Building infrastructure and capacity to enable genomic research

Nicola Mulder
University of Cape Town

H3ABioNet
Pan African Bioinformatics Network for H3Africa

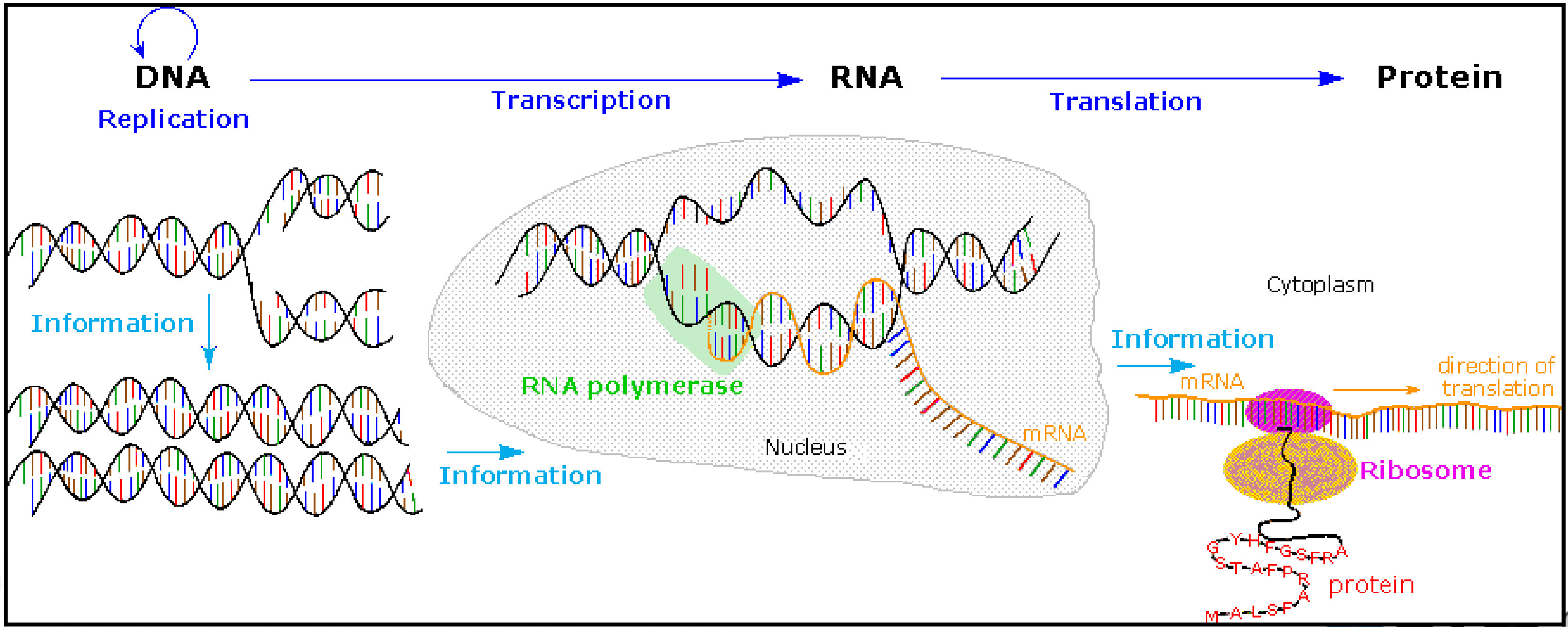
CBIO
Computational Biology @ UCT

Outline

- What is genomics?
- Local efforts -BSP
- Africa-wide efforts -H3Africa and H3ABioNet
- World-wide efforts –Global Alliance
- Enabling genomic research

What is genomics?

- http://en.wikipedia.org/wiki/Genomics: **Genomics** is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes (the *complete* set of DNA within a single cell of an organism)



Human genome

- 3 billion base pairs of DNA -requires >3 gigabytes of computer storage space
- Full genome done by NGS: 1 BAM = 50-80GB
- Humans vary by 0.1% -1 in every 1000bp (3 mill SNPs)
- We study genetic variation to:
  - Learn about human migration patterns and history
  - Improve power to identify and localize disease genes
  - Leading to personalized medicine

eResearch challenges in genomics

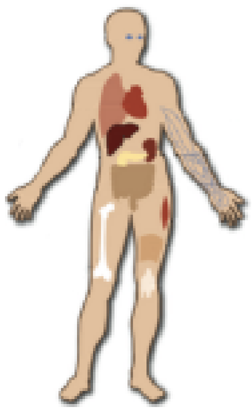
- Data generation is expensive and time consuming therefore effective data management and exploitation is essential
- Data is getting bigger to store
- Processing is getting more intense
- Data security is vital
- Results can save lives!

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Genomics experiment workflow

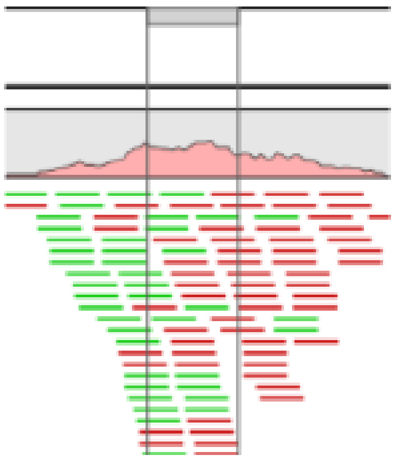
Patient cohorts



Collect samples

Extract DNA

Biorepository



Generate data

Data analysis and storage

Data repository

Computing Infrastructure	**Research/tool development**	**User support**	**Training**

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Bioinformatics support for genomics

- Local: DST funded National Bioinformatics Service Platform (BSP) –supporting SA researchers, industry and national projects
- Africa-wide: H3ABioNet –supporting H3Africa
- World-wide: Global Alliance for Genomics and Health (G4GH) –world-wide sharing of data

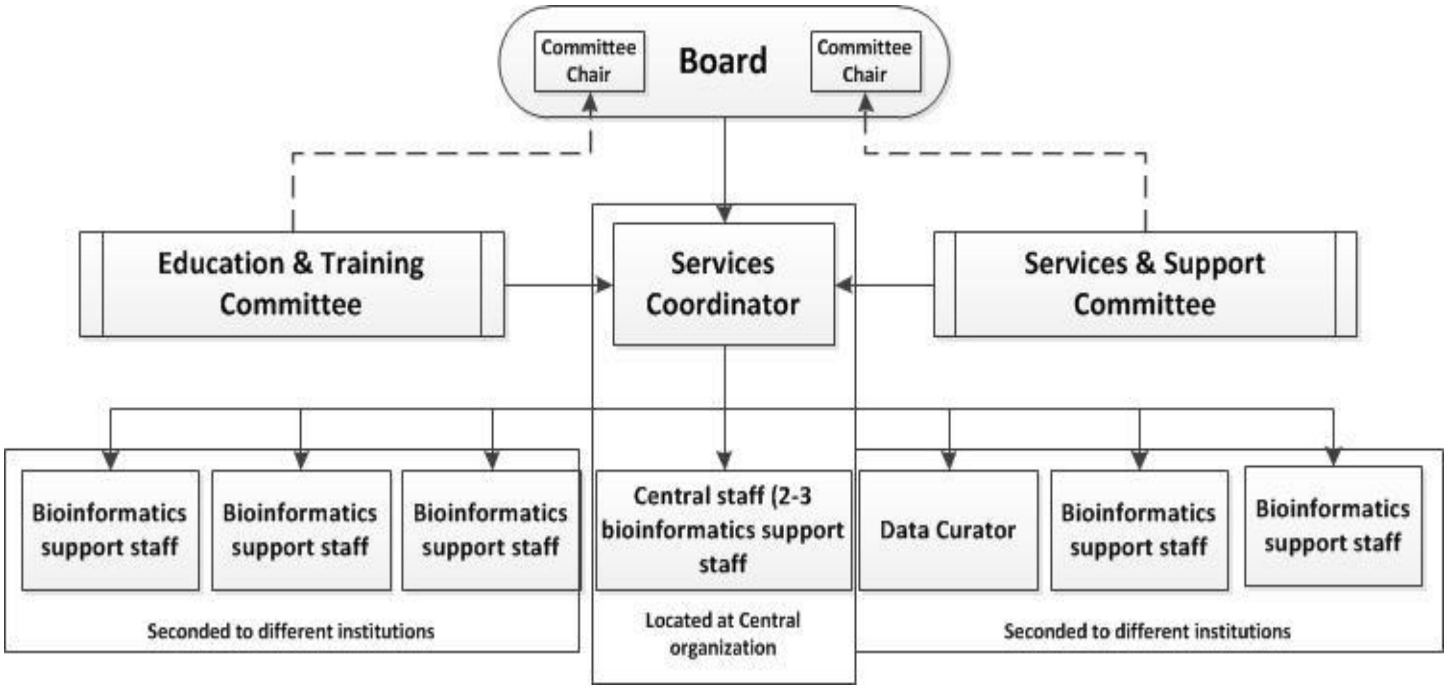
H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

National Bioinformatics Service Platform

- Funded by Dept of Science and Technology, managed through CSIR-CHPC
- Objectives:
  - Provision of bioinformatics services and support to researchers at universities and companies who require custom bioinformatics services
  - Provision of technical and intellectual support for bioinformatics specialists working within lab-based research groups
  - Development and maintenance of new tools and resources relevant to the needs of the biotech industry and local researchers
  - Development of computing infrastructure accessible to SA life scientists
  - Provision of an annual national training program for postgraduate students in bioinformatics
  - Provision of specialized bioinformatics training to life science researchers in industry and academia

BSP structure



Institutions bid to host a BSP Bioinformatician

Life Scientists submit requests for support –central coordination

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Progress to date

- Proposal approved by DST
- Steering Committee established
- Training and Services committees established
- BSP Manager hiring in progress
- Next:
  - Bid to host
  - Hiring of support staff

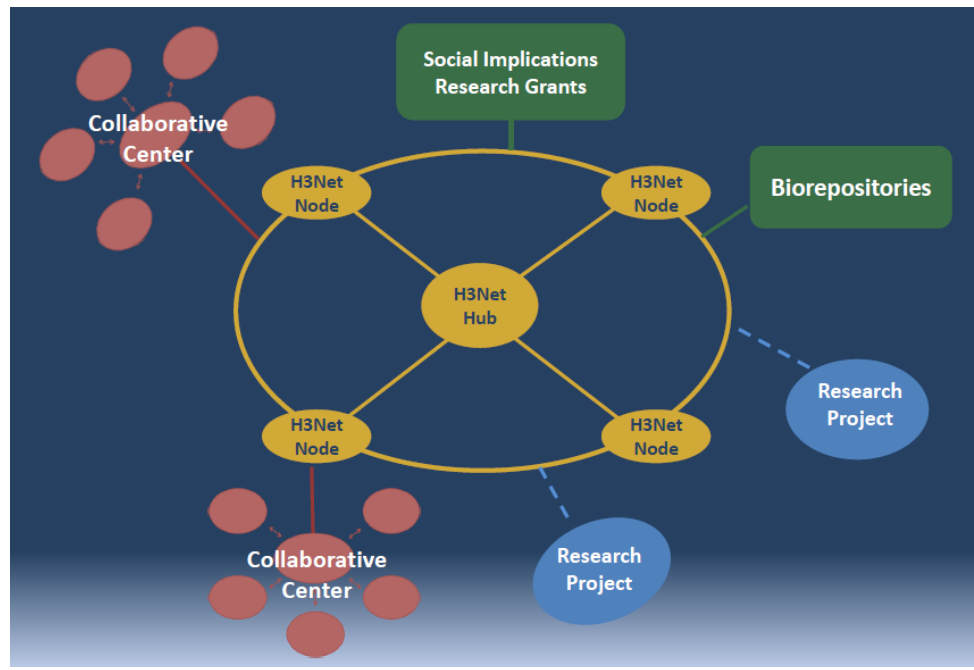
H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

H3Africa Initiative

- Human Heredity and Health in Africa -Aim is to encourage genomic research in Africa on relevant health issues

Projects funded: Obesity, Diabetes, Heart Disease, Stroke, Kidney Disease, Sickle Cell (Ethics), Microbiome (2), Trypanosomes (2), TB (2), Schizophrenia, Rare neurological disorders, Fevers of unknown origin



- Types of projects:
  - Collaborative Centres (NIH)
  - Research Projects (NIH and WT)
  - Biorepositories (NIH)
  - Societal Implications Research (NIH)
  - **Bioinformatics Network (NIH)**

H3ABioNet
Pan African Bioinformatics Network for H3Africa

www.h3africa.org

CBIO
Computational Biology @ UCT

H3ABioNet

- Aim: to build H3ABioNet -- a sustainable African Bioinformatics Network -- to provide bioinformatics infrastructure and support for the H3Africa consortium
- Administrative hub at UCT, 34 partner institutions, 32 in 15 African countries, 2 in USA, >200 consortium members
- Expect >50,000 samples, >500TB data (NGS and arrays)

www.h3abionet.org

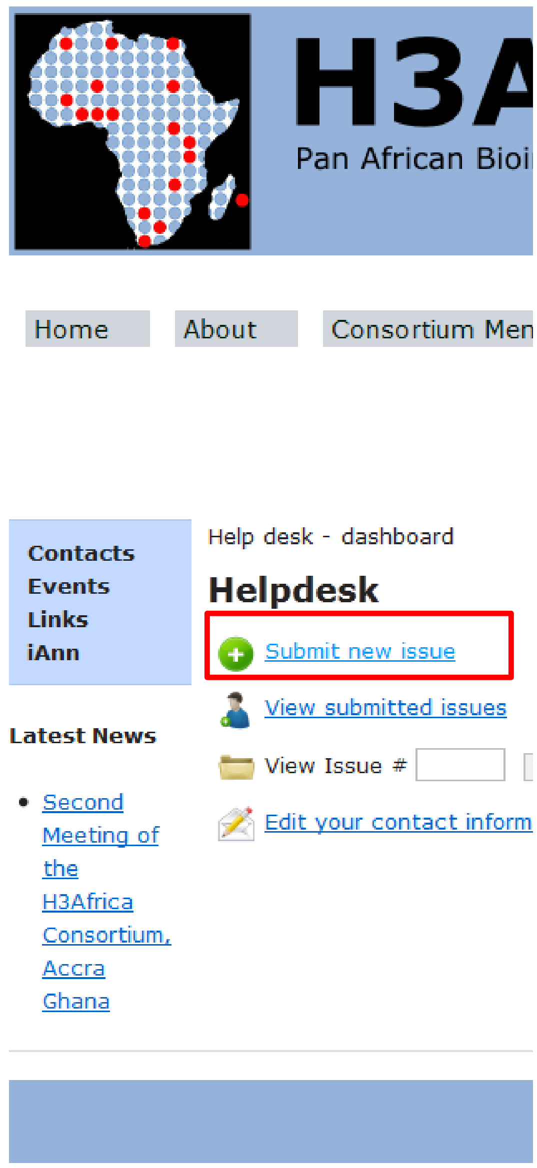
H3ABioNet
Pan African Bioinformatics Network for H3Africa

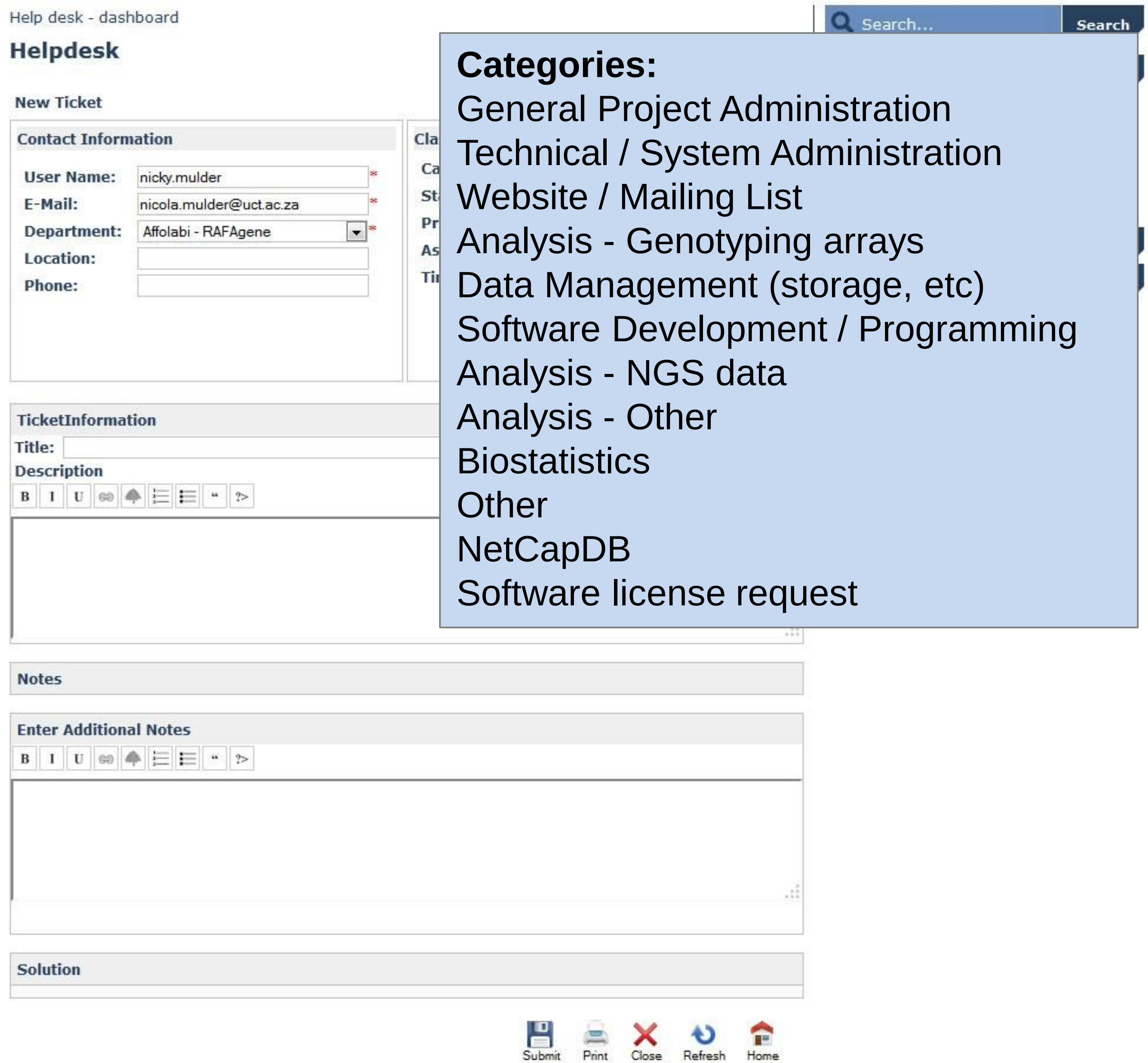
Website

H3ABioNet
Pan African Bioinformati

CBIO Computational Biology @ UCT

H3ABioNet help desk

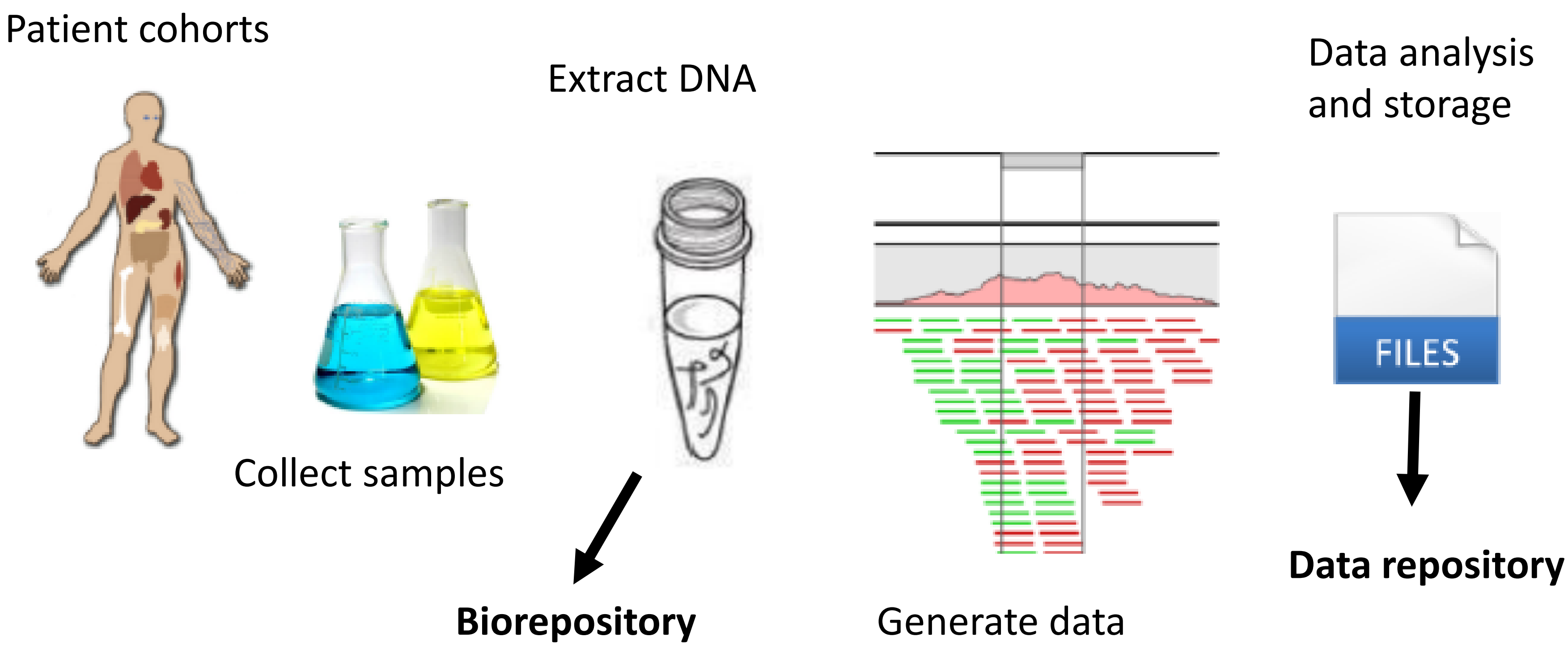




Categories:

General Project Administration
Technical / System Administration
Website / Mailing List
Analysis - Genotyping arrays
Data Management (storage, etc)
Software Development / Programming
Analysis - NGS data
Analysis - Other
Biostatistics
Other
NetCapDB
Software license request

Genomics experiment workflow



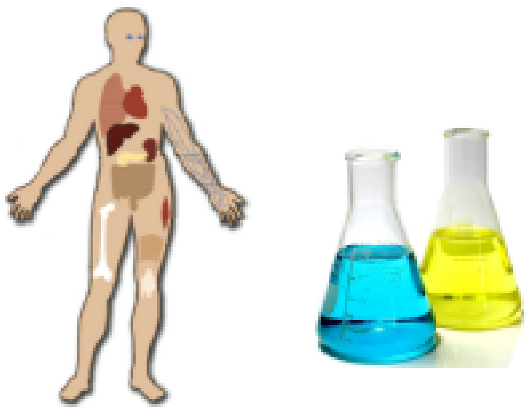
Computing Infrastructure **Research/tool development** **User support** **Training**

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Genomics experiment workflow

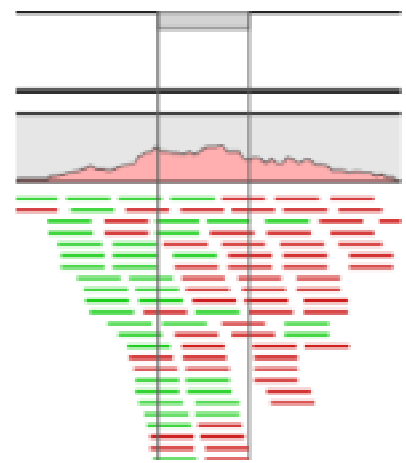
Patient cohorts



Collect samples

Extract DNA

Data analysis
and storage



Patient databases, LIMS,
data integration

Biorepository

Generate data

Data repository

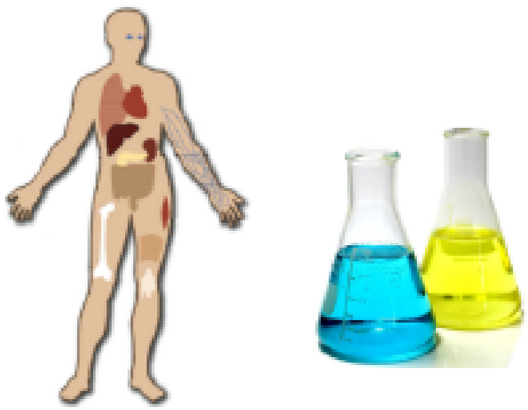
Computing Infrastructure | **Research/tool development** | **User support** | **Training**

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Genomics experiment workflow

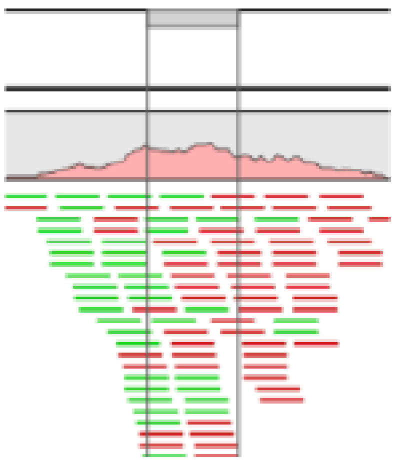
Patient cohorts



Collect samples

Extract DNA

Biorepository



Generate data

Data analysis
and storage

Data repository

Computing Infrastructure **Research/tool development** **User support** **Training**

H3ABioNet
Pan African Bioinformatics Network for H3Africa

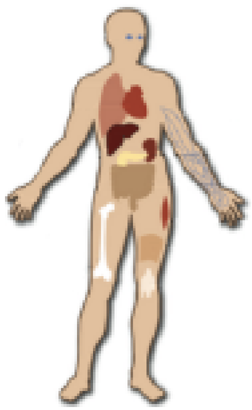
Chip design
Data management and analysis

CBIO
Computational Biology @ UCT

Genomics experiment workflow

Ontologies, data integration, data archive

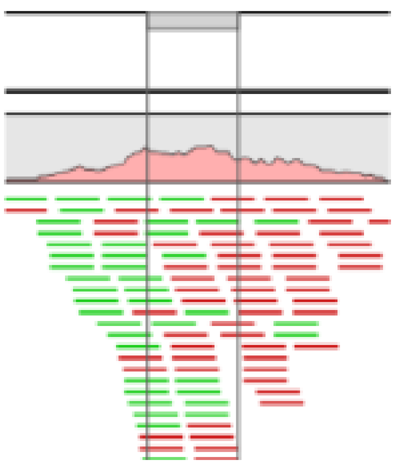
Patient cohorts



Collect samples

Extract DNA

Biorepository



Generate data

Data analysis and storage

Data repository

Computing Infrastructure | **Research/tool development** | **User support** | **Training**

Computing infrastructure: Systems Administrator Assistance

- Developed preferred list of bioinformatics applications based on survey
- Developed a workflow to determine the hardware required based on the nodes research
- Rolled out hardware across African sites
- Developed server installation documentation
- Developing server security and best practices documentation
- Future development of a server monitoring best practice documentation
- Provided training in sys admin, HPC, etc

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

H3ABioNet

Pan African Bioinformatics Network for H3Africa

Home | Training & Education | Research | H3ABioNet Helpdesk | Working Groups | Tools and R

H3ABioNet Tools | SOPs | Useful documents | External tools

Useful documents

System Administrator HowTo Guides

The system administrator taskforce has developed a series of howto documents to assist H3ABioNet system administrators with installing, configuring and securing, monitoring and managing their server hardware.

The howto documentation has been developed to be usable by both non-IT individuals and those just new to Linux by providing detailed text based step by step instructions supported by images where possible. These howto's have been divided into three levels: level1, level 2 and level 3.

Level 1

The level one documentation focuses on helping you get started by installing a Linux operating system on your server, the howto covers the below subjects:

- Howto configure the BIOS on the Dell C6145 server
- Howto setup and configure a hardware and software RAID
- Howto install a Linux operating system
- Howto navigate a Linux system via the command line
- Howto edit files via the command line
- Howto setup file sharing
- Howto schedule redundant tasks via cron
- Provide a list of useful commands

Download a PDF copy of the level 1 howto guide here >> Linux: Getting Started HowTo Guide

Level 2

The level two documentation is intended to be used in conjunction with the level 1 documentation and assumes the knowledge covered in the level 1. The Level 2 howto covers the below subjects:

x: Getting Started HowTo Guide

A technical howto document presented to H3ABioNet

H3ABioNet

Pan African Bioinformatics Network for H3Africa

Created by

The System Administrator Task-force

Prepared for

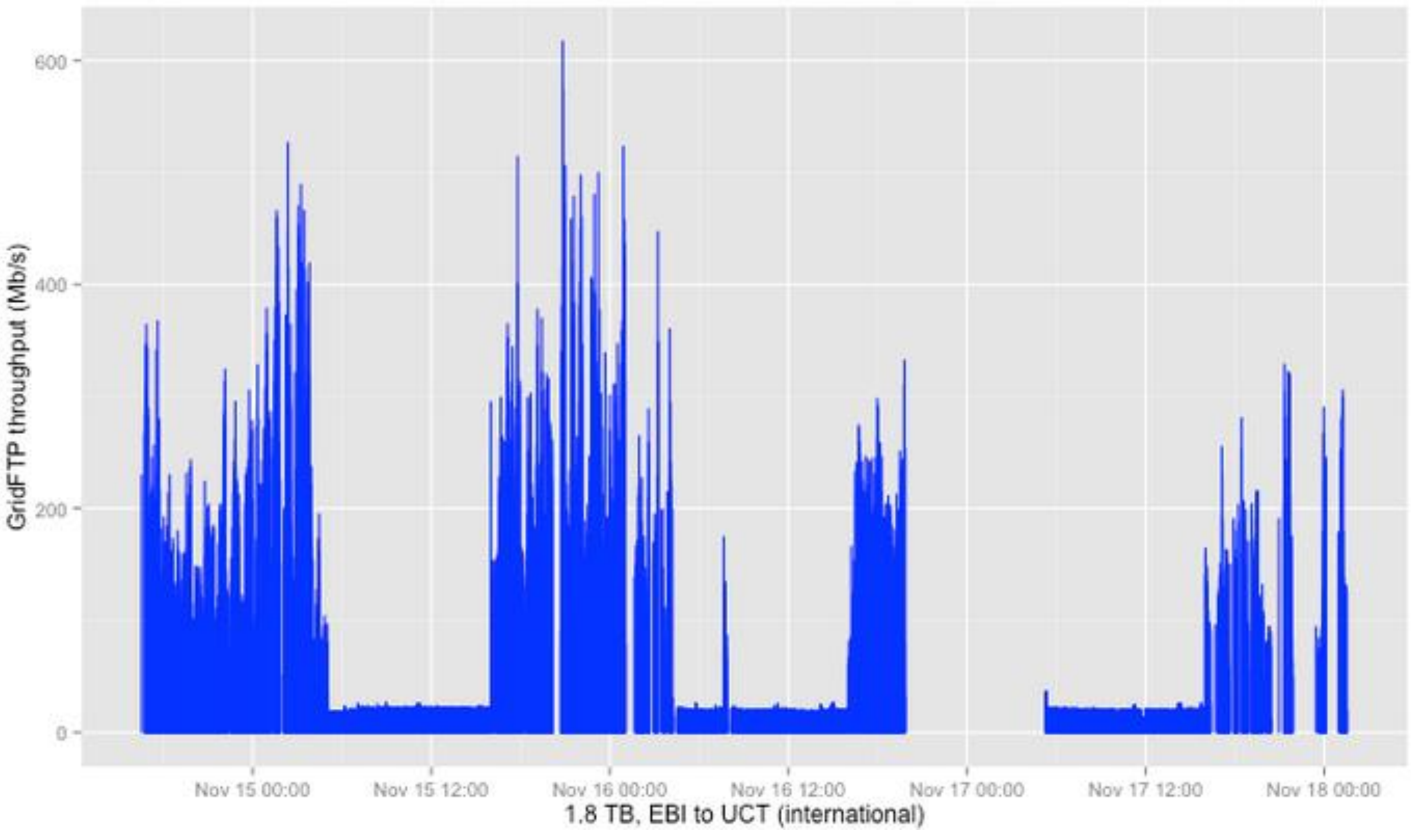
The greater H3ABioNet and H3Africa Consortium community

Computing infrastructure: Data transfer

- Traditional transfer methods such as FTP are not optimal for large datasets
- Globus Online (GridFTP) and Aspera use different protocols, and are more reliable and secure
- E.g. 16GB dataset from EBI (UK) to UCT: FTP 22 Mb/s; GridFTP 220 Mb/s
- Assisting H3ABioNet nodes with Globus installation
- Recent transfer experiences
  - Genome and exome datasets from 3 different sites in the US (e.g. 1 50X genome, 350GB @ 39Mbps, took 20 hours)
  - 24 full genomes (2TB) between UCT and WITS
  - Microbiome data from JCVI (USA)
  - RNASeq data from Qatar to UCT to Bulgaria
  - 1000 genomes subset (7TB)

Example international transfer

Globus results from a transfer of 1000 genomes data from EBI to UCT's landing area using the UCT international bandwidth



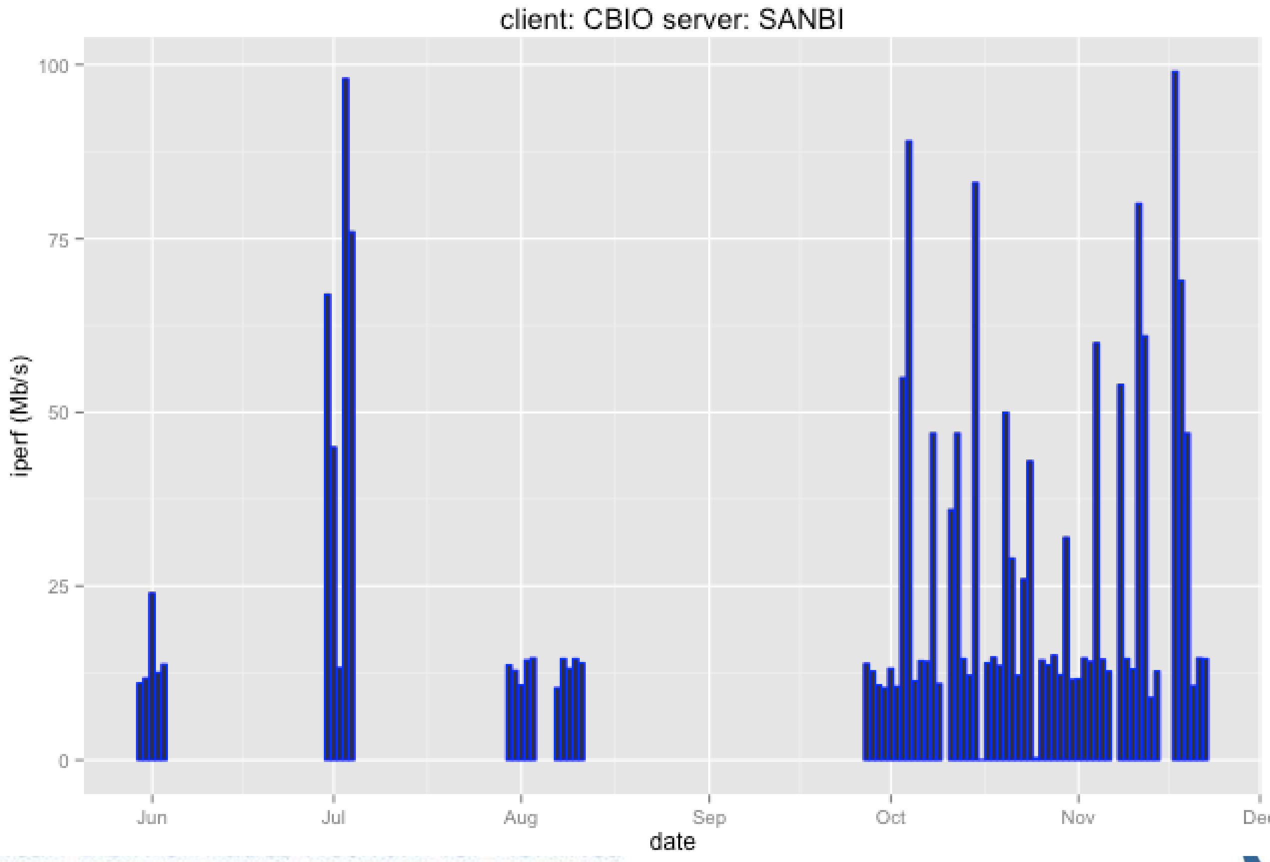
Note, transfer to DMZ is quicker than to CBIO endpoint

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Computing infrastructure: Netmap project

- Documenting network topology/bandwidth between H3ABioNet nodes and non-African sites (JCVI, NCSA, EBI, NCBI)
- Netmap results from one iperf test pair (CBIO to SANBI) -missing data points are due to iperf server on the remote end being down or iperf client not running the test



H3A
Pan Af

CBIO
Computational Biology @ UCT

Computing infrastructure: H3Africa Archive

- Submission of H3Africa data to European Genome-phenome archive (EGA) is a funder requirement
- Estimated overall storage capacity of 500TB
- Architecture modeled on EGA system, components:
  - Landing area
    - datasets are encrypted by submitter
  - Vault area
    - Focus on security, data only ever decrypted in Vault
    - All access and operations are logged
    - Analyses are limited to
      - QC validation
      - Checking EGA file format requirements
      - H3A metadata validation
  - Archive area
    - Purely for storage, no processing
    - encrypted files are mirrored to a separate physical location

Other useful documents and policies

- Data sharing, access and release policy
- Data access agreement
- SOPs for pipelines with computing requirements
  - GWAS
  - NGS
  - 16srRNA

Other useful documents

- Data sharing, access and relea
- Data access agreement
- SOPs for pipelines with comput
  - GWAS
  - NGS
  - 16srRNA

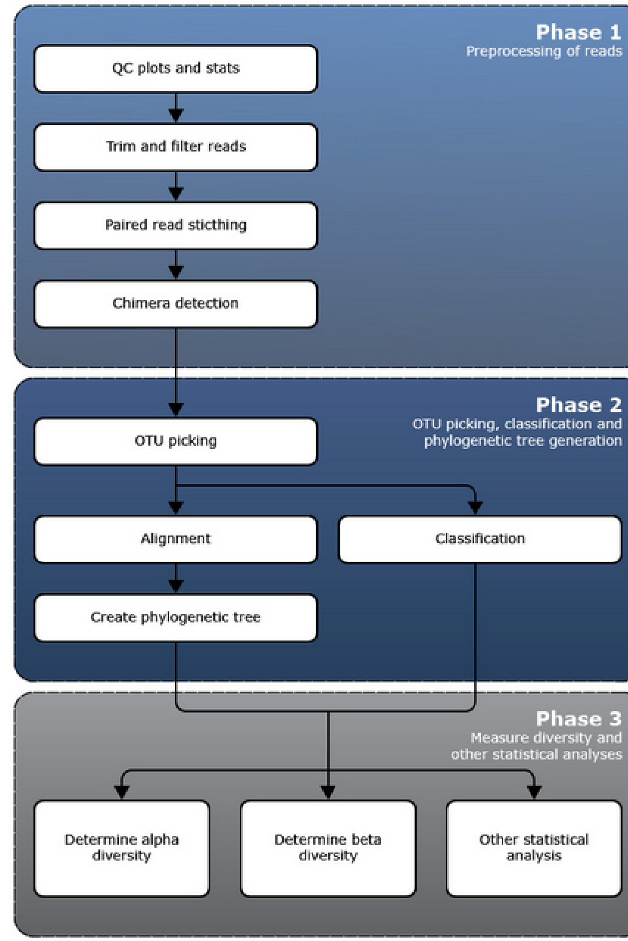
Standard operating procedure for 16S rRNA diversity analysis

Introduction

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. Nevertheless, 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. The analytical process is known as "16S rDNA diversity analysis", and is the focus of the present SOP.

The procedure and tools are only recommendations and it is up to the user to evaluate what works best for their needs.

Schematic workflow of the analysis



Definition of terms used

Phase 1: Preprocessing of reads

Phase 2: OTU picking, classification and phylogenetic tree generation

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Bioinformatics training for genomics

- Need to train bioinformaticians and data scientists to provide support
  - BSP: National courses for postgrad Bix students
  - H3ABioNet: postgrad degree curriculum development and course work –African Bioinformatics Education committee
- Need to train researchers to manage and analyse their data
  - Short specialised courses (BSP & H3ABioNet) for researchers
- Internships

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

H3ABioNet training activities

- Run 12 workshops over 2 years, **trained >230 people directly**
- Placed and trained 5 interns
- NGS Train-the-trainer program with EBI
- Wordpress sites available for all courses –many have recorded lectures
- Evaluating existing online courses

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

H3ABioNet training activities

Workshop Name	Period / location	Number of people trained
Grants management	May 2013 – South Africa	19
Technical (Sys admin) workshop	June 2013 – South Africa	18
Train the Trainer workshop	July 28th – Kenya	21 (6 participants from H3Africa projects)
eBioKits workshop	August 2013 – Kenya	26 (some local student participants from ICIPE)
NABDA visual analytics workshop	August 2013 – Nigeria	~ 20 participants (10 H3ABioNet, 5 H3Africa)
Curriculum Development workshop	March 2014 – Botswana	25 (6 H3Africa participants)
Introductory Bioinformatics workshop	March 2014 – Ghana	36 participants
GWAS workshop (part funded in conjunction with AWI-GEN)	April 2014 – South Africa	28 participants from various funding streams (H3ABioNet had 5 participants)
Postgraduate workshop	April 2014 – Nigeria	26 (3 H3Africa participants)
Data management workshop	June 2014 - South Africa	33 participants (19 H3Africa)
Intermediate Bioinformatics workshop	July 2014 – Ghana	36 participants
Metagenomics data analysis workshop	July 2014 – Nigeria	21 participants

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Trainee /Trainer Database

- Monitor training workshops held
- Track number of people trained and impact
- Monitor applicants and trainees over time
- Track course evaluations
- Keep record of trainers

Shakuntala Baichoo, Zahra Mungloo-Dilmohamud

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Types of data to be collected

- Demographic and other personal details from H3Africa members & non-members
  - name, gender, dob, address, institution details, language, email, role in H3Africa working groups
- Outputs/outcomes from H3Africa members every 6 months
  - Ongoing research project, list of publications, list of awards and fellowships, details about postgraduate students supervised
- Survey training needs and challenges being faced by H3Africa members
  - Outcome of training attended, how will the training help members to enhance their career

eResearch-related H3Africa projects

- Recruitment ‘database’
- Biobank data integration
- **Ontologies for metadata**
- **Custom chip design**

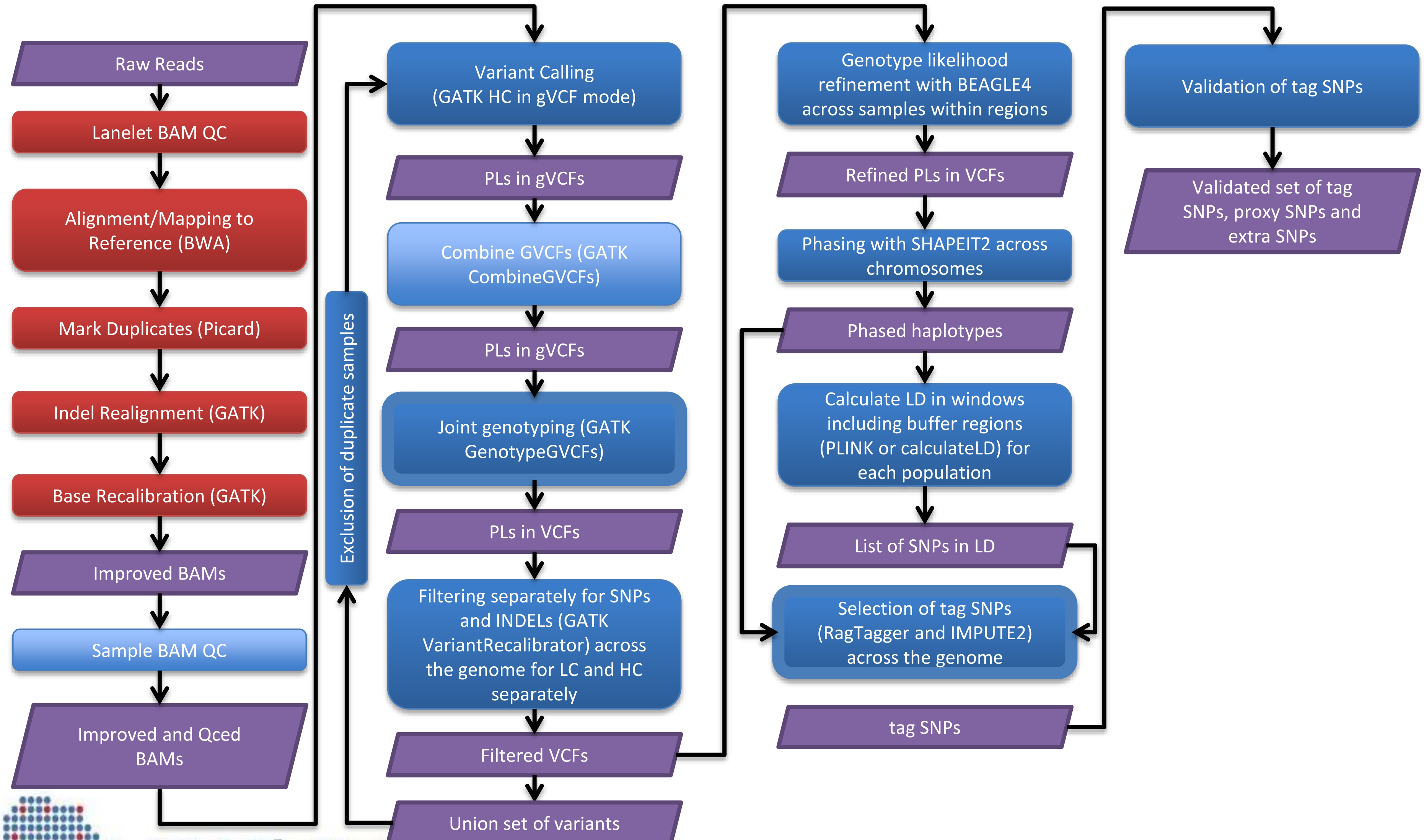
Ontologies project

- 24 questions from the Phenotype Harmonization working group plus domain specific questions e.g. CVD
- Not always a standard way of reporting these –need ontologies
- EGA data submission will require different ontologies e.g
  - experimental factor ontology -NGS/array platform, clinical measurements
  - disease ontology for phenotypes for diseases
- Mapping 24 phenotypes to ontologies
- Curator appointed to assist with the mapping and curation of the ontologies to the phenotypes and CRFs
- Aim is to integrate these ontologies as metadata for the Data archive so the contents can easily be searched
- Funding to run pilot project on SCD ontologies

H3Africa custom chip design

- Genotyping can be done by sequencing or by genotyping arrays –arrays are cheaper
- Existing arrays are biased to non-African populations
- Designing a new chip/array more appropriate for African populations
- Develop a set of markers with optimal coverage of H3Africa populations
- Platform agnostic as far down the pipeline as possible
- Negotiate for good pricing based on bulk purchase
- Task force set up through through Genome Analysis WG, in collaboration with Wits, Sanger and many other experts
- Project will include 3935 human genomes from ~50 African populations from 18 countries sequenced at coverage from 4-40X

Workflow for data processing



H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Chip design computing requirements

- Data: ~4000 genomes of varying coverage
- Raw files for cold storage 110TB
- Processed files for analysis 220TB
- Additional analysis files 12-15TB
- Total cpu days required if done on a single core = 4600 days
- Therefore need ~230TB “fast storage” + 1000 cores to complete the project in Africa……

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

H3ABioNet and H3Africa ideals

- Build capacity to do research in Africa
- Ensure data is held in Africa
- Ensure analysis is done in Africa
- Ensure publications are driven from Africa!

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Global Alliance for Genomics & Health

- "International coalition dedicated to improving human health by maximizing the potential of genomic medicine through effective and responsible data sharing."
- Need access to genomic data world-wide to compare millions of human genome sequences and thus coordination to ensure this
- Working groups:
  - Clinical
  - Regulatory and ethics
  - Security
  - Data

Global Alliance for Genomics & Health

http://genomicsandhealth.org/

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT

G4GH Data Expert WG

- Data is kept in different places, each set is not enough on its own

GA4GH APIs being developed to allow sharing without moving data –metadata, coordination, collaboration are key!

How are these efforts enabling genomic research for health?

Challenges of health-related research

- Most datasets are not big enough to answer all questions on diseases
  - Global Alliance is enabling access to diverse data sets
  - H3Africa consortium is ensuring some common phenotypes for meta analysis, facilitated by ontologies
- Population diversity and admixture
  - H3ABioNet is developing tools for admixture analysis, studying background populations -> custom chip design
- Data processing and analysis
  - H3ABioNet is developing SOPs and facilitating access to compute facilities
- Data transfer and storage
  - H3ABioNet is developing infrastructure for long term storage and transfer mechanisms
- Lack of skills in health-related genomics research
  - BSP and H3ABioNet are providing training through courses, internships and collaborative projects

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Acknowledgements

CBIO group

- **STAFF**
  - Ayton Meinjies
  - Gerrit Botha
  - Sumir Panji
  - Suresh Malsamoney
  - Vicky Nembaware
- **POSTDOCS**
  - Richard Akinola
  - Emile Chimusa
- **MSC STUDENTS**
  - Jacquiline Mugo
  - Joel Defo
- **PHD STUDENTS**
  - Kenneth Opap
  - Holifidy Arisoa Rapanoel
  - Gustavo Salazar
  - Jon Ambler
  - Chacha Issarow
  - Twaha Mlwilo

Funding: NIH Common Fund

UCT ICTS: HPC, support

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

H3ABioNet Consortium

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT