Enabling reproducible data analysis for metagenomics

eResearch Africa Conference 2017

Gerrit Botha | CBIO | H3ABioNet

3 May 2017

Outline

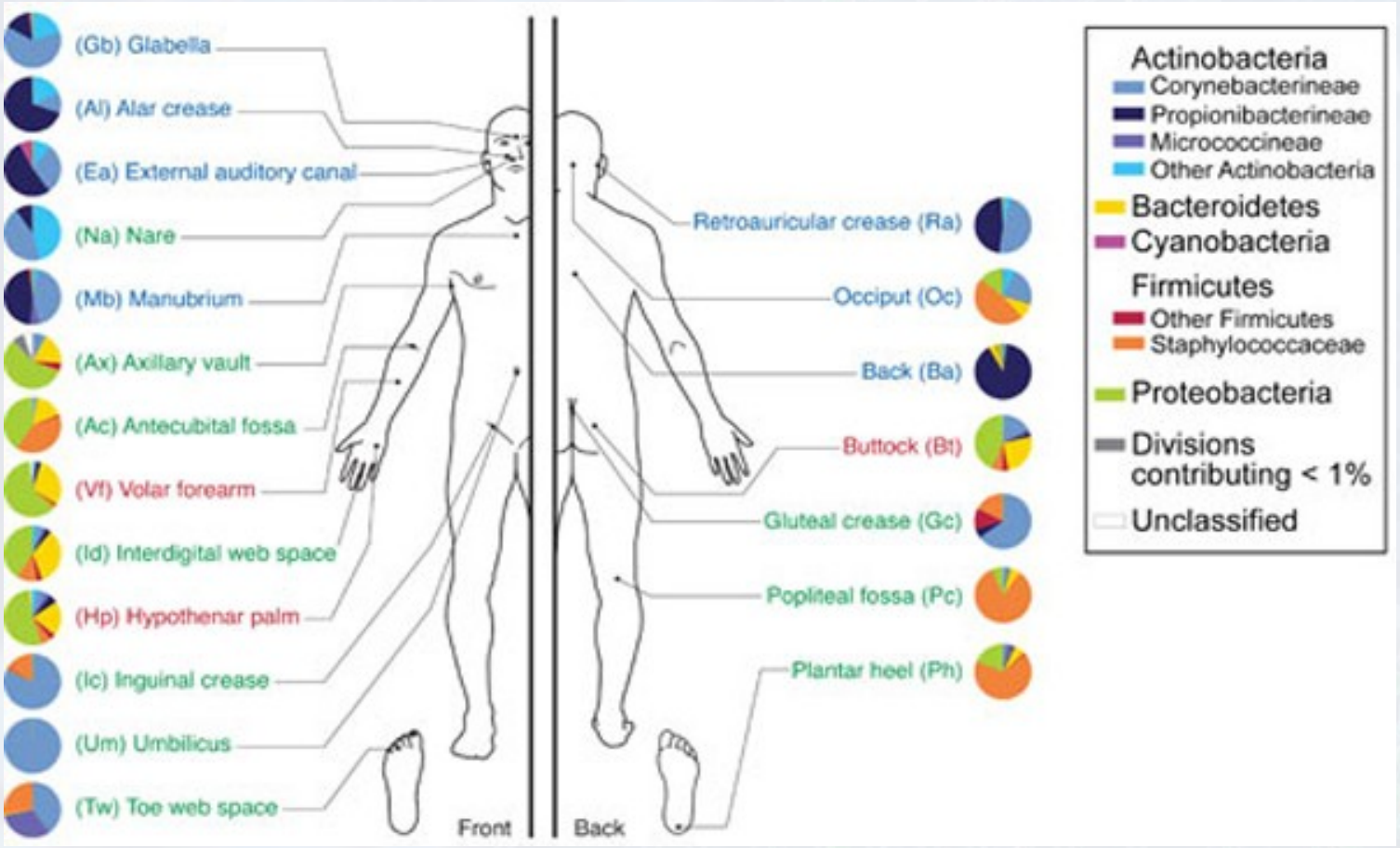
- 16S rRNA analysis
- Current CBIO 16S rRNA analysis setup
- H3ABioNet hackathon
- New h3abionet16S package setup
- Run stats on h3abionet16S setup
- Advantages of h3abionet16S setup
- Future work

What is in a microbiome?

Human microbiome numbers

- 100 trillion microbes
- 3% human body mass
- 1-10X microbes : human cells
- Largest # microbes – GI tract

Human bacterial distribution



Grice et al, Science 2009
http://blogs.discovermagazine.com/notrocketscience/2009/05/28/the-bacterial-zoo-living-on-your-skin/#.VzX2f5MrIdV

Types of analysis

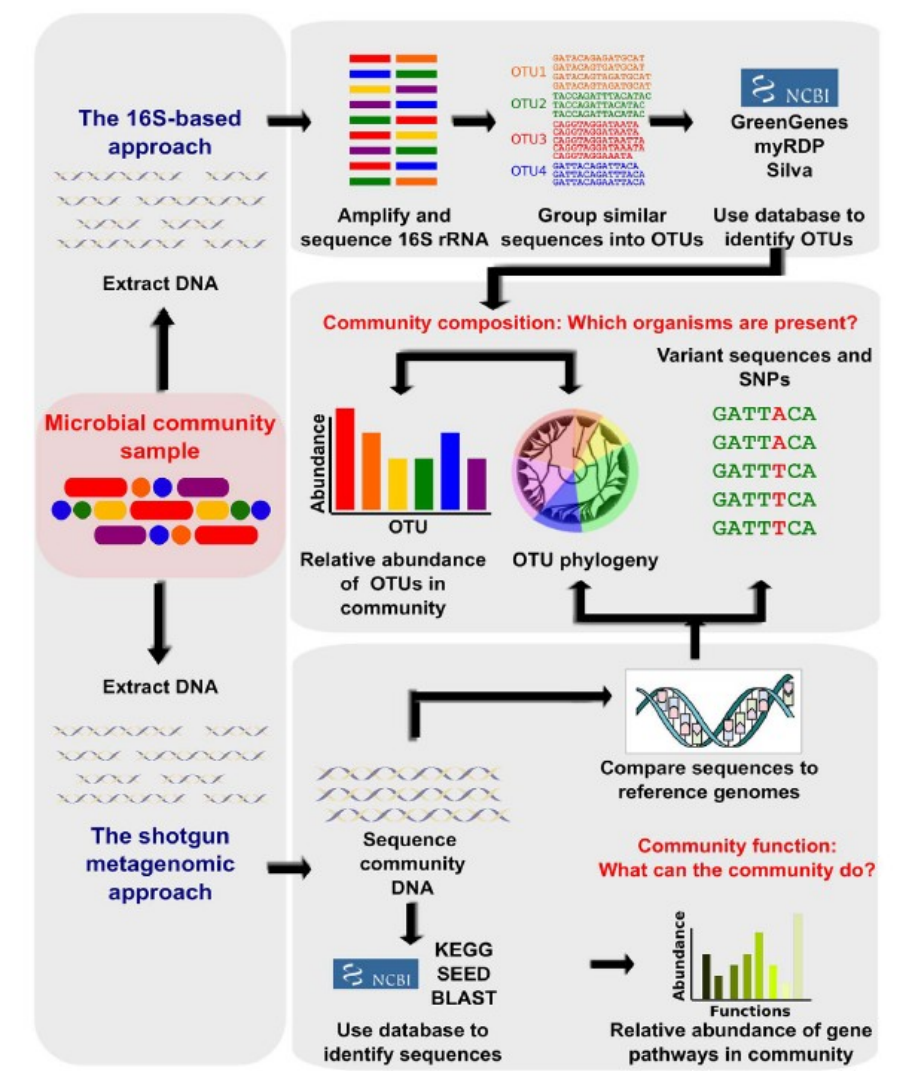
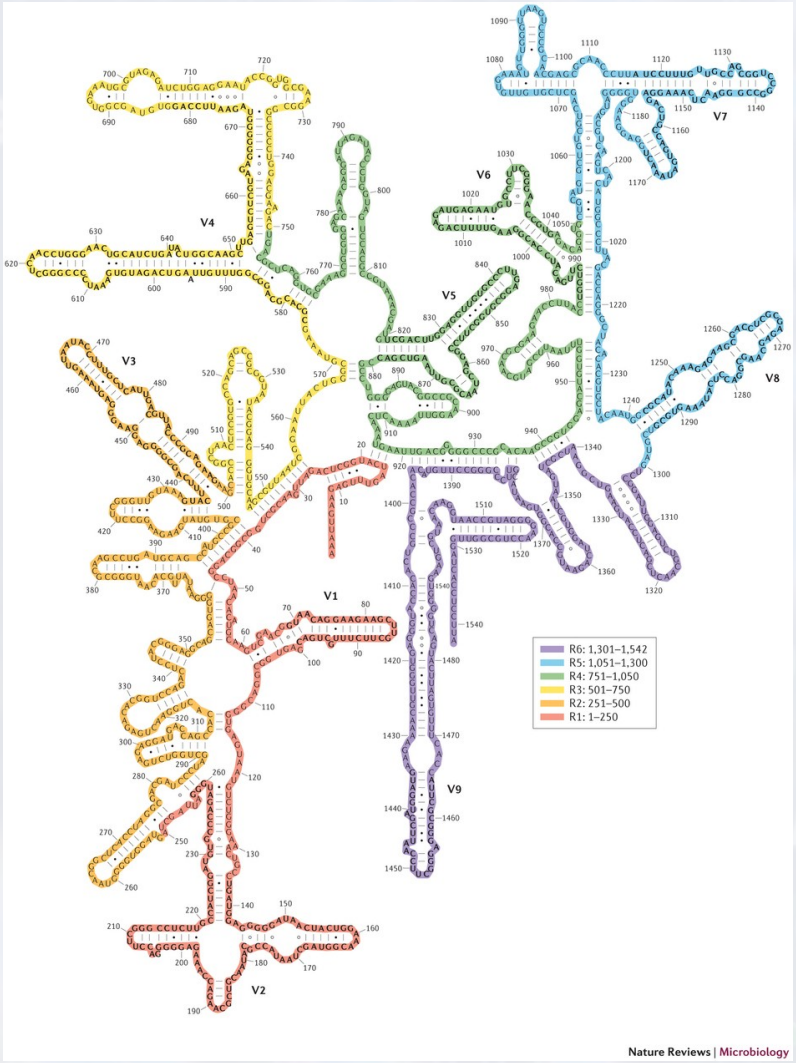


Figure 1. Bioinformatic methods for functional metagenomics. Studies that aim to define the composition and function of uncultured microbial communities are often referred to collectively as "metagenomic," although this refers more specifically to particular sequencing-based assays. First, community DNA is extracted from a sample, typically uncultured, containing multiple microbial members. The bacterial taxa present in

From: Morgan XC, Huttenhower C (2012) Chapter 12: Human Microbiome Analysis. PLoS Comput Biol 8(12): e1002808. doi:10.1371/journal.pcbi.1002808

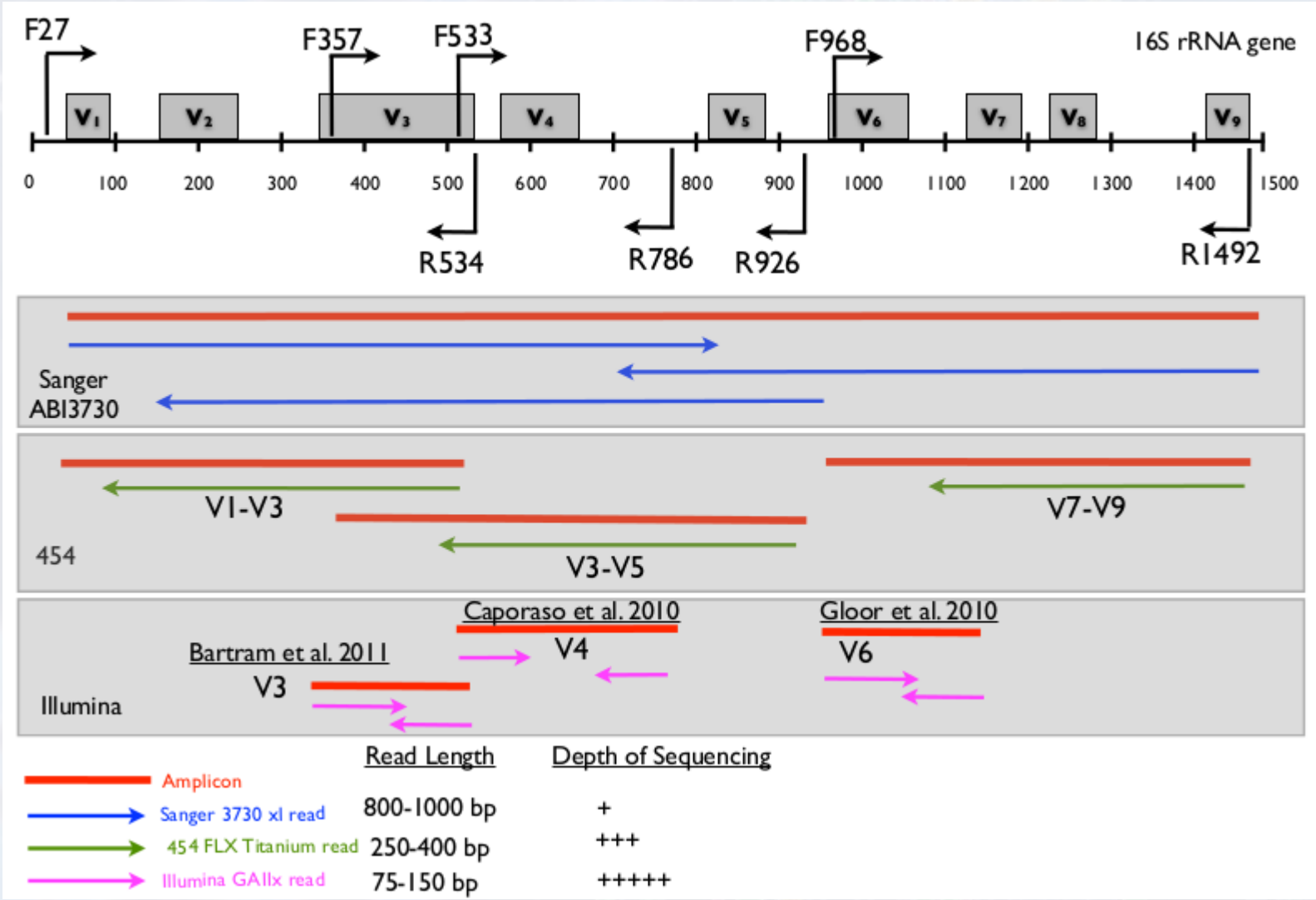
16S rRNA analyses



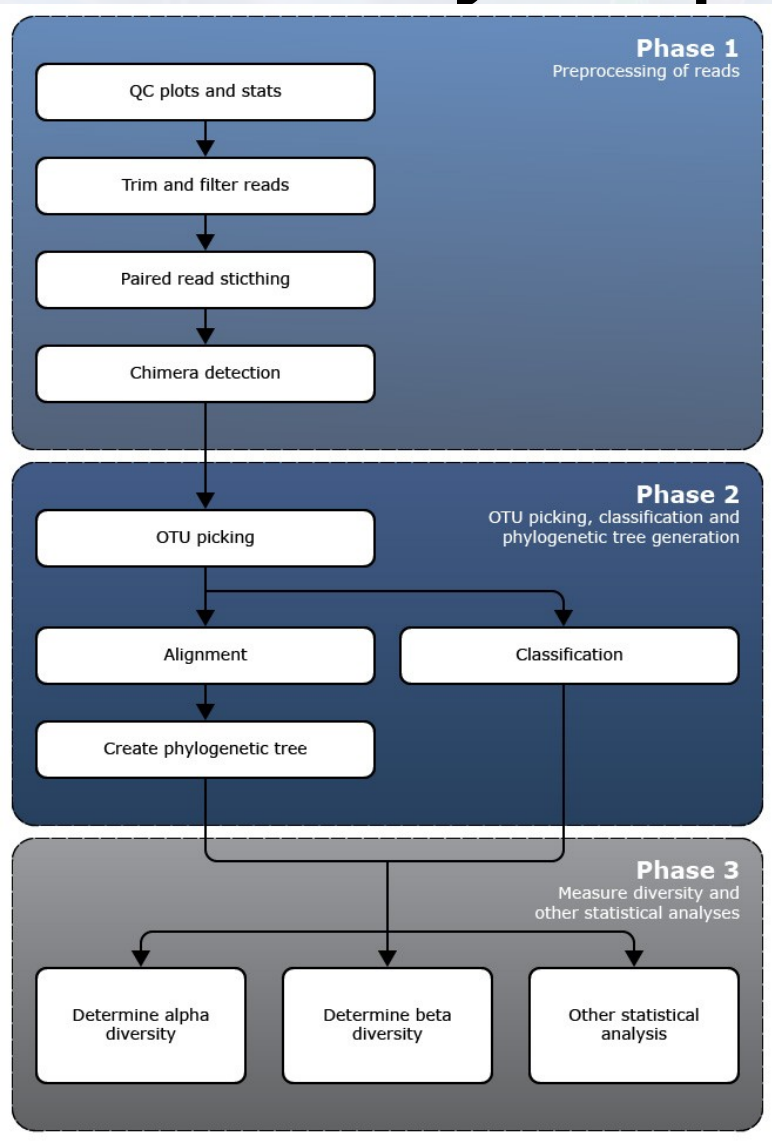
- The genes encoding the RNA component of the small subunit of ribosomes, commonly known as the 16S rRNA in bacteria and archaea, are among the most conserved across all kingdoms of life.
- They contain regions that are less evolutionarily constrained and whose sequences are indicative of their phylogeny.
- Amplification of these genomic regions by PCR from an environmental sample and subsequent sequencing of a sufficiently large number of individual amplicons enables the analysis of the diversity of clades in the sample and a rough estimate of their relative abundance.
- Cheap, with many standard analysis tools.

Nature Reviews Microbiology 12, 635–645 (2014)

16S rRNA sequencing



16S rRNA analysis pipeline



Available pipelines

- UPARSE: http://www.drive5.com/uparse
- IM Tornado: https://github.com/pjeraldo/imtornado2
- QIIME: http://qiime.org
- Mothur: http://www.mothur.org
- VSEARCH: https://github.com/torognes/vsearch
- FROGS: https://github.com/geraldinepascal/FROGS
- DADA2: https://github.com/benjjneb/dada2

Current setup

- CBIO has been providing support in 16S rRNA diversity analysis to many UCT projects across different fields, including, amongst others oceanography, medical microbiology and immunology.

- UCT Hex cluster
  - Enough resources in terms of storage and processing for average scale 16S runs.

  - UCT researchers can easily obtain access and start processing their data.

  - A combination of automated PBS submission scripts that are driven by configuration files containing tool and run specific settings (https://bitbucket.org/grbot/cbio-pipelines)

- Where things are lacking
  - Only staff and students at UCT have access to the cluster to make use of the pipeline. 3Rd party access can be cumbersome.

  - Updates to the pipeline (e.g. upgrading software packages) are time consuming and can cause run conflicts.

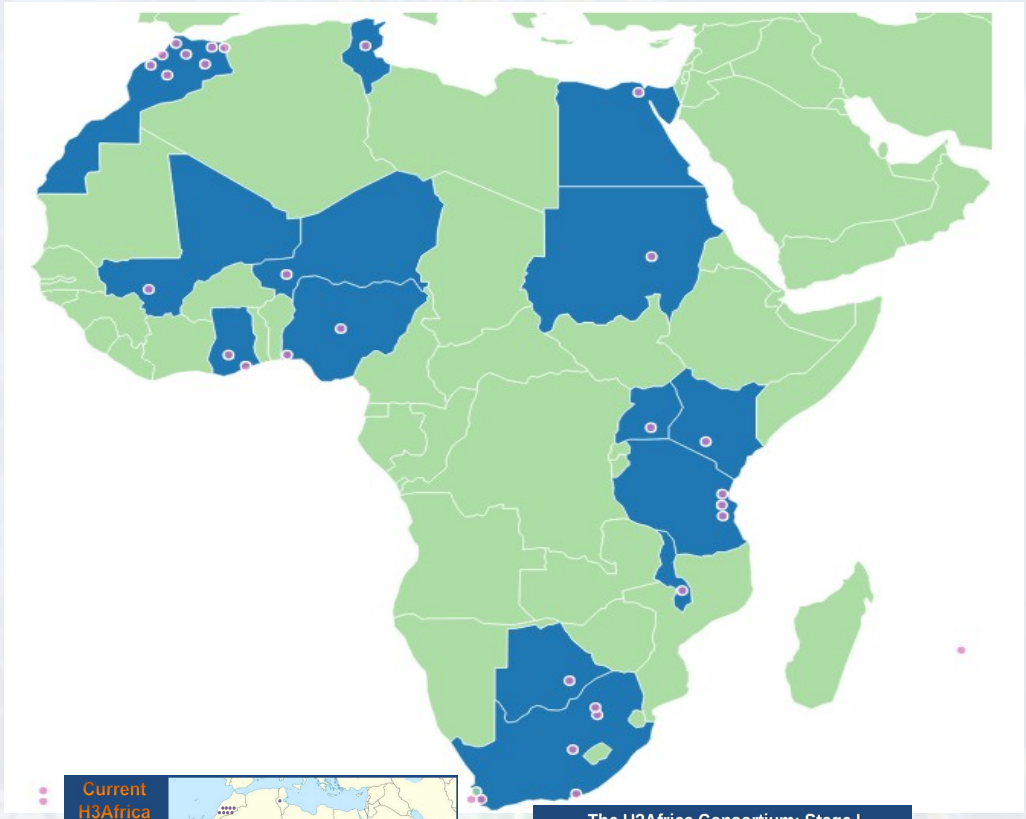
  - The code base was very specific to the UCT Hex cluster and this made portability to other compute clusters or single machines difficult.

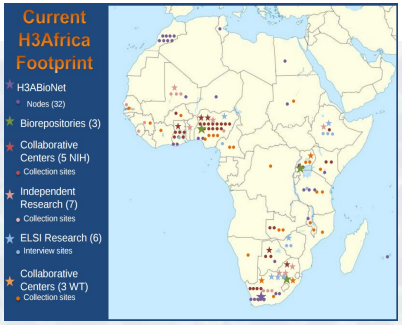
H3ABioNet hackathon

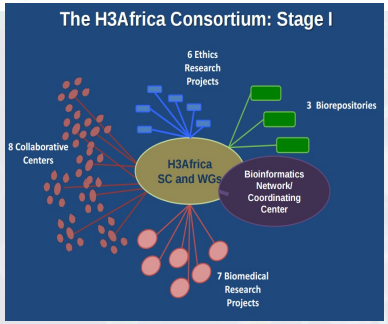
- Pan African bioinformatics network
  - 30 nodes
  - 15 African countries
  - 2 partner institutions in the USA and UK.

- Aim: To support H3Africa projects and researchers while building bionformatics capacity in Africa.

- A week-long hackathon was organized in August 2016 to builld H3Africa usable pipelines in
  - genome-wide association studies
  - human variant calling
  - 16S rRNA analysis and
  - genotype chip imputation

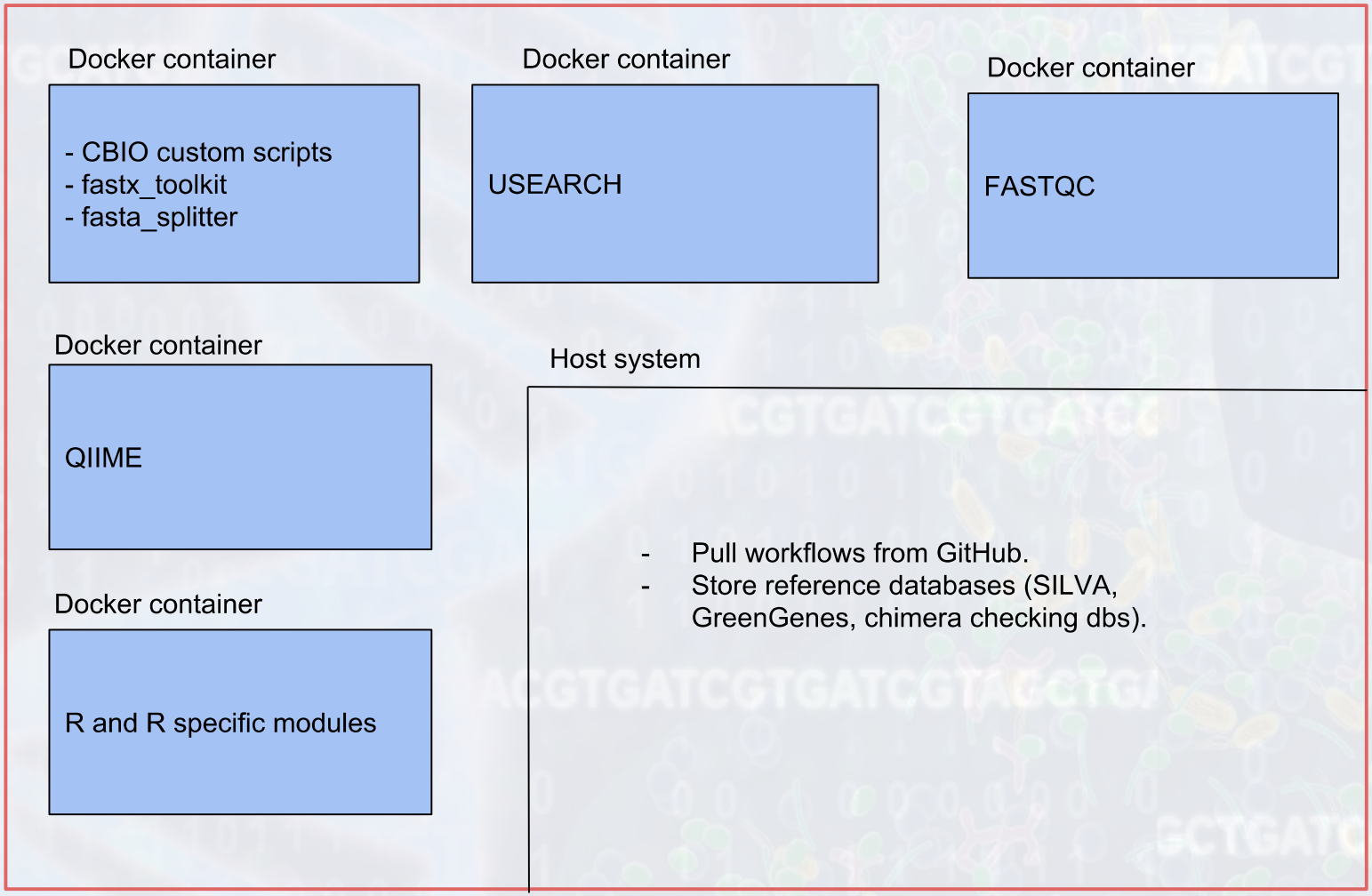






Package setup

Single machine / Cluster environment



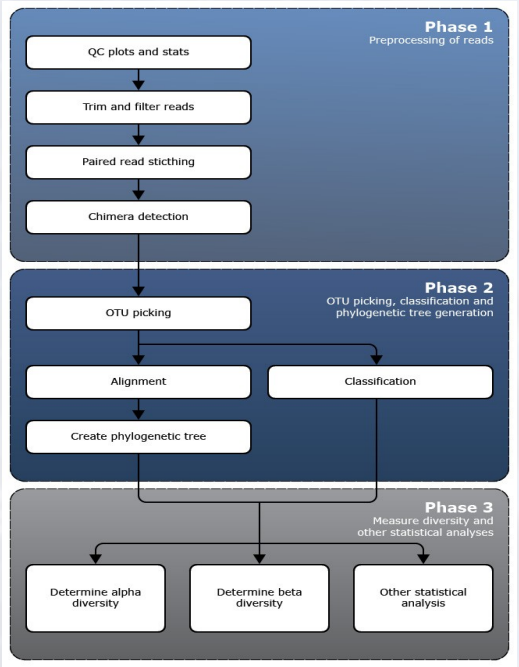
https://quay.io/user/longyee

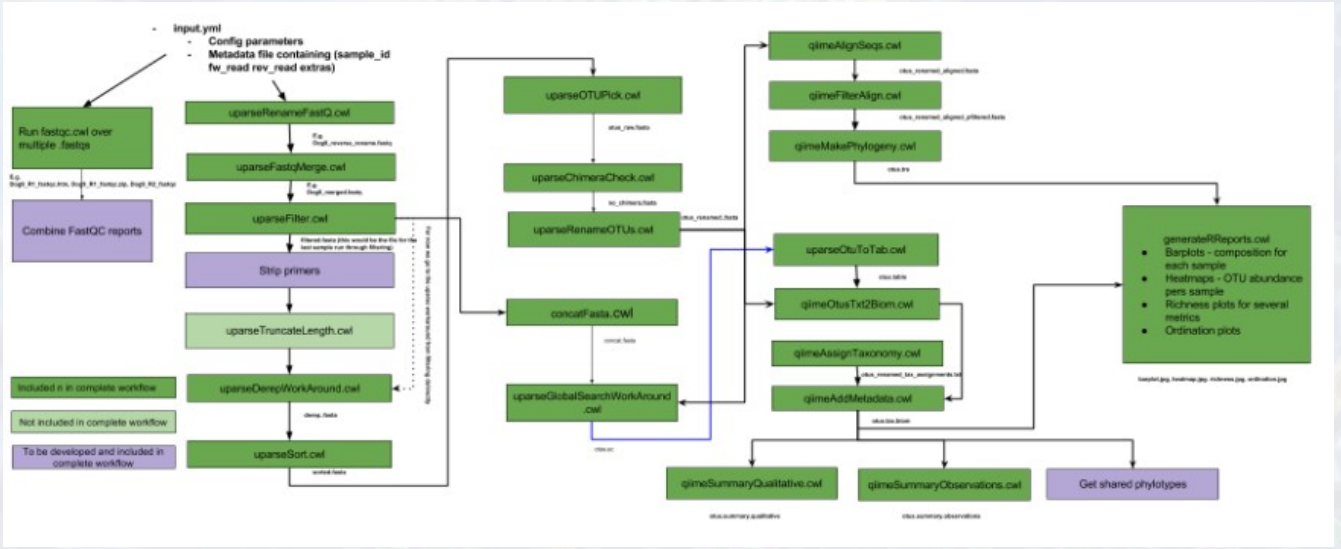
docker

QUAY

Building the pipeline

- CWL (Common Workflow Language)
  - A specification for describing analysis workflows
  - Workflows are portable and scalable across different hardware and software environments (workstation, cluster, cloud)
  - Creating reproducible workflows
  - We used cwltool – reference implementation





https://github.com/h3abionet/h3abionet16S

Run stats

- New packages 80 minutes vs 40 minutes on the original CBIO pipeline (based on 14 MiSeq samples). CWL reference implementation is single threaded we need to look into Toil for a multi-threaded implementation to speed up the process.
- Tested on local computers (Linux and macOS) with and without Docker, EC2 and Azure VMs with and without Docker, a SGE cluster with Docker support (SANBI) and a PBS cluster without Docker support (CHPC).

Advantages of package

- It is now portable:
  - We can run the package on a Docker enabled machine or cluster. E.g. Azure, Hex (700series), AWS or local machine.
  - We can also run it on a non-Docker enabled machine or cluster and just drive the pipeline with CWL. E.g. CHPC.
- Changes to the workflow can be made more easily in CWL.
- Software updates only require an update in a specific Docker container.
- In terms of reproducibility running an analysis on the one compute platform should produce the exact same results as when running the same analysis on another platform.
- Simple run command:

```
cwltool --cachedir /scratch/cache --outdir /scratch/workflow_output completeWorkflow.cwl input.yml
```

Future work

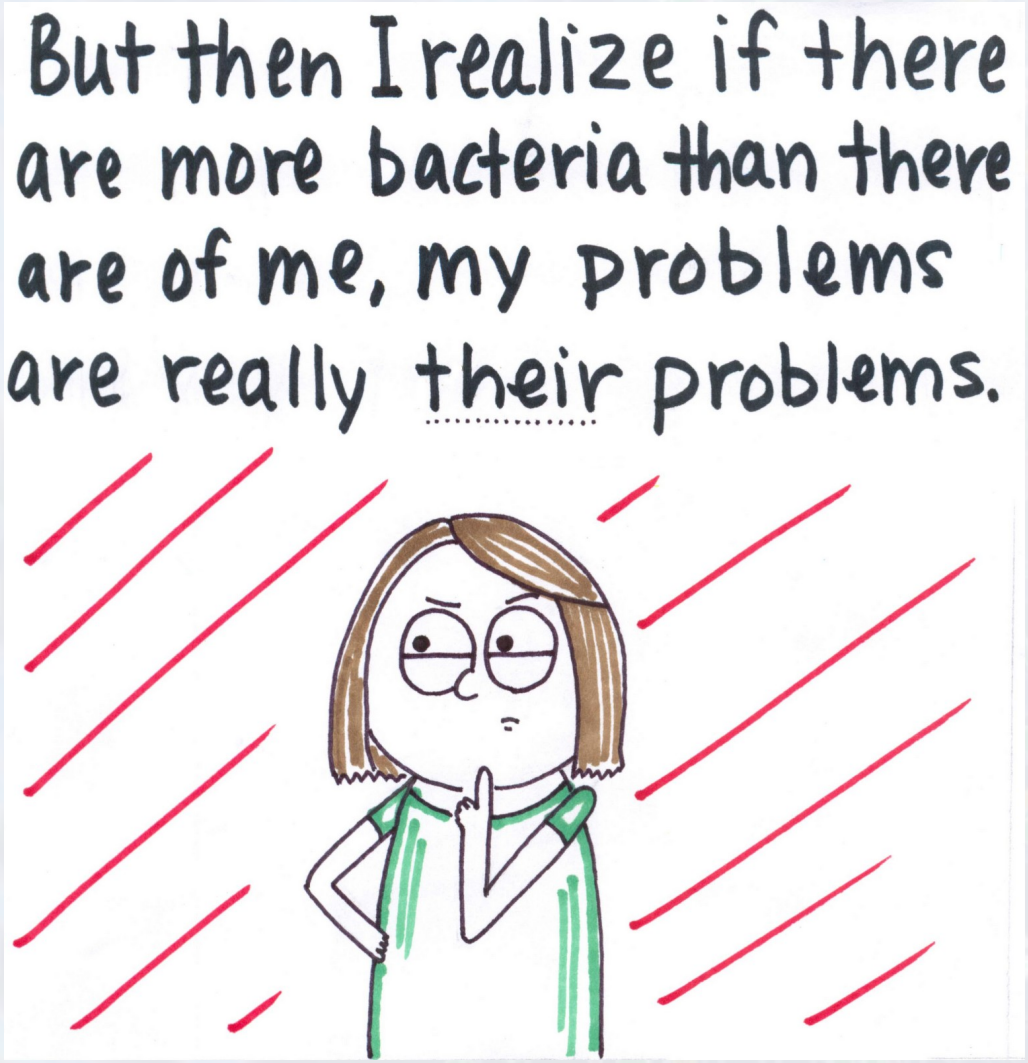
- Fix demultiplexing issue and to start using it in production runs on the 700 series.
- Add additional quality control steps (primer trimming and truncation).
- Consider modifying workflows to be Toil compatible to allow for multi-threaded runs.
- Investigate Singularity containers. This might allow us to run things as containers on CHPC.
- Improve documentation on how to analyse data using the package.

Acknowledgements

- Shakuntala Baichoo
- Michael Crusoe
- Milt Epstein
- Souiai Oussama
- Sumir Panji
- Long Yi
- Nicola Mulder
- Peter van Heusden

- Timothy Carr
- Andrew Lewis

Thank you


But then I realize if there are more bacteria than there are of me, my problems are really their problems.