Session 2: Data workflows for analysis and interpretation

Instructors for Session:

Stanford Kwenda, George Githinji, Fatma Guerfali, Kirsty Lee Garson, Aquillah Kanzi, Leigh Jackson, Amadou Diallo

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Session outline

- Provide an overview of pathogen genomics applications in public health
- Describe bioinformatics workflows including examples of existing workflows
- Give a brief description of pathogen types
- Group activity 1
- Workflow managers
- Introduce reporting and data integration
- Group activity 2

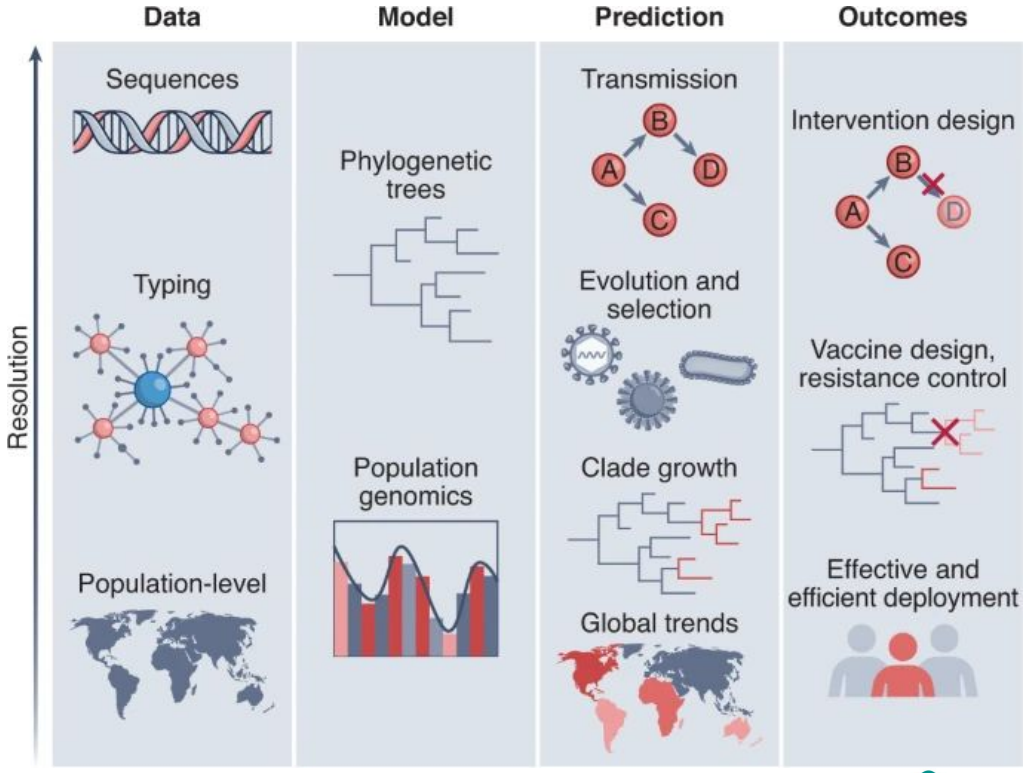
Session outcomes

At the end of this session participants will be able to:

- Describe common applications of pathogen genomics in public health & surveillance
- Outline common steps of bioinformatics analysis workflows
- Understand the use of workflow managers in pipeline development and automation
- Identify existing pipelines and workflows for analysis of different pathogen types
- Introduce data integration tools

Common applications of pathogen genomics in public health & surveillance

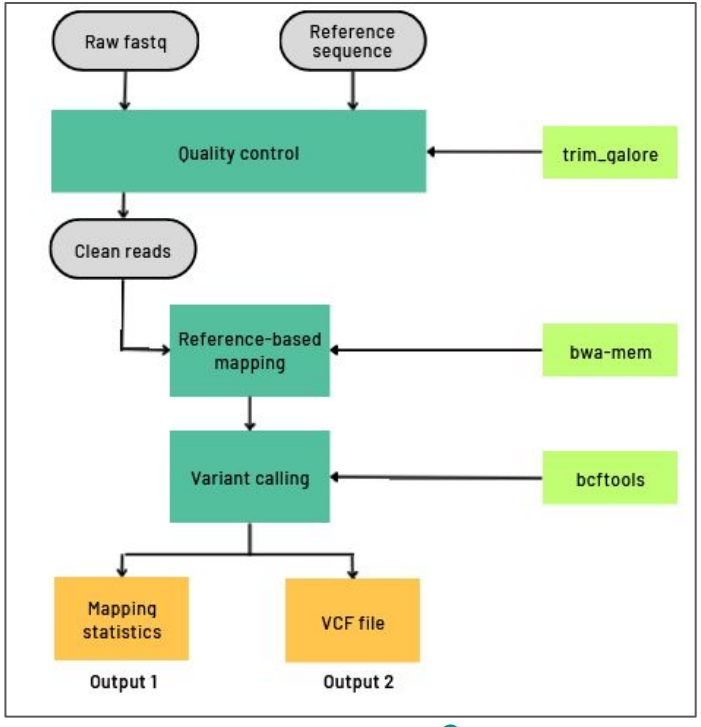
Genomic data may comprise Sequences, to help model sequence types or genotypes, to forecast global disease trends and thereby design an efficient resource deployment strategy.



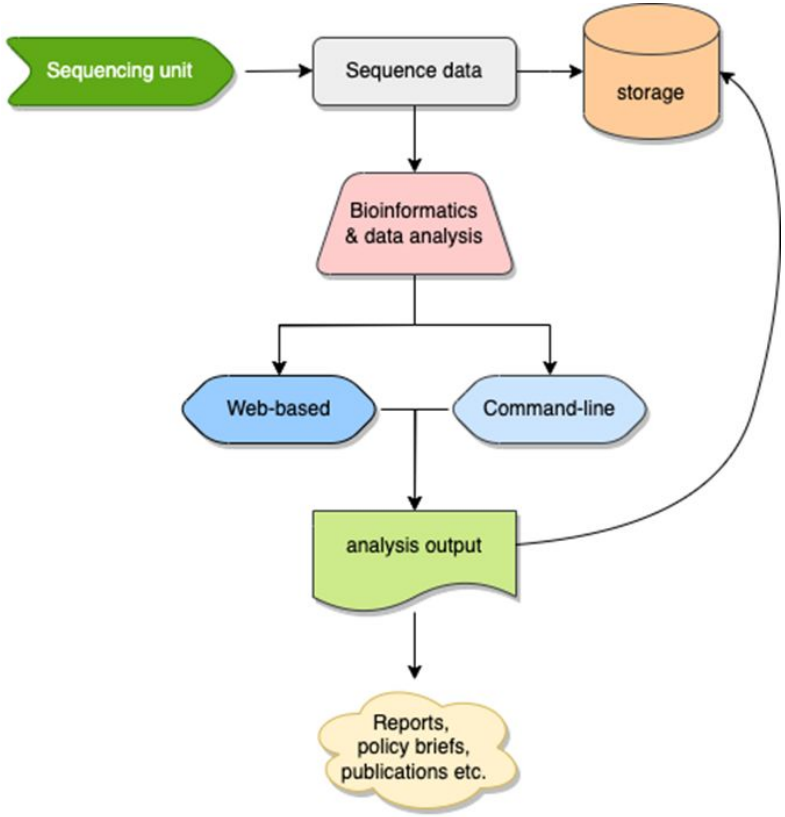
Stockdale et al. *Nat Microbiol* 7, 1736–1743 (2022).

Bioinformatics data analysis

- Standard analysis tasks often involve running a large number of tools
  - Transforming data into information
  - Can employ several open-source standalone tools
- Each tool can be executed individually
- Multiple tools can be chained together into pipelines
  - Developed using custom Bash scripts or Make files



Standard bioinformatics Pipeline - overview

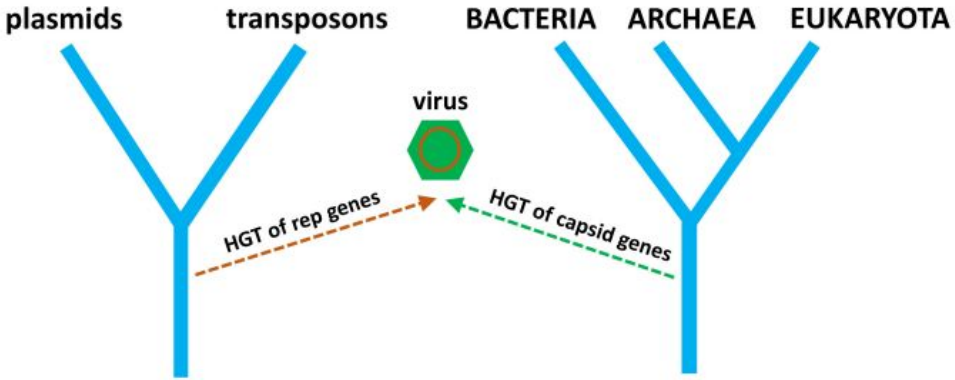


Things to consider:

- Sequence data format
  - e.g. different raw data formats from different sequencing platforms
- Computing infrastructure and resources
- Level of bioinformatics expertise
  - Web-based tools vs command-line based tools
- Genetic characteristics of the sequenced organism
- Purpose of the analysis
- Reports format
  - Spreadsheets, dashboards

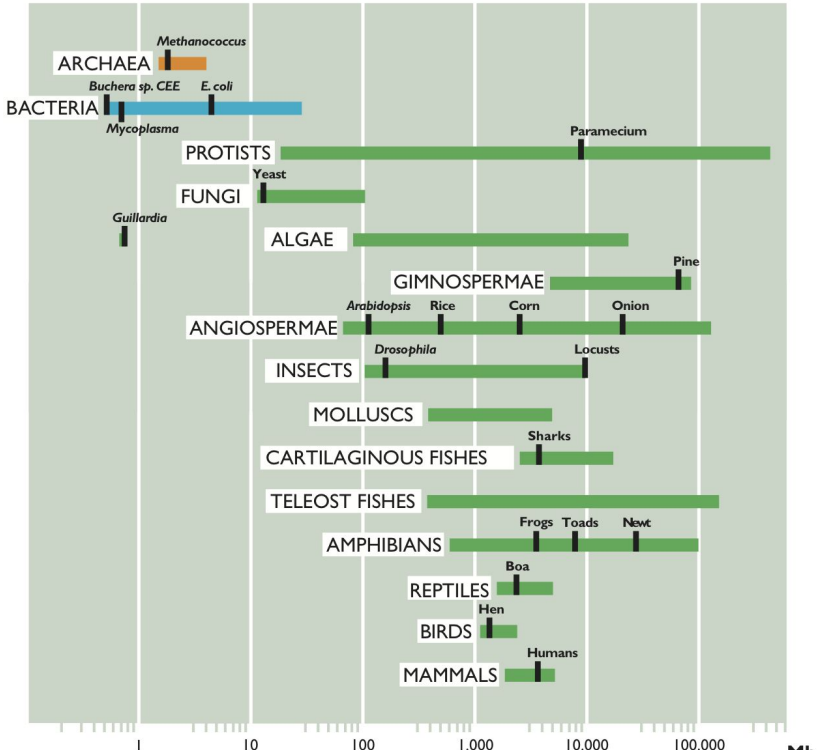
A note on genome sizes and complexity

Domains of life (Tree of Life)



Harris & Hill (2021) 10.3389/fmicb.2020.604048

What is a small or a big genome?



https://metode.org/issues/monographs/the-size-of-the-genome-and-the-complexity-of-living-beings.html

Components of a standard analysis pipeline/ workflow

Contamination check and classification

Input:
fastq or contigs

Tools:
fastQC, MultiQC, kraken2, sourmash, bactInspector, Confindr, speciesFinder

Typing and/or seotyping

Input:
fastq or contigs

Tools:
mlst, SeroTypeFinder, SeqSero2*, SISTR*, Kleborate*,

*Some tools are species specific

AMR and virulence prediction

Input:
fastq or contigs

Tools:
resfinder, AMRFnderPlus, ABRicate, staramr, pointfinder,ARIBA

*Some tools only take contigs as input

Phylogenetic analysis based on cgMLST, wgMLT, SNP analysis

Input:
fastq or contigs

Tools:
ska, kSNP3, PopPUNK , kraken2, rapidnj, fastree, iqtree, RAxML, BEAST

*Sequence alignments can be reference-free or reference based

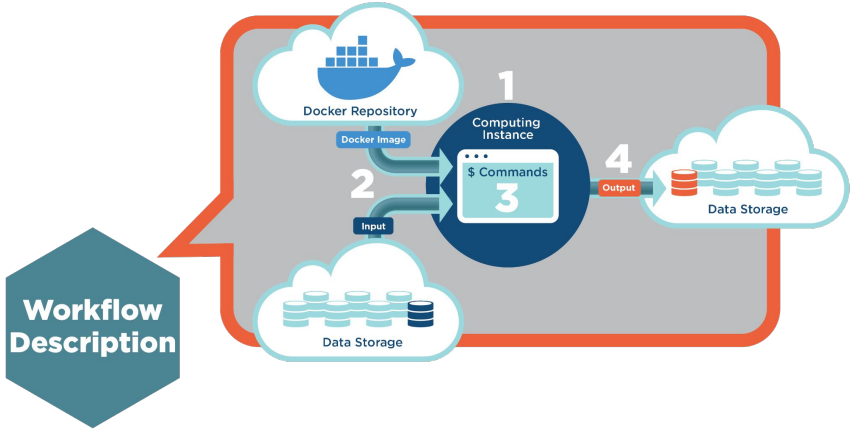
Things to consider:

- Choice of tool
  - Sequencing platform used
  - Processing time and available options
- Choice of databases
  - AMR prediction databases
  - Classification of isolates
  - Typing e.g MLST profile to use
- Specialized typing
  - Serotyping
  - Phylogroup detection
  - Plasmid detection
- Phylogenetic analysis
  - Different models and methods

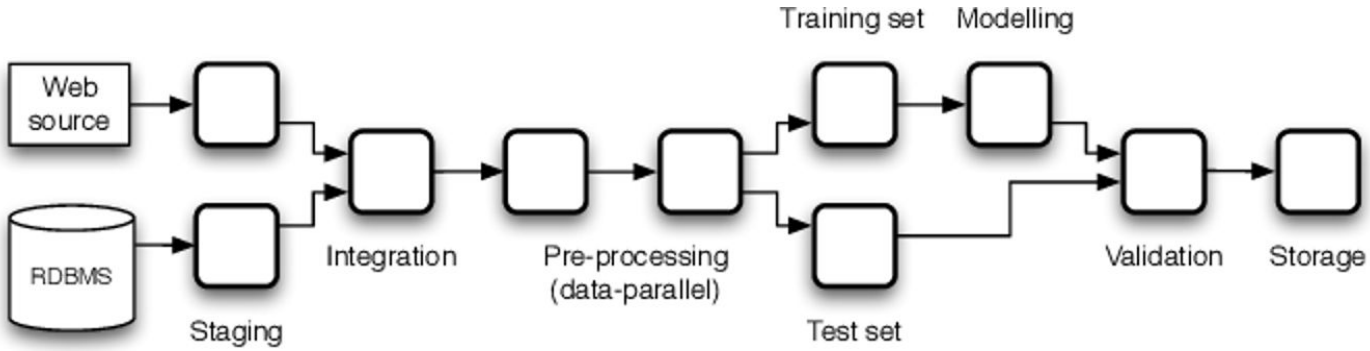
Lists of tools not exhaustive, only showing some popular examples for bacterial genomics

What are workflows?

- A system for managing repetitive processes and tasks which occur in a particular order.



- Example Workflow



Examples of pipelines/workflows by pathogen group

Command line interface (CLI)

Pathogen group	Workflow name	Workflow manager	Repository
Viruses	WhoFlu IRMA	SnakeMake	https://github.com/ammaraziz/wfi
	Nextstrain	SnakeMake	https://docs.nextstrain.org/en/latest/index.html
	Pangolin	-	https://cov-lineages.org/resources/pangolin.html
Bacteria	Bactmap	Nextflow	https://nf-co.re/bactmap
	Bactopia	Nextflow	https://bactopia.github.io/v2.2.0/
	Jekesa	-	https://github.com/stanikae/jekesa
	rMAP	-	https://github.com/GunzIvan28/rMAP
	TORMES	-	https://github.com/nmquijada/tormes
Fungi	MycoSNP	Nextflow	https://github.com/CDCgov/mycosnp-nf
	FungiPhyloGen#	Nextflow	https://github.com/stanikae/FungiPhyloGen
	NASP	-	https://github.com/TGenNorth/NASP
Parasites	ConTest	SnakeMake	https://github.com/sdune/contest
	sch_man_nwinvasion	-	https://github.com/nealplatt/sch_man_nwinvasion/releases/tag/v0.2
	LGAAP	Snakemake	https://github.com/hatimalmutairi/LGAAP

Key:
- Traditional analysis pipeline
Not yet open access

Examples of pipelines/workflows by pathogen group

Cloud/Web based tools

EDGE Bioinformatics (Web-based + Local)	https://edgebioinformatics.org/	- Environmental surveillance - Infectious diseases
Terra Bioinformatics (Web-based)	https://terra.bio/	- Genomics & Transcriptomics - Infectious diseases
CZ ID (CZ BioHub) (Web-based)	https://czid.org/	- Metagenomics - Genomic epidemiology
PathogenWatch	https://pathogen.watch/	- Species and taxonomy prediction for bacteria, viruses and fungi
Center for Genomic Epidemiology	http://www.genomicepidemiology.org/services/	- Various bacterial typing and phylogenetic analysis tools

Key:
- Traditional analysis pipeline
Not yet open access

Suggestions for more resources :

PHA4GE pipelines resources :
https://github.com/pha4ge/pipeline-resources

Nf-core (NextFlow) community pipelines :
https://github.com/nf-core

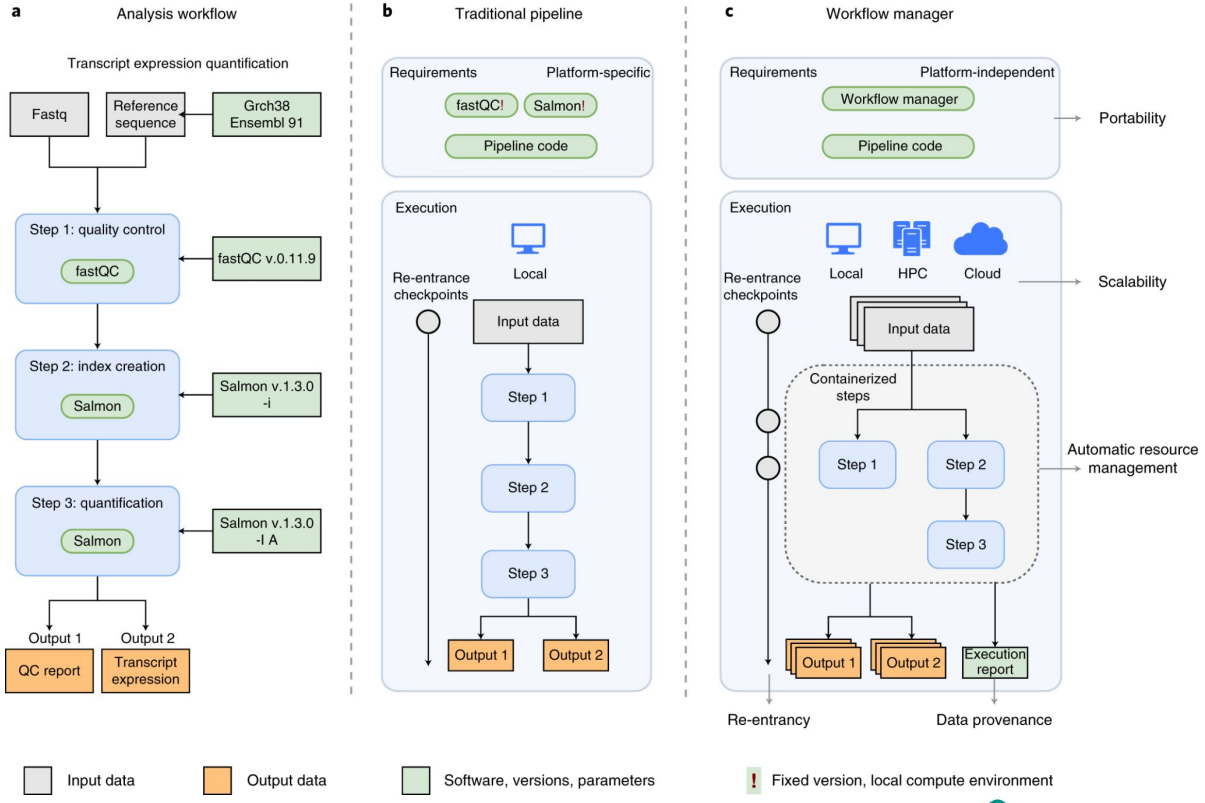
Theiagen PHB :
https://github.com/theiagen/public_health_bioinformatics

Dockstore:
https://dockstore.org/

Workflowhub:
https://workflowhub.eu/

Pipeline vs Workflow managers

Illustration of analysis performed using workflow managers or traditional pipelines



Wratten, et al. *Nat Methods* 18, 1161–1168 (2021)

Executable program vs Pipeline vs Workflow managers

Executable program

- Large array of open-source bioinformatics tools
- Most individual tools carry out a single specialized analysis

Pipeline

- Allows the automatic chaining of multiple tools for recurrent data analysis
- Heavily impacted by local infrastructure, documentation, tool versioning, and installation processes on other devices: difficulty for sharing, maintaining, and ensuring reproducibility

Workflow managers

- Create a framework for the “creation, execution, and monitoring of pipelines”
- Automates the software installation process, ensures the portability across platforms, and improves reproducibility using package managers and containerization

Wratten, et al. *Nat Methods* 18, 1161–1168 (2021);
https://albertvilella.substack.com/p/workflow-management-software-in-bioinformatics

Group activity 1

1. Given a set of sequences e.g. raw fastq files or assembled contigs, list any 3 things that you would consider when deciding/planning your analysis. Also provide reasons for your answers above.
2. Your lab recently acquired a sequencer, and completed its first successful sequencing run and generated data for 5 *E. coli* isolates. You have been tasked with analyzing this data, however, your lab doesn't have any computing infrastructure e.g. servers or HPCs. How will you go about analyzing this data? Think about the analyses you would need to perform to characterize these isolates? State the resources you would use to achieve this task?
3. Assuming that your institution has a well established single node computing server, how will you analyze the following data sets:
   a. 100 SARS-CoV-2 samples
   b. 55 *Vibrio cholerae* isolates
   c. 100 *Plasmodium falciparum* amplicons

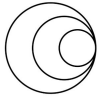
Pro tip: Consider the different CLI tools/workflows provided in the lecture - not all of them will be applicable.

Bioinformatics workflow managers

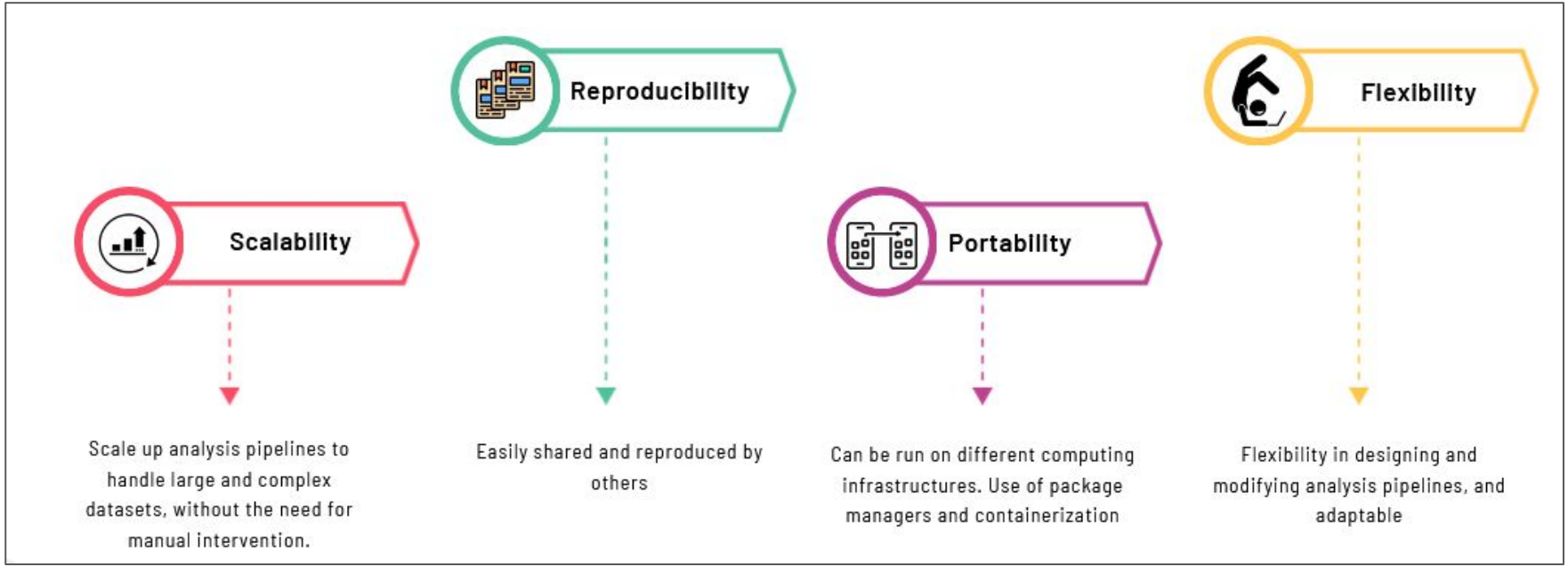
- Provide integration with:
  - Containers e.g. Docker and Singularity
  - Package managers e.g. Conda
  - Cloud computing
- Automatic resource management
  - Tasks parallelization
  - Handling dependencies
  - Scheduling
- Major examples include:
  - Galaxy, Nextflow & SnakeMake

Tool	Class	Ease of use[a]	Expressiveness[b]	Portability[c]	Scalability[d]	Learning resources[e]
Galaxy	Graphical	●●●	●○○	●●●	●●●	●●●
KNIME	Graphical	●●●	●○○	○○○	●●◐	●●●
Nextflow	DSL	●●○	●●●	●●●	●●●	●●●
Snakemake	DSL	●●○	●●●	●●◐	●●●	●●○

Wratten et. al 2021. https://doi.org/10.1038/s41592-021-01254-9

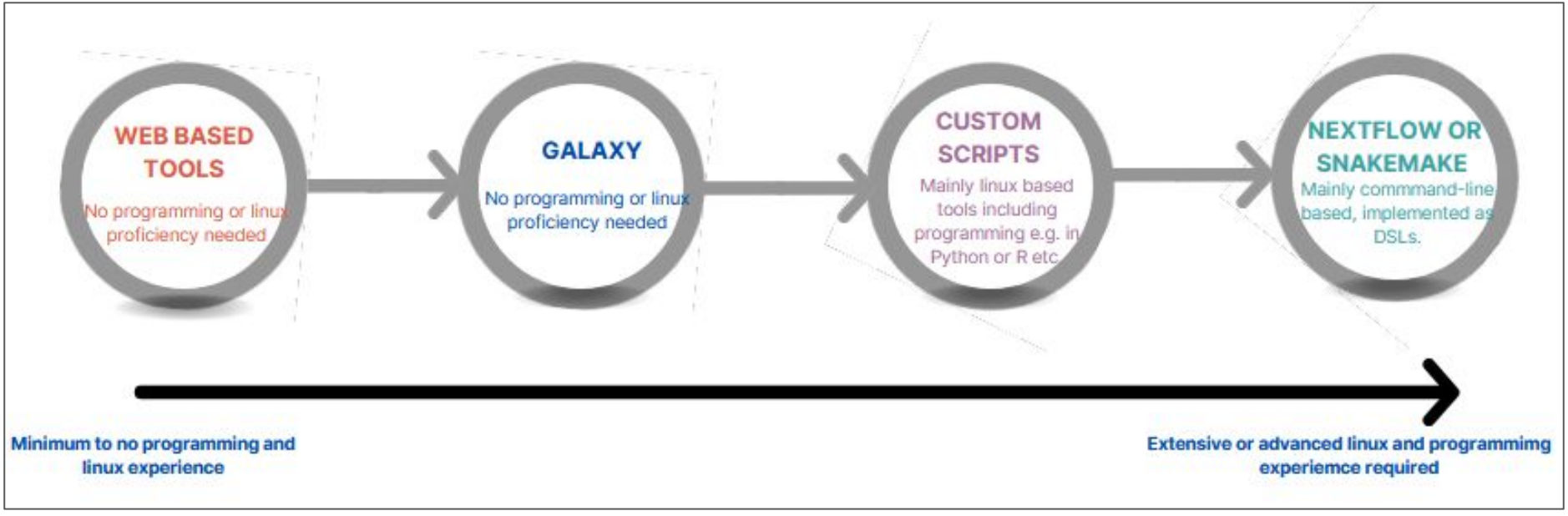


Key advantages of using workflow management tools



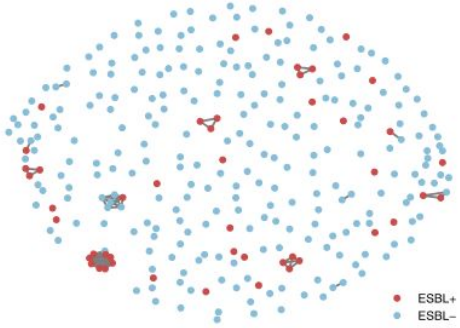
https://albertvilella.substack.com/p/workflow-management-software-in-bioinformatics

Choice of tool or workflow?



Data integration and reporting

- Integration of SNP information and epidemiology data
  - Transmission networks
    - Clonal vs plasmid mediated outbreak
    - Single strain mediated outbreak vs multiple introductions
  - Phylogenetic trees
  - SNP cluster analysis
    - E.g. using unsupervised ML methods
- AMR mechanisms
  - Antimicrobial susceptibility testing (AST) results vs predicted AMR results
- Identify high risk clones/strains
  - Strain typing
- Source of infection/outbreak
  - Phylogenetic analysis/ Transmission networks
  - Clinical/Epidemiological data



Gorrie et al (2022).
https://doi.org/10.1038/s41467-022-30717-6

Data integration and reporting

- Examples of reporting tools:
  - Microreact (https://microreact.org/)
  - iTOL (https://itol.embl.de/)
  - PathogenWatch (https://pathogen.watch/)
  - Center for Genomic Epidemiology (http://www.genomicepidemiology.org/services/)
  - NextStrain (https://nextstrain.org/)
  - Figtree (http://tree.bio.ed.ac.uk/software/figtree/)
  - Various R packages and/or python libraries
    - E.g. APE, ggtree

Web based

Microreact

Pathogenwatch

ggtree

MEGA Molecular Evolutionary Genetics Analysis

Nextstrain

Center for Genomic Epidemiology

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Group Activity 2: Working with data in a surveillance scenario

You are the data management group for a response to an outbreak of Marburg virus

- Work with your groups to design an analysis workflow system
- You have to process 10 samples per day and diagnosis is critical
- You have access to human resources and budget for equipment
  - But not infinite budgets! You will be audited after
- Simplicity and speed are priorities