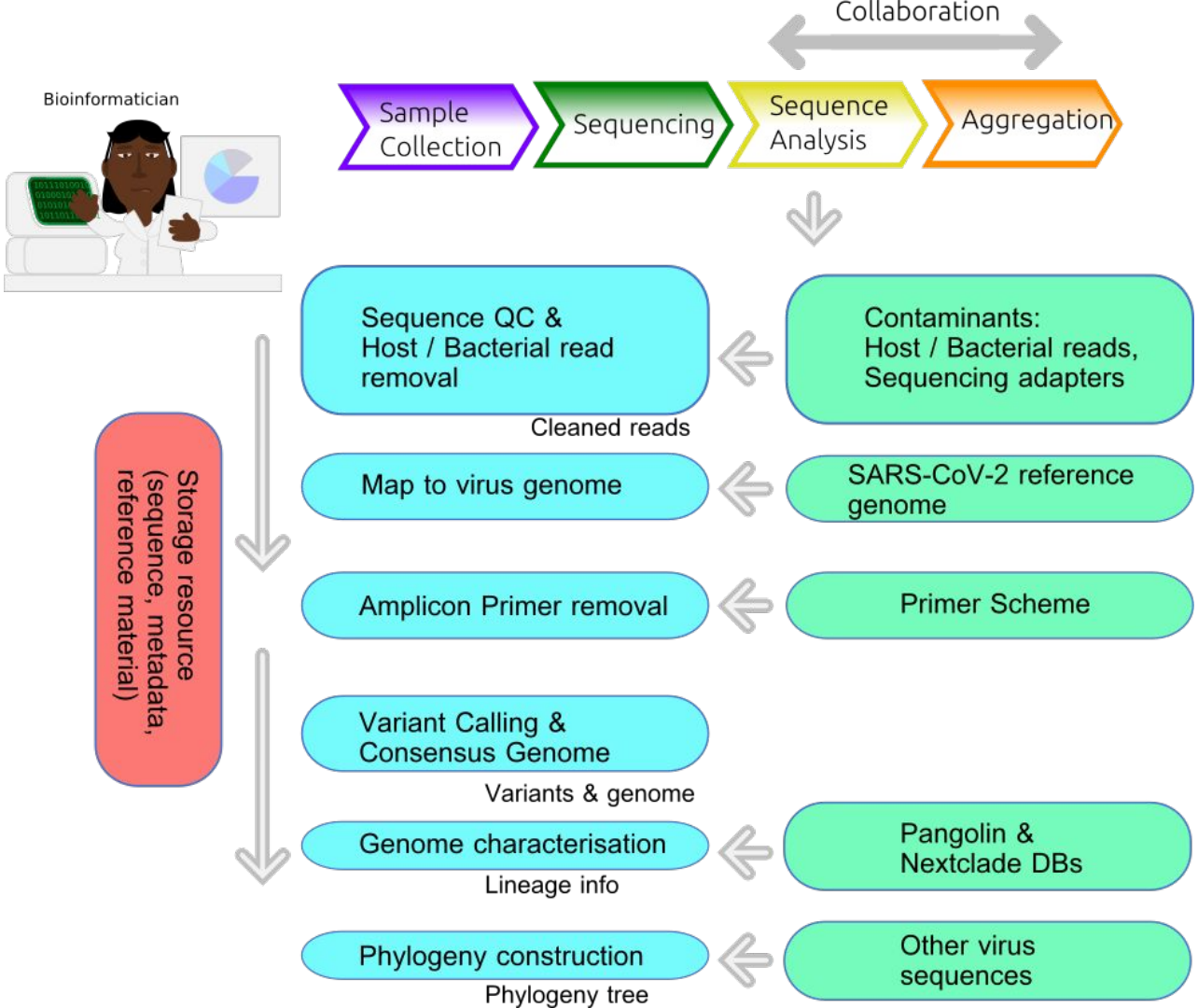
UNIVERSITY *of the*
WESTERN CAPE

SANBI
South African National
Bioinformatics Institute

SARS-CoV-2 sequence analysis workflows

Peter van Heusden pvh@sanbi.ac.za,

South African National Bioinformatics Institute


Collaboration
Bioinformatician
Sample Collection
Sequencing
Sequence Analysis
Aggregation
Sequence QC & Host / Bacterial read removal
Contaminants: Host / Bacterial reads, Sequencing adapters
Cleaned reads
Map to virus genome
SARS-CoV-2 reference genome
Amplicon Primer removal
Primer Scheme
Storage resource (sequence, metadata, reference material)
Variant Calling & Consensus Genome
Variants & genome
Genome characterisation
Pangolin & Nextclade DBs
Lineage info
Phylogeny construction
Other virus sequences
Phylogeny tree

SARS-CoV-2 workflow platforms

The analysis workflow can be run as:

- Software as a Service
- Platform as a Service
- Self-management infrastructure

Software as a Service (SaaS)

- Software & infrastructure is provided
- No user input on workflow
- Data management as part of platform
  - E.g. IdSeq, EDGE (open source), Genome Detective, Exatype (proprietary)

Self-managed SaaS

- Open Source SaaS can be locally installed
  - EDGE (from Los Alamos National Labs)
  - COMBAT Sars-CoV-2 Workbench
    - from SANBI
    - Manages metadata & sample together
    - Executes Galaxy workflows (“behind the scenes”)

Platform as a Service (PaaS)

- Platform for running user-specified workflows
- More flexible, more complex than SaaS
- Data management part of platform

PaaS example: Galaxy

- Galaxy workflow platform
  - Public servers: usegalaxy.eu, usegalaxy.org
  - Can be installed locally
  - Visual workflow editor
- Data management
  - Data stored on servers in “histories”
  - Not designed for long term data storage

PaaS example: Broad Terra

- Broad Terra platform
  - Maintained by Broad Institute
  - Running on Google Cloud Platform
  - Workflows are WDL run using Cromwell

- Data management
  - Storage as spreadsheets & files on GCP

Self managed infrastructure

Self managed infrastructure is diverse

- “On prem” servers
- Shared (e.g. regional or national) HPC
- Commercial or research cloud (IaaS)

- Typically command line oriented

Workflow execution self-managed infrastructure

- Command line workflow systems
  - Nextflow
    - Nextflow Tower
  - Snakemake
  - WDL

Data management on self-managed systems

- Typically on “shared storage”
  - Structured as directories and files
- Storage during analysis vs
- Storage of samples & analysis results

Beyond the Genome

- Genotyping
  - Nextclade
  - Pangolin

- Visualising genotyping results
  - Sampling strategy
  - Turnaround time
  - Data sharing vs data publication

Building the tree

- Phylogenies:
  - Collect genomes, align, build tree
    - Mafft & IQtree, Nextstrain augur, BEAST
  - Which tree? Outbreak? Regional? Global?

Building the tree

- Phylogeny gains value with context
  - “Related” samples
    - UShER - place your sample in a tree
  - Metadata
    - SARS-CoV-2 context data specification
- Visualisation:
  - Nextstrain, Microreact, MicrobeTrace

PHA4GE

- PHA4GE is the Public Health Alliance for Genomic Epidemiology
- Established in 2019 and open to all working in the field of bioinformatics & genomic epidemiology with a Public Health focus
- https://pha4ge.org
- Pipelines & Visualisations Working group docs:
  - https://github.com/pha4ge/pipeline-resources

Any questions?