Building pipelines to support African bioinformatics

The H3ABioNet Pipelines Project

Scott Hazelhurst
University of the Witwatersrand, Johannesburg
on behalf of the
H3A Bioinformatics Network Consortium

H3ABioNet
Pan African Bioinformatics Network for H3Africa

SDIMB
SYDNEY BRENNER INSTITUTE FOR MOLECULAR BIOSCIENCE

AWI-Gen
AfricaWits-INDEPTHPartnership

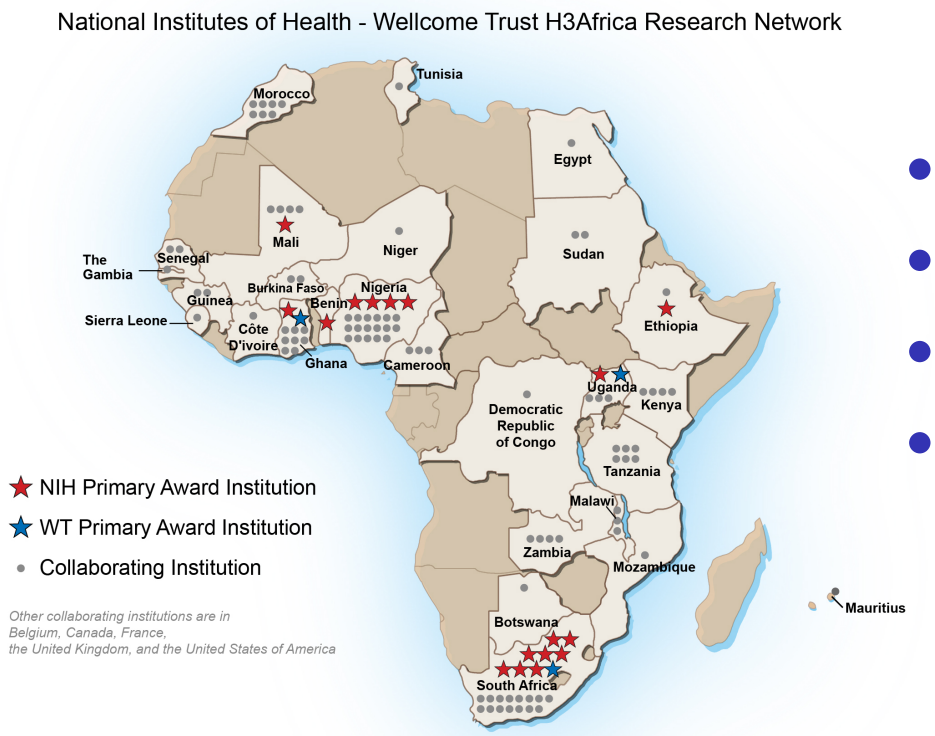
H3A: Human Heredity and Health in Africa

Initiative of ASHG, NIH and Wellcome Trust

- need to build capacity
- Africa's genetic diversity
- rapid epidemiological transition
- health challenges: infectious disease and NCDs
- under-researched

10th Consortium Meeting, Gaborone, Botswana
May 2017

H3ABioNet
Pan African Bioinformatics Network for H3Africa



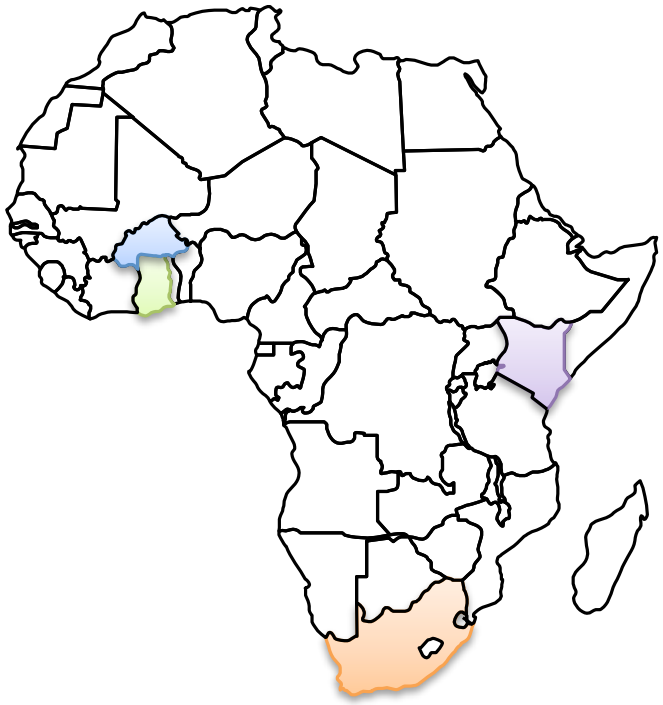
- 8 collaborative centres
- 13 research projects
- Biorepositories
- Pan-African Bioinformatics Network for H3Africa

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Example project: AWI-Gen

Genetic & environmental factors in cardio-metabolic disorders in African populations – hub at Wits

- Soweto, Johannesburg
- Wits Agincourt Research Unit
- Dikgale HDSS, University of Limpopo
- APHRC, Nairobi, Kenya
- Navrongo Health Research Centre, Ghana
- CRUN, Nanoro, Burkina Faso



Data to be collected

12000 participants in all – data collected at site

- Extensive personal histories
- Measured, weighed, scanned, blood, urine samples
- DNA extracted

Data to be collected

12000 participants in all – data collected at site

- Extensive personal histories
- Measured, weighed, scanned, blood, urine samples
- DNA extracted

Genome-wide association test: relating genotype data to the phenotype data

Why a pipeline?

- large data set

Why a pipeline?

- large data set
- data can be sliced in many ways

Why a pipeline?

- large data set
- data can be sliced in many ways
- different phenotypes, population structure, covariates

Why a pipeline?

- large data set
- data can be sliced in many ways
- different phenotypes, population structure, covariates
- quality control crucial

Why a pipeline?

- large data set
- data can be sliced in many ways
- different phenotypes, population structure, covariates
- quality control crucial
- needs to be reproducible

Why a pipeline?

- large data set
- data can be sliced in many ways
- different phenotypes, population structure, covariates
- quality control crucial
- needs to be reproducible
- needs to be portable

Why a pipeline?

- large data set
- data can be sliced in many ways
- different phenotypes, population structure, covariates
- quality control crucial
- needs to be reproducible
- needs to be portable
- under tight deadlines

Other projects – ditto

AWI-Gen just one project

- Similar needs – range of different workflows

Pan-African Bioinformatics Network for H3Africa

H3ABionet created by NIH in 2012 to complement H3Africa research projects:

- bioinformatics key for the projects
- bioinformatics capacity in Africa sparse

Goals:

- Support H3A projects
- Build capacity in bioinformatics

H3ABioNet
Pan African Bioinformatics Network for H3Africa

- Central node at UCT – Overall PI, Nicola Mulder
- $>$ 30 nodes in 15 African countries; US and UK partners

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Overall goal: improve capacity through training and infrastructure development

Working groups

- Education & Training
- Research
- Infrastructure
- User Support
- Accreditation

H3A Bionet Pipelines Project

1. Strategic decision by Bionet to explore

H3A Bionet Pipelines Project

1. Strategic decision by Bionet to explore
   - "cloud computing"

H3A Bionet Pipelines Project

1. Strategic decision by Bionet to explore
   - "cloud computing"
   - build skills in pipelines

H3A Bionet Pipelines Project

1. Strategic decision by Bionet to explore
   - "cloud computing"
   - build skills in pipelines
2. Bionet partners at the University of Illinois/NCSA involvement in NIH"s Big Data to Knowledge (BD2K) initiative & the *Knowledge Engine for Genomics*.

H3A Bionet Pipelines Project

1. Strategic decision by Bionet to explore
   - "cloud computing"
   - build skills in pipelines
2. Bionet partners at the University of Illinois/NCSA involvement in NIH"s Big Data to Knowledge (BD2K) initiative & the *Knowledge Engine for Genomics*.
3. Some prior work (Lerato Magosi and Rob Clucas)

Overview of project

- Decision to start at H3A Consortium meeting in Dakar in May 2016.
- Involves about 30 people from over 10 institutions, led by Sumir Panji.
- Identified key people, planning starts
- Ran 5-day "Cloud Hackathon" at the University of Pretoria in August 2016
- Sept– pipelines being completed
- January – pipelines published, paper written

Goals

1. Develop production-quality pipelines for key workflows
    - Direct support for stated needs
    - Position Bionet strategically
2. Develop human capacity for building pipelines
3. Explore different technologies

Constraints

1. Must be highly portable, scalable
   - Ideally laptop to CHPC
   - Support cloud environments
   - Container support essential
2. Must have skills within the network : both workflow and technology
3. Limited resources : start with focus and explore other technologies and workflows later

Identified four workflows, two workflow technologies

Identified four workflows, two workflow technologies

CWL		
Nextflow		

Identified four workflows, two workflow technologies

CWL	NGS data	16S metagenomics
Nextflow	GWAS	Imputation

Nextflow

Developed by the Comparative Bioinformatics group at the Barcelona Center for Genomic Regulation (CRG)

- General purpose workflow system
- DSL based on Groovy
- Portable
- Scalable
- Very easy to install
- Container support
- Supports a range of scheduling systems, cloud

Common Workflow Language

Language specification rather than a tool – several tools support it.

- Community-driven, Multi-vendor
- Supports Docker, parallelism
- Language based on YAML
- Extensible

Came out of the bioinformatics community (BOSC) but general purposes

- has buy in from major players

Bionet has good collaborators with CWL experience

```
class commandLineTool
inputs:
  fastqFile:
    type: File
    inputBinding:
      position: 1
baseCommand: [ fastqc, "--outdir", "res" , "--extract" ]
outputs:
  zippedFile:
    type: File
    outputBinding:
      glob: "*.zip"
  report:
    type: Directory
    outputBinding:
      glob: "res"
```

```
process fastQC {
  input:
     file fastqFile
  output:
     set file("*.zip"), file(res) into output_ch
  script:
    "fastqc $fastqFile --outdir res --extract"
}
```

Pros & Cons

Differences in syntax, semantics, capabilities, performance in tools

- Community & standards based language
- Small group driving architecture and system

For whom?

- Biologists with bioinformatics expertise
- Bioinformaticists
- Specialist pipeline builders

Pipeline building process

Building capacity an important goal

- Team distributed across 4 continents, many countries, timezones
- Variable quality internet

Make heavy use of electronic communication

- mconf
- Slack
- Trello
- GitHub

But personal interaction key

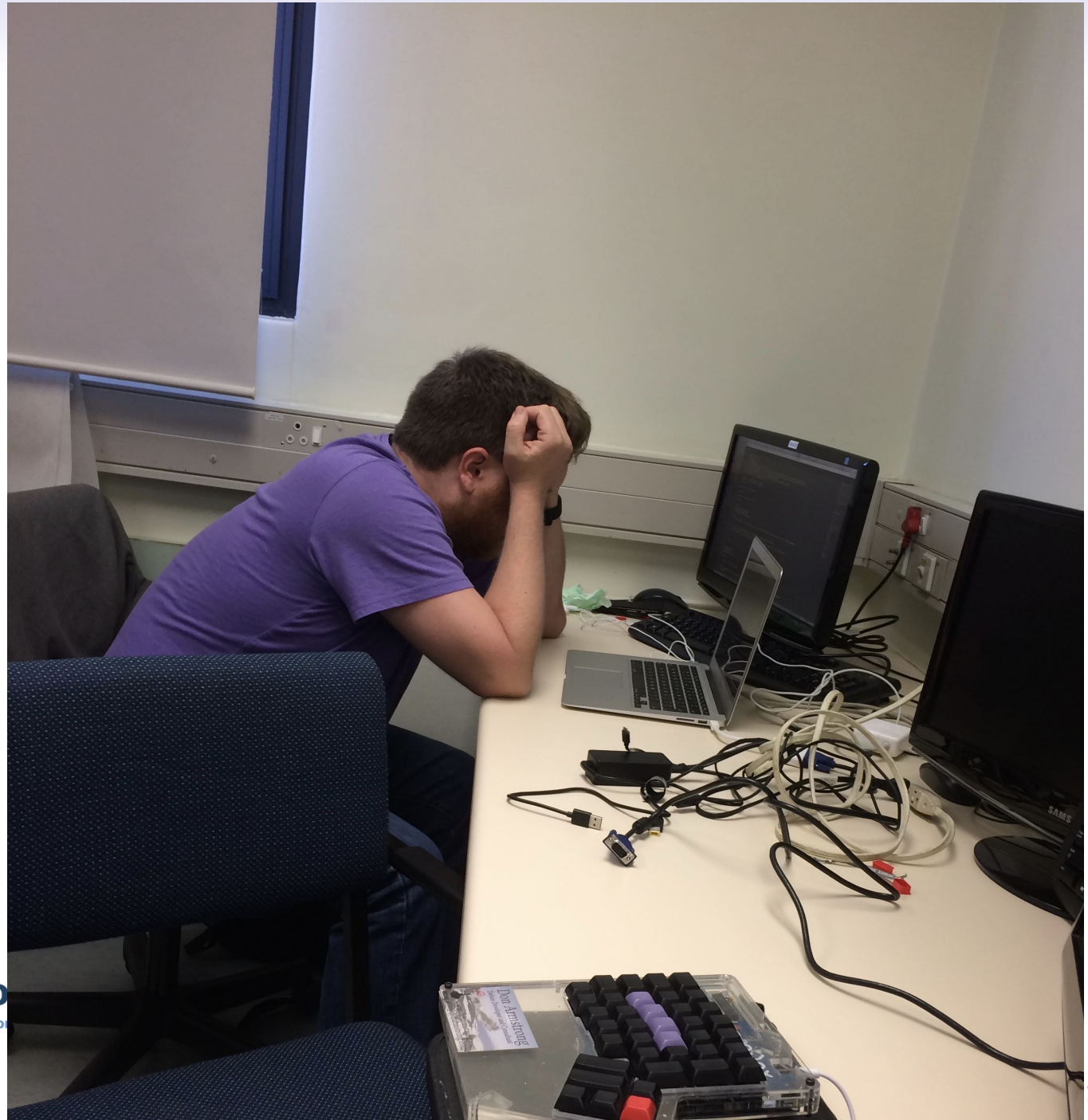
Bionet Cloud Hacakthon

Planning crucial for production meeting

- Identified key participants but also advertised widely – selection process
- Pre-workshop preparation
- 5-days concentrated work – some participation remotely
- Some training, a few talks but mainly work

2014 HPC-ISC
STUDENT CLUSTER COMPETITION

SAMSUNG



H3ABio
Pan African Bioinfo

Status of project

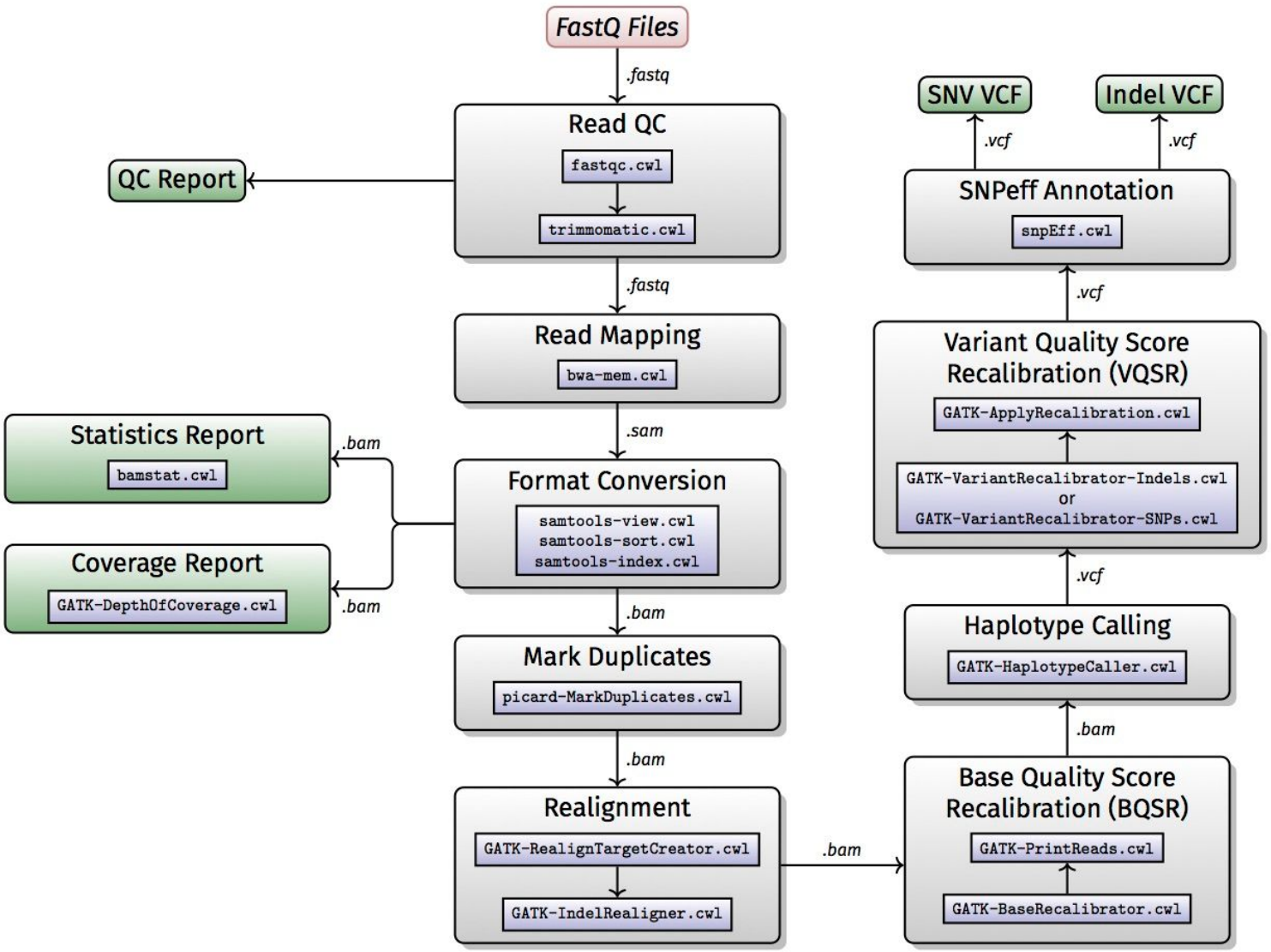
- Developed skills in pipeline creation
- Almost done!
  - `github.com/h3abionet/h3agtk`
  - `github.com/h3abionet/h3abionet16s`
  - `github.com/h3abionet/h3agwas`
  - `github.com/h3abionet/chipimputation`

Whole genome/exome analysis

Based on GATK's *Best Practices* SOP – Done in CWL

- sequence adapter trimming
- QC of fastsq files, BAM files
- Indel and SNV annotation

Dockerised (except for GATAnalysisTK.jar)



Pan African Bioinformatics Network for H3Africa

16S rRNA Diversity

Done in CWL – want to re-implement in Nextflow
Use key tools

- FastQC
- USEARCH and UPARSE for clustering and more QC
- QIIME for demultiplexing, taxonomic reconstruction, diversity analysis
- Statistical analysis (scripts using R)

Helper scripts to put together

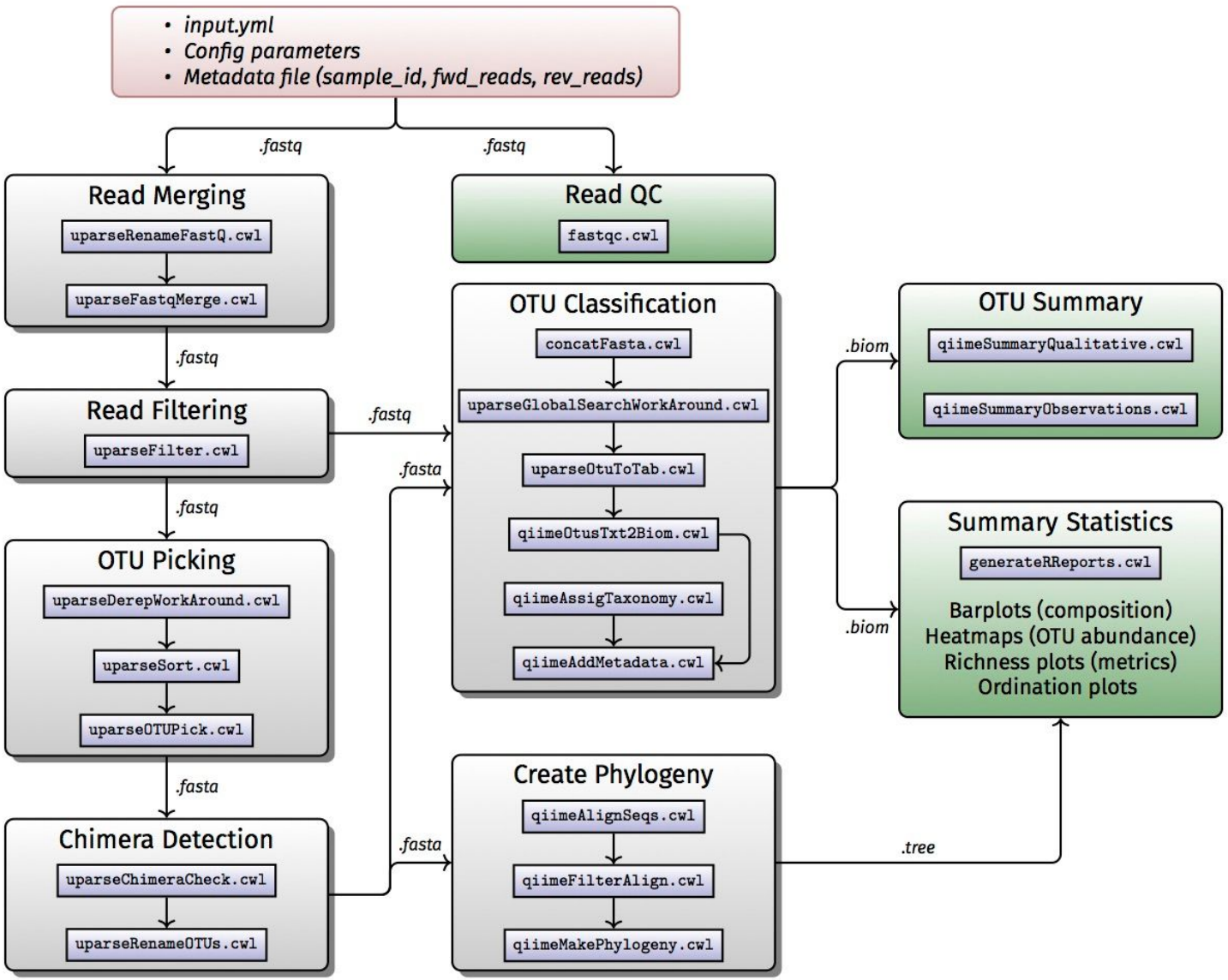


Figure 2 - Workflow B: 16S rRNA Diversity Analysis

GWAS Pipeline

Three separate pipelines – implemented in Nextflow

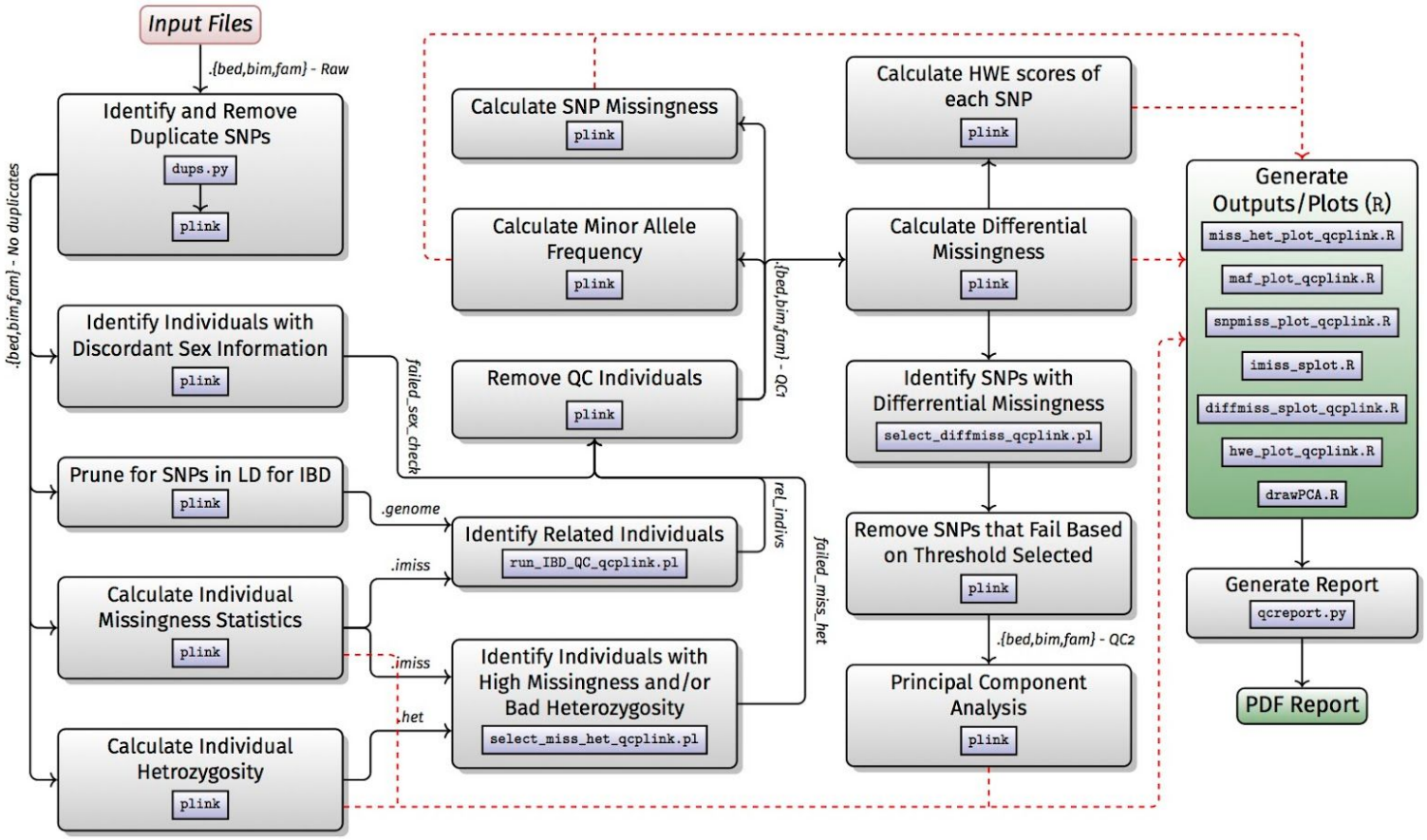
- Convert from Illumina calls to PLINK format
- Quality control: PLINK to PLINK
- GWAS

- SNP, individual missingness
- heterozygosity (individual, SNP)
- relatedness
- case-control differences
- basic PCA

Workflows very portable

Used

- local computer (with or without Docker)
- local cluster (with or without Docker)
- Amazon EC2 AMI
- Docker Swarm
- OpenNebula (NCSA, ARC)



H3ABioNet
Pan African Bioinformatics Network for H3Africa

Imputation

Developed in Nextflow

- Assume genotype data as input
- Converted to Impute format
- Chromosomes split
- Check for correct strandedness
- Pre-phased using Shapeit
- Use IMPUTE2 to do imputation
- Recombine dataa

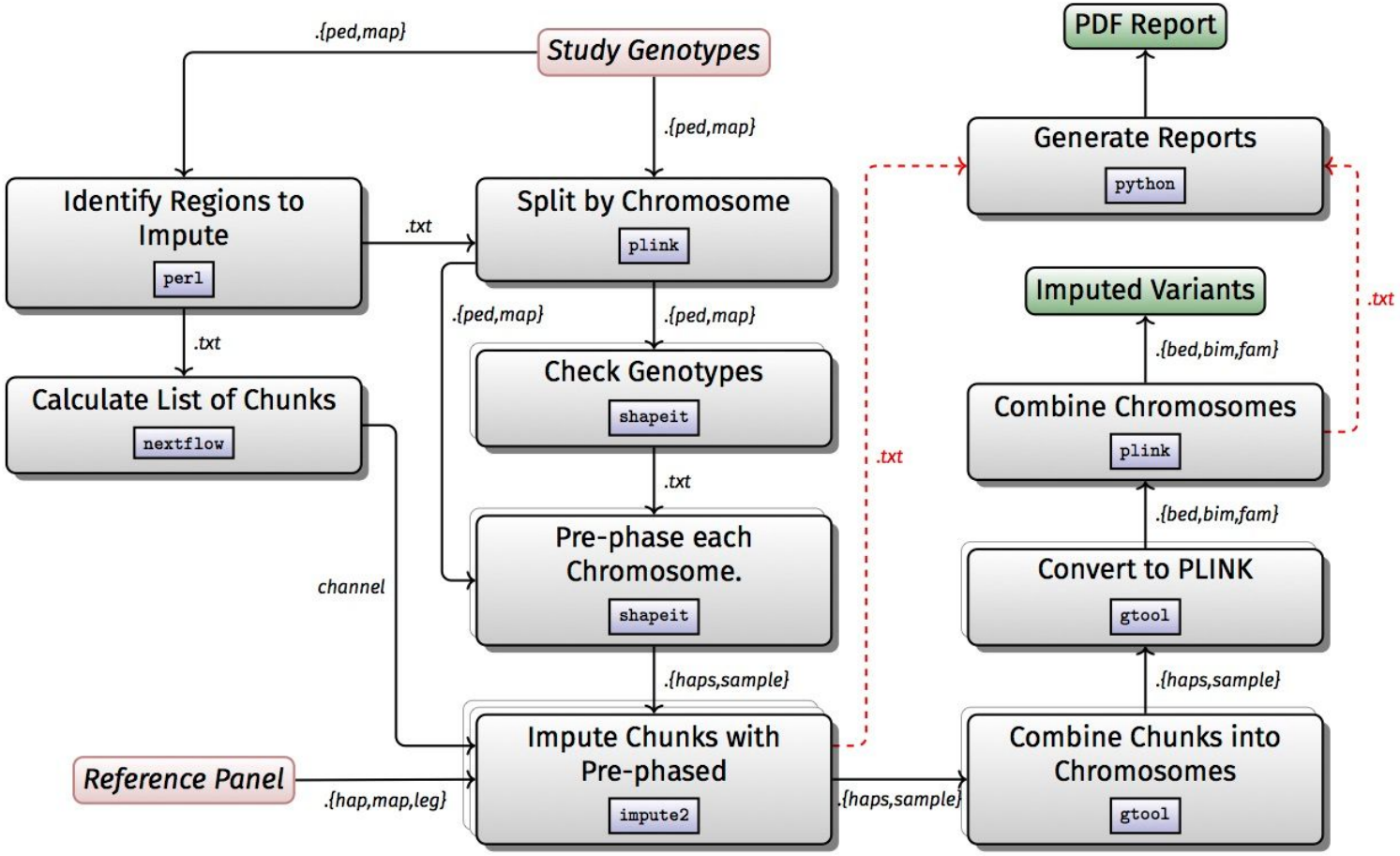


Figure 4 - Workflow D: SNPs Imputation

Experiences

Both Nextflow and CWL worked well

- Both have responsive communities
- ?? Nextflow has an easier learning curve
- ?? When we started Nextlow was maturer
- ?? may be momentum behind CWL.
- ?? CWL may have an advantage in packaging workflows

Future work

1. Extend, maintain existing pipelines
2. Rigorous comparison
3. Develop new pipelines – major set of activities for the next five years
4. Look at other systems, e.g., JMS

Funded by NIH NHGRI grants U41HG006941, HG006938. Work at different institutions and individuals.

Shakuntala Baichoo
Yassime Souilmi
Sumir Panji
Ayton Meintjes
Brian O'Connor
Don Armstrong
Eugene de Beste
Gerrit Botha

Hocine Bendou
Lerato Magosi
Luda Mainzer
Mustafa Alghali
Michael Crusoe
Nicola Mulder
Oussema Souiai
Peter van Heusden

Phelelani Mpangase
Rob Clucas
Scott Hazelhurst
Shaun Aron
Shakuntala Baichoo