H3ABioNet

Pan African Bioinformatics Network for H3Africa

The H3ABioNet Reproducible WorkflowsProject

www.h3abionet.org

International Data Week

SciDataCon 2018

Botswana

H3ABioNet 2.0 Specific Aims

① To implement a Pan African informatics infrastructure

② To develop an H3Africa data coordinating center

③ To provide high quality informatics support to H3Africa

④ To enable and enhance innovative translational research

⑤ To address outreach, development and sustainability

H3ABioNet 2.0 Specific Aims

① To implement a Pan African informatics infrastructure

② To develop an H3Africa data coordinating center

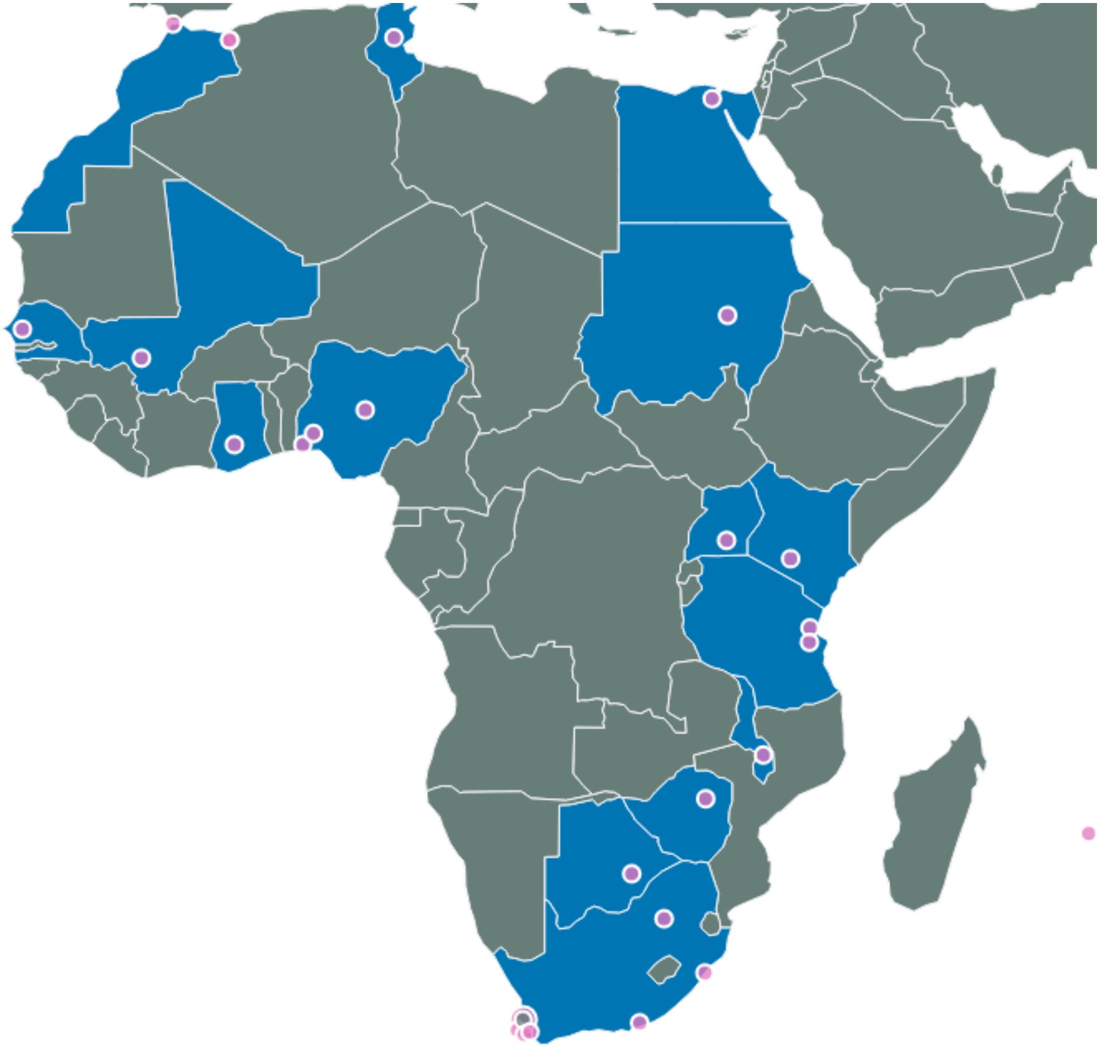
③ To provide high quality informatics support to H3Africa

④ To enable and enhance innovative translational research

⑤ To address outreach, development and sustainability

H3Africa Bioinformatics Network (H3ABioNet)

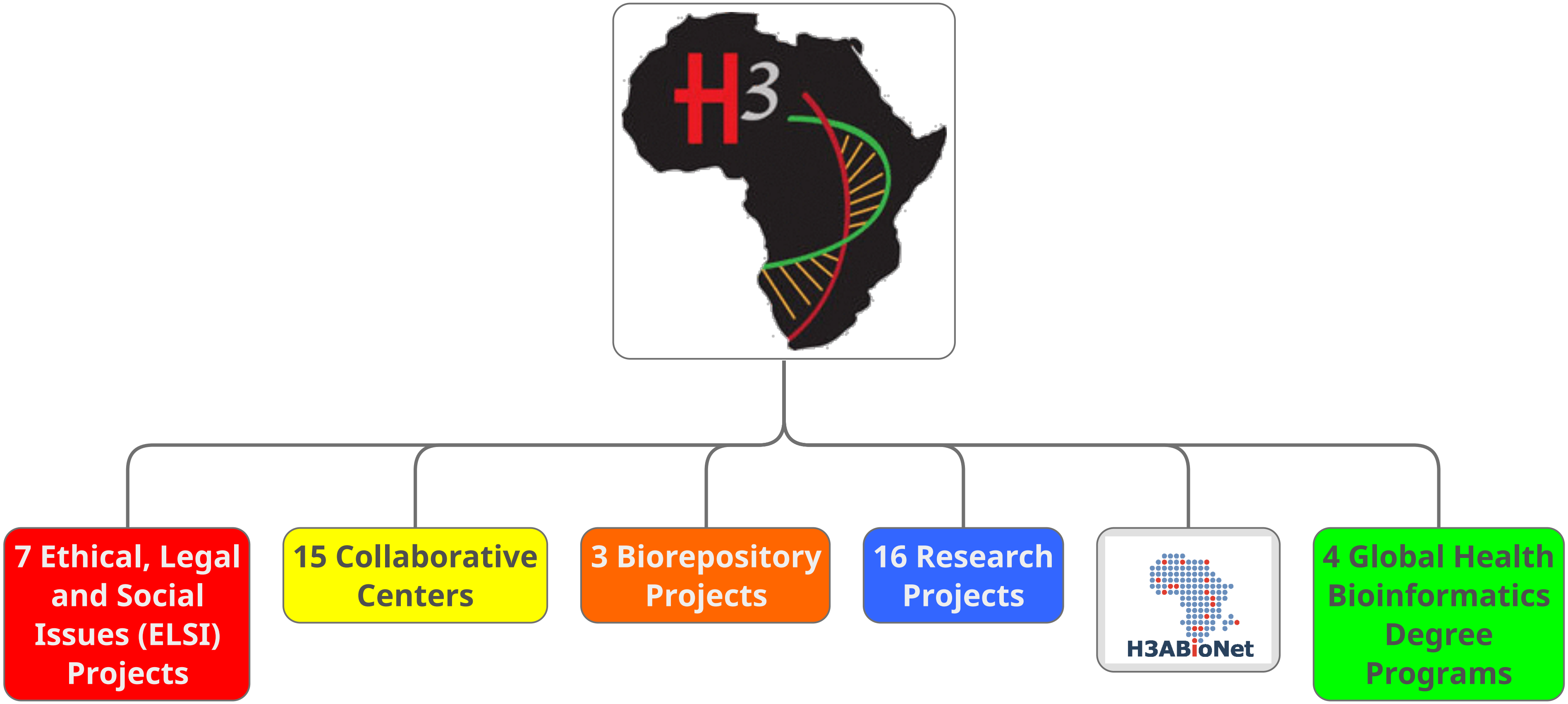
- Pan African Bioinformatics Network to develop bioinformatics capacity in Africa and support the H3Africa research projects



H3ABioNet

Pan African Bioinformatics Network for H3Africa

H3Africa Consortium



7 Ethical, Legal and Social Issues (ELSI) Projects

15 Collaborative Centers

3 Biorepository Projects

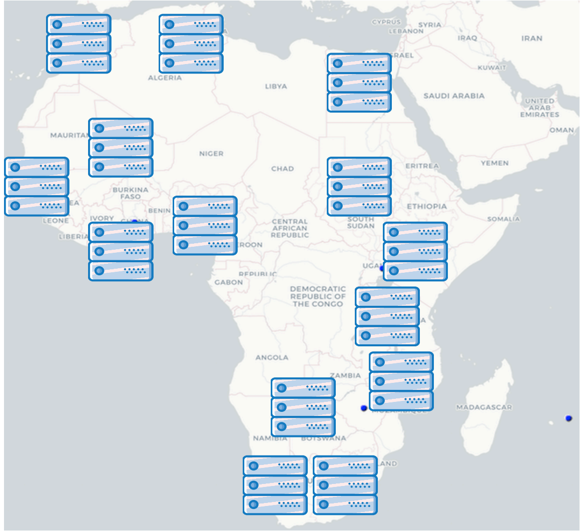
16 Research Projects

H3ABioNet

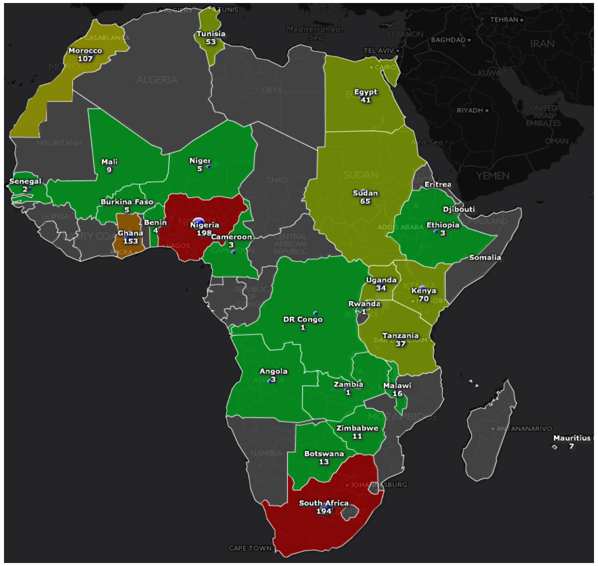
4 Global Health Bioinformatics Degree Programs

http://h3africa.org/consortium/projects

H3ABioNet Workflows project

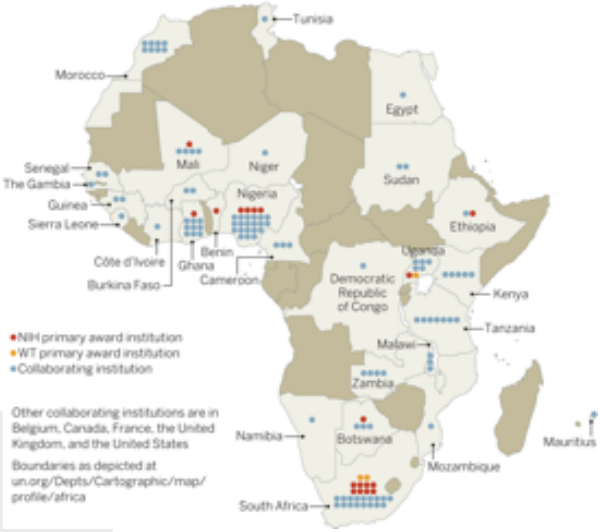


Compute infrastructure provided



Training provided (sys admin, NGS)

National Institutes of Health
Wellcome Trust H3Africa Research Network


Tunisia
Morocco
Egypt
Senegal
The Gambia
Guinea
Sierra Leone
Mali
Niger
Nigeria
Sudan
Ethiopia
Côte d'Ivoire
Burkina Faso
Ghana
Benin
Cameroon
Uganda
Democratic Republic of Congo
Kenya
Tanzania
Malawi
Zambia
Namibia
Botswana
Mozambique
Mauritius
South Africa
● NIH primary award institution
● WT primary award institution
● Collaborating institution
Other collaborating institutions are in Belgium, Canada, France, the United Kingdom, and the United States
Boundaries as depicted at un.org/Depts/Cartographic/map/profile/africa

Expertise by Nodes to run analysis to support H3Africa projects in region?

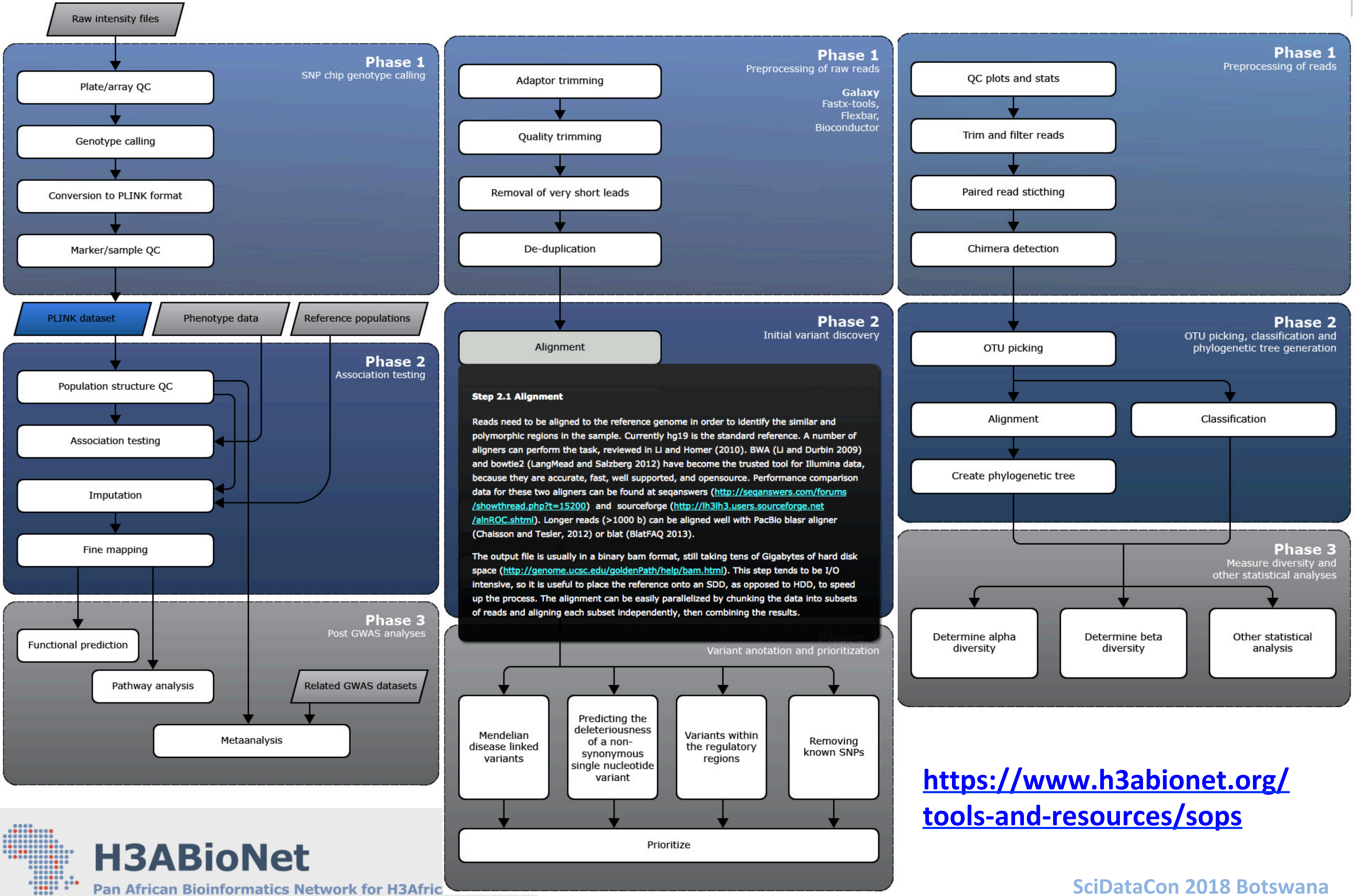
H3ABioNet
Pan African Bioinformatics Network for H3Africa

H3ABioNet Node Accreditation Task force

Create a process for H3ABioNet Nodes to be assessed on the ability to undertake bioinformatics analyses of expected H3Africa genomics data

- Constituted by Victor Jongeneel to:
  - Create Standard Operating Procedures for various “omics” analysis
  - Create practice datasets for Nodes to work with
  - Create assessment datasets
  - Setup process for administrating a Node assessment
  - Constitute external board of reviewers to assess final report submitted

H3ABioNet SOPs and Accreditation



Step 2.1 Alignment

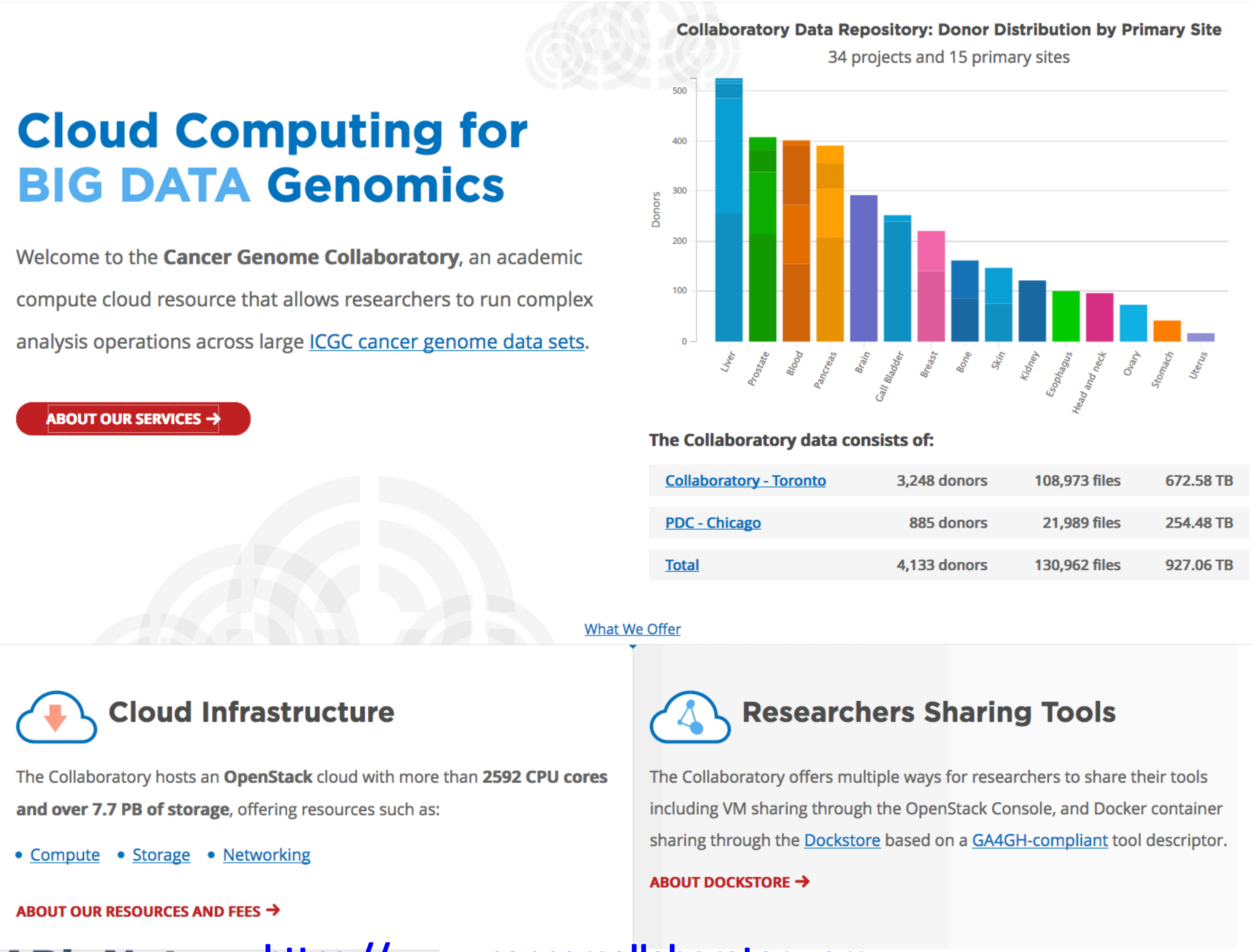
Reads need to be aligned to the reference genome in order to identify the similar and polymorphic regions in the sample. Currently hg19 is the standard reference. A number of aligners can perform the task, reviewed in Li and Homer (2010). BWA (Li and Durbin 2009) and bowtie2 (LangMead and Salzberg 2012) have become the trusted tool for Illumina data, because they are accurate, fast, well supported, and opensource. Performance comparison data for these two aligners can be found at seqanswers (http://seqanswers.com/forums/showthread.php?t=15200) and sourceforge (http://lh3lh3.users.sourceforge.net/alnROC.shtml). Longer reads (>1000 b) can be aligned well with PacBio blasr aligner (Chaisson and Tesler, 2012) or blat (BlatFAQ 2013).

The output file is usually in a binary bam format, still taking tens of Gigabytes of hard disk space (http://genome.ucsc.edu/goldenPath/help/bam.html). This step tends to be I/O intensive, so it is useful to place the reference onto an SDD, as opposed to HDD, to speed up the process. The alignment can be easily parallelized by chunking the data into subsets of reads and aligning each subset independently, then combining the results.

https://www.h3abionet.org/tools-and-resources/sops

H3ABioNet
Pan African Bioinformatics Network for H3Afric

H3ABioNet Workflows project – Cloud



Cloud Computing for BIG DATA Genomics

Welcome to the **Cancer Genome Collaboratory**, an academic compute cloud resource that allows researchers to run complex analysis operations across large ICGC cancer genome data sets.

ABOUT OUR SERVICES →

The Collaboratory data consists of:

Collaboratory - Toronto	3,248 donors	108,973 files	672.58 TB
PDC - Chicago	885 donors	21,989 files	254.48 TB
Total	4,133 donors	130,962 files	927.06 TB

What We Offer

Cloud Infrastructure

The Collaboratory hosts an **OpenStack** cloud with more than **2592 CPU cores and over 7.7 PB of storage**, offering resources such as:

- Compute
- Storage
- Networking

ABOUT OUR RESOURCES AND FEES →

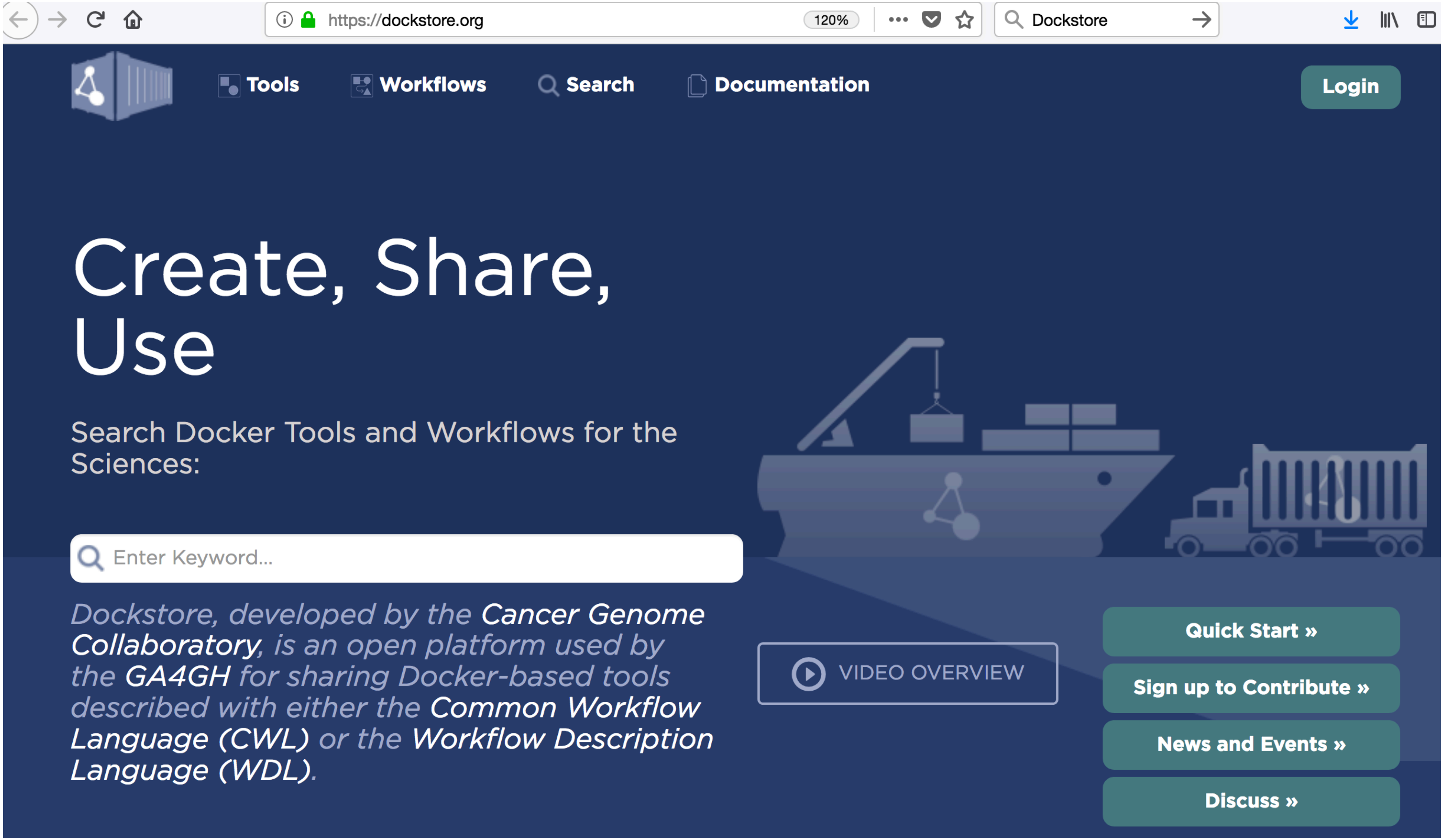
Researchers Sharing Tools

The Collaboratory offers multiple ways for researchers to share their tools including VM sharing through the OpenStack Console, and Docker container sharing through the Dockstore based on a GA4GH-compliant tool descriptor.

ABOUT DOCKSTORE →

https://www.cancercollaboratory.org

H3ABioNet Workflows project - Containers



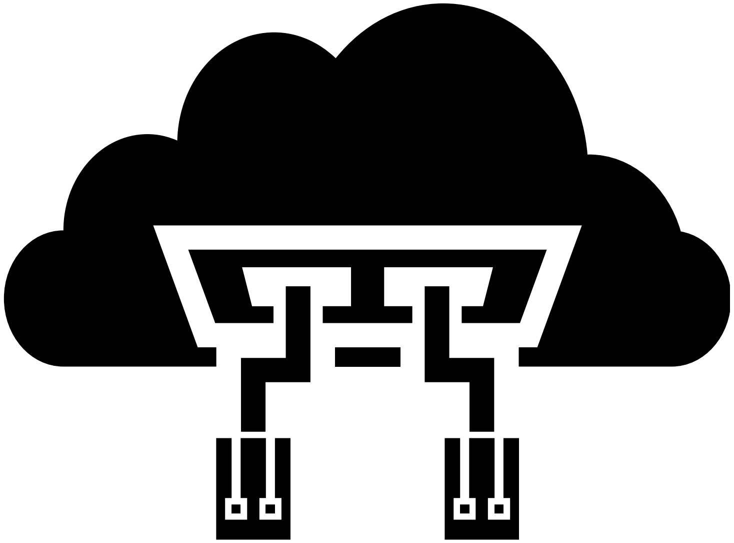
Create, Share, Use

Search Docker Tools and Workflows for the Sciences:

Dockstore, developed by the Cancer Genome Collaboratory, is an open platform used by the GA4GH for sharing Docker-based tools described with either the Common Workflow Language (CWL) or the Workflow Description Language (WDL).

https://dockstore.org

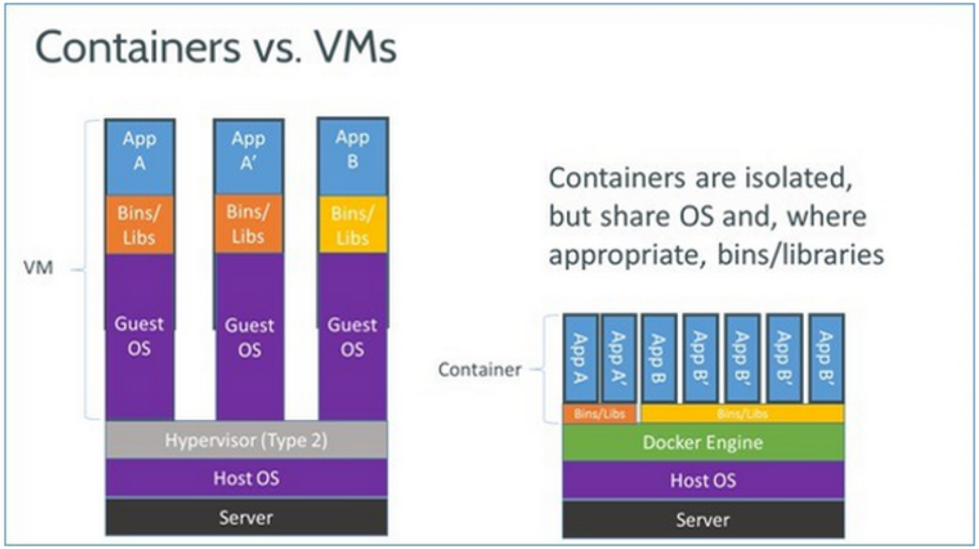
H3ABioNet Workflows project

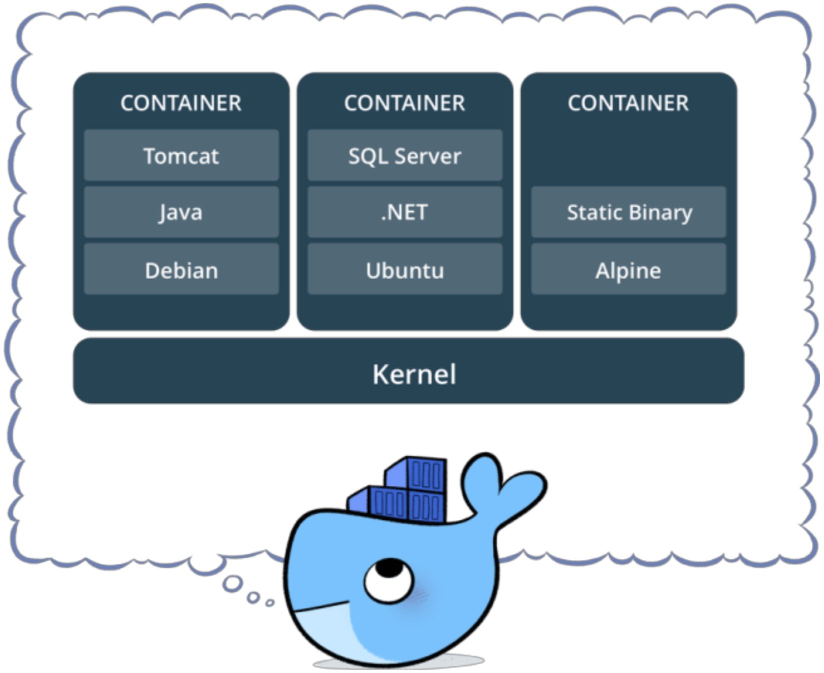


- Cloud computing task force constituted to look at potential applications in March 2016
- Mainly learning with some groups having experience
- Brainstormed different ideas – convert pipelines to workflows and containerize for easy deployment

H3ABioNet Workflows project - Containers

- Containers – means to package software, tools, dependencies





- Docker – containerization technology widely adopted by groups

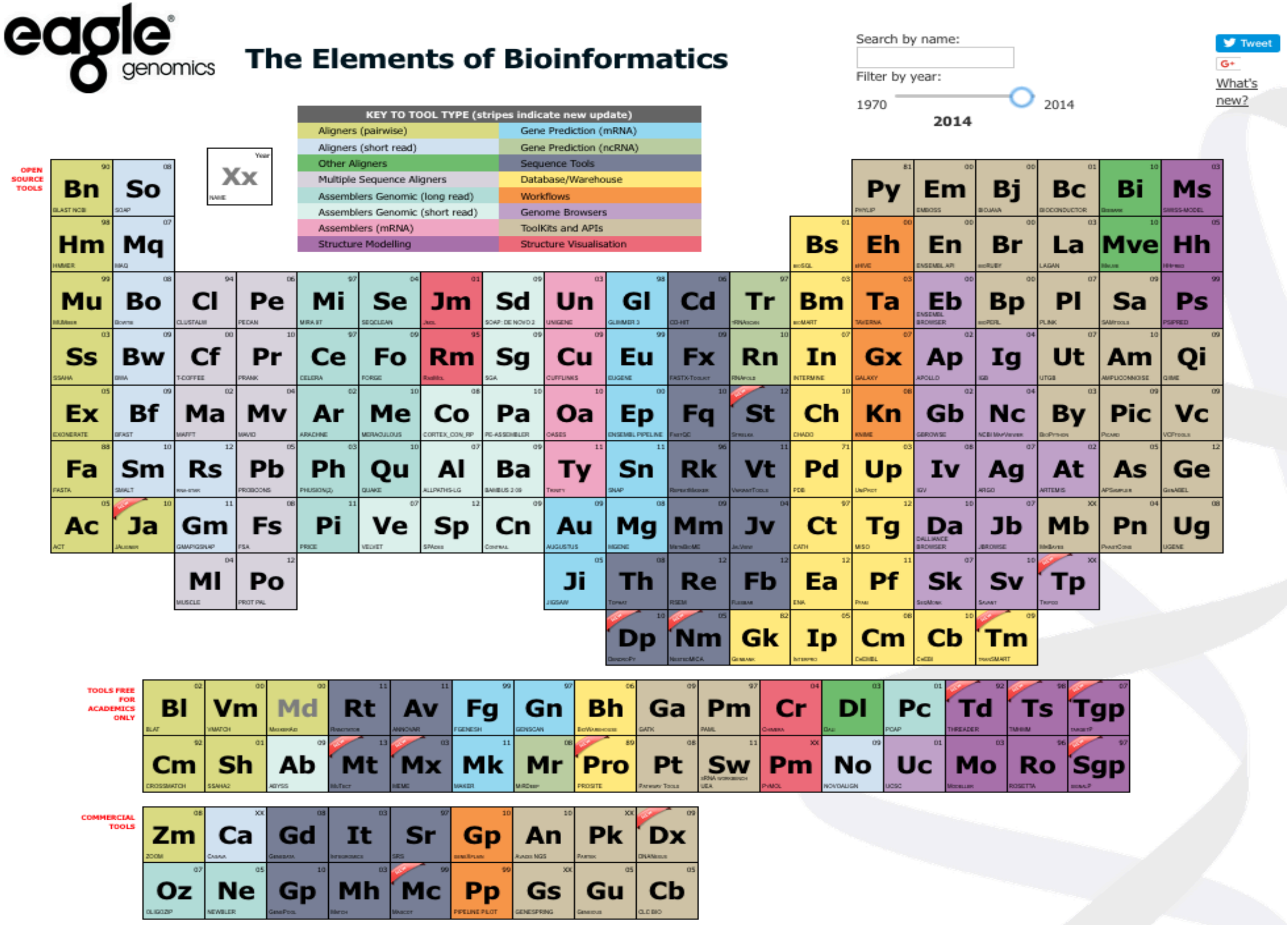
- Portable from various Linux OS to Linux OS

Image credits:

https://www.zdnet.com/article/what-is-docker-and-why-is-it-so-darn-popular

https://blog.docker.com/2017/08/docker-101-introduction-docker-webinar-recap

H3ABioNet Workflows project - Software



eagle genomics

The Elements of Bioinformatics

Search by name:

Filter by year:

1970 2014

2014

Tweet

G+

What's new?

KEY TO TOOL TYPE (stripes indicate new update)	
Aligners (pairwise)	Gene Prediction (mRNA)
Aligners (short read)	Gene Prediction (ncRNA)
Other Aligners	Sequence Tools
Multiple Sequence Aligners	Database/Warehouse
Assemblers Genomic (long read)	Workflows
Assemblers Genomic (short read)	Genome Browsers
Assemblers (mRNA)	ToolKits and APIs
Structure Modelling	Structure Visualisation

Year Xx NAME

OPEN SOURCE TOOLS

Bn 90 BLAST NCBI	So 08 SOAP												Py 81 PHYLIP	Em 00 EMBOSS	Bj 00 BIOJAVA	Bc 01 BIOCONDUCTOR	Bi 10 Bismark	Ms 03 SWISS-MODEL
Hm 98 HMMER	Mq 07 MAQ											Bs 01 bioSQL	Eh 00 eHIVE	En 00 ENSEMBL API	Br 00 bioRUBY	La 03 LAGAN	Mve 10 Mauve	Hh 05 HHpred
Mu 99 MUMmer	Bo 08 Bowtie	Cl 94 CLUSTALW	Pe 06 PECAN	Mi 97 MIRA BT	Se 04 SEQCLEAN	Jm 01 Jmol	Sd 09 SOAP: DE NOVO 2	Un 03 UNIGENE	Gl 98 GLIMMER 3	Cd 06 CD-HIT	Tr 97 tRNAscan	Bm 03 bioMART	Ta 03 TAVERNA	Eb 00 ENSEMBL BROWSER	Bp 00 bioPERL	Pl 07 PLINK	Sa 09 SAMtools	Ps 99 PSIPRED
Ss 03 SSAHA	Bw 09 BWA	Cf 00 T-COFFEE	Pr 10 PRANK	Ce 97 CELERA	Fo 09 FORGE	Rm 95 RasMol	Sg 09 SGA	Cu 09 CUFFLINKS	Eu 99 EUGENE	Fx 09 FASTX-Toolkit	Rn 10 RNAfold	In 07 INTERMINE	Gx 07 GALAXY	Ap 02 APOLLO	Ig 04 IGB	Ut 07 UTGB	Am 10 AMPLICONNOISE	Qi 09 QIIME
Ex 05 EXONERATE	Bf 09 BFAST	Ma 02 MAFFT	Mv 04 MAVID	Ar 02 ARACHNE	Me 10 MERACULOUS	Co 08 CORTEX_CON_RP	Pa 10 PE-ASSEMBLER	Oa 10 OASES	Ep 00 ENSEMBL PIPELINE	Fq 10 FastQC	St 12 Strelka	Ch 10 CHADO	Kn 08 KNIME	Gb 02 GBROWSE	Nc 04 NCBI MapViewer	By 03 BioPython	Pic 09 Picard	Vc 09 VCFtools
Fa 88 FASTA	Sm 10 SMALT	Rs 12 rna-star	Pb 05 PROBCONS	Ph 03 PHUSION(2)	Qu 10 QUAKE	Al 07 ALLPATHS-LG	Ba 09 BAMBUS 2.09	Ty 11 Trinity	Sn 11 SNAP	Rk 96 RepeatMasker	Vt 11 VariantTools	Pd 71 PDB	Up 03 UniProt	Iv 08 IGV	Ag 07 ARGO	At 02 ARTEMIS	As 05 APSampler	Ge 12 GenABEL
Ac 05 ACT	Ja 10 JAligner	Gm 11 GMAP/GSNAP	Fs 08 FSA	Pi 11 PRICE	Ve 07 VELVET	Sp 12 SPAdes	Cn 09 Contrail	Au 09 AUGUSTUS	Mg 08 MGENE	Mm 09 MetaBioME	Jv 04 JalView	Ct 97 CATH	Tg 12 MISO	Da 10 DALLIANCE BROWSER	Jb 07 JBROWSE	Mb XX MrBayes	Pn 04 PhastCons	Ug 08 UGENE
		Ml 04 MUSCLE	Po 12 PROT PAL					Ji 05 JIGSAW	Th 08 TopHat	Re 12 RSEM	Fb 12 Flexbar	Ea 12 ENA	Pf 11 Pfam	Sk 07 SeqMonk	Sv 10 Savant	Tp XX Tripod		
									Dp 10 DendroPy	Nm 05 NestedMICA	Gk 82 GenBank	Ip 05 InterPro	Cm 08 ChEMBL	Cb 10 ChEBI	Tm 09 tranSMART			

TOOLS FREE FOR ACADEMICS ONLY

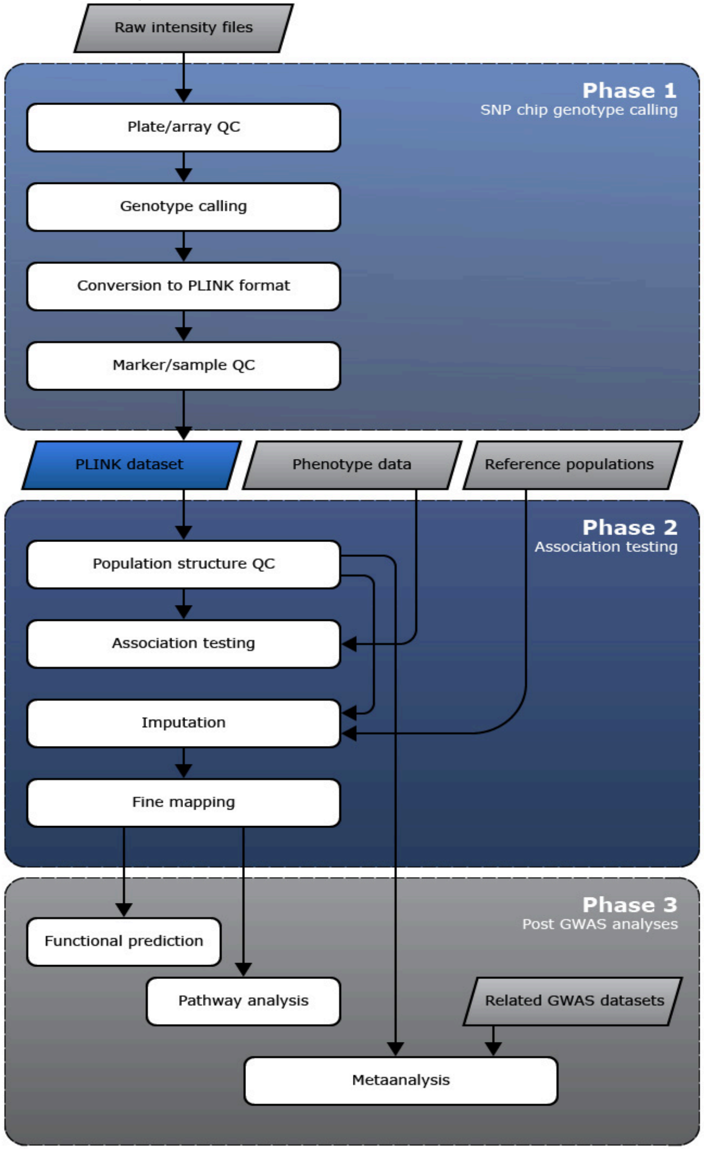
Bl 02 BLAT	Vm 00 VMATCH	Md 00 MosaikAligner	Rt 11 [illegible]	Av 11 ANNOVAR	Fg 99 FGENESH	Gn 97 GENSCAN	Bh 06 BioWarehouse	Ga 09 GATK	Pm 97 PAML	Cr 04 Chimera	Dl 03 Dali	Pc 01 PCAP	Td 92 THREADER	Ts 98 TMHMM	Tgp 07 targetP
Cm 92 CROSSMATCH	Sh 01 SSAHA2	Ab 09 ABYSS	Mt 13 MuTect	Mx 03 MEME	Mk 11 MAKER	Mr 08 MiRDeep	Pro 89 PROSITE	Pt 08 Pathway Tools	Sw 11 sRNA workbench UEA	Pm XX PyMOL	No 09 NOVOALIGN	Uc 01 UCSC	Mo 03 Modeller	Ro 96 ROSETTA	Sgp 97 signalP

COMMERCIAL TOOLS

Zm 08 ZOOM	Ca XX CASAVA	Gd 08 Genedata	It 03 Integromics	Sr 97 SRS	Gp 10 geneXplain	An 10 Avadis NGS	Pk XX Partek	Dx 09 DNAnexus
Oz 07 OLIGOZIP	Ne 05 NEWBLER	Gp 10 GenePool	Mh 03 Mirch	Mc 99 Mascot	Pp 99 PIPELINE PILOT	Gs XX GENESPRING	Gu 05 Geneious	Cb 05 CLC BIO

http://elements.eaglegenomics.com/

H3ABioNet Workflows project - Pipelines



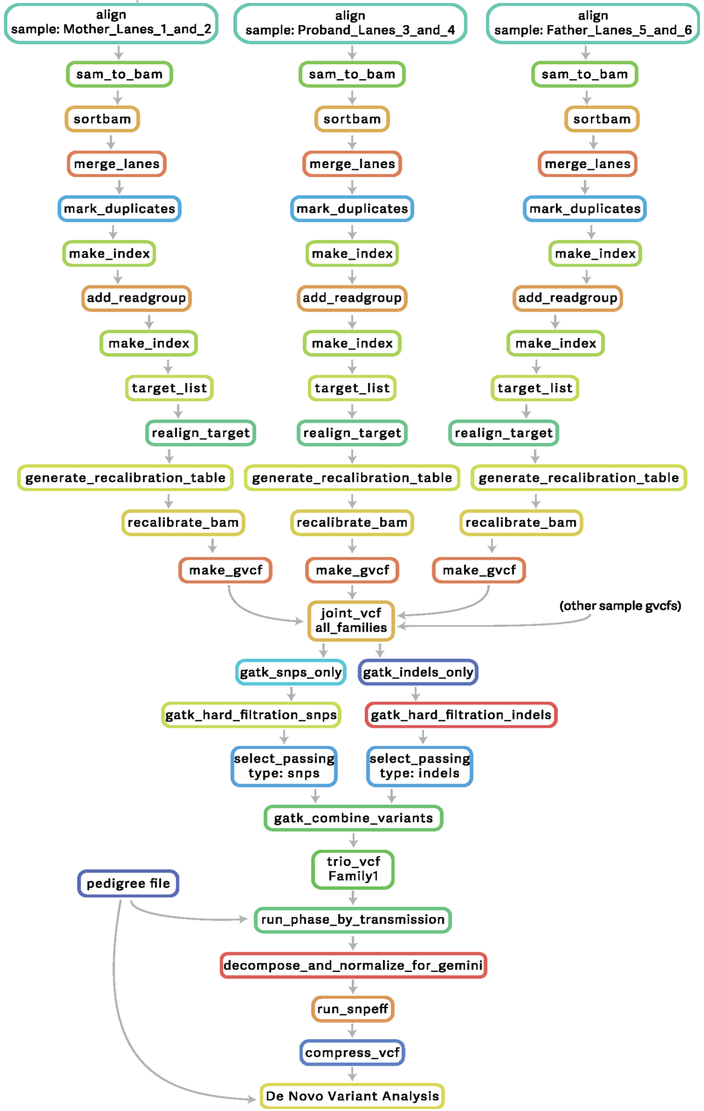
- Bioinformatics analyses → directing files through a series of transformations and programs to a final output i.e. a computational pipeline

- Transformations typically done by third-party executable command line software written for Unix-compatible operating systems

- Manually started when previous transformation step completed e.g. qsub command run

H3ABioNet
Pan African Bioinformatics Network for H3Africa

H3ABioNet Workflows project - Workflows



align
sample: Mother_Lanes_1_and_2

align
sample: Proband_Lanes_3_and_4

align
sample: Father_Lanes_5_and_6

sam_to_bam

sortbam

merge_lanes

mark_duplicates

make_index

add_readgroup

make_index

target_list

realign_target

generate_recalibration_table

recalibrate_bam

make_gvcf

joint_vcf
all_families

(other sample gvcfs)

gatk_snps_only

gatk_indels_only

gatk_hard_filtration_snps

gatk_hard_filtration_indels

select_passing
type: snps

select_passing
type: indels

gatk_combine_variants

trio_vcf
Family1

pedigree file

run_phase_by_transmission

decompose_and_normalize_for_gemini

run_snpeff

compress_vcf

De Novo Variant Analysis

- A workflow is a description of a process (pipeline) that consists of a series of tasks connected in the form a directed graph
- Tasks can be defined as single units of work e.g. split files
- A workflow comprises of an initial unique task and ends with a unique terminal task
- Completion of a task can initiate one or more tasks
- Enables automation e.g. run a pipeline from start to finish without manual input (pipeline)

Image credit: Brief Bioinform. 2016;18(3):530-536. doi:10.1093/bib/bbw020

Workflow languages and Workflow systems

COMMON
WORKFLOW
LANGUAGE

- CWL is a workflow language that is explicit
- Provides a specification for describing workflows making them portable and scalable
- Used by Galaxy, International Cancer Genome Consortium, GATK

nextflow

- Nextflow is a workflow language and system that integrates to with common resource management systems (work on CWL Toil and Cromwell integrations has progressed)
- Supports parallelism and has clear execution blocks
- Used by the Center for Genomic Regulation, Wellcome Trust Sanger Institute, Berkley, Wits' Bioinformatics, UCT CBIO

H3ABioNet Workflows project - planning

- Workflow languages to use? nextflow COMMON WORKFLOW LANGUAGE

- Containerization technology? docker

- Heterogeneous African compute environments (portability) – HPC

- What workflows to develop?

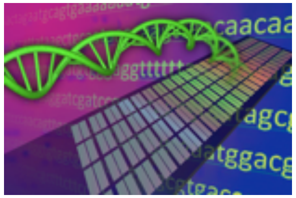


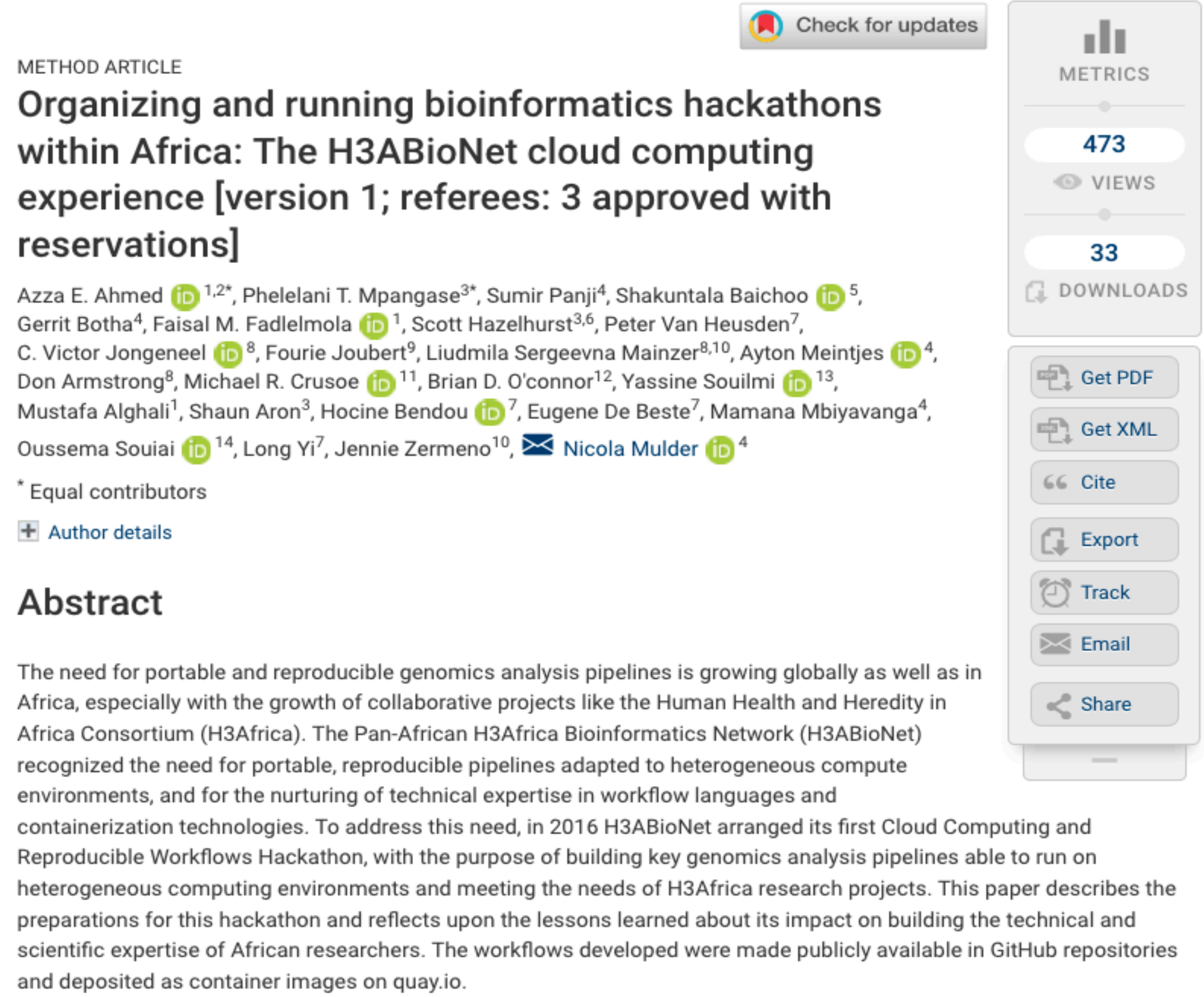
Image credits: National Human Genome Research Institute (https://www.genome.gov/imagegallery/)

H3ABioNet Workflows project - execution

- Hackathon held at University of Pretoria (Prof. Fourie Joubert's lab)

- Four streams devised (Variant calling, 16S, GWAS and Imputation)
- Each stream had a mixture of skills from bioinformaticists, developers, sys admin, knowledge of tools for the pipeline and expertise in CWL (Michael R. Crusoe) and Nextflow (Prof. Scott Hazelhurst)
- Expert in Docker containerization shared between all streams (Dr. Brian O'Connor)

H3ABioNet Workflows project - execution



METHOD ARTICLE

Organizing and running bioinformatics hackathons within Africa: The H3ABioNet cloud computing experience [version 1; referees: 3 approved with reservations]

Azza E. Ahmed[1,2*], Phelelani T. Mpangase[3*], Sumir Panji[4], Shakuntala Baichoo[5], Gerrit Botha[4], Faisal M. Fadlelmola[1], Scott Hazelhurst[3,6], Peter Van Heusden[7], C. Victor Jongeneel[8], Fourie Joubert[9], Liudmila Sergeevna Mainzer[8,10], Ayton Meintjes[4], Don Armstrong[8], Michael R. Crusoe[11], Brian D. O'connor[12], Yassine Souilmi[13], Mustafa Alghali[1], Shaun Aron[3], Hocine Bendou[7], Eugene De Beste[7], Mamana Mbiyavanga[4], Oussema Souiai[14], Long Yi[7], Jennie Zermeno[10], ✉ Nicola Mulder[4]

[*] Equal contributors

+ Author details

Abstract

The need for portable and reproducible genomics analysis pipelines is growing globally as well as in Africa, especially with the growth of collaborative projects like the Human Health and Heredity in Africa Consortium (H3Africa). The Pan-African H3Africa Bioinformatics Network (H3ABioNet) recognized the need for portable, reproducible pipelines adapted to heterogeneous compute environments, and for the nurturing of technical expertise in workflow languages and containerization technologies. To address this need, in 2016 H3ABioNet arranged its first Cloud Computing and Reproducible Workflows Hackathon, with the purpose of building key genomics analysis pipelines able to run on heterogeneous computing environments and meeting the needs of H3Africa research projects. This paper describes the preparations for this hackathon and reflects upon the lessons learned about its impact on building the technical and scientific expertise of African researchers. The workflows developed were made publicly available in GitHub repositories and deposited as container images on quay.io.

https://doi.org/10.12688/AASOPENRES.12847.1

H3ABioNet Workflows project - outputs

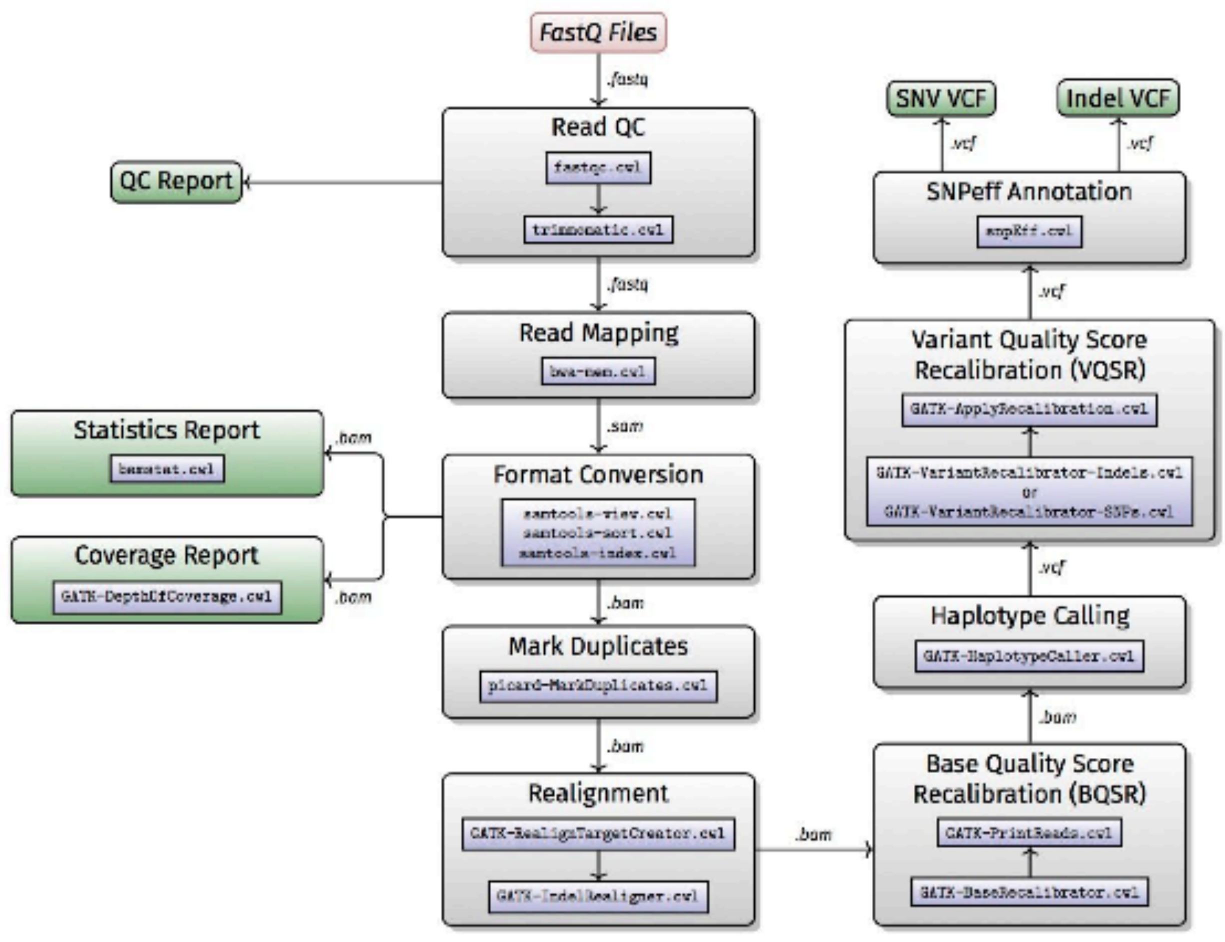


Figure 1 - Workflow A: Whole Genome/Exome NGS Data Analysis

① Sequencing adaptor, barcode and base QC trimming with *Trimmomatic*
② Quality control (QC) of the input fastq files with FastQC
③ Short reads mapping: BWA-MEM is used to perform paired-end Illumina reads
④ Quality control of the aligned reads using BAMstats
⑤ Quality control of the aligned reads using GATK's DepthOfCoverage to check the observed depth of coverage meets expected yield values.
⑥ Indels and single nucleotide variant (SNV) annotation: *SnpEff* extends the VCF file containing the variants with information relevant for downstream analysis. The information includes ranges from the SNP rsID, to clinically relevant variants from ClinVar.

Availability:
https://github.com/h3abionet/h3agatk

https://dockstore.org/workflows/h3abionet/h3agatk

Image credits: Phelelani Mpangase

H3ABioNet Workflows project - outputs

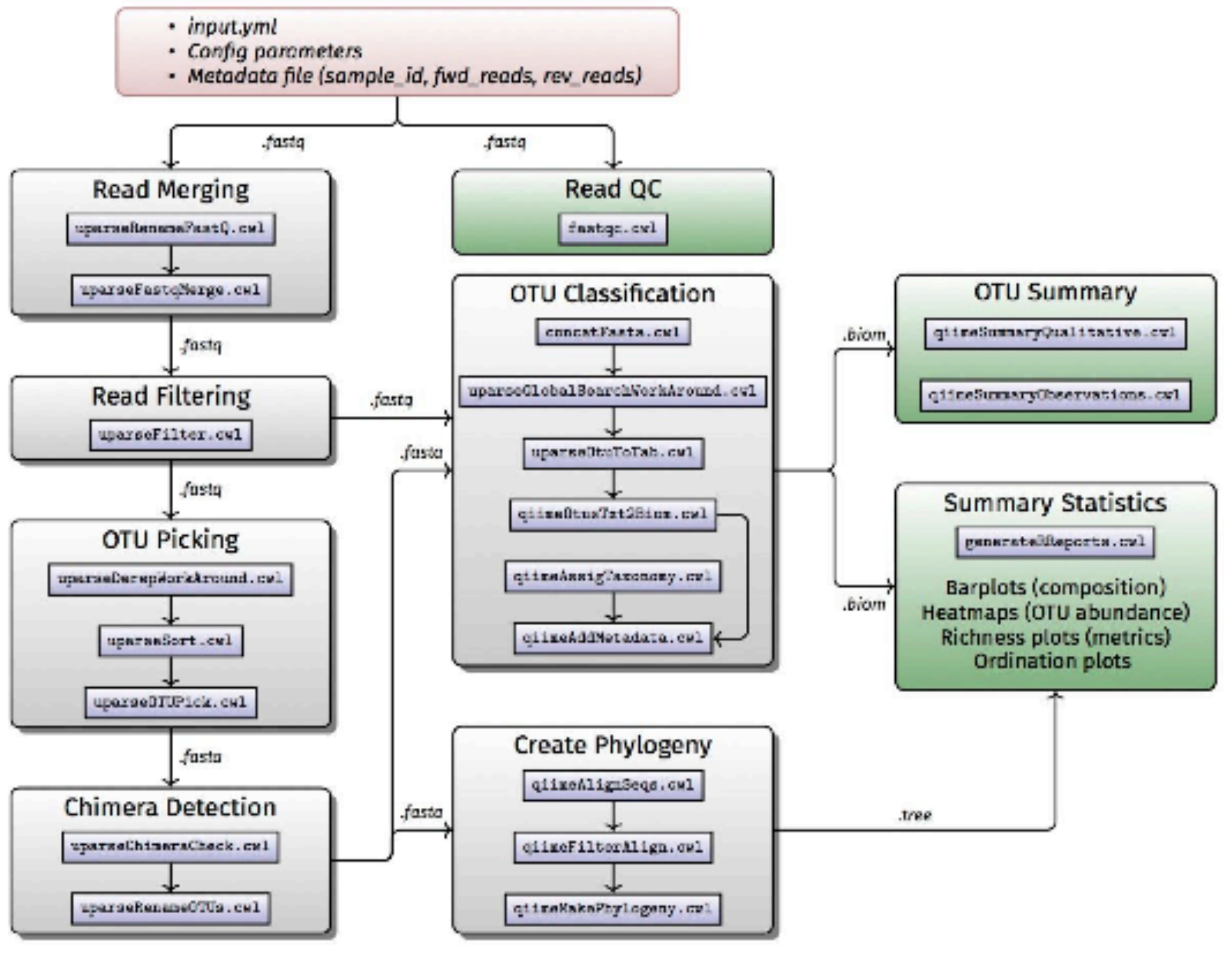


Figure 2 - Workflow B: 16S rDNA Diversity Analysis

1. FastQC for QC reporting of the data quality
2. In house scripts: “UPARSE fastq renamer” to rename the FastQ files for compatibility with UPARSE scripts and “UPARSE derep” workaround and mapping back to OTUs
3. Helper scripts: “Combine fastqc reports”, “fastq renamer” for UPARSE compatibility and “Fasta splitter” for splitting files
4. usearch for QC and OTU clustering
5. QIIME modules for demultiplexing, quality filtering, OTU picking, taxonomic assignment, phylogenetic reconstruction, diversity analyses of the data and the generation of OTU summary
6. R and relevant packages for statistical analysis using QIIME results, namely PhyloSeq

Availability:
https://github.com/h3abionet/h3abionet16S

https://quay.io/repository/h3abionet_org/h3a16s-qiime

Image credits: Phelelani Mpangase

H3ABioNet Workflows project - outputs

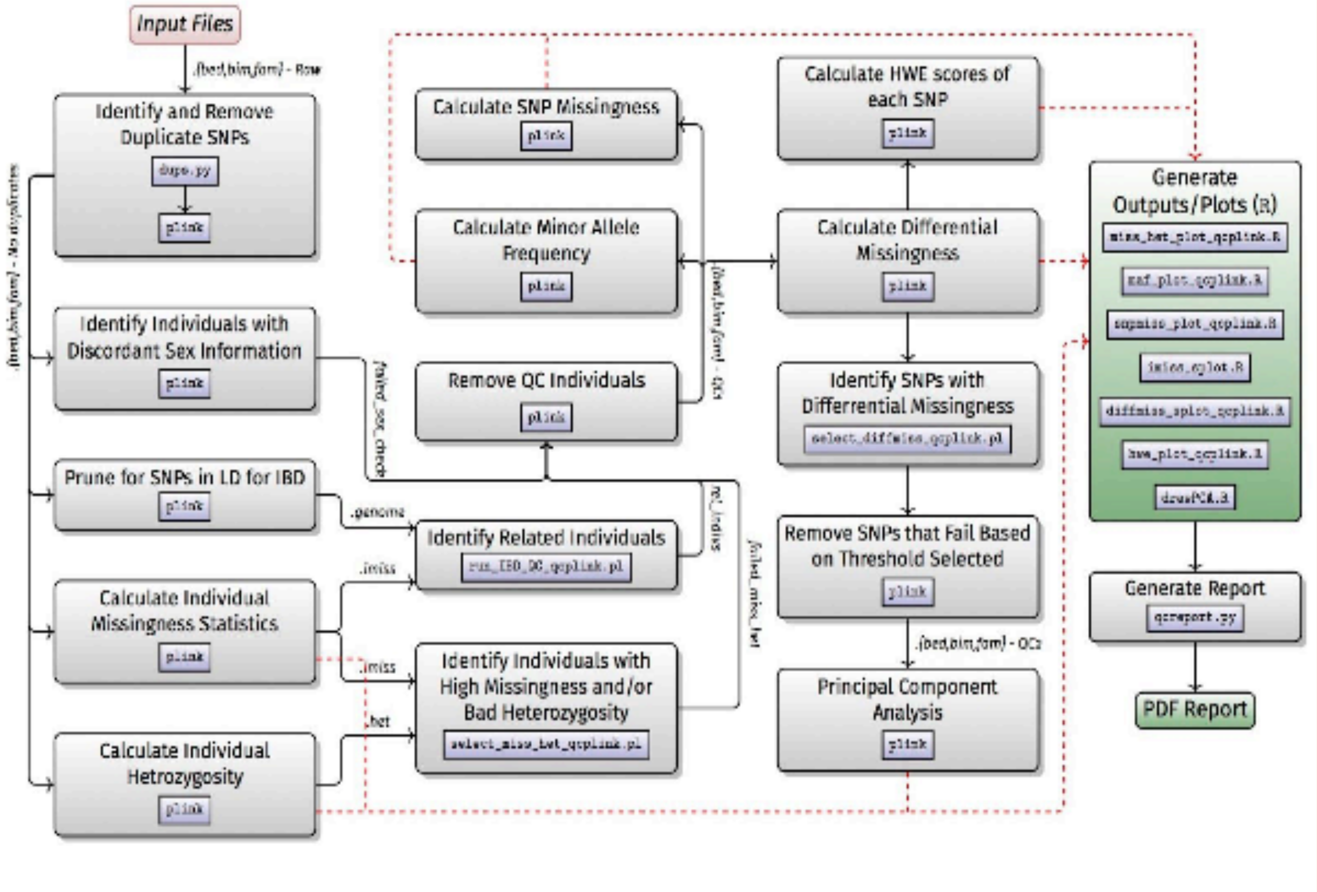

Input Files
.{bed,bim,fam} - Raw
Identify and Remove Duplicate SNPs
dups.py
plink
.{bed,bim,fam} - No duplicates
Identify Individuals with Discordant Sex Information
plink
Prune for SNPs in LD for IBD
plink
.genome
Calculate Individual Missingness Statistics
plink
.imiss
.imiss
Calculate Individual Hetrozygosity
plink
.het
failed_sex_check
Calculate SNP Missingness
plink
Calculate Minor Allele Frequency
plink
Remove QC Individuals
plink
Identify Related Individuals
run_IBD_QC_qcplink.pl
Identify Individuals with High Missingness and/or Bad Heterozygosity
select_miss_het_qcplink.pl
.{bed,bim,fam} - QC1
rel_indivs
failed_miss_het
Calculate HWE scores of each SNP
plink
Calculate Differential Missingness
plink
Identify SNPs with Differential Missingness
select_diffmiss_qcplink.pl
Remove SNPs that Fail Based on Threshold Selected
plink
.{bed,bim,fam} - QC2
Principal Component Analysis
plink
Generate Outputs/Plots (R)
miss_het_plot_qcplink.R
maf_plot_qcplink.R
snpmiss_plot_qcplink.R
imiss_qplot.R
diffmiss_qplot_qcplink.R
hwe_plot_qcplink.R
drawPCA.R
Generate Report
qcreport.py
PDF Report

Figure 3 - Workflow C: Genome Wide-association studies

The workflow (Figure 3) consists of 3 modules, which can be swapped in and out depending on the analysis needs:

1. Conversion from Illumina TOP/ BOTTOM call format to PLINK format.
2. The core workflow carries out a set of QC steps, starting with standard PLINK files and resulting in quality controlled PLINK files.
3. Basic association testing and structure analysis.

In addition, we expect many researchers will use the imputation workflow after QC and before association testing.

Availability:

http://github.com/h3abionet/h3agwas

https://quay.io/organization/h3abionet_org

Image credits: Phelelani Mpangase

H3ABioNet Workflows project - outputs

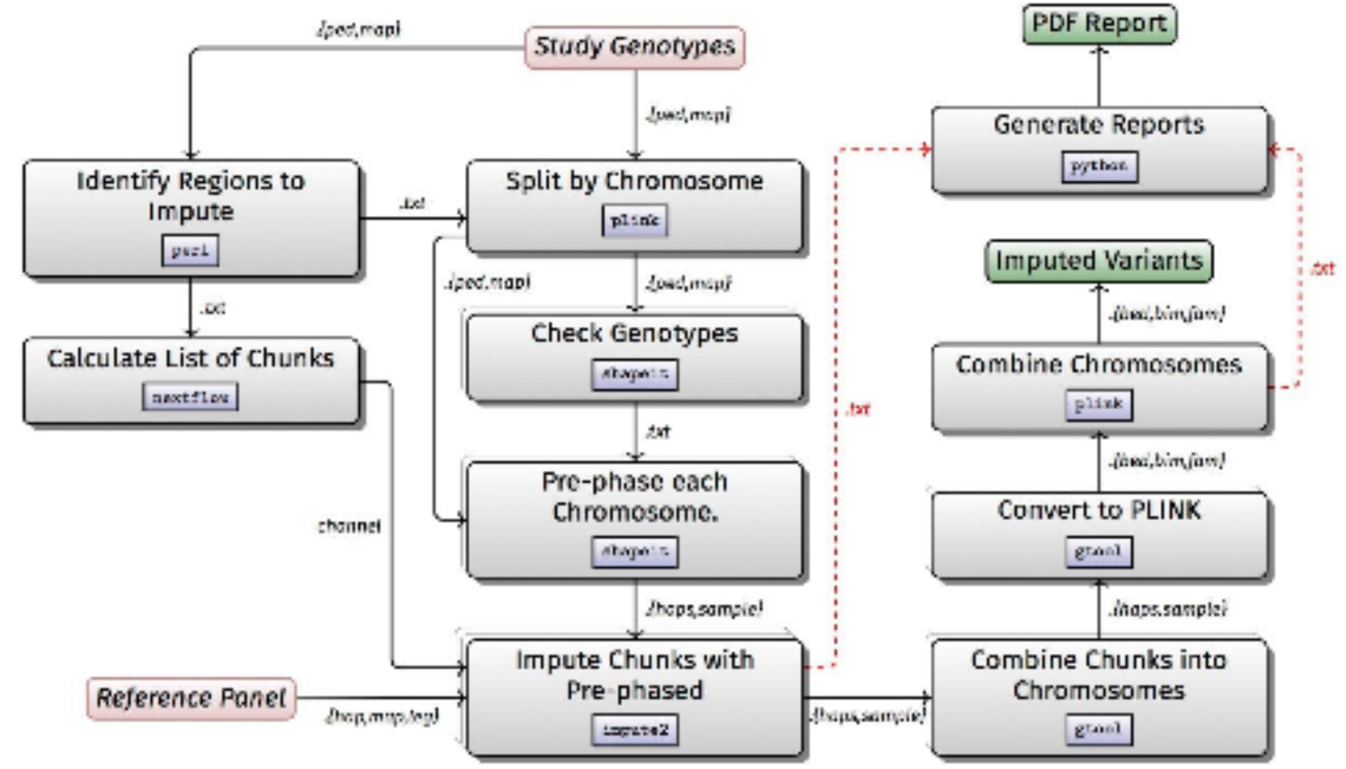


Figure 4 - Workflow D: SNPs Imputation: Boxed subgraphs indicate pathways which are executed in parallel (per chromosome and per region within each chromosome) as computational resources permit.

① Identify regions for imputation based on PLINK format an input file, output produced in IMPUTE haplotype format

② Ped and map input files split by chromosome using PLINK, chromosome extents are identified using a combination of awk and grep

③ Genotyped positions on individual chromosomes checked for strand flipping errors, improperly stranded positions are excluded using SHAPEIT

④ Genotyped positions prephased using SHAPEIT in parallel on each chromosome

⑤ IMPUTE2 run in parallel across all 500kB blocks in the entire genome

⑥ Imputed blocks combined into a single compressed haplotype file using custom perl script provided with the workflow.

⑦ File converted back to a PLINK dataset for integration back into the GWAS workflow

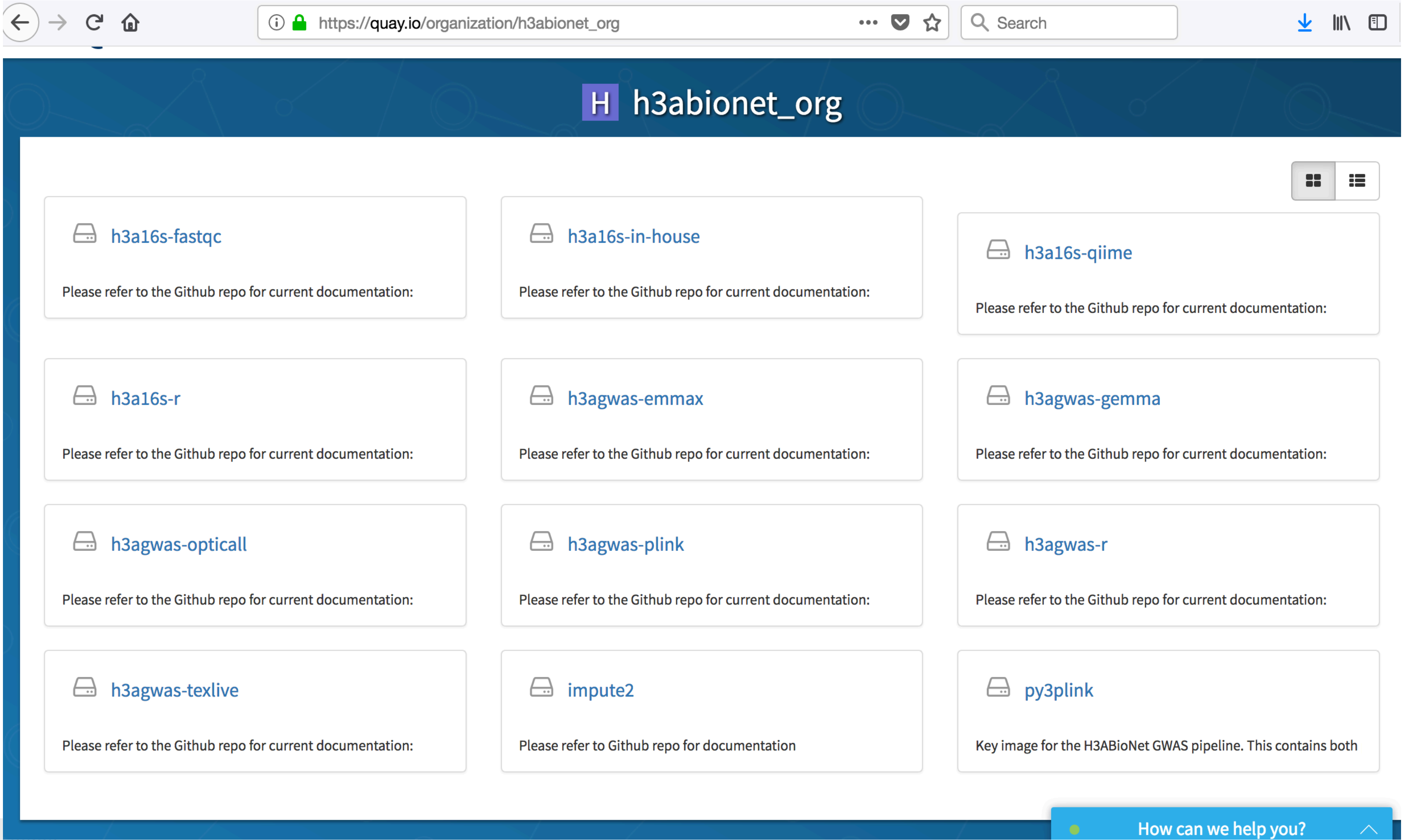
Availability:

https://github.com/h3abionet/chipimputation/

https://quay.io/repository/h3abionet_org/impute2

Image credits: Phelelani Mpangase

H3ABioNet Workflows project - outputs



https://quay.io/organization/h3abionet_org

h3abionet_org

h3a16s-fastqc

Please refer to the Github repo for current documentation:

h3a16s-in-house

Please refer to the Github repo for current documentation:

h3a16s-qiime

Please refer to the Github repo for current documentation:

h3a16s-r

Please refer to the Github repo for current documentation:

h3agwas-emmax

Please refer to the Github repo for current documentation:

h3agwas-gemma

Please refer to the Github repo for current documentation:

h3agwas-opticall

Please refer to the Github repo for current documentation:

h3agwas-plink

Please refer to the Github repo for current documentation:

h3agwas-r

Please refer to the Github repo for current documentation:

h3agwas-texlive

Please refer to the Github repo for current documentation:

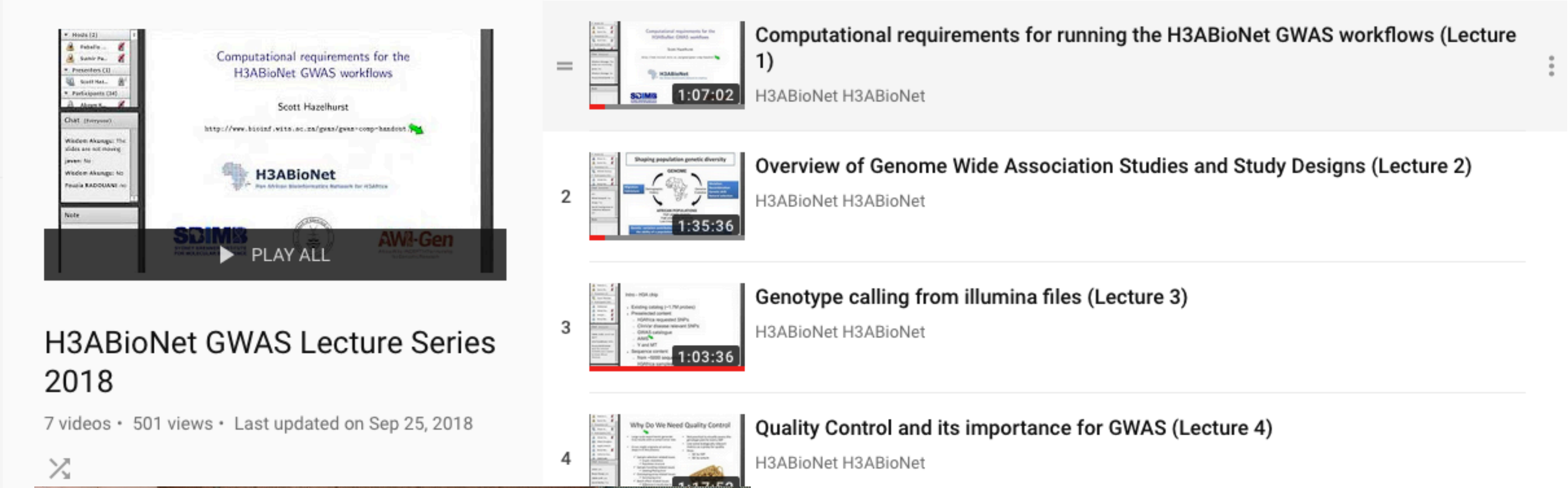
impute2

Please refer to Github repo for documentation

py3plink

Key image for the H3ABioNet GWAS pipeline. This contains both

H3ABioNet Workflows project - training



H3ABioNet GWAS Lecture Series 2018

7 videos • 501 views • Last updated on Sep 25, 2018

1. Computational requirements for running the H3ABioNet GWAS workflows (Lecture 1)
   H3ABioNet H3ABioNet
2. Overview of Genome Wide Association Studies and Study Designs (Lecture 2)
   H3ABioNet H3ABioNet
3. Genotype calling from illumina files (Lecture 3)
   H3ABioNet H3ABioNet
4. Quality Control and its importance for GWAS (Lecture 4)
   H3ABioNet H3ABioNet

Population structure in GWAS (Lecture 5)
H3ABioNet H3ABioNet

Imputation and its importance in GWAS (Lecture 6)
H3ABioNet H3ABioNet

Statistical models used for GWAS (Lecture 7)
H3ABioNet H3ABioNet

https://www.youtube.com/playlist?list=PLcQ0XMykNhCQJPz0amnbz9BPM4Bu0Nkgf

H3ABioNet Workflows project - FAIR

Findable: https://github.com/h3abionet

Accessible: https://github.com/h3abionet

Interoperable: https://quay.io/organization/h3abionet_org

Reusable (Reproducible): https://quay.io/organization/h3abionet_org

Acknowledgements

- Prof Nicky Mulder and H3ABioNet members

H3ABioNet Consortium Members 2018

H3Africa
Human Heredity and Health in Africa