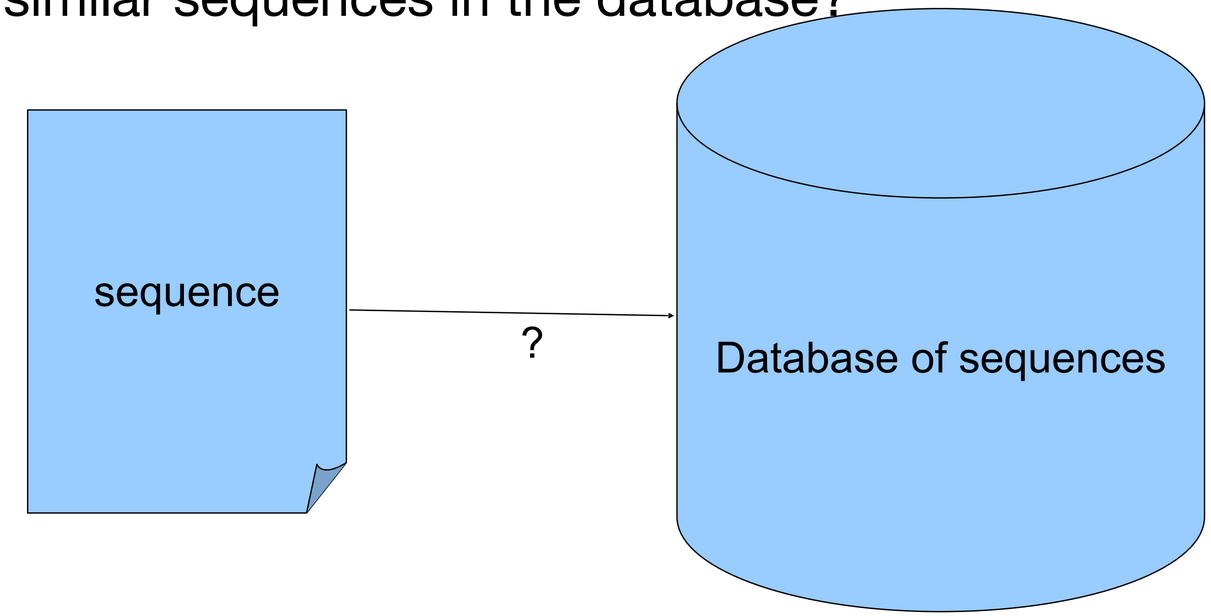
Databases

- Need to store and query data
- Biological data is highly structured
- Relational database systems (postgres, mysql)
- Database schemas - normalization
- Non-relational systems



Identify similar sequences

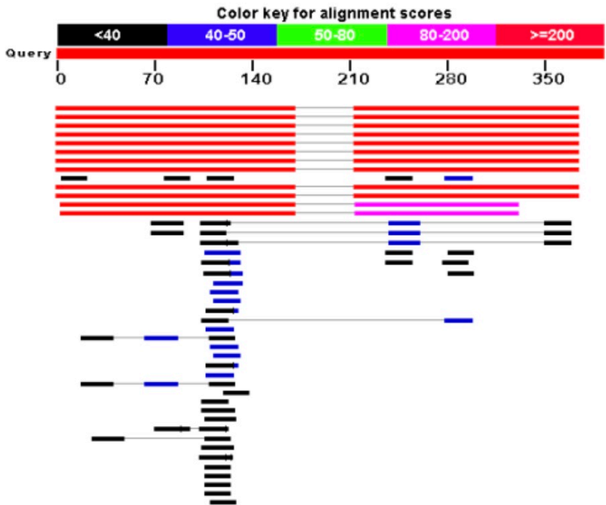
- Is the sequence (query) contained within a list of known sequences (database)?
- Are the similar sequences in the database?



- Compare the query sequence to each sequence in the database and calculate a similarity score, report sequences above a certain cutoff.

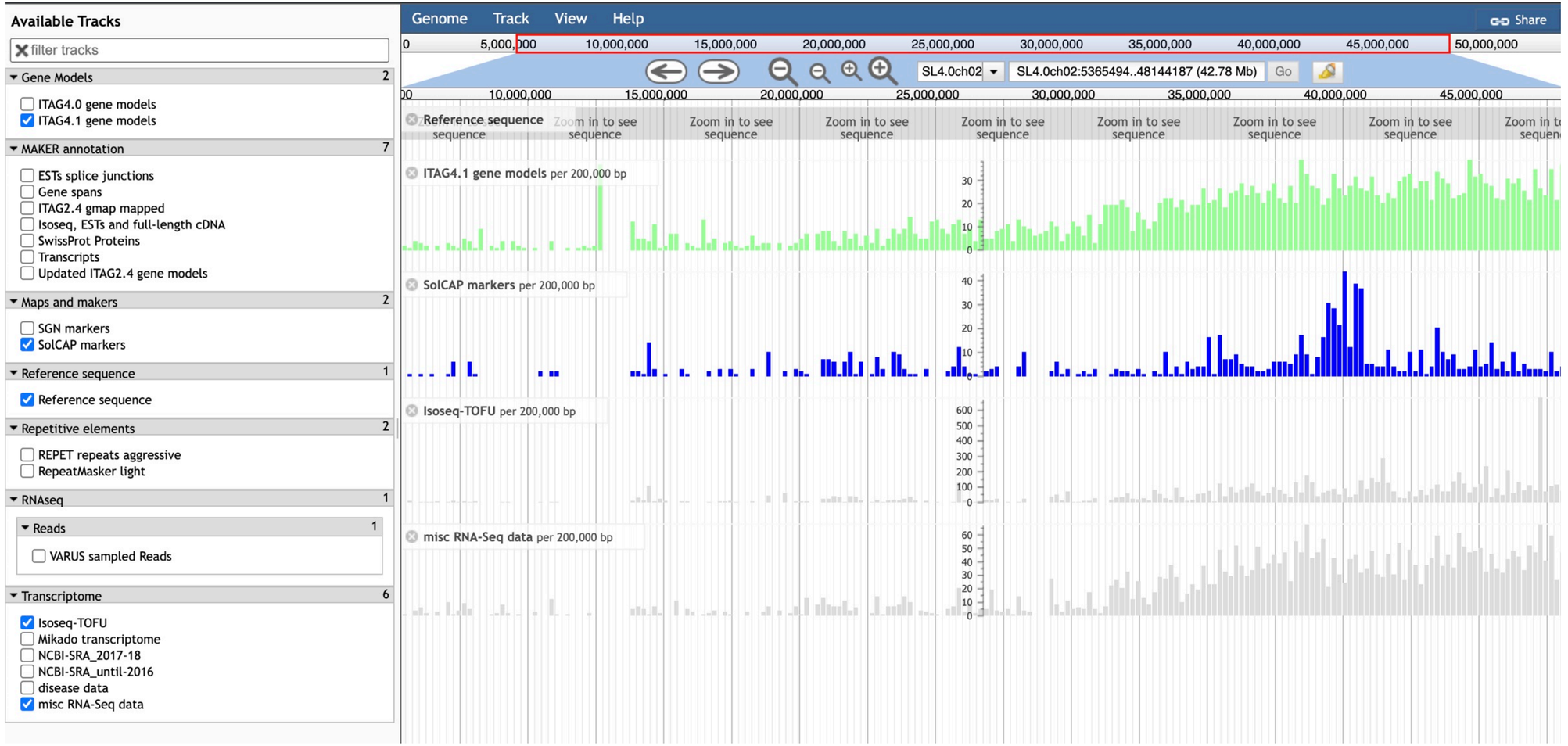
BLAST

- Basic Local Alignment and Search Tool
- Created by Altschul et al. at NCBI
- Homology search
  - Find similar sequences
  - Ranked by homology
  - Reports an e-value (expected number of hits by chance)
- A sequence *query* is run against many other sequences which form the *database*
- Offered by many websites



Name	Type	Web links
GenBank	Database	http://www.ncbi.nih.gov/entrez/query.fcgi?db=protein
RefSeq	Database	https://www.ncbi.nlm.nih.gov/refseq/
nr	Database	http://www.ncbi.nlm.nih.gov/BLAST/
UniProt	Database	http://www.pir.uniprot.org/
UniRef	Database	http://www.pir.uniprot.org/database/nref.shtml
UniParc	Database	http://www.pir.uniprot.org/database/archive.shtml
TrEMBL	Database	http://kr.expasy.org/sprot/
SwissProt	Database	http://kr.expasy.org/sprot/
PIR	Database	http://pir.georgetown.edu/
OWL	Database	http://www.bioinf.man.ac.uk/dbbrowser/OWL/

Genome browsers



4. Operating systems used in bioinformatics

Why use Unix?

- Free to use.
- Stable and secure
- Easily maintained
- Can edit and enhance source code (Open Source)
- World-wide community of developers
- Can be installed on just about anything
- Many bioinformatics programs will not run on Windows
- Many analysis servers run Linux

Linux Distributions

- Distributions have been created around the Linux kern
- Many distros available (distrowatch.com)
- Examples:
  - RedHat
  - Debian (we will use this)
  - Ubuntu

Fall 2020

Page Hit Ranking

Data span: Last 6 months Go

Rank	Distribution	HPD*
1	MX Linux	3823▼
2	Manjaro	2640▼
3	Mint	2368▼
4	Ubuntu	1670▼
5	Pop!_OS	1398▲
6	Debian	1377▼
7	elementary	1341▼
8	Solus	1050▼
9	Fedora	1000▼
10	Zorin	921▼
11	KDE neon	904▼
12	deepin	893▼
13	openSUSE	782▬
14	EndeavourOS	764▲
15	Ubuntu Kylin	723▬
16	Arch	701▼
17	antiX	662▼
18	CentOS	599▼
19	Linuxfx	563▲
20	ArcoLinux	552▼
21	PCLinuxOS	531▬
22	Puppy	519▲
23	Kali	509▬

Fall 2021

Page Hit Ranking

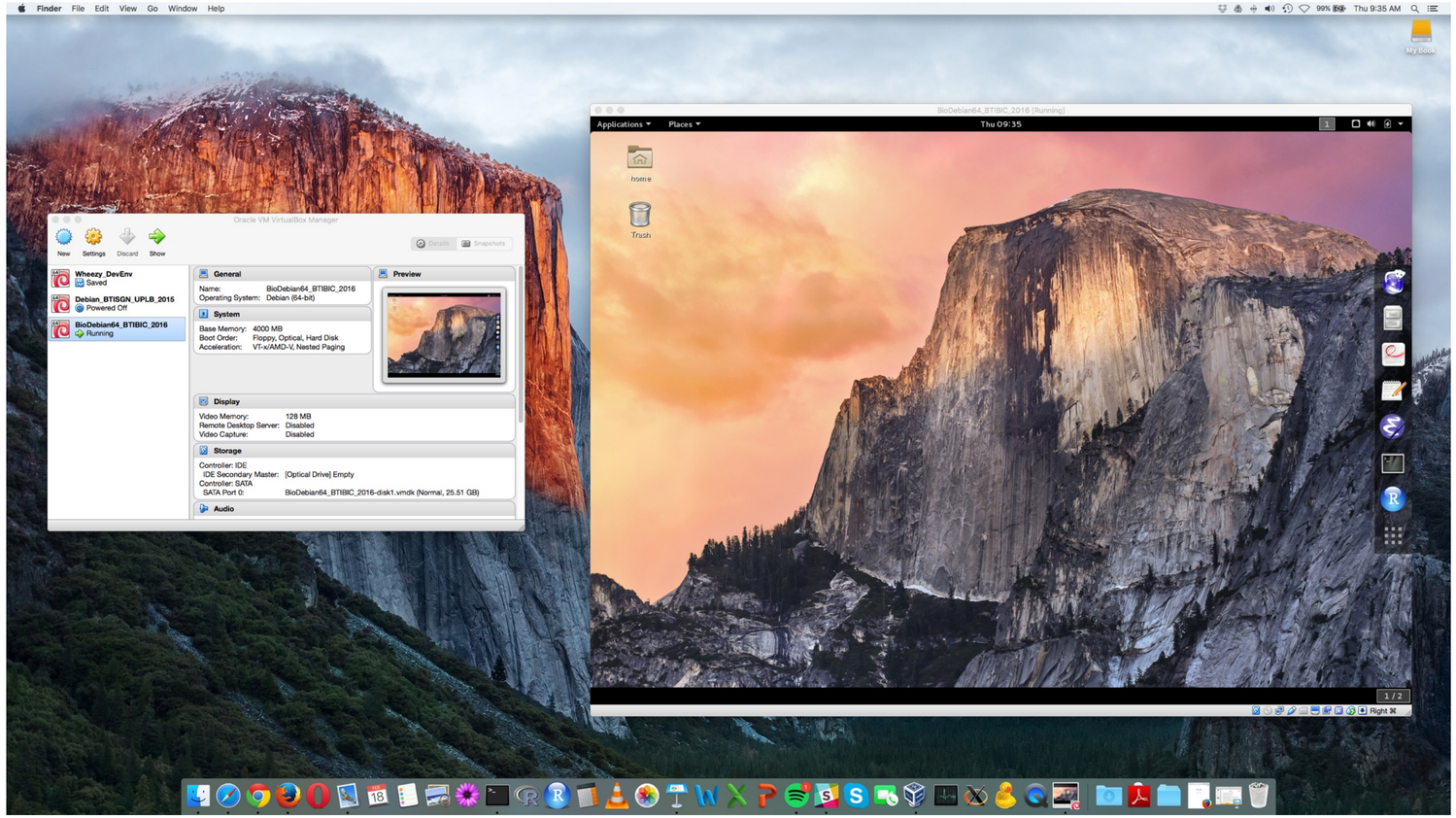
Data span: Last 6 months Go

Rank	Distribution	HPD*
1	MX Linux	3161▼
2	EndeavourOS	2627▲
3	Manjaro	2229▼
4	Mint	1928▲
5	Pop!_OS	1542▬
6	Ubuntu	1327▼
7	Debian	1259▬
8	Garuda	1152▲
9	elementary	1129▬
10	Fedora	958▲
11	Zorin	859▲
12	openSUSE	819▼
13	KDE neon	692▲
14	Solus	607▼
15	antiX	543▲
16	Arch	503▼
17	Slackware	463▲
18	Lite	459▼
19	Artix	438▬
20	PCLinuxOS	429▬
21	Kali	412▬
22	Puppy	404▬
23	deepin	398▼

https://distrowatch.com/

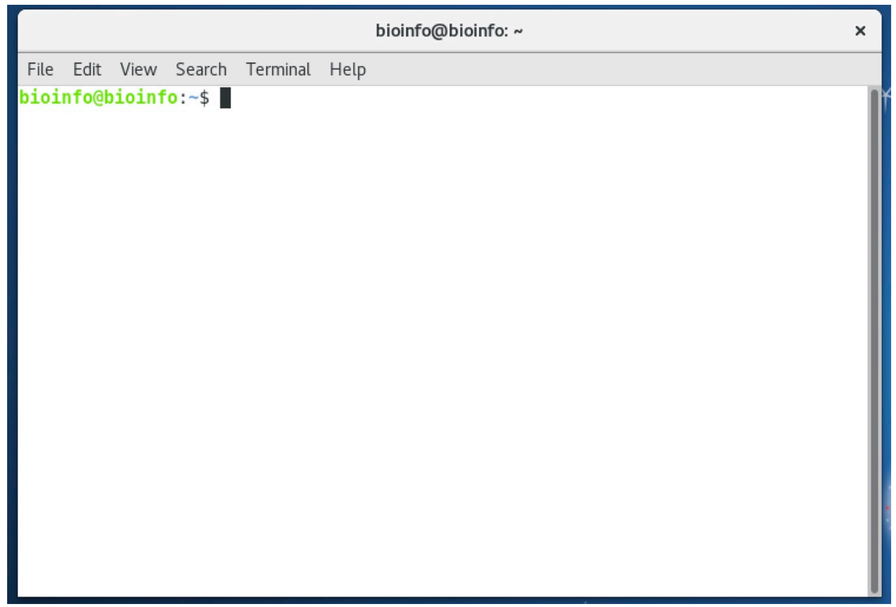
What is a Virtual Machine?

- A VM is an operating system which has been installed inside a simulated environment.
- Unlike emulators, virtual machines interface with real hardware.



What is a Terminal?

- A terminal is a textual interface for interacting with a computer.
- Using the terminal, one can issue powerful and concise command-line instructions for the computer to follow.

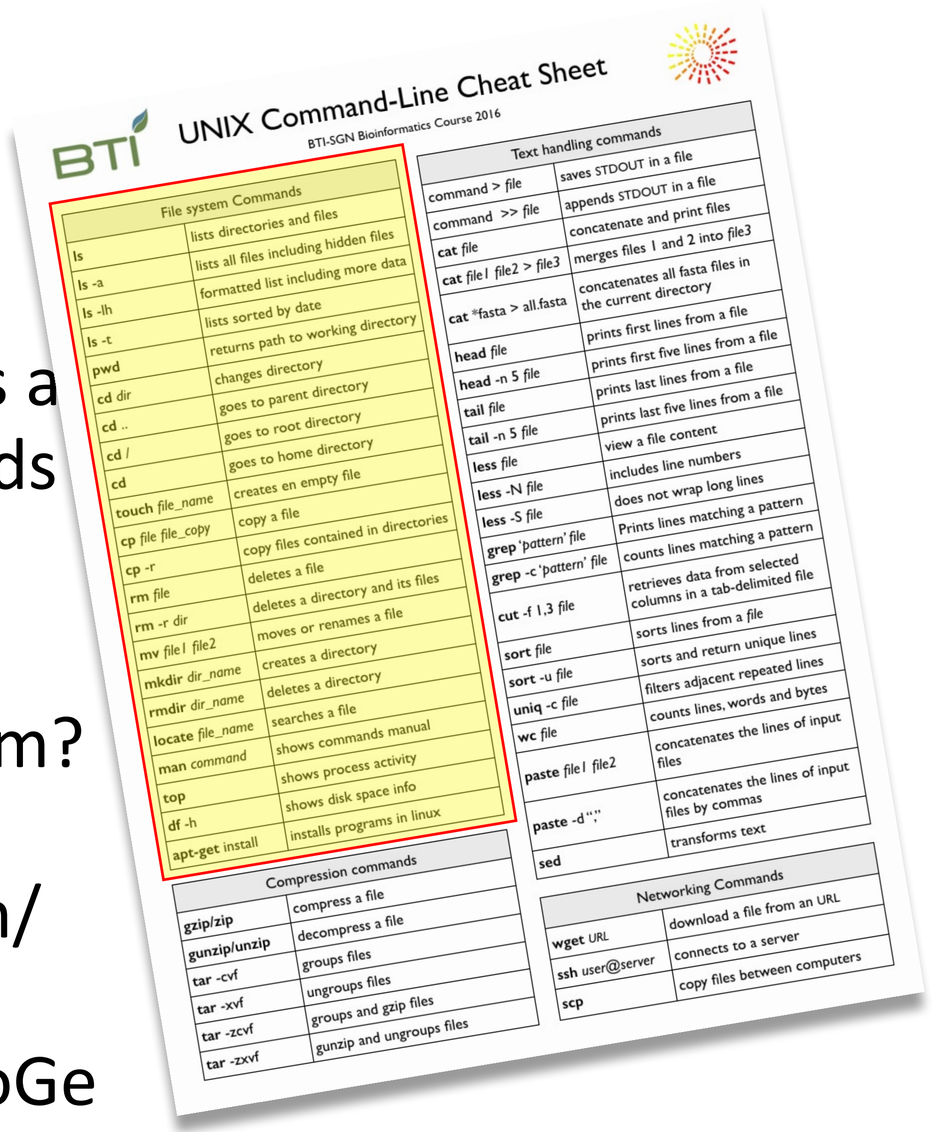


Why use the command-line (terminal)?

- Most software for biological data analysis is used through UNIX command-line operations.
- Most of the servers for biological data analysis use Linux/ Unix as their operating system.
- Data analysis on calculation servers are much faster since we can use more CPUs and RAM than in a PC or laptop (*e.g.* BTI's "Boyce" server has 64 cores and 1TB RAM)
- Large NGS data files can not be opened or loaded in most graphical software and web sites.

Command-line File System Navigation.

- The cheat sheet contains a list of common commands for navigating the file system.
- But what *is* the file system?
- https://drive.google.com/file/d/1Z8ryxpO-xyYhulfrWUfVsCS0YuZHoGeR/view



BTI

UNIX Command-Line Cheat Sheet

BTI-SGN Bioinformatics Course 2016

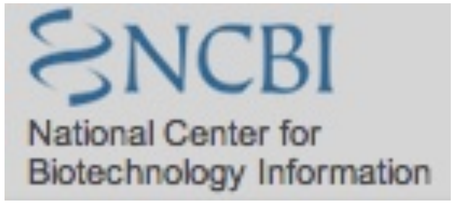
File system Commands	
ls	lists directories and files
ls -a	lists all files including hidden files
ls -lh	formatted list including more data
ls -t	lists sorted by date
pwd	returns path to working directory
cd dir	changes directory
cd ..	goes to parent directory
cd /	goes to root directory
cd	goes to home directory
touch file_name	creates en empty file
cp file file_copy	copy a file
cp -r	copy files contained in directories
rm file	deletes a file
rm -r dir	deletes a directory and its files
mv file1 file2	moves or renames a file
mkdir dir_name	creates a directory
rmdir dir_name	deletes a directory
locate file_name	searches a file
man command	shows commands manual
top	shows process activity
df -h	shows disk space info
apt-get install	installs programs in linux

Text handling commands	
command > file	saves STDOUT in a file
command >> file	appends STDOUT in a file
cat file	concatenate and print files
cat file1 file2 > file3	merges files 1 and 2 into file3
cat *fasta > all.fasta	concatenates all fasta files in the current directory
head file	prints first lines from a file
head -n 5 file	prints first five lines from a file
tail file	prints last lines from a file
tail -n 5 file	prints last five lines from a file
less file	view a file content
less -N file	includes line numbers
less -S file	does not wrap long lines
grep 'pattern' file	Prints lines matching a pattern
grep -c 'pattern' file	counts lines matching a pattern
cut -f 1,3 file	retrieves data from selected columns in a tab-delimited file
sort file	sorts lines from a file
sort -u file	sorts and return unique lines
uniq -c file	filters adjacent repeated lines
wc file	counts lines, words and bytes
paste file1 file2	concatenates the lines of input files
paste -d ","	concatenates the lines of input files by commas
sed	transforms text

Compression commands	
gzip/zip	compress a file
gunzip/unzip	decompress a file
tar -cvf	groups files
tar -xvf	ungroups files
tar -zcvf	groups and gzip files
tar -zxvf	gunzip and ungroups files

Networking Commands	
wget URL	download a file from an URL
ssh user@server	connects to a server
scp	copy files between computers

Web-based bioinformatics

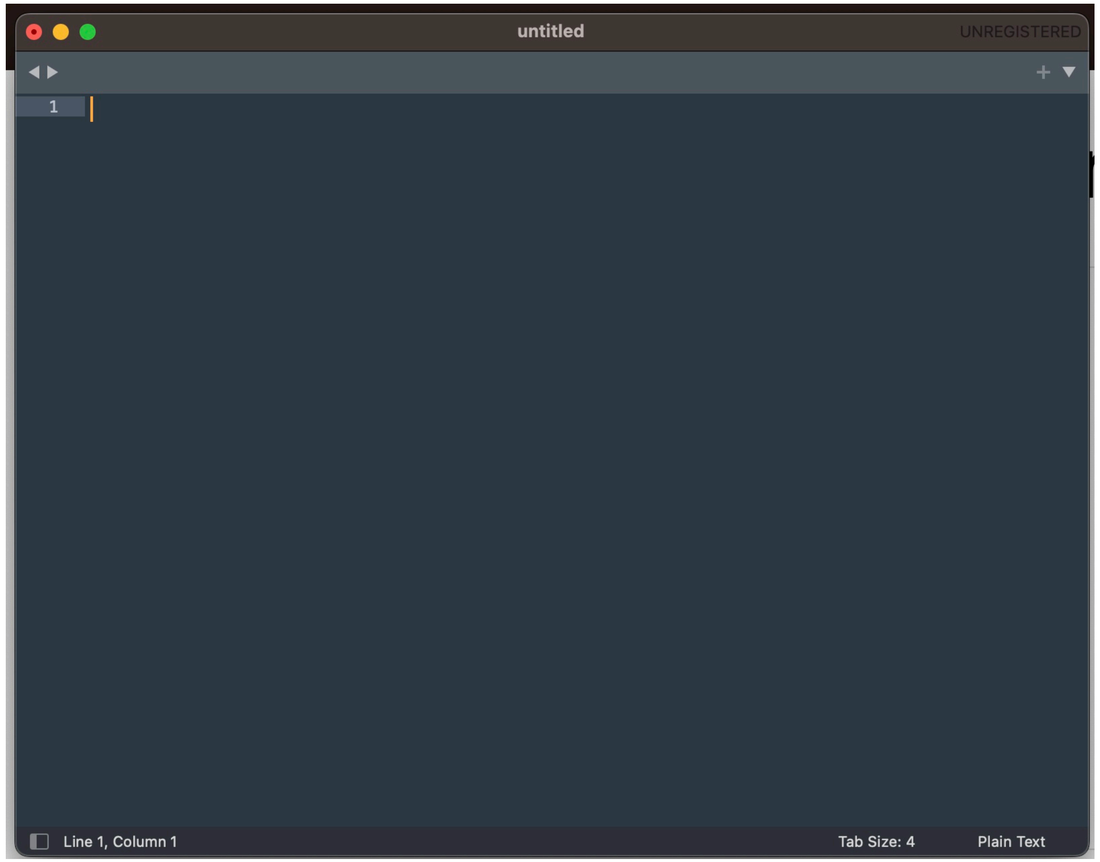


***Web-based not as powerful**

5. Programming languages used in bioinformatics

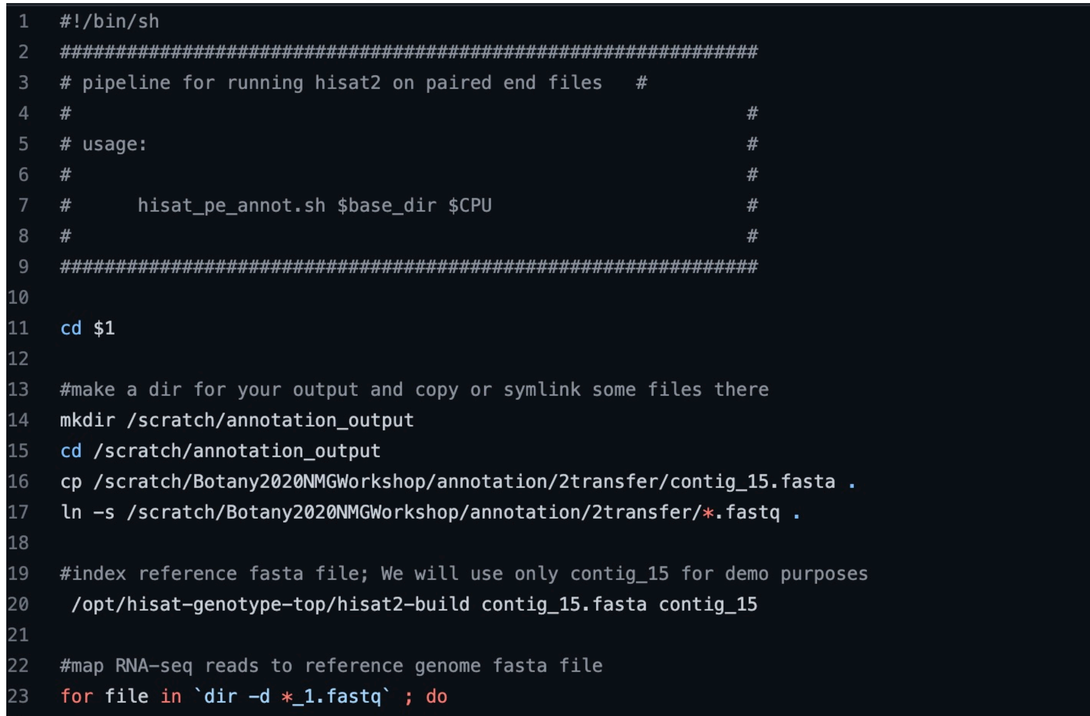
Text Editors for Programming

- Don’t write code in Word!
- Vi, Vim, Emacs
- Sublime
- Visual Studio



What are scripts?

- A text document that contains instructions to be executed by a program
- Usually combine existing components
- Usually use an interpreter instead of compiler to translate to machine code
  - **Bash**, **Python**, Perl, JavaScript, PHP, Ruby



```
#!/bin/sh
############################################################
# pipeline for running hisat2 on paired end files   #
#                                                          #
# usage:                                                   #
#                                                          #
#      hisat_pe_annot.sh $base_dir $CPU                    #
#                                                          #
############################################################

cd $1

#make a dir for your output and copy or symlink some files there
mkdir /scratch/annotation_output
cd /scratch/annotation_output
cp /scratch/Botany2020NMGWorkshop/annotation/2transfer/contig_15.fasta .
ln -s /scratch/Botany2020NMGWorkshop/annotation/2transfer/*.fastq .

#index reference fasta file; We will use only contig_15 for demo purposes
 /opt/hisat-genotype-top/hisat2-build contig_15.fasta contig_15

#map RNA-seq reads to reference genome fasta file
for file in `dir -d *_1.fastq` ; do
```

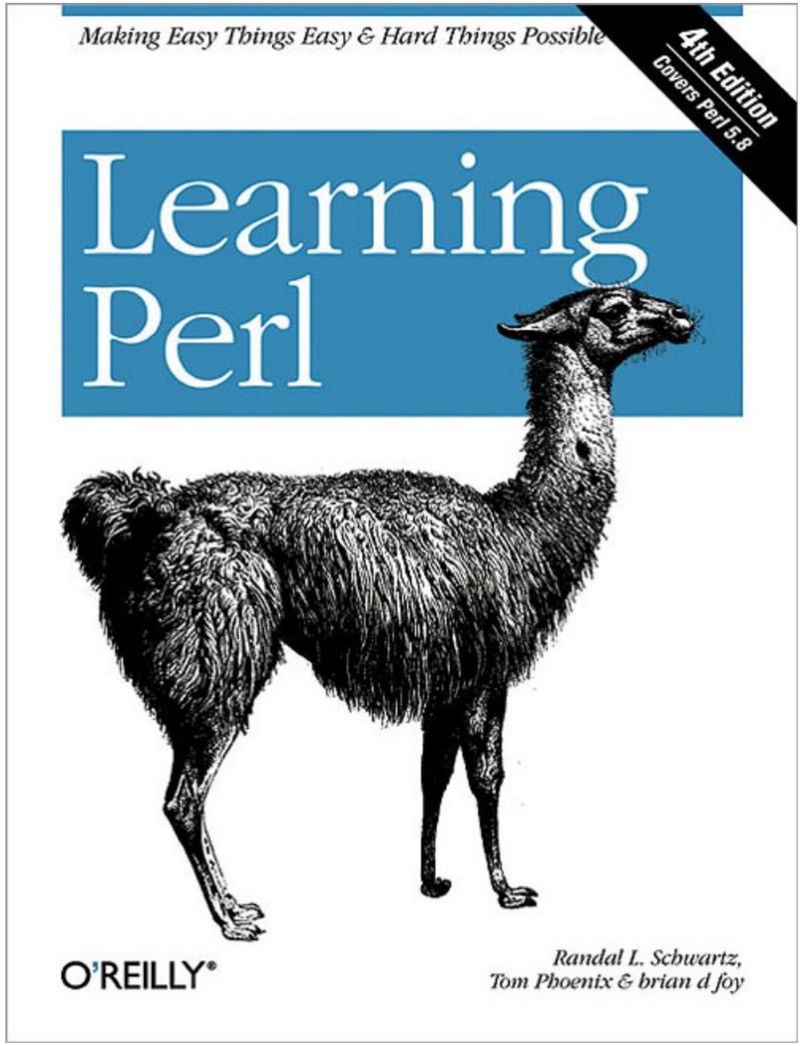
Why use bash?

- It is useful for chaining other programs together.
- This allows one to automate tasks and build pipelines.

```
#!/bin/sh

#move to annotation directory
cd /opt/annotation

#get some files
cp /scratch/Botany2020NMGWorkshop/annotation/2transfer/list.txt .
cp /scratch/Botany2020NMGWorkshop/annotation/2transfer/uniprot_sprot_plants.fasta .

#run Portcullis: https://bioconda.github.io/recipes/portcullis/README.html
#docker pull maplesond/portcullis:stable #already done with Adrian
docker run --rm -v /scratch/annotation_output:/data maplesond/portcullis:stable portcullis full -t 7 -v /data/contig_15.fasta /data/SRR5046448_contig15.sort.bam

#run Mikado pipeline: https://bioconda.github.io/recipes/mikado/README.html
#docker pull cyverseuk/mikado #already done with Adrian
docker run --rm -v /scratch/annotation_output:/data cyverseuk/mikado mikado configure --list list.txt --reference data/contig_15.fasta --mode permissive --scori
docker run --rm -v /scratch/annotation_output:/data cyverseuk/mikado mikado prepare --json-conf /data/configuration.yaml

makeblastdb -in uniprot_sprot_plants.fasta -dbtype prot
blastx -max_target_seqs 5 -num_threads 7 -query mikado_prepared.fasta -outfmt 5 -db uniprot_sprot_plants.fasta -evalue 0.000001 2> blast.log | sed '/^$/d' | gzi

screen -L /opt/TransDecoder-TransDecoder-v5.5.0/TransDecoder.LongOrfs -t mikado_prepared.fasta

screen -L hmmscan --cpu 7 --domtblout pfam.domtblout Pfam-A.hmm mikado_prepared.fasta.transdecoder_dir/longest_orfs.pep

blastp -query mikado_prepared.fasta.transdecoder_dir/longest_orfs.pep -db uniprot_sprot_plants.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 7 >

/opt/TransDecoder-TransDecoder-v5.5.0/TransDecoder.Predict -t mikado_prepared.fasta --retain_blastp_hits blastp.outfmt6 --cpu 7  --retain_pfam_hits pfam.domtblo

docker run --rm -v /scratch/annotation_output:/data cyverseuk/mikado mikado serialise --json-conf /data/configuration.yaml --xml /data/mikado.blast.xml.gz --orf
```

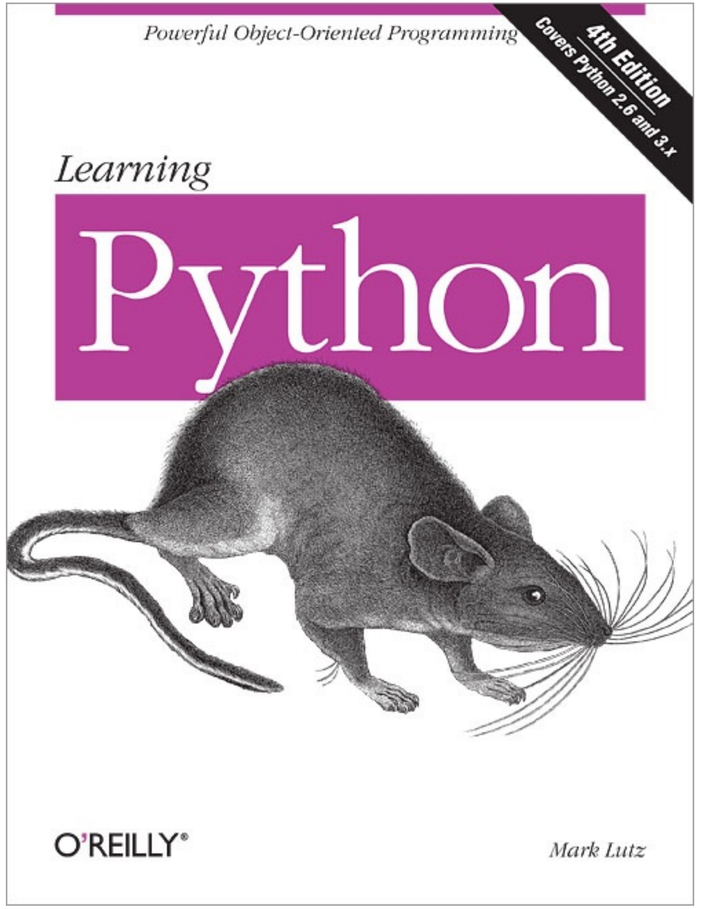
Perl

- Developed since 1980s by Larry Wall
- Useful for bioinformatics and web development
- Support for *objects*
- Excellent integration of *regular expressions* (text handling language)
- Vast open source code library (http:/cpan.org/)
  - BioPerl (http://bioperl.org/)
- Easy to learn
- http://www.perl.org/



Python

- Created by Guido van Rossum in 1989
- Very elegant language
- BioPython libraries
- The “new” popular language
- Many frameworks (Django for web etc.)



Virtual Environments

- Virtual environments are isolated from system directories
- Each has its own Python binary and has its own independent set of packages
- Example:

```
(base) bioinfo@bioinfo:~/Scripts/python$ conda create --name testpy3 python=3.8
```

```
(base) bioinfo@bioinfo:~/Scripts/python$ conda activate testpy3
```

```
(testpy3) bioinfo@bioinfo:~/Scripts/python$ less ~/.conda/environments.txt
```

Python Package managers

- pip

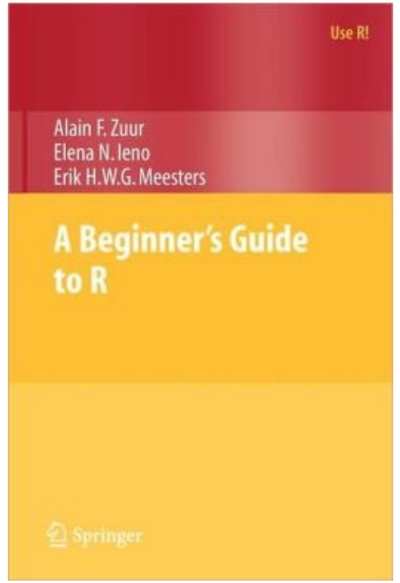
```
pip install "SomeProject==1.4"
```

- Conda

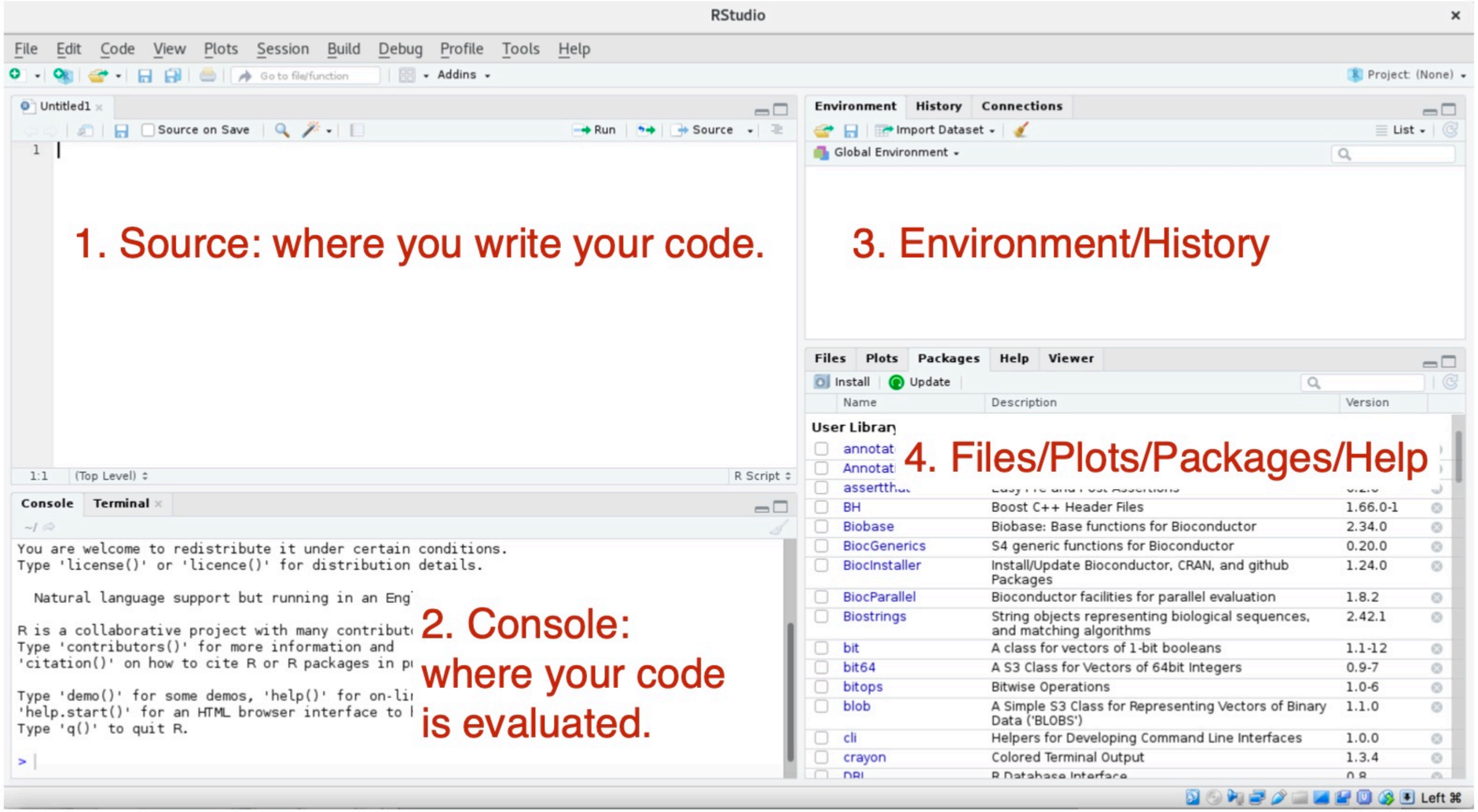
```
conda install package-name=2.3.4
```

  - Virtual environments

- Language designed for statistics
- Support for matrix calculations, graphics
- Expression analysis, Next-Gen sequence analysis, Graphics, genome annotation statistics, phylogeny
- Interactive
- “Bioconductor” package



R Studio

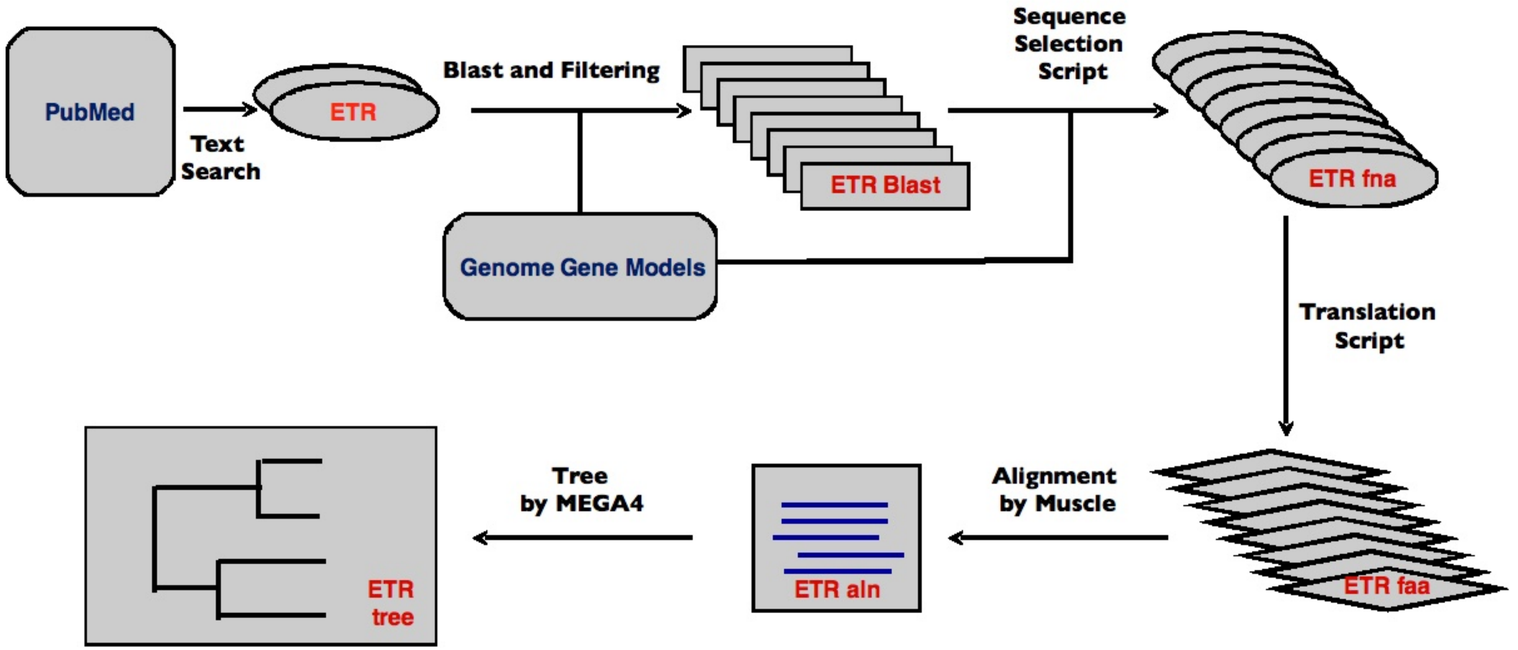

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1. Source: where you write your code.
Environment History Connections
Import Dataset
List
Global Environment
3. Environment/History
Files Plots Packages Help Viewer
Install
Update
Name
Description
Version
User Library
4. Files/Plots/Packages/Help
annotat
Annotat
BH Boost C++ Header Files 1.66.0-1
Biobase Biobase: Base functions for Bioconductor 2.34.0
BiocGenerics S4 generic functions for Bioconductor 0.20.0
BiocInstaller Install/Update Bioconductor, CRAN, and github Packages 1.24.0
BiocParallel Bioconductor facilities for parallel evaluation 1.8.2
Biostrings String objects representing biological sequences, and matching algorithms 2.42.1
bit A class for vectors of 1-bit booleans 1.1-12
bit64 A S3 Class for Vectors of 64bit Integers 0.9-7
bitops Bitwise Operations 1.0-6
blob A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS') 1.1.0
cli Helpers for Developing Command Line Interfaces 1.0.0
crayon Colored Terminal Output 1.3.4
1:1
(Top Level)
R Script
Console
Terminal
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an Eng
R is a collaborative project with many contribut
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in p
Type 'demo()' for some demos, 'help()' for on-li
'help.start()' for an HTML browser interface to
Type 'q()' to quit R.
>
2. Console: where your code is evaluated.
Left ⌘

6. Data Best Practices

Data Best Practices

- Data should be stored in the appropriate repository
  - EX: SRA/Genbank, TreeBase
- Organizing schemas for your own data
- Backup storage, redundancy
- Compress large files to save disk space
- rsync, md5sum, diff, zless

Running “pipelines”



Code Libraries

- Modularizing code
  - Breaking apart problem into individual units
  - Easier to maintain
  - Reusable
- Modules, Packages - grouping of modules, Library - a collection of packages
- Libraries we will use:
  - BioPython, NumPy, Pandas in later classes
- Package Repos
  - Python Package Index (PyPi) and Anaconda

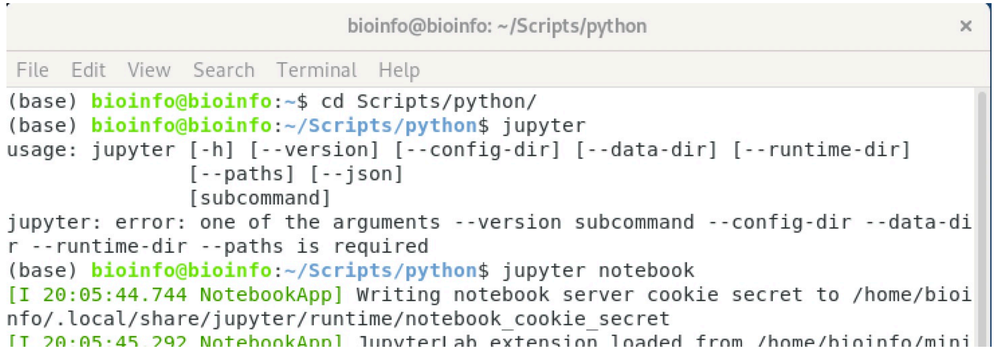
Jupyter Notebooks

- Starting the notebook server

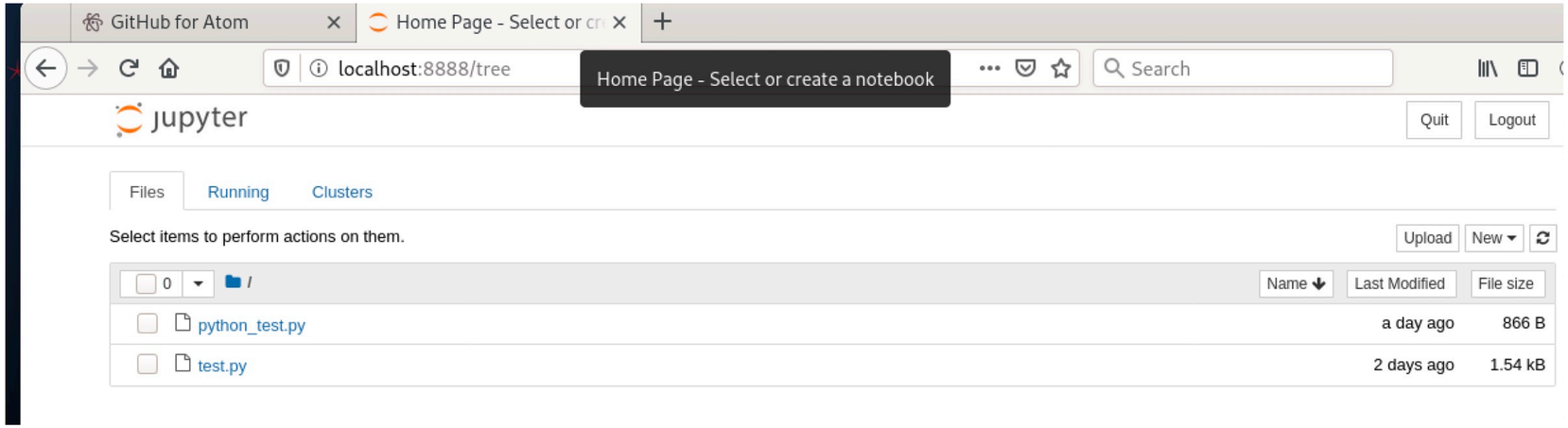
mkdir ~/Scripts/python

cd Scripts/python/

jupyter notebook

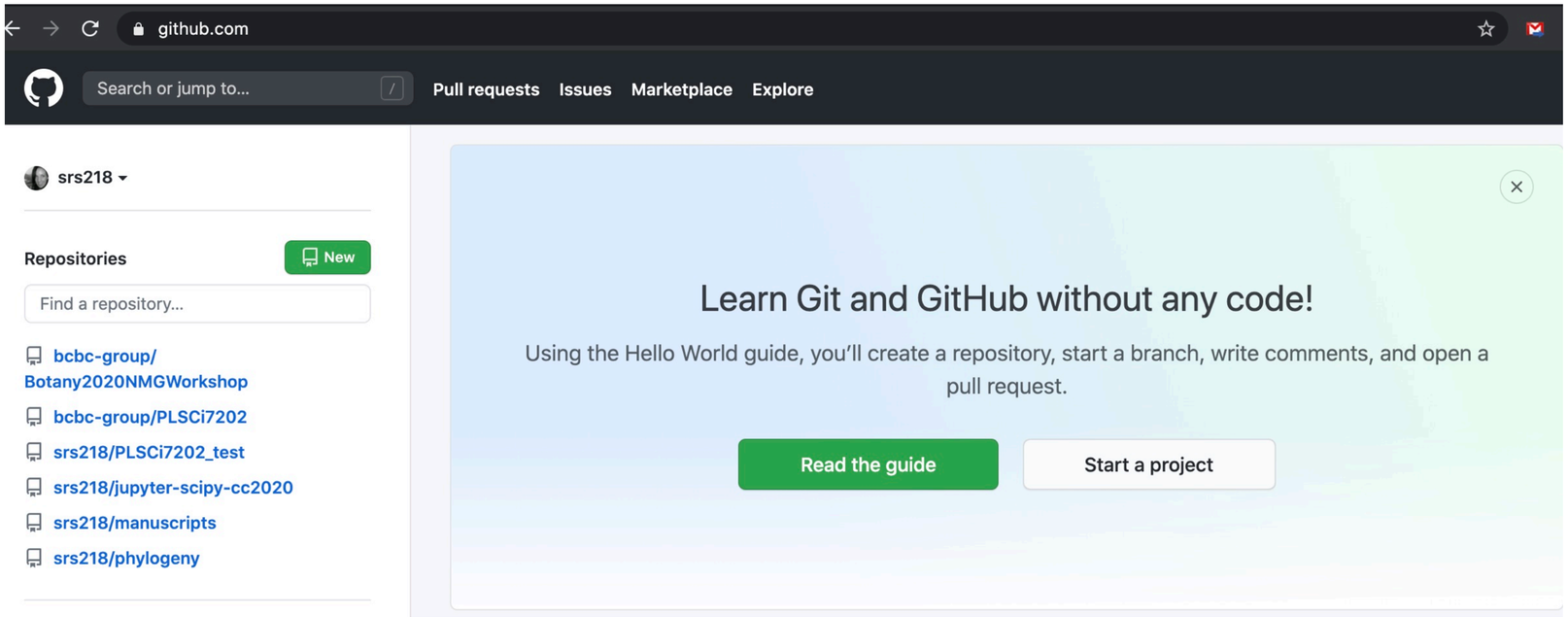


- Notebook Dashboard

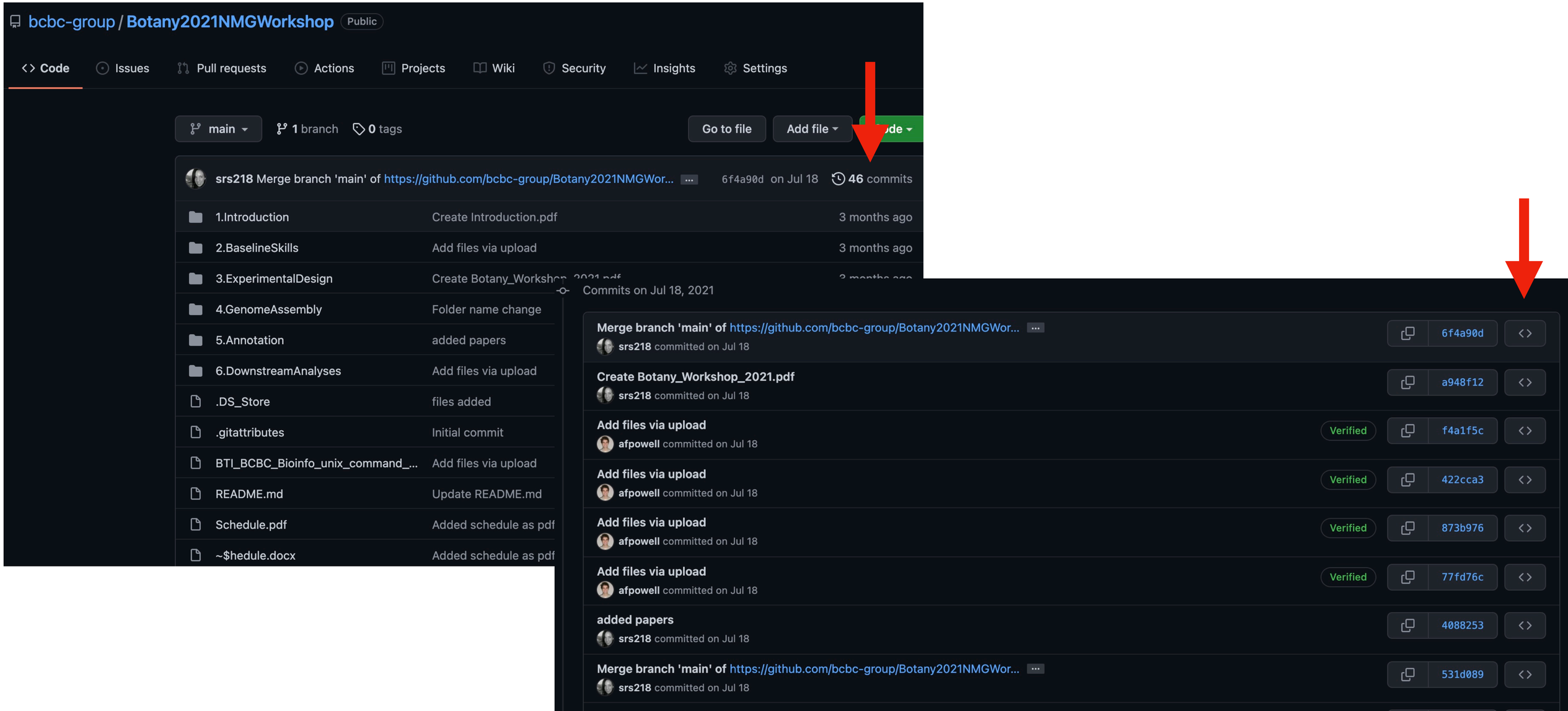


GitHub

- GitHub is a repository hosting service.
- Has a command line tools and GUI options
- Wikis can be used for software documentation
- Collaborative coding
- Version control



Versioning


bcbc-group / Botany2021NMGWorkshop
Public
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
main
1 branch
0 tags
Go to file
Add file
srs218 Merge branch 'main' of https://github.com/bcbc-group/Botany2021NMGWor...
6f4a90d on Jul 18
46 commits
1.Introduction
Create Introduction.pdf
3 months ago
2.BaselineSkills
Add files via upload
3 months ago
3.ExperimentalDesign
4.GenomeAssembly
Folder name change
5.Annotation
added papers
6.DownstreamAnalyses
Add files via upload
.DS_Store
files added
.gitattributes
Initial commit
BTI_BCBC_Bioinfo_unix_command_...
Add files via upload
README.md
Update README.md
Schedule.pdf
Added schedule as pdf
~$hedule.docx
Added schedule as pdf
Commits on Jul 18, 2021
Merge branch 'main' of https://github.com/bcbc-group/Botany2021NMGWor...
srs218 committed on Jul 18
6f4a90d
Create Botany_Workshop_2021.pdf
srs218 committed on Jul 18
a948f12
Add files via upload
afpowell committed on Jul 18
Verified
f4a1f5c
Add files via upload
afpowell committed on Jul 18
Verified
422cca3
Add files via upload
afpowell committed on Jul 18
Verified
873b976
Add files via upload
afpowell committed on Jul 18
Verified
77fd76c
added papers
srs218 committed on Jul 18
4088253
Merge branch 'main' of https://github.com/bcbc-group/Botany2021NMGWor...
srs218 committed on Jul 18
531d089

Reproducible Research

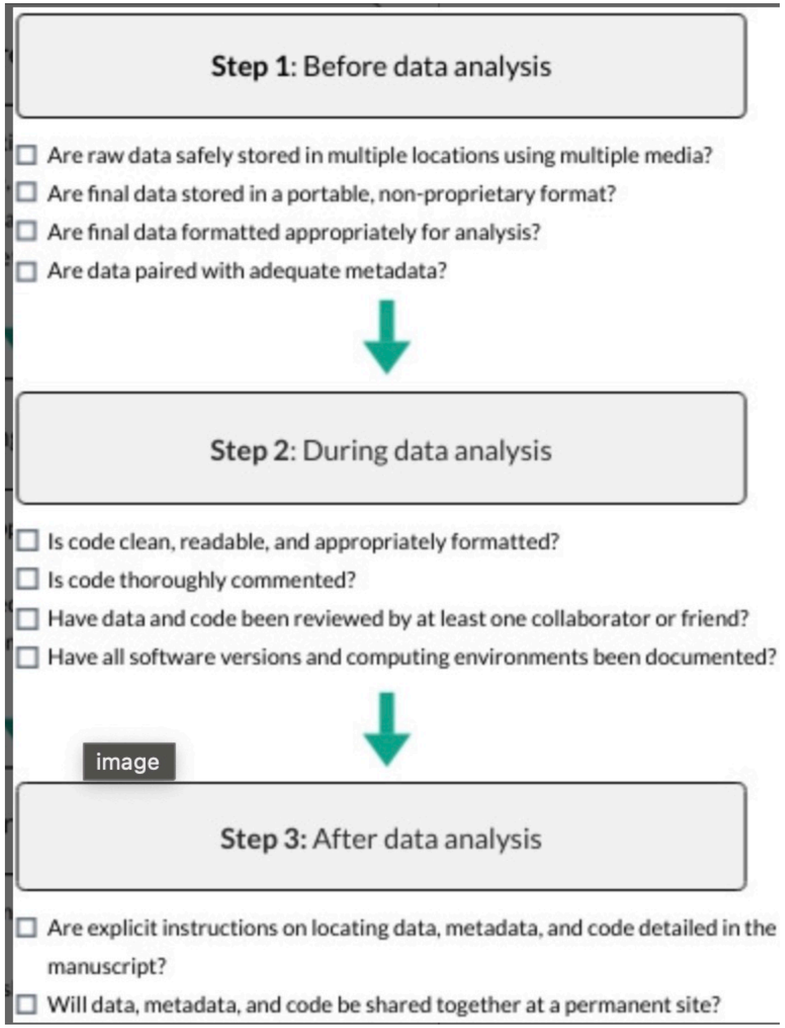
- No “custom scripts”
- Versioning of software and datasets
- System architecture and software
- Making code publicly available
- Raw data is available
- Everything is provided to produce a consistent result
- Reproducible research increases dependability of results and make it easier and more efficient to continue the research.

Reproducible Research

	Free	Open source	Website
Data and code management			
Version control			
GitHub	Y†	N	https://github.com
BitBucket	Y†	N	https://bitbucket.com
GitLab	Y†	Y	https://www.gitlab.com
Make			
GNU Make	Y	Y	https://www.gnu.org/software/make/
Software containers and virtual machines			
Docker	Y	Y	https://docker.com
Singularity	Y†	Y	https://syslabs.io
Oracle VM VirtualBox	Y	Y	https://virtualbox.org
Sharing research			
Preprint servers			
ArXiv	Y		https://arxiv.org/
bioRxiv	Y		https://www.biorxiv.org/
Preprint servers			
ArXiv	Y		https://arxiv.org/
bioRxiv	Y		https://www.biorxiv.org/
EcoEvoRxiv	Y		https://ecoevorxiv.org/
Manuscript creation			
Overleaf	Y†	Y	https://overleaf.com
TeXstudio	Y	Y	https://www.texstudio.org/
Rstudio	Y	Y	https://rstudio.org
Data repositories			
Dryad	N		https://datadryad.org/
Figshare	Y†		https://figshare.com/
Zenodo	Y		https://zenodo.org/
Open Science Framework	Y		https://osf.io/

Alston and Rick, 2021

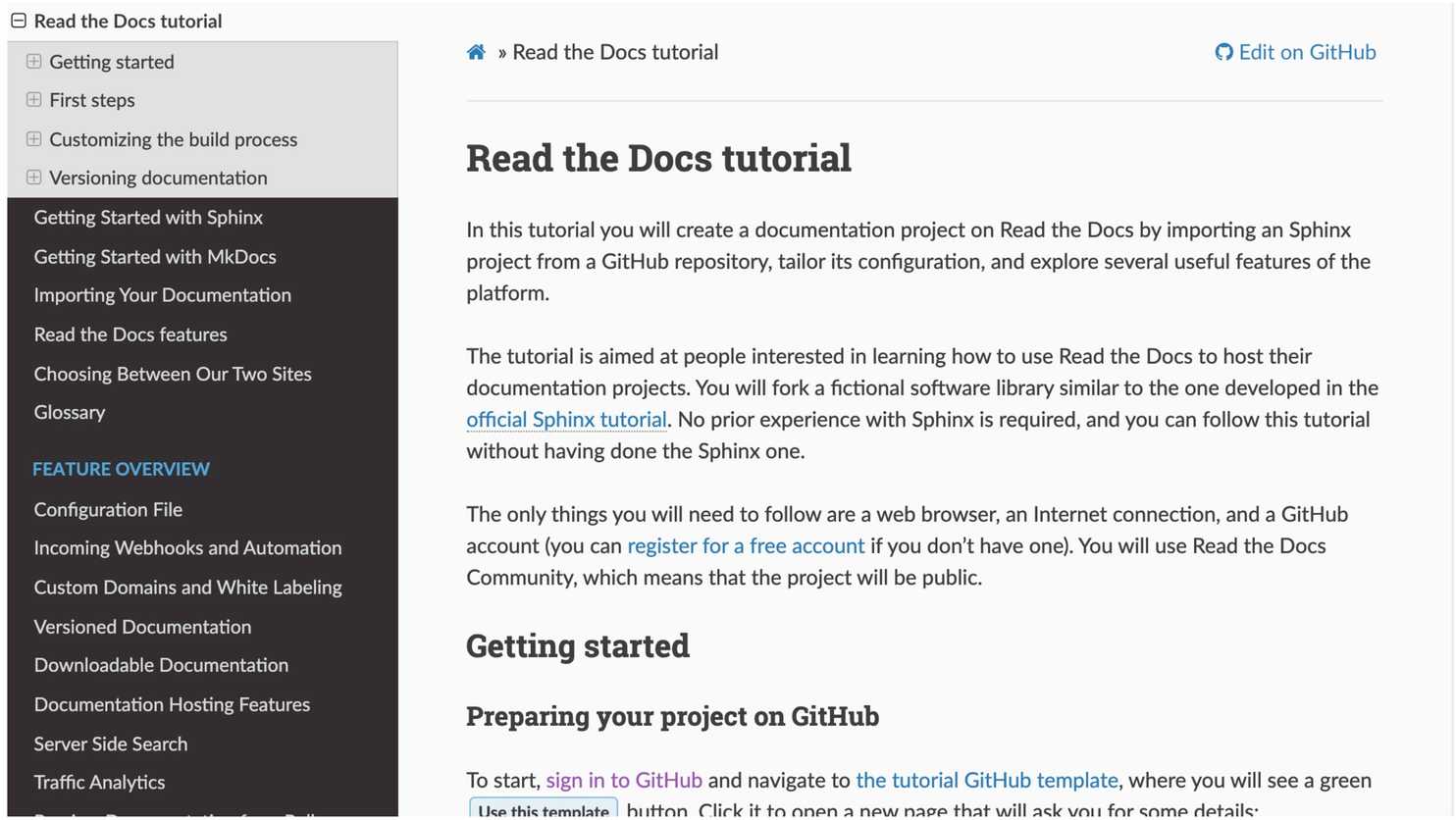
Reproducible Research



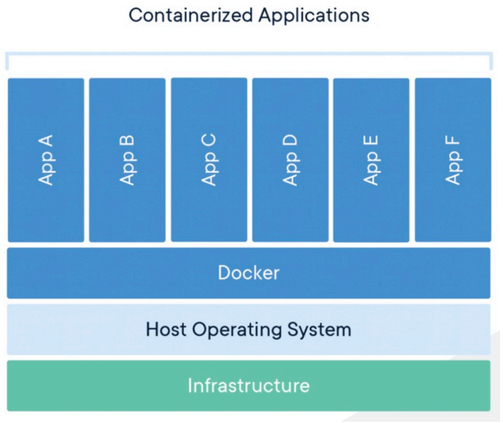
Alston and Rick, 2021

Documentation

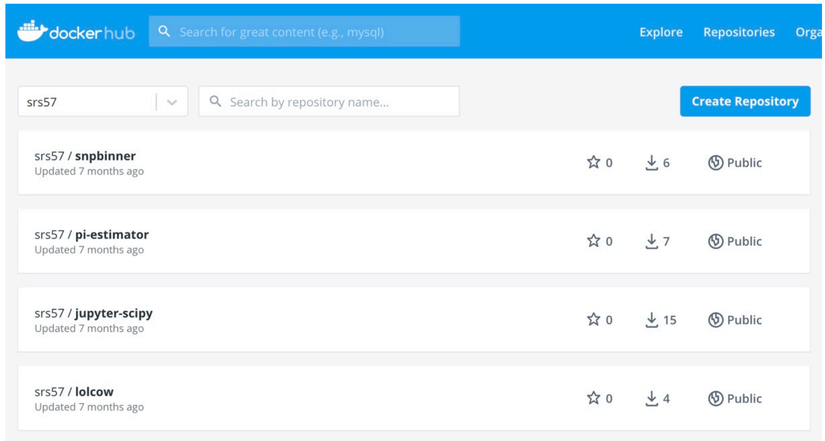
- ReadTheDocs
- Jupyter / Markdown
- Commenting code
- README
- Documenting data - where it came from



Containers



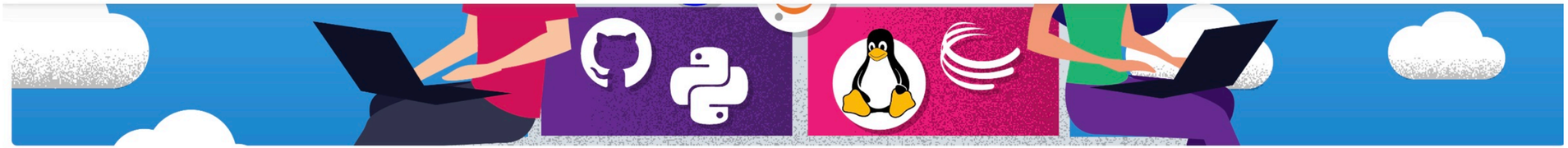
- What is a container?
  - A standardized unit of software that packages code and all of its dependencies so the software can run reliably across platforms (https://www.docker.com/resources/what-container)
- docker
  - Cheatsheet: https://www.docker.com/sites/default/files/d8/2019-09/docker-cheat-sheet.pdf
- dockerhub



Additional Resources

CYVERSE®

Schedule Instructors Contact Us Testimonials



Foundational Open Science Skills

CyVerse's 8-week virtual workshop teaches you the principles, practices, and how-tos for doing collaborative open science using cutting-edge, open source cyberinfrastructure, in a collaborative, hands-on setting. To see how our FOSS workshop can support your work, check out the curriculum.

September 7 - November 2, 2023

8 weekly virtual sessions on Thursdays; 11 am - 1 pm Pacific Daylight Time (or MST)
September 7 is an optional, pre-session brush-up on Git and Unix with an intro to ChatGPT

REGISTER NOW

7. Computational infrastructure

- Powerful servers with large amounts of memory, compute cores, and disk

8. Careers

- Academia/Industry/Goverment
- BS/MS/PhD
- Bioinformatics core
- Professor
- Scientist
- Biomedical/Pharma
- Plant Breeders
- Botanic Gardens ;)

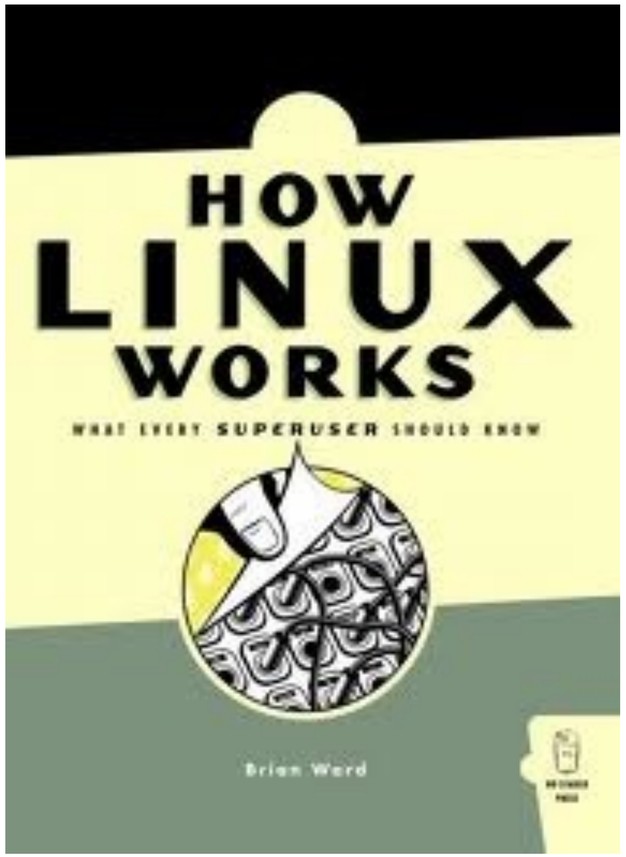
- Tends to have higher pay than other biology jobs
- Many positions offer remote work option
- https://www.biology.pitt.edu/sites/default/files/publication-images/BINF%20Career.pdf

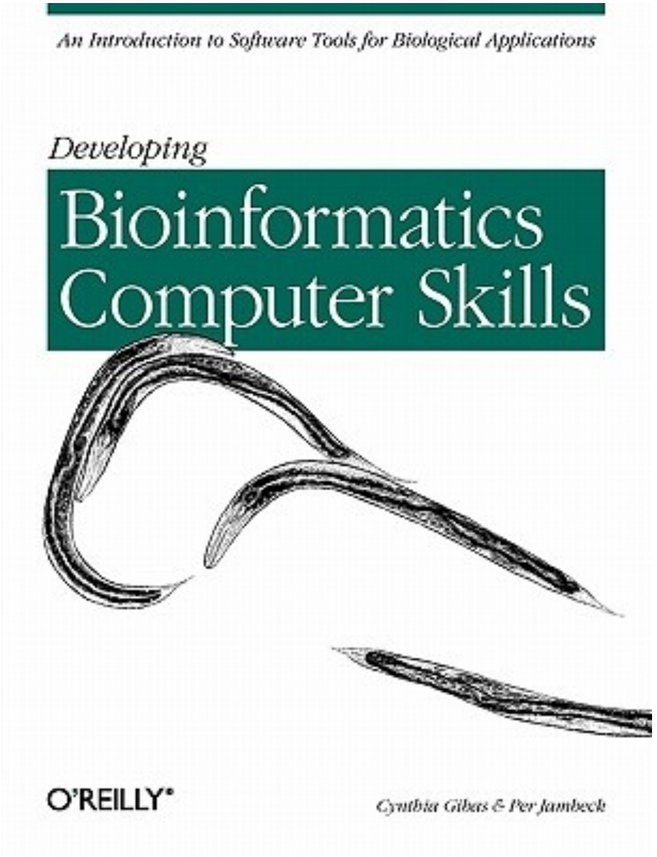
More on Bioinformatics

CYVERSE®

- https://btiscience.org/our-research/computational-biology/bcbc-symposium/bcbc-2018-symposium/

Further Reading





Empowering Beginners in Bioinformatics with ChatGPT

Evelyn Shue,[1] Li Liu,[2,3] Bingxin Li,[4] Zifeng Feng,[5] Xin Li,[6] and Gangqing Hu[1,*]

Author information Copyright and License information Disclaimer

The complete version history of this preprint is available at bioRxiv.

Next time

- Who would like to attend?
- You will need a Linux Virtual Machine
    - Instructions for installing it here: https://btiscience.org/our-research/computational-biology/educationtraining/intern-2023-bcbc-bioinformatics-course/
    - Please try installing it today and let me know by end of day if it doesn’t work for you