Building Bioinformatics Capacity in Africa

Nicky Mulder
CBIO Group, UCT

H3ABioNet
Pan African Bioinformatics Network for H3Africa

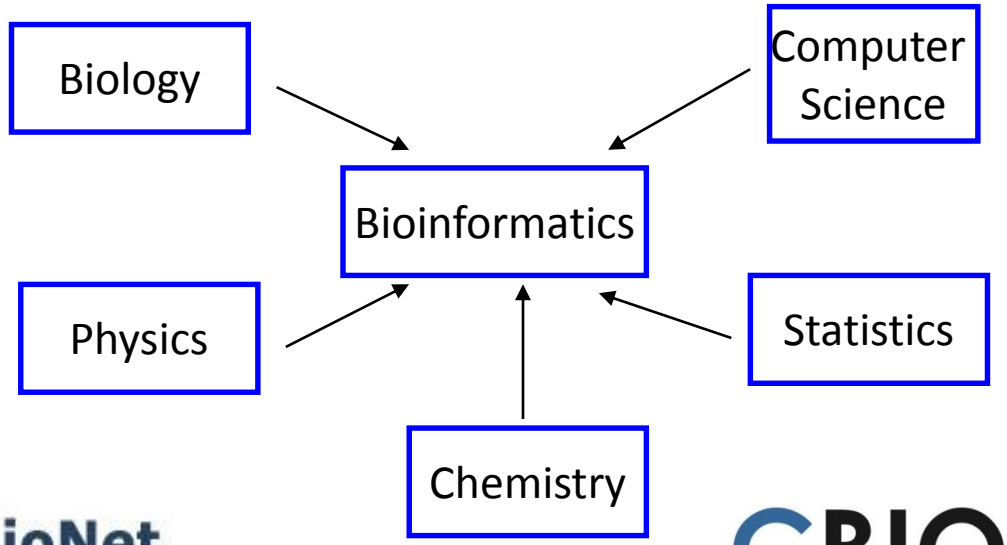
CBIO
Computational Biology @ UCT

Outline

- What is bioinformatics?
- Why do we need IT infrastructure?
- What e-infrastructure does it require?
- How we are developing this

What is Bioinformatics?

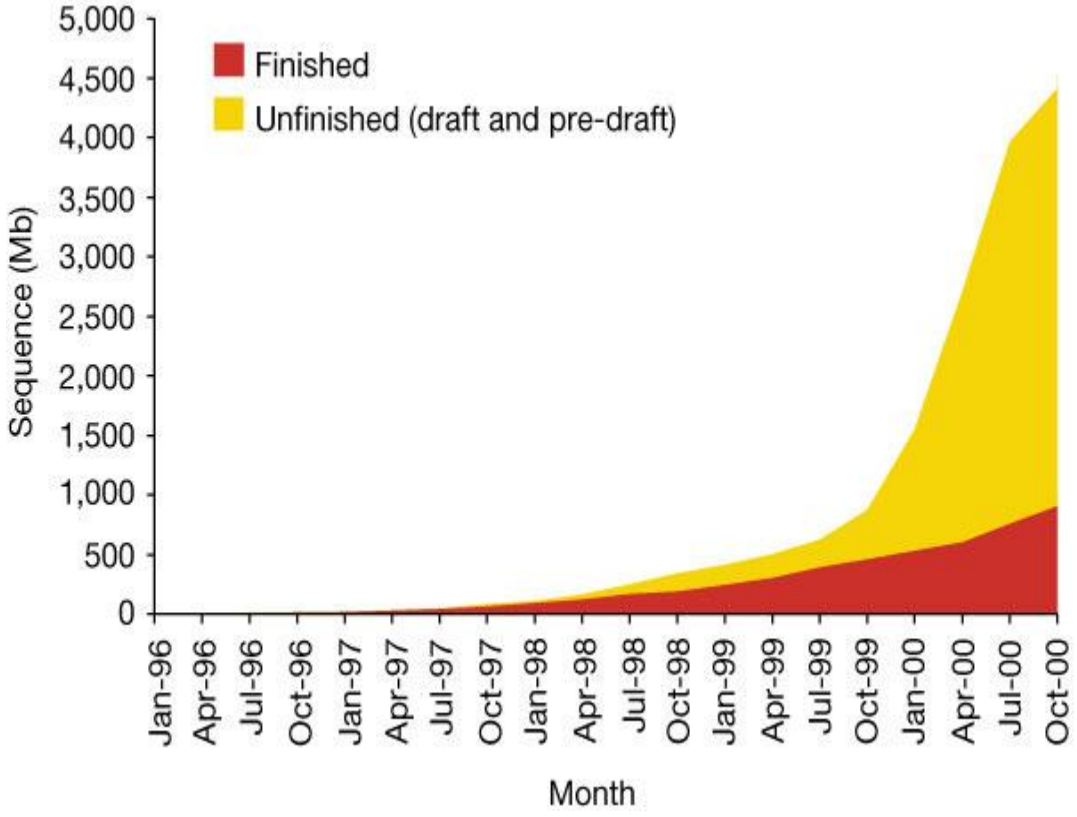
- The analysis of biological information using computers and statistical techniques; the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. www.niehs.nih.gov/nct/glossary.htm



Why do we need IT Infrastructure?

- Collection and storage of biological information
- Manipulation of biological information
- Small- and large-scale biological analyses
- New laboratory technologies, move towards big data generation
- Example is genome sequencing

Sequencing the human genome



International Human Genome Sequencing Consortium 2001. Nature 409, 860 – 921.

First human genome took ~5 years and cost ~$3 billion

Now, can sequence in a few weeks for ~$5,000

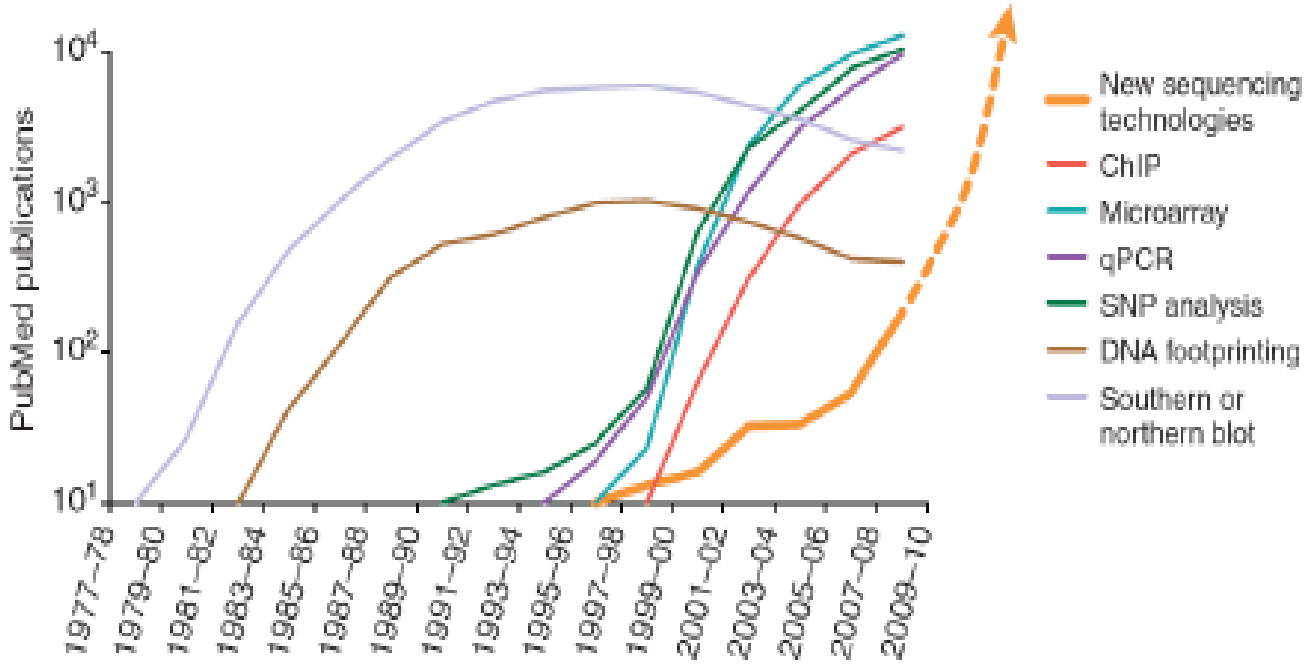
BUT: doesn't consider cost and time for data analysis!

human genome, sequenced at 30x coverage = ~ 1 billion raw reads of about 100bp = ~250Gb of raw data, when processed = x 10-15

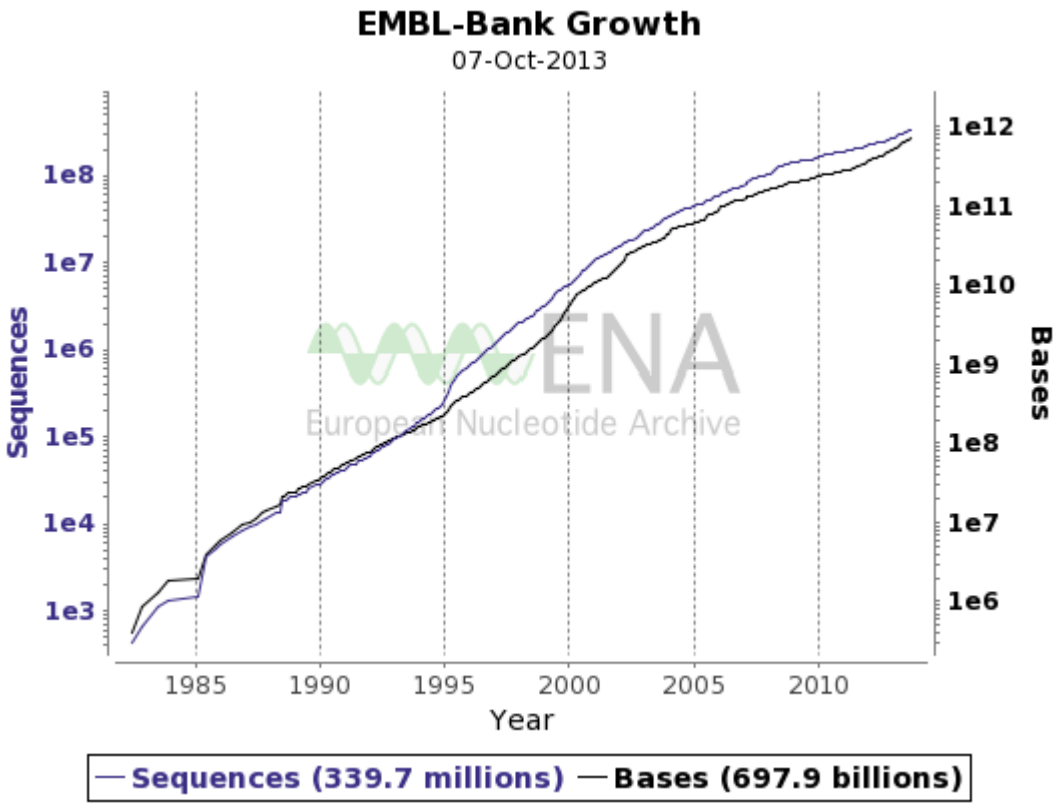
H3ABioNet
Pan African Bioinformatics Network for H3Africa

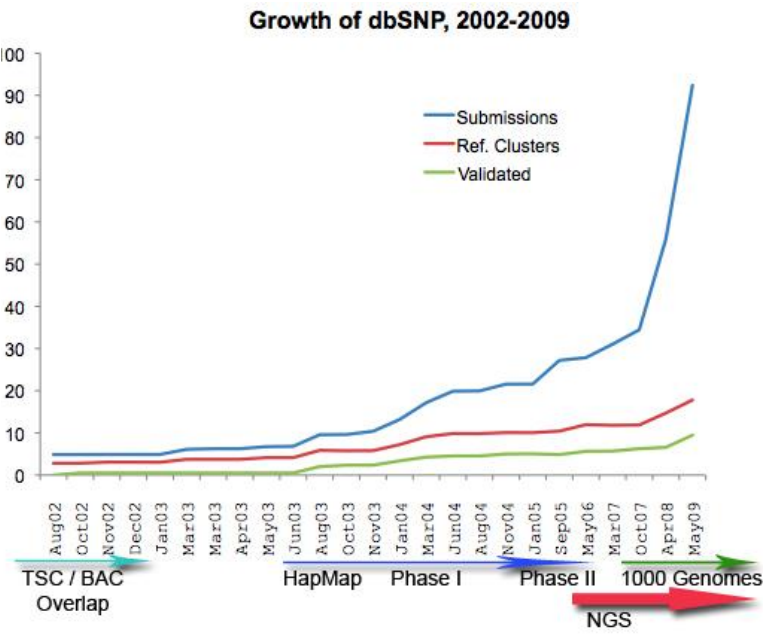
CBIO Computational Biology @ UCT

Technologies used based on publications



Increase in biological data





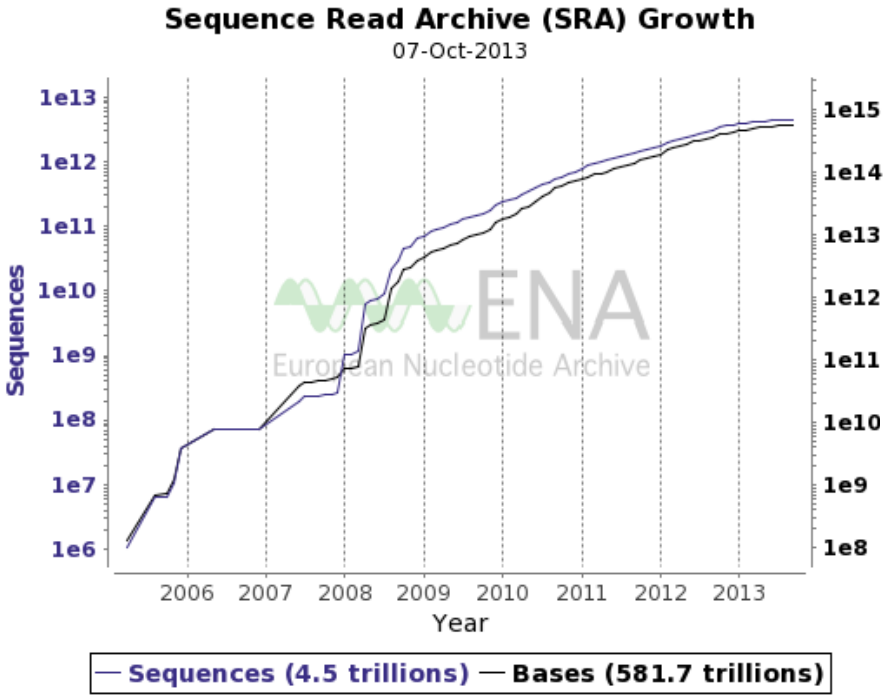
http://www.ebi.ac.uk/ena/about/statistics -Last release of ENA >330 mill seqs, 1.6TB

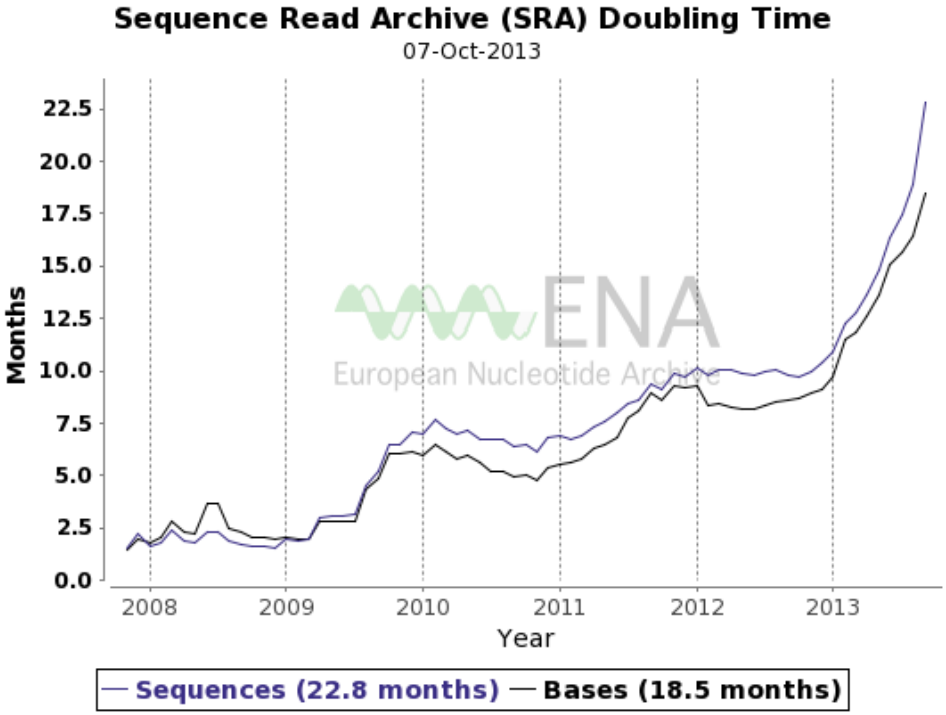
Biological data growth has exceeded Moore's law

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

New types of data



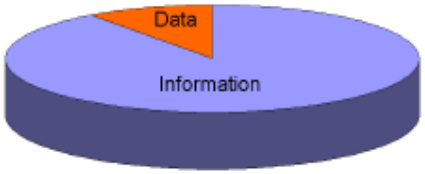


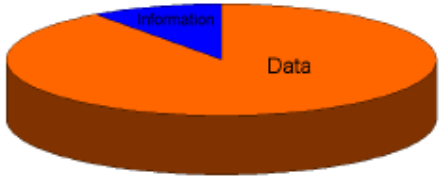
http://www.ebi.ac.uk/ena/about/statistics

H3ABioNet
Pan African Bioinformatics Network for H3Africa

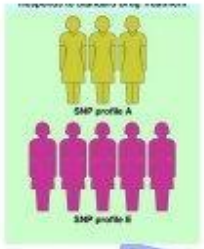
CBIO
Computational Biology @ UCT

Data versus information





Genes in the DNA...



…produces the final phenotype

…code for proteins...

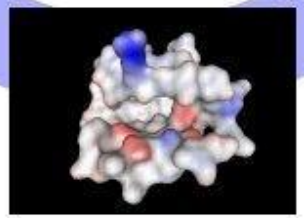
>protein kunase

acctgttgatggcgacagggactgt
atgctgatctatgctgatgcatgcatg
ctgactactgatgtgggggctattga
cttgatgtctatc....

From genotype to phenotype.

…whose structure accounts for function...

…plus the environment...



H3AB Pan African

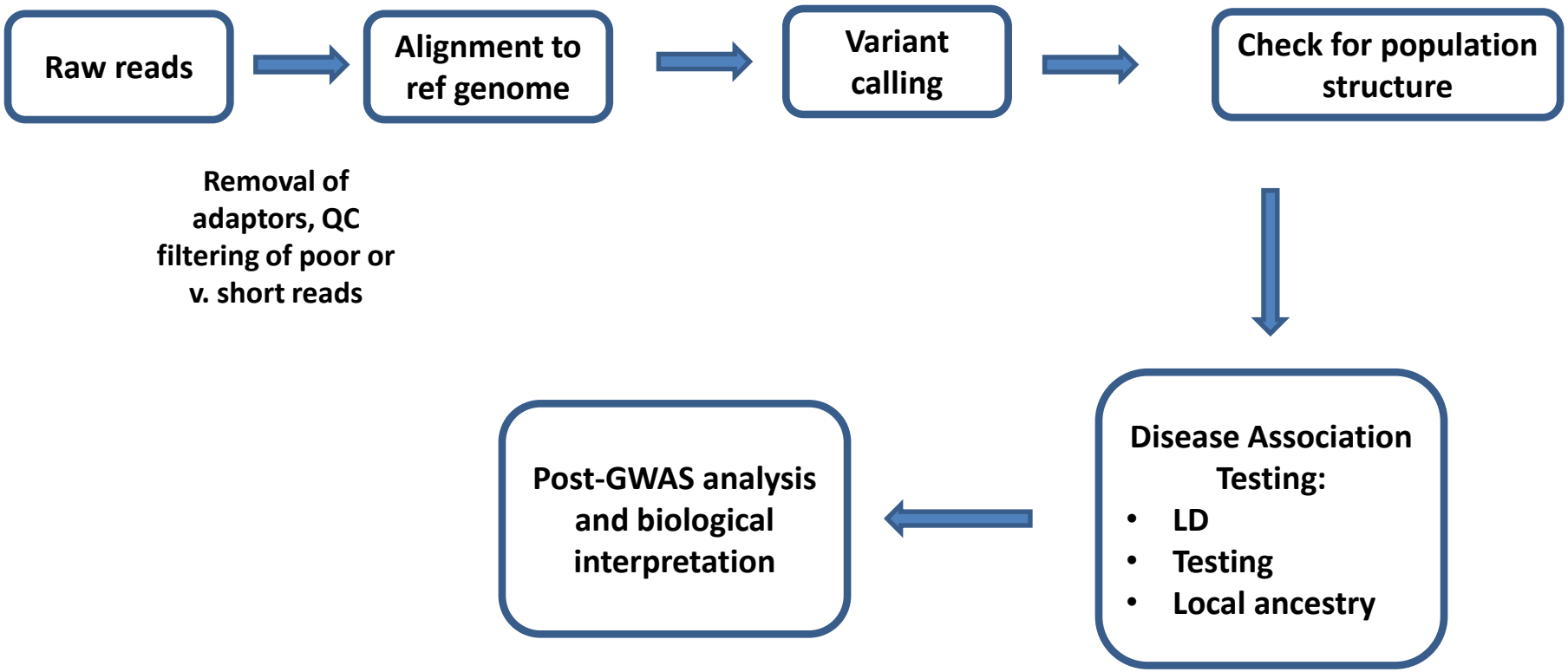
What infrastructure is required?

- Data
  - Storage and management –flatfile archive versus mineable database
  - Processing –single nodes, HPC, Cloud
  - Interpretation –analysis and visualization tools
- Learning how to work with it
  - IT/technical skills –developing tools and data management
  - Bioinformatics skills –using tools and interpreting results

Data requirements for genomes

- Raw files for NG sequencing are large and increase 5-10 fold after analysis
- Data needs to be stored and processed
- Raw data:
  - Storage requirements are large –some people are working on compression formats for NGS data
  - Data needs to be backed up
  - Also need access to public data
- Data needs to be processed, and made available in a user-friendly format
  - Need compute hardware for QC, alignment, imputation, and tool/visualization development

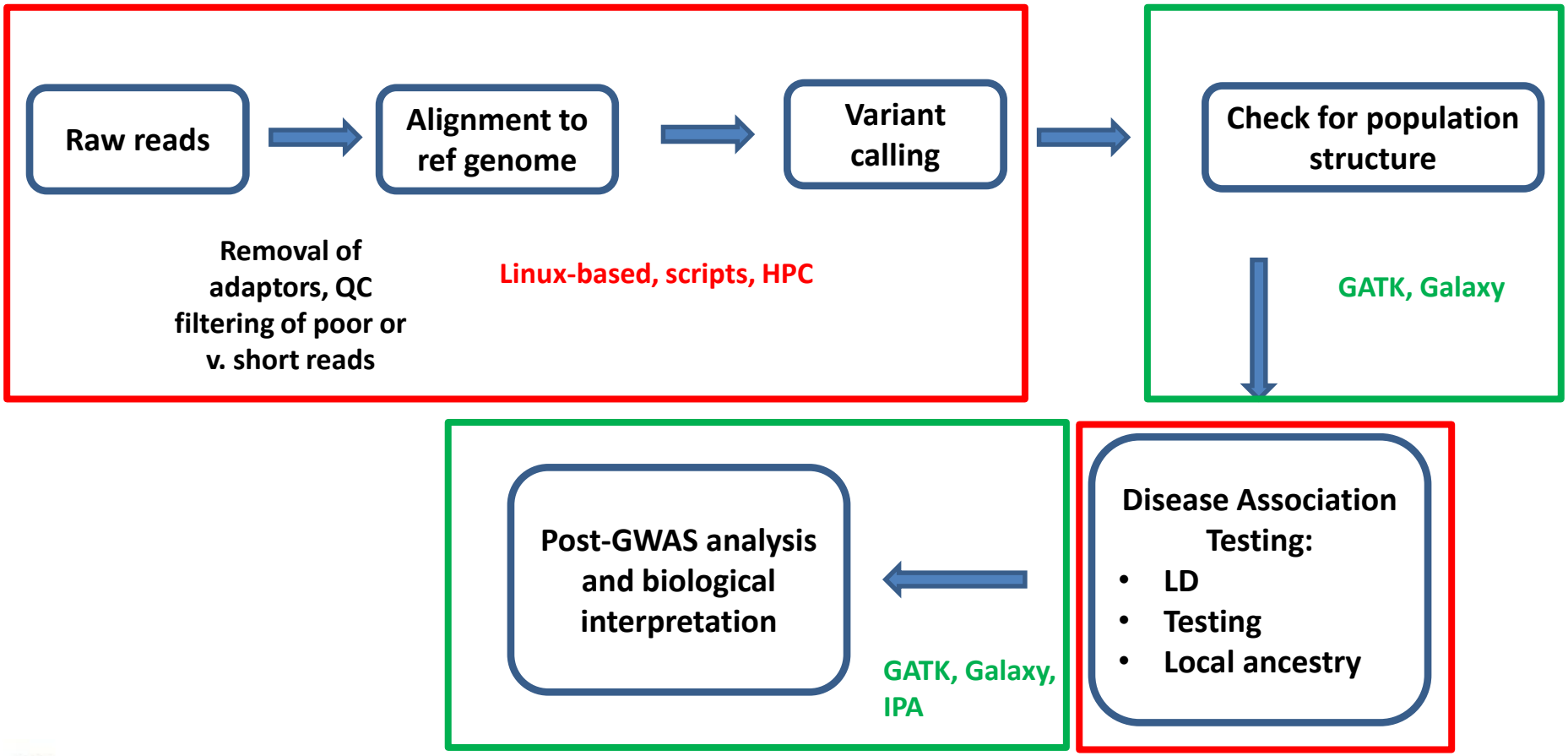
Workflow for genotyping by NGS



H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT

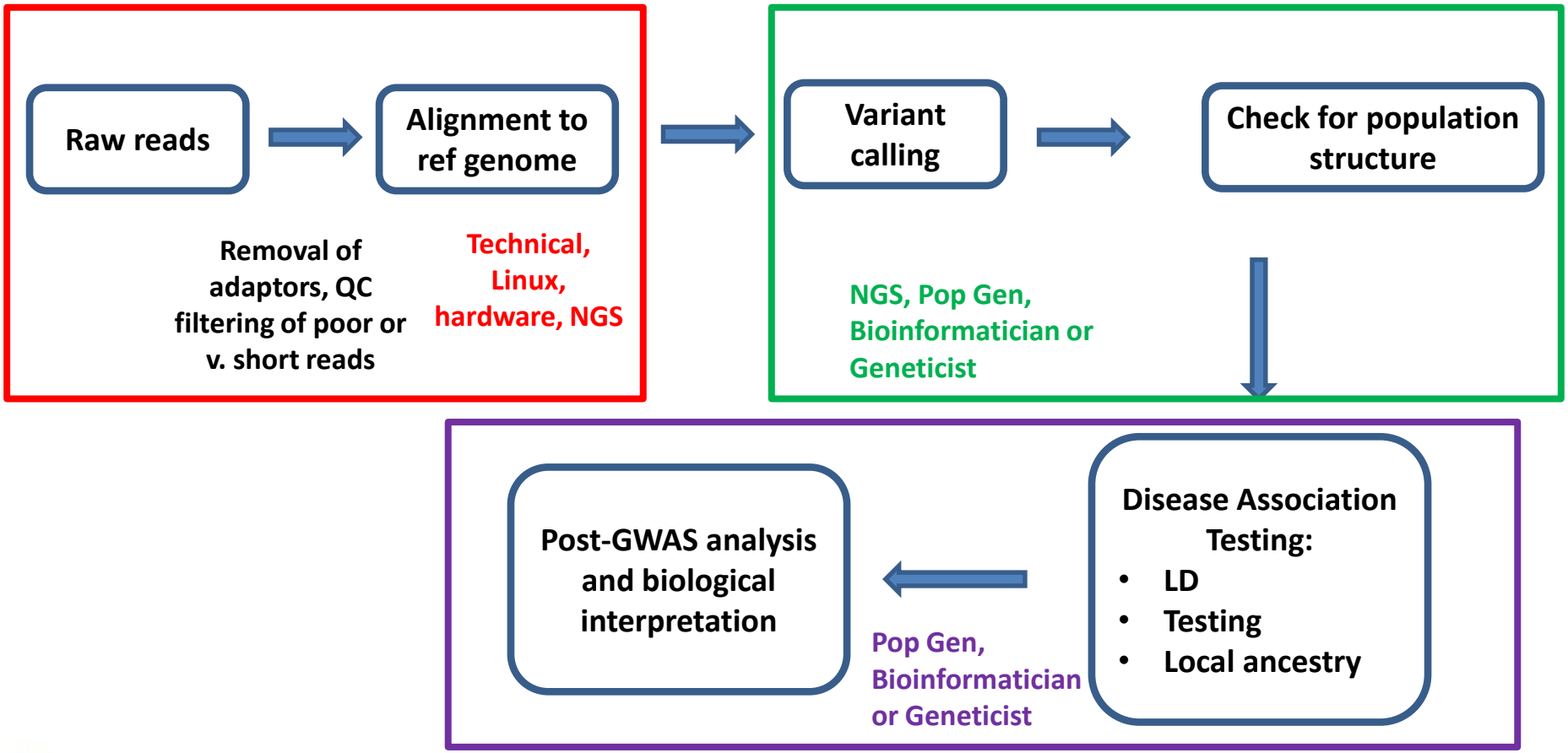
Workflow for genotyping by NGS -IT



H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT

Workflow for genotyping by NGS -HCD



H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT

Human Capacity Development

Training of 2 groups of scientists:

- Bioinformaticians
  - Programming, data management
  - Data processing
  - Biostatistics
  - Specialised skills (NGS, GWAS, PopGen)
- Researchers –data analysers
  - Basic file manipulation (Galaxy)
  - Biostatistics
  - Specialised skills (NGS, GWAS, PopGen)

How we are developing infrastructure

- Work on national & Africa-wide initiatives
- Local:
  - Support researchers at UCT with data analysis and training for this
  - Working with DST on a national Bioinformatics support platform
- Africa-wide –H3ABioNet

H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT

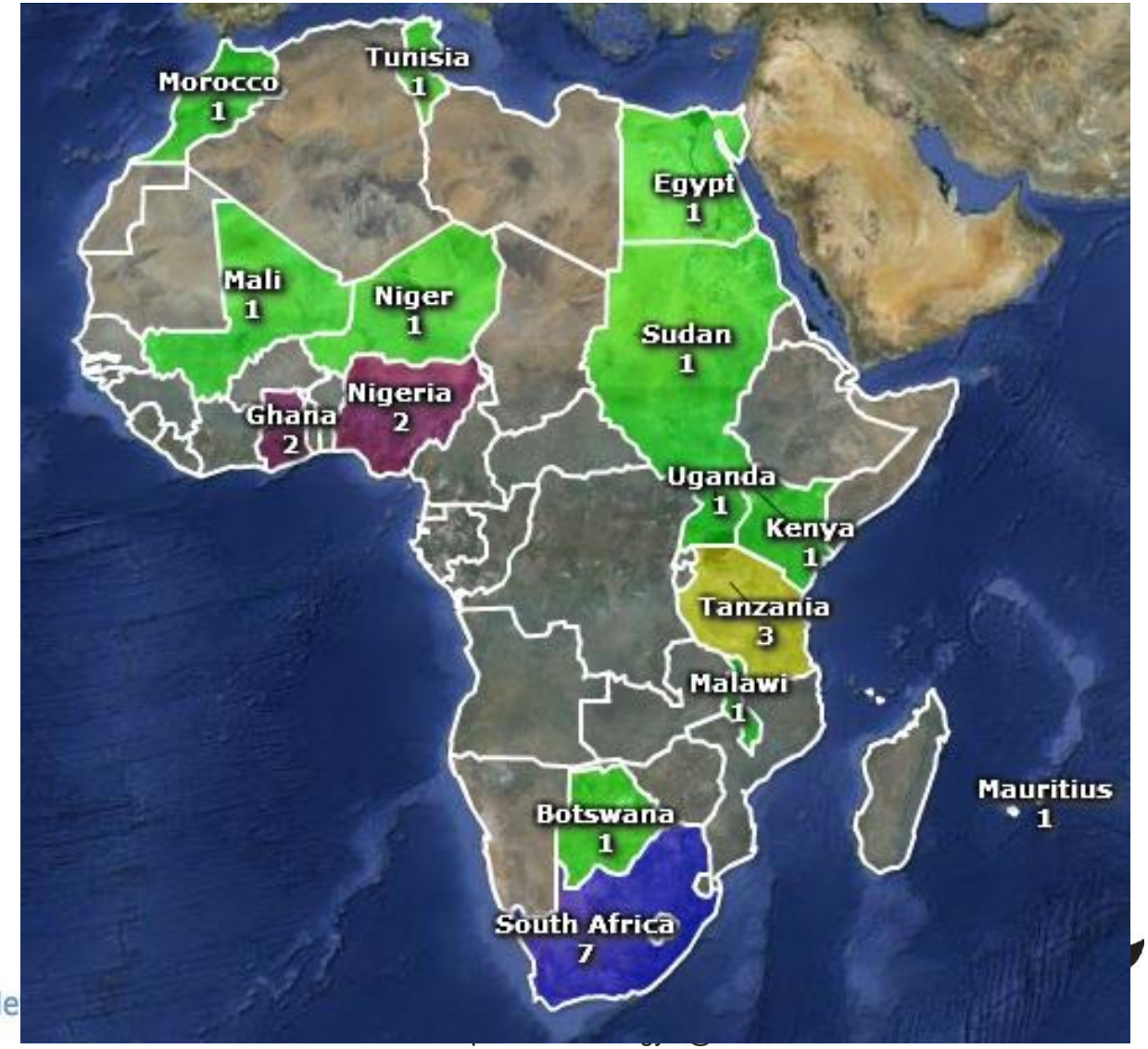
H3ABioNet Project Goal

- To build H3ABioNet, a sustainable African Bioinformatics Network, to provide bioinformatics infrastructure and support for the H3Africa consortium.
- H3Africa (Human Heredity and Health in Africa) is an initiative to develop genomics research in Africa on diseases of importance in Africa

Partner institutions

Administrative hub at UCT

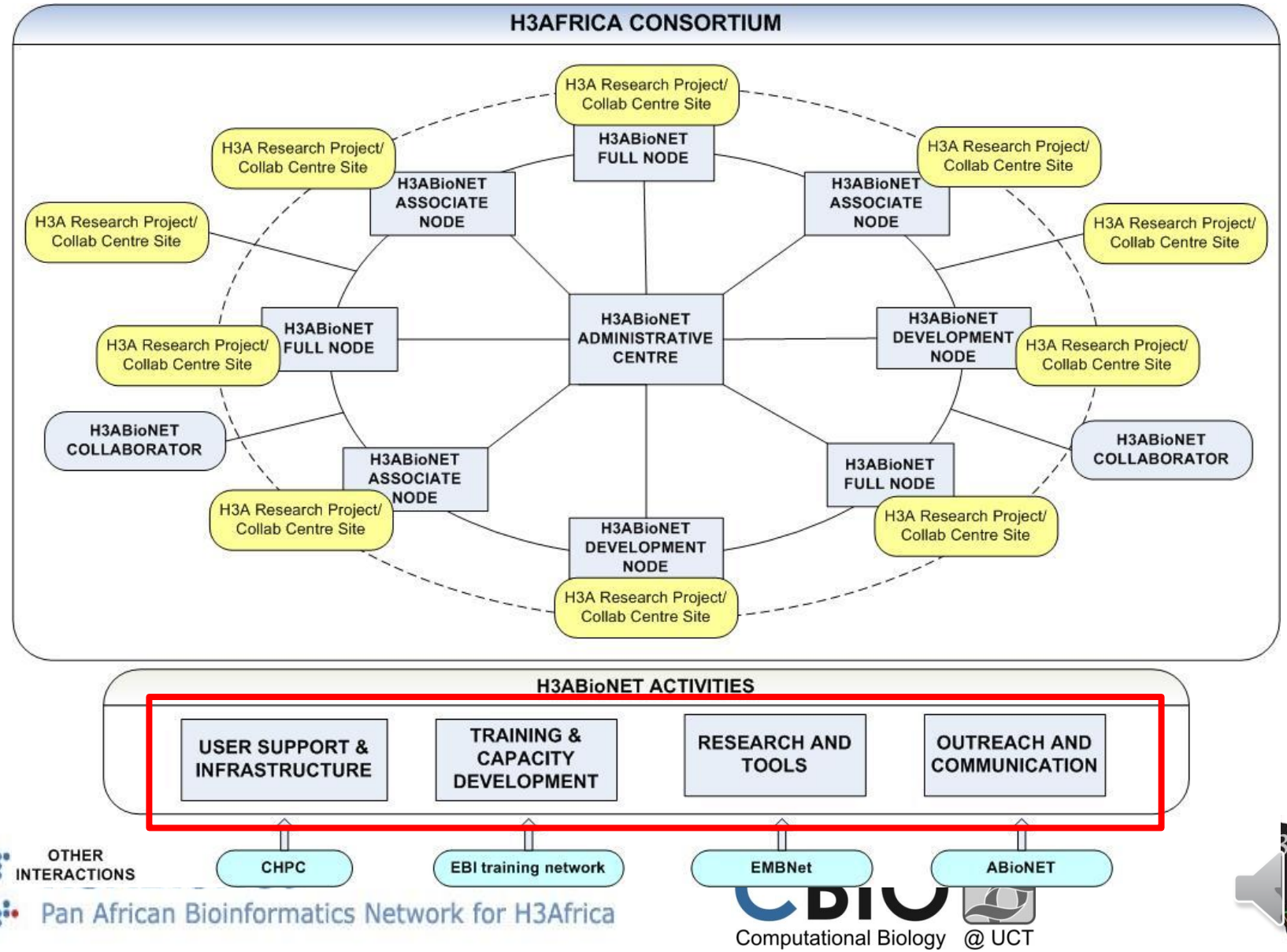
34 partner institutions, 32 in 15 African countries, 2 in USA



H3ABioNet

Pan African Bioinformatics Ne

Network structure



Pan African Bioinformatics Network for H3Africa

Computational Biology @ UCT

Infrastructure development & support

- Node server purchases
- Investigating access to HPC, Cloud
- Internet connectivity measurement
- Communication structure established: website, surveys and mailing list setup
- Set up help desk

Hardware Infrastructure development

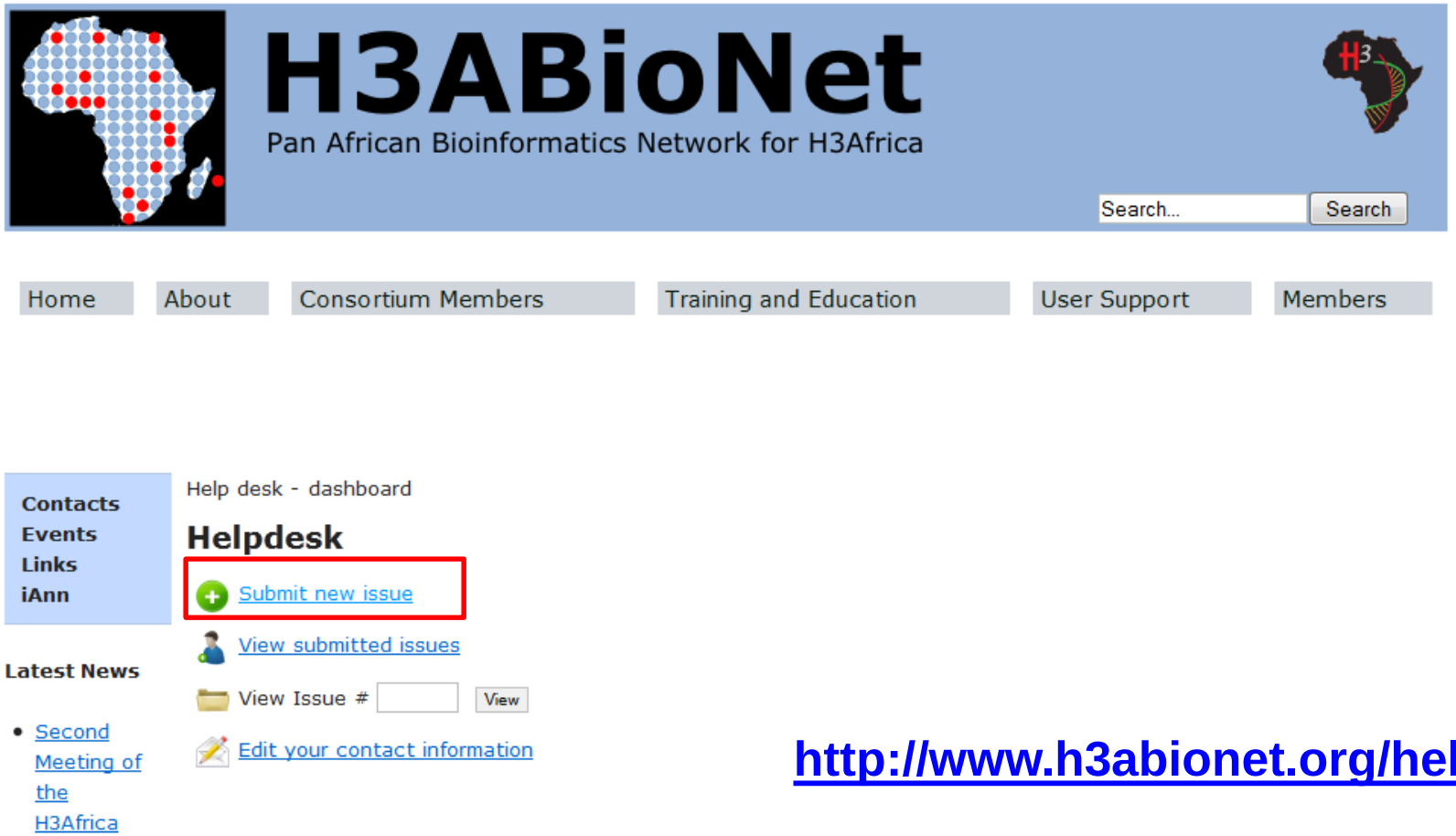
- Process followed:
  - 34 nodes were surveyed via email to determine existing infrastructure and expected needs
  - 18 nodes requested funding to purchase infrastructure
  - One on one skype follow-up calls with nodes who requested assistance
    - 12 nodes requested assistance
      - 10 bought the recommended hardware
      - 1 bought HP based on recommended hardware
      - 1 bought own configuration

Hardware Infrastructure development

- 3 server builds developed based on survey feedback
  - Option A – Recommended for nodes with an existing infrastructure (single server with max cores and high RAM)
  - Option B – Recommended for nodes needing two physical servers (smaller HPC server and one database server)
  - Option C – Recommended for nodes with no infrastructure (Smaller server but included a rack, switch and UPS)
- Hardware Vendor
  - Approached Dell, IBM and HP
    - Dell was the preferred provider as they have a presence across Africa
    - Provided a 5 year next business day warranty / support to nodes
    - Offered best value for money
    - Negotiated a discount of up to 45% per hardware device

Iperf tests

H3ABioNet help desk



http://www.h3abionet.org/helpdesk

H3ABioNet help desk

Home | About | Consortium Men

Contacts
Events
Links
iAnn

Latest News

- Second Meeting of the H3Africa Consortium, Accra Ghana

Help desk - dashboard

Helpdesk

Submit new issue

View submitted issues

View Issue #

Edit your contact inform

Help desk - dashboard

Helpdesk

New Ticket

Submit | Print | Close | Refresh | Home

Contact Information

User Name:	nicky.mulder *
E-Mail:	nicola.mulder@uct.ac.za *
Department:	Affolabi - RAFAgene *
Location:	
Phone:	

Classification

Category:	Select Category
Status:	
Priority:	
Assigned To:	
Time Spent:	(minutes)

TicketInformation

Title:

Description

B I U

Notes

Enter Additional Notes

B I U

Solution

Submit | Print | Close | Refresh | Home

Search... Search

Helpdesk user guide

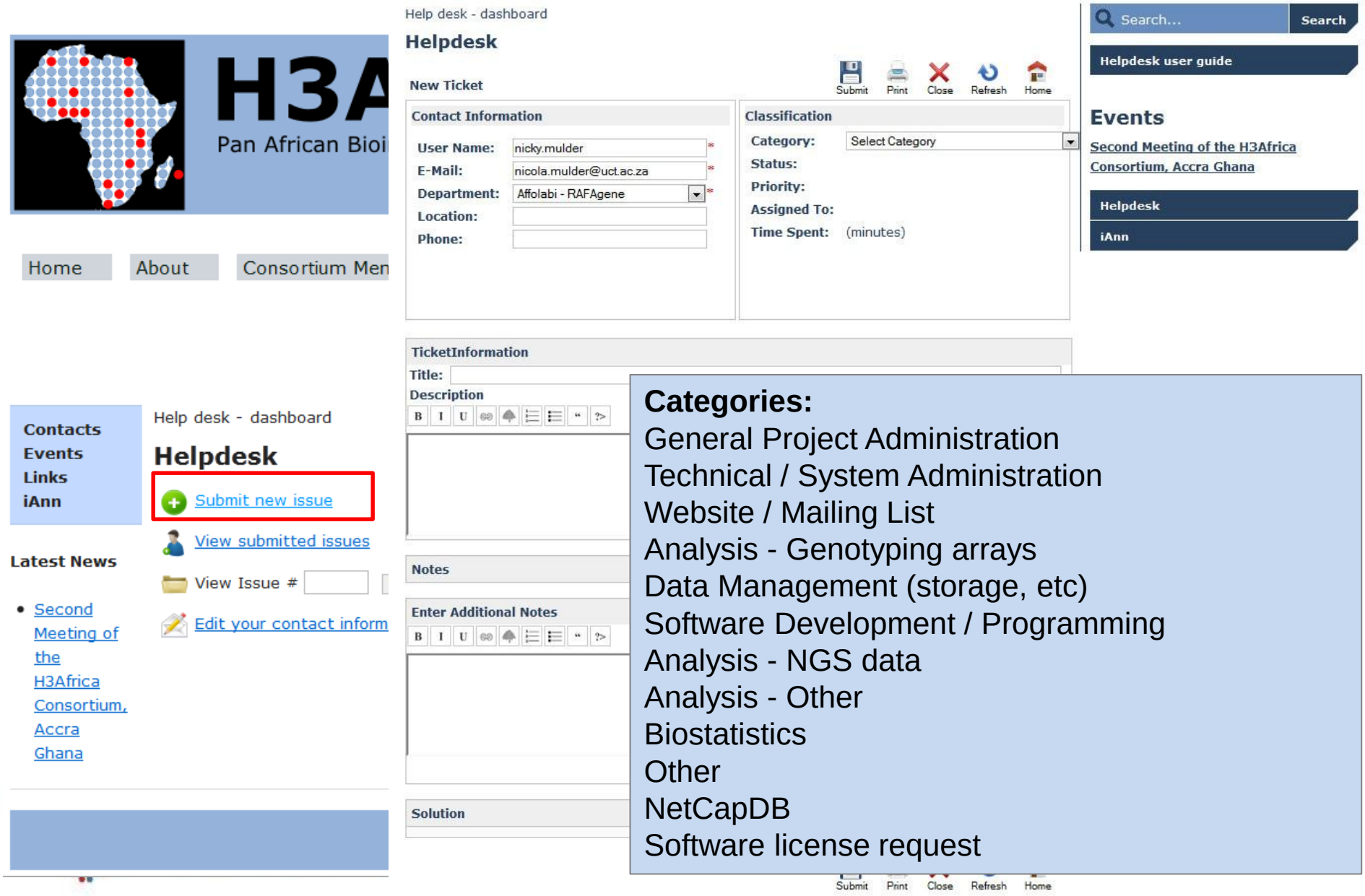
Events

Second Meeting of the H3Africa Consortium, Accra Ghana

Helpdesk

iAnn

H3ABioNet help desk

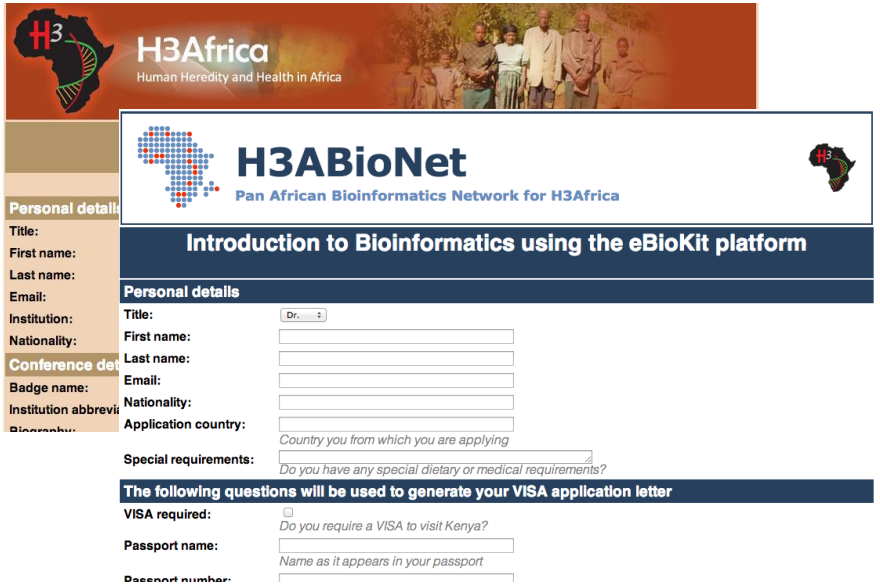


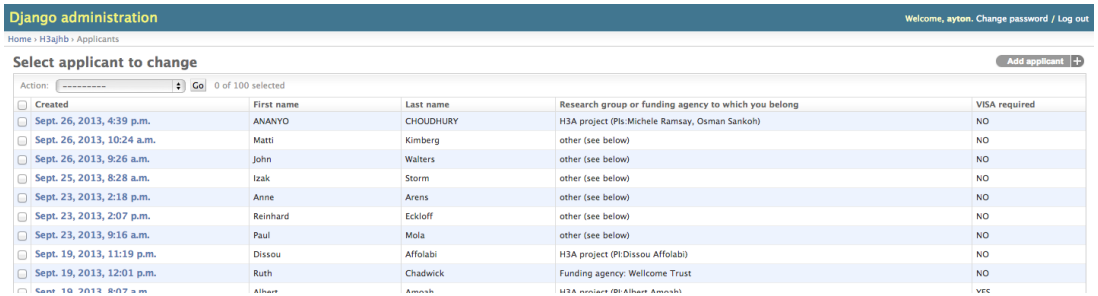
Categories:

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

Event registration

- Field validation
- Email notifications
- Admin backend
- Excel export
- 10 events, 461 registrations



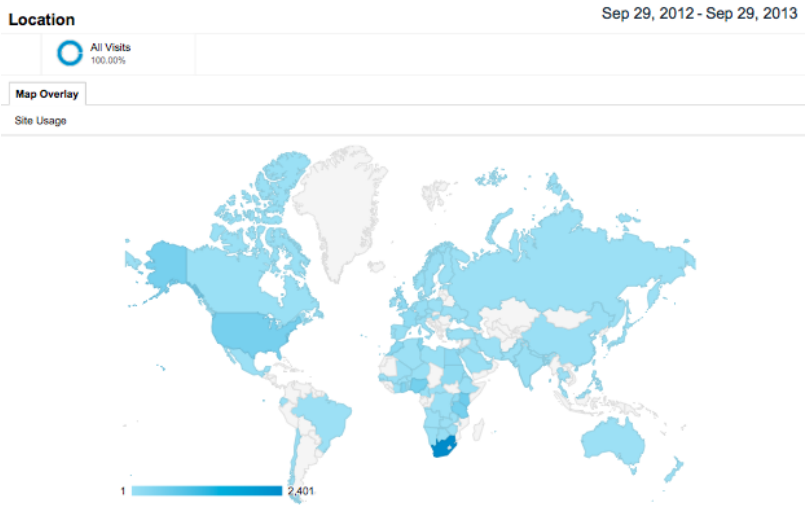


H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Website monitoring

- Google analytics
  - traffic flow
  - traffic sources
- Uptime monitored with pingdom
- Amazon failover server (manual switch)



Country / Territory	Visits	Pages / Visit	Avg. Visit Duration	% New Visits	Bounce Rate
	7,044 % of Total: 100.00% (7,044)	3.67 Site Avg: 3.67 (0.00%)	00:04:10 Site Avg: 00:04:10 (0.00%)	34.64% Site Avg: 29.40% (17.82%)	52.81% Site Avg: 52.81% (0.00%)
1. South Africa	2,401	6.08	00:07:08	18.70%	35.61%
2. Kenya	572	2.13	00:02:16	46.50%	61.19%
3. Nigeria	477	2.26	00:03:26	43.19%	57.65%
4. Tanzania	463	1.91	00:02:23	16.20%	73.87%
5. United States	458	3.16	00:02:21	59.83%	49.78%
6. (not set)	428	2.72	00:03:00	25.47%	70.09%
7. Ghana	311	2.29	00:03:08	26.37%	67.85%
8. Tunisia	259	2.31	00:02:33	47.88%	61.78%
9. Morocco	210	3.58	00:03:21	55.24%	50.48%
10. Egypt	165	2.42	00:02:37	49.70%	50.91%

Rows 1 - 10 of 85

© 2013 Google

H3ABioNet
Pan African Bioinformatics Network for H3Africa

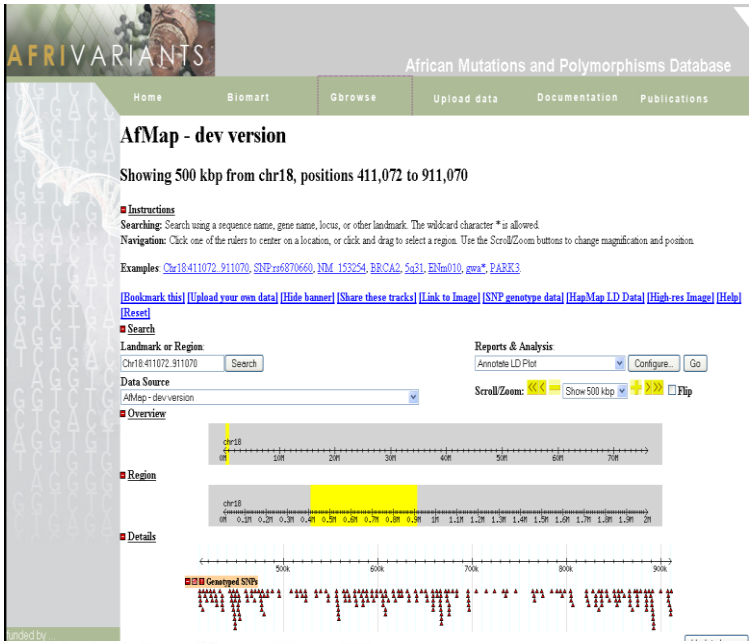
CBIO Computational Biology @ UCT

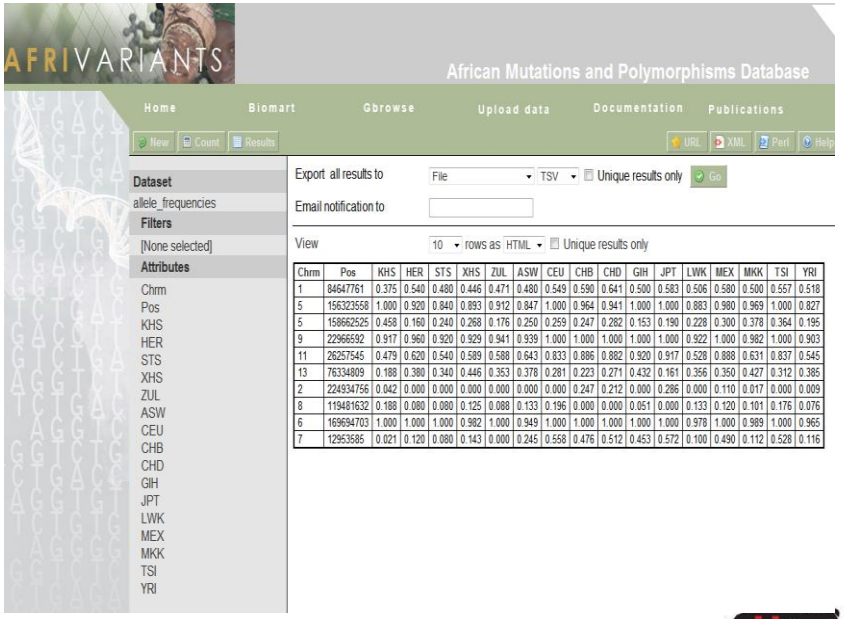
Research and tool development

- Data management and storage
  - Data integration platforms, e.g. BioMart
- Data analysis tools
  - Developing pipelines in Galaxy
  - Further developing eBioKits
  - Visualization tools

BioMart database

- Converting relevant data from flatfiles into mineable format
- Includes user-friendly GUI

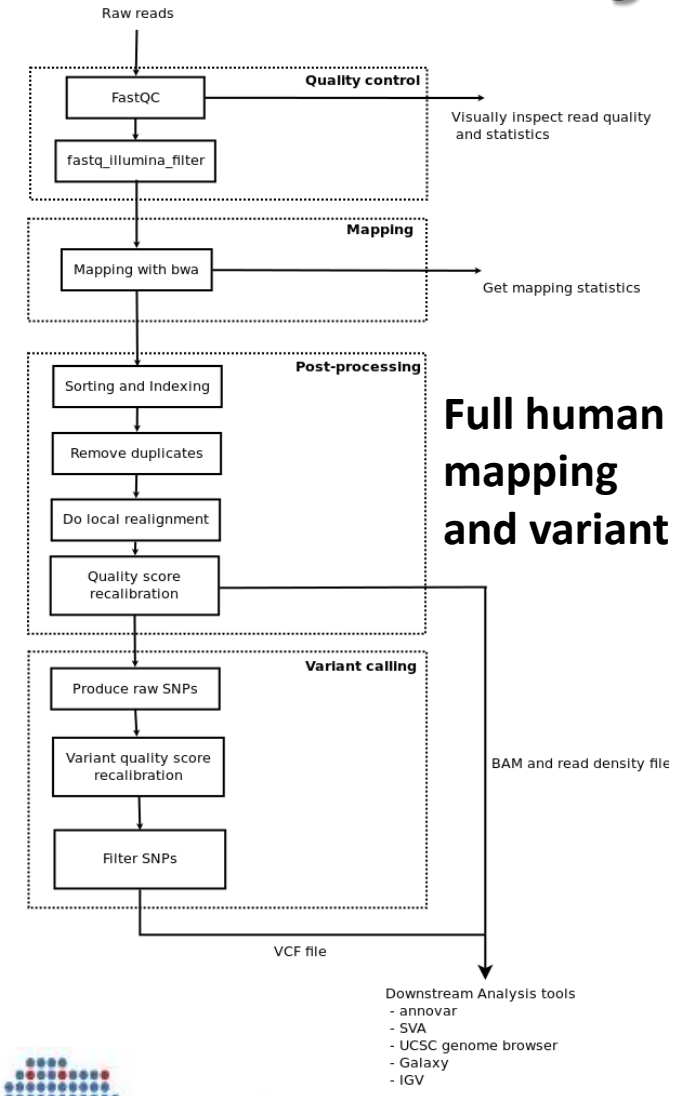




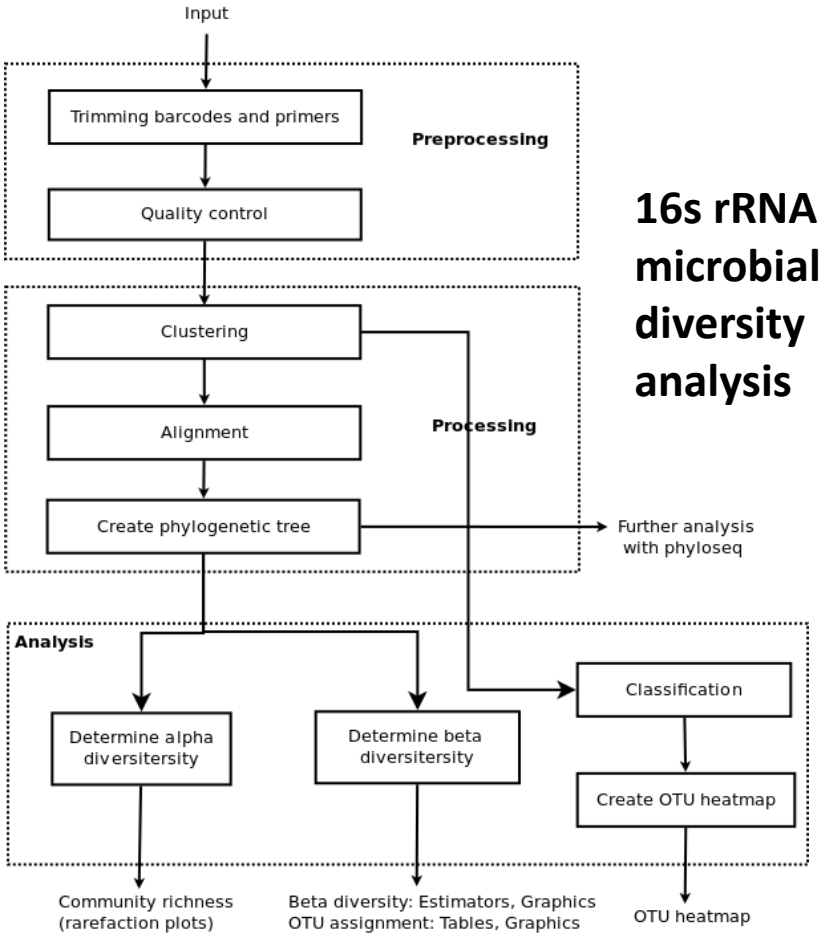
H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO
Computational Biology @ UCT

Analysis workflows



Full human genome mapping and variant analysis

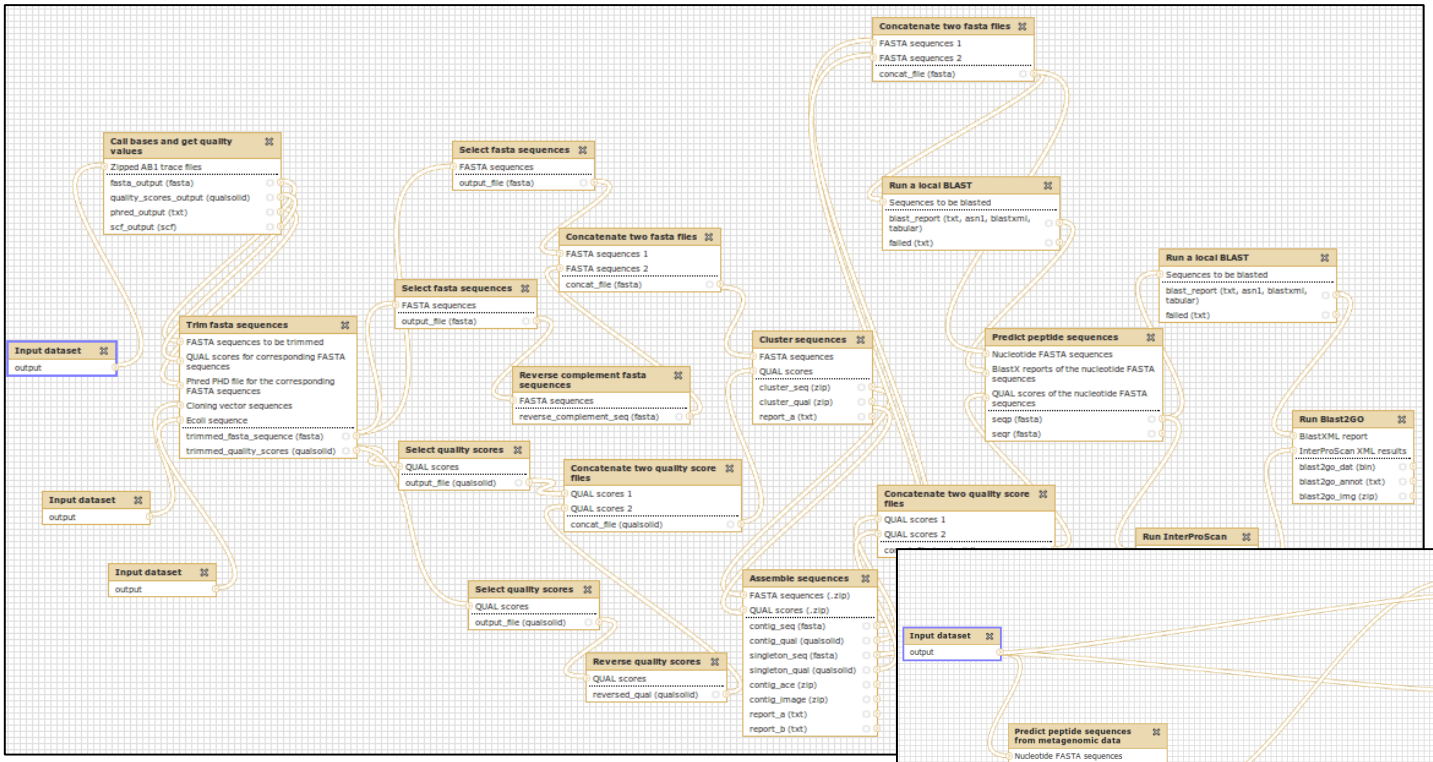


16s rRNA microbial diversity analysis

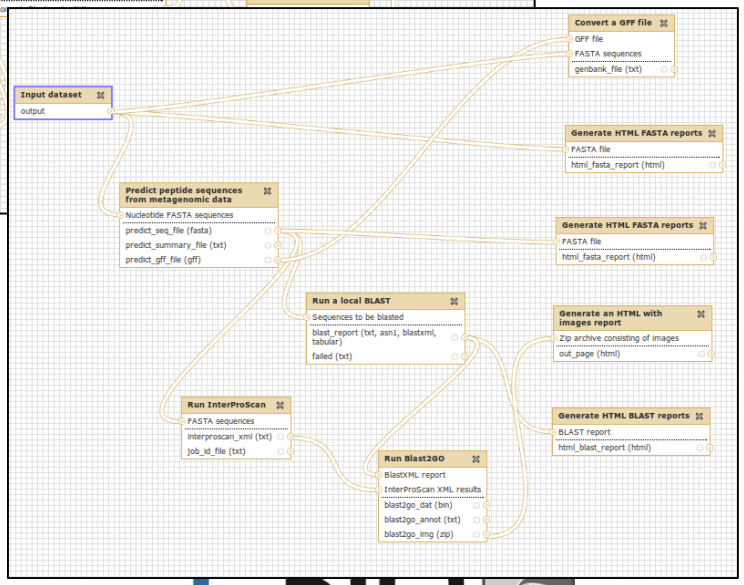
H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT

Galaxy workflows



EST assembly and annotation pipeline



Metagenomics functional annotation pipeline

H3ABioNet
Pan African Bioinformatics Network for H3Africa

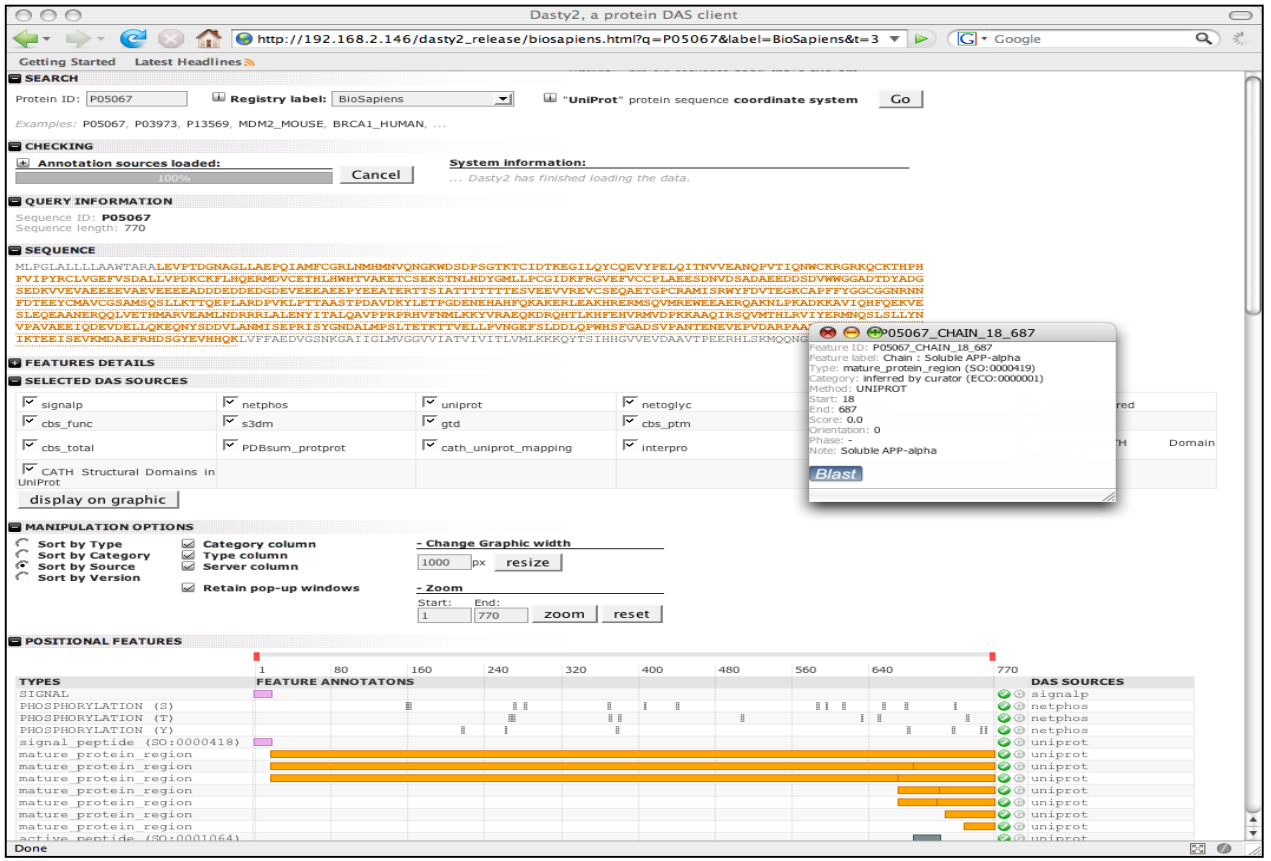
Computational Biology @ UCT

eBiokits

- Standalone hardware without the need for internet
- Includes databases, tools, tutorials
- Based on Mac hard drives

Visualization tools

- DAS-based tools –avoid integrating data locally

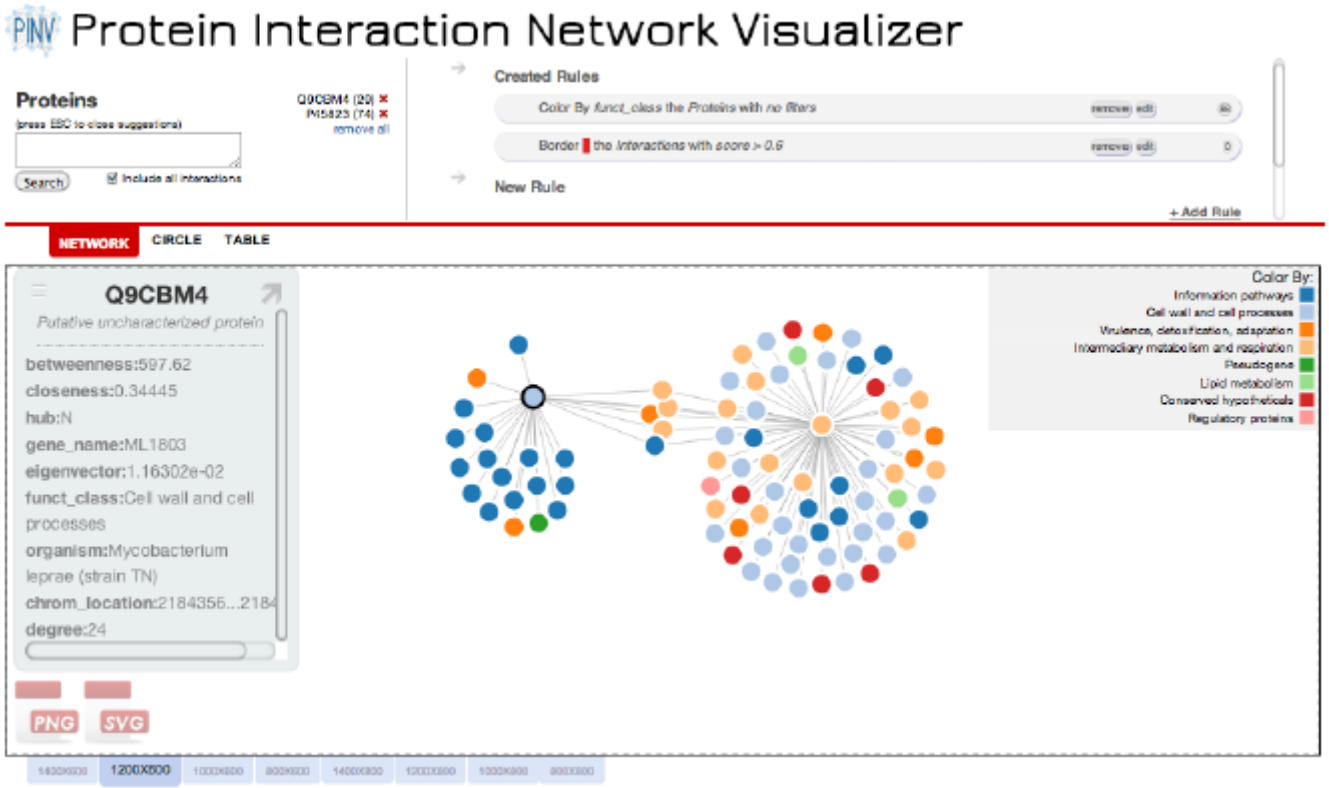


H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT

Visualization tools

- Interactive web-based network visualization tool



H3ABioNet
Pan African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT

H3ABioNet training plan

- Short specialised courses
- Research driven workshops
- Regional workshops
- Career development workshops
- Internships, KTP
- Train the trainer program
- Mentorship
- Postgrad degree curriculum development
- Facility for live-streaming to other classrooms

Training activities

- Organized training courses:
  - Technical course (Linux, Cloud, data security, HPC)
  - Train-the-trainer course (Biostats, GWAS, NGS)
  - eBiokit course (NGS)
- Courses live streamed to 2 class rooms
- Set up online application system
- Set up online evaluation system
- Set up training course website
- Developed training course planning document

Summary

- Bioinformatics is required to convert data into information
- As biology goes more high-throughput so the need for mathematics, statistics and computing increases
- As well as ensuring adequate IT infrastructure is in place, we need to train people at different levels on the use of it
- If we can build these in Africa we will stop the movement of data off the continent!