H3ABioNet

Pan African Bioinformatics Network for H3Africa
H3ABioNet Project Goal

• To build H3ABioNet -- a sustainable African Bioinformatics Network -- to provide bioinformatics infrastructure and support for the H3Africa consortium.
Partner institutions

Administrative hub at UCT

34 partner institutions, 32 in 15 African countries, 2 in USA
Partner institutions - projects

PI sites, other sites around Africa too
H3Africa’s needs

- Large-scale data analysis for:
  - Genotyping by arrays
  - Next generation sequencing
  - GWAS

- Support for analysis:
  - General questions
  - Access to computing resources
  - Technical computing support

- Data access and visualization (public and new)
- Data storage, backup and transfer
- Data submission
- Training
H3Africa data expected (1)

Types of Data To Be Generated H3Africa Projects

Percentage of H3Africa Projects (N=9)

SNP Arrays
NGS Data (Exome, RNA Seq, Full Genome)
Other (Biorepository)

Types of Data Generated

Types of Sequence Data Generated by H3Africa Projects

Exome
Full genome
RNA-Seq
Other (Metagenomics)

Sequencing Platform Used

Percentage of H3Africa Projects Using a Sequencing Platform (N=9)

Illumina
454
Other

2nd H3Africa Consortium Meeting, Accra
H3Africa data expected (2)

- **GWAS arrays:**
  - 30Mb/cell file (per sample), 150Gb per 96 well plate processed.
  - Samples planned ~33,380
  - Storage required for processed data: **52TB**

- **Sequencing NGS:**
  - Exome = illumina, ~45 depth, 90bp paired end, 17GB raw fastq,
    Full = Illumina, ~33 depth, 101 bp paired end, 250GB raw fastq
    (per sample)
  - Samples planned in total 12, 580 (exome 11,330, full genome 1,250)
  - Total storage for raw reads: 505TB, processed **2.5PB**
Network activities

H3Africa Projects
GENERALIZED PROJECT WORKFLOW

PATIENT SAMPLE COLLECTION CLINICAL DATA
Phenotyping

GWAS, MICROARRAYS, PROTEOMICS, SEQUENCING
Genotyping, Functional Genomics

ASSOCIATIONS, DATA ANALYSIS AND INTERPRETATION
Biological interpretation

PUBLICATION, DATA SUBMISSION

H3ABioNet
BIOMETRICS REQUIREMENTS

EXPERIMENTAL DESIGN
Patient database

LIMITS, DATA STORAGE AND MANAGEMENT

ANALYSIS TOOLS AND PIPELINES

DATA SUBMISSION TOOLS

INFRASTRUCTURE REQUIREMENTS

R1. Protocols & standards for databases
U1. Data transfer mechanisms
U2. Data storage facilities, data management protocols
R2. Tool development, pipeline development
U3. Data analysis support, access to public databases
U4. Support and tools for submission

HARDWARE REQUIREMENTS

DATABASE SERVERS
Large databases, GRID storage, portable hard drives, internet access
Development and analysis computing resources (local and Cloud). Internet access, eBioKits

TRAINING/CAPACITY BUILDING REQUIREMENTS
T1, T2. Biostatistics, Patient DB
T1, T2. Data management, Cloud computing, HPC
T1, T2. Data analysis (GWAS, NGS, arrays), Biostatistics, Galaxy
T1. Trained bioinformaticians: postgrad positions, mid career development

H3ABioNet
Pan African Bioinformatics Network for H3Africa

2nd H3Africa Consortium Meeting, Accra
Major activities

- **Infrastructure Development**
  - Develop the hardware and software infrastructure capacity of African Bioinformatics laboratories in preparation for genomic scale studies.

- **User Support**
  - Provide user support to H3Africa researchers and Bioinformatics capacity development through a pool of specialists and supported tools & resources.

- **Node Assessment and Accreditation**
  - Provide an accreditation programme to help prepare and develop African Bioinformatics’ laboratories in line with current international standards.

- **Research and Tool Development**
  - Provide research and development of Bioinformatics and genomics tools, standardised workflows and pipelines for H3Africa research projects.

- **Education and Training**
  - Develop Bioinformatics capacity of African scientists through co-ordinated education and training programmes.
Infrastructure development

- Technical support – programming, pipeline development, sys admin
- Access to HPC, Cloud, eBioKits
- Access to public data
- Standard operating procedures and guidelines for data analysis
- Tools/support for data submission
- Storage, movement & management of data
Infrastructure activities and plans (1)

- Working group met 6 times
- Developed specifications of types of machines for different types of Bioinformatics workflows
- Negotiations for purchasing of ~ 20 high end servers capable of doing NGS analysis for various nodes
- One-on-one consultations with development and associate nodes for advice on hardware
- Purchasing of e-BioKits for training and processing at nodes
Infrastructure activities and plans (2)

• Conducting Iperf (internet bandwidth tests) between servers from all H3ABioNet nodes to determine:
  – True internet bandwidth across Africa
  – Best place to host mirror sites
  – How to move data around

• Researching data storage and processing options:
  – Getting Galaxy running at CHPC
Infrastructure activities and plans (2)
Infrastructure activities and plans (2)

• Conducting Iperf (internet bandwidth tests) between servers from all H3ABioNet nodes to determine:
  – True internet bandwidth across Africa
  – Best place to host mirror sites
  – How to move data around

• Researching data storage and processing options:
  – Getting Galaxy running at CHPC
User support

- H3Africa research projects provide support on:
  - Small bioinformatics queries (scripts etc)
  - Data management
  - Large-scale analysis
  - Access to computing infrastructure (data & hardware)
  - Submission to EGA

- Support provided through:
  - Helpdesk
  - Training
H3ABioNet help desk

http://www.h3abionet.org/helpdesk
H3ABioNet help desk

Help desk - dashboard

Helpdesk

New Ticket

Contact Information

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

Classification

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

Ticket Information

Title: 
Description: 

Notes

Enter Additional Notes: 

Solution

Search...
Helpdesk user guide

Events

Second Meeting of the H3Africa Consortium, Accra Ghana

Helpdesk
iAnn
H3ABioNet help desk

Categories:
- Analysis - Genotyping arrays
- Analysis - NGS data
- Analysis - Other (Not arrays or sequence data)
- Biostatistics, Data Management (storage, etc)
- Software Development/Technical/System Admin
- Website / Mailing List
- General Project Administration.
Help desk continued

- Help desk system has been tested
- Created documentation & technical specs
- Created pool of 20 people to answer queries

Personnel Contributing to H3ABioNet Helpdesk Categories

- Data Management / Storage: 43%
- General Project Management: 19%
- Software Development / Programming: 19%
- Technical Systems Administration: 9%
- Website / Mailing List: 5%
- Other: 5%
Node assessment & accreditation

- Assessing internet capacity at each node
- Node assessment exercises
- Set of workflows with simulated data covering:
  - Analysis of NGS data (exome and full genome)
  - Variant calling
  - Genotyping array data analysis
  - GWAS analysis
Node assessment & accreditation

• Accreditation board established with international experts:
  – Brad Chapman, Harvard University School of Public Health, USA
  – Noah Zaitlen, University of California San Francisco, USA
  – Fran Lewitter and George Bell, Whitehead Institute, USA

• Getting datasets together (simulated)

• Nodes can choose when they are ready for accreditation
Research and tools

- Working group met 5 times
- Aim to facilitate collaborative projects
- Working group has established project and collaboration documents
- Nine collaborative projects set up so far
- Some overlap with the User Support and Infrastructure working groups
Research and tools

• Data management and storage
  – Patient databasing protocols
  – **BioMart for genotyping data**
  – Grid-based tool for data storage

• Data analysis tools
  – **Galaxy for NGS and Genotyping analysis**
  – Genome assembly pipeline
  – Functional SNP calling pipeline
  – **Admixture mapping and network tools**
  – Data visualization tools
  – Recombination tools
  – Structural SNP analysis tool

• Joint research projects
BioMart for genotyping data

- BioMart – suite of tools for integration of data in RDBMS
- Widely use for public databases, e.g. Ensembl, InterPro
- Has web front-end, BioMart library for R or import into Galaxy
- Have developed BioMart for local genotype data and interface for searching
- Can develop BioMart for public data
- Can query across BioMarts
- Can protect data in BioMart
Example of BioMart interface 1

AfMap v2.0
**Example of BioMart interface 2**

AfMap v2.0

### Dataset: AfMap v2.0

**Attributes:**
- **RSID**
- **Chrm**
- **Pos**
- **KHS**
- **HER**
- **STS**
- **XHS**
- **ZUL**
- **ASW**
- **CEU**
- **CHB**
- **GIH**
- **JPT**
- **LWK**
- **MEX**
- **MKK**
- **TSI**
- **YRI**

**Gene id:** %BRCA%
- **Genesymbol:** %BRCA%

### Filters

- **Allele A frequencies**
  - Select all
  - Deselect all

- **Marker:**
  - **RSID**
  - **Chrm**
  - **Pos**
  - **Allele A**
  - **Allele B**

- **Allele A frequencies:**
  - **KHS**
  - **HER**
  - **STS**
  - **XHS**
  - **ZUL**
  - **ASW**
  - **CEU**
  - **CHB**

**Sample counts:**
- **KHS n**
- **HER n**
- **STS n**
- **XHS n**
- **ZUL n**
- **ASW n**
- **CEU n**
- **CHB n**

---

**Export all results to:**
- **TSV**

**Email notification to:**

**View:**
- **10 rows as:** HTML
- **Unique results only**

---

### Example of BioMart interface 2

- **AfMap v2.0**

- **Gene id:** %BRCA%
- **Genesymbol:** %BRCA%

- **Marker:**
  - **RSID**
  - **Chrm**
  - **Pos**
  - **Allele A**
  - **Allele B**

- **Allele A frequencies:**
  - **KHS**
  - **HER**
  - **STS**
  - **XHS**
  - **ZUL**
  - **ASW**
  - **CEU**
  - **CHB**

- **Sample counts:**
  - **KHS n**
  - **HER n**
  - **STS n**
  - **XHS n**
  - **ZUL n**
  - **ASW n**
  - **CEU n**
  - **CHB n**

---

**H3AB Pan African**
Example of BioMart interface 3
Galaxy for NGS and arrays

- Galaxy – web-based framework for building workflows for data analysis
- Contains modules for searching and extraction of human data and many NGS analyses
- Can plugin in own modules
- Workflows and histories can be customized, saved and shared
- Can export data to visualization tools
- Different levels of protection
- Can integrate with HPC solutions
Galaxy example
Ensembl view
Admixture mapping tools

- Most tools were developed for 2-way admixture, although some can handle 4-way admixture.
- Some local populations are at least 5-way admixed.
- Developing new tools for multi-way admixture mapping.
- Developing tools for network analysis of admixed populations.
GWAS network tools
Education and training

• Working group has met 7 times
• Researching current programs:
  – Collected Bioinformatics curriculums from different Universities in Africa –very few!
• Developing application and selection procedures for courses and internships
• Developing training courses
Training – H3ABioNet

Training Requests of H3ABioNet Nodes

- Other
- GWAS / Population genetics
- Metagenomics
- Genotyping
- Variant calling
- NGS analyses
- Software development
- Databases
- Development of tools
- Application of tools
- Systems administration

Percentage of Respondant Requesting for Training (N=33)
H3ABioNet training plan

- Train-the trainer
- Graduate training MSc/PhD/Postdocs
  - Shared course curriculum
  - Co-supervision across nodes
  - Project approval by advisory board
  - Specialised courses
- Technical staff
  - Computing/sys admin courses
  - HPC, Cloud, data management
- Internships with external partners
Training – H3Africa

Types of Training Required from H3ABioNet

- Microarray analyses (SNP)
- NGS tools and analyses
- Metagenomics / microbiome analyses
- GWAS / population genetics
- Application of tools
- Development of tools
- Databasing / data mining
- Network / pathway analysis
- Statistical analyses
- Other

Percentage of H3Africa Projects (N=9)

H3ABioNet
Pan African Bioinformatics Network for H3Africa

2nd H3Africa Consortium Meeting, Accra
H3Africa training

• Specialised courses (NGS, genotyping data analysis, pop genetics, databasing)

• Internships for H3-researchers to visit bioinformatics nodes

Estimated Year H3Africa Projects Require Bioinformatics Training

<table>
<thead>
<tr>
<th>Year</th>
<th>Percentage of H3Africa Projects Requesting Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>40%</td>
</tr>
<tr>
<td>Year 2</td>
<td>70%</td>
</tr>
<tr>
<td>Year 3</td>
<td>80%</td>
</tr>
<tr>
<td>Year 4</td>
<td>60%</td>
</tr>
<tr>
<td>Year 5</td>
<td>20%</td>
</tr>
</tbody>
</table>
First training course

- NIH grants management workshop 27\textsuperscript{th} – 28\textsuperscript{th} May, UCT
- Anticipated number 25 – 30
- Topics covered:
  - NIH and recipient roles and responsibilities, roles and responsibilities of the grantee and consortium sites, understanding the Notice of Award, award budgets, costs and allowable expenditures, award Delays, Progress Reports/Completing the PHS 2590, Consortium/Subcontract Policy requirements, Payment Management System training, Federal Financial Report reporting, Post Award Actions
Second training course

- Pretoria for systems administrators, developers
  3rd – 14th June
- Anticipated number 25
- Topics covered:
  - HPC, Grid Computing, Linux admin and set up of purchased servers, backups, Cloud computing, Data management and visualisation and Data Security
- Students will learn to install and manage machines they purchase
Third training course

- Bioinformatics training of H3ABioNet trainers to be held in Central location –ICIPE, Nairobi, Kenya
- 25 participants from H3ABioNet nodes
- Criteria – must be staff, have some bioinformatics experience, involved in teaching and provide support and training for their nodes when they return
- Course from 8th to 26th July and covers:
  - Unix / Linux, biostatistics, programming, data management, NGS pipelines, GWAS and genotyping.
Fourth training course

• eBioKit workshop ICIPE 29th July – 1st August
• Anticipated number 25
• Some will be from bioinformatics course
• Topics covered:
  – UNIX, Biological databases/ENSEMBL NGS (Galaxy mainly), phylogenetics or wEMBOSS
Training courses

- Technical course will be webcast to some sites using USTREAM
- Train-the trainer course will have live Vidyo streaming to 2 sites Tunisia and Nigeria
- In total >150 people all over Africa anticipated to have access to training
- All lectures will be recorded
- This will enable us to reach a wider audience
Knowledge transfer program - CPGR

- KTP brings together experts from around the world to Africa to train upcoming researchers on specialized projects.
- Transfer of knowledge from higher concentration to lower concentration for longer period of time e.g. 3-6 months.
- Faster turnaround time of effective transfer of knowledge.
- The programme cuts running costs as more people are trained within their local environment.
- KTP allows training within the resource deprived environment and provides a wider experience for visiting experts.
- You can take part in the programme as an expert, associate or host of a project.
- Online registration will be available at www.ktp.cpgr.org.za
Future training plan

- Work with H3Africa projects to set up suitable training plan for years 2 and 3
- Ensure courses reach wide audience
- Coordinate with other training programs – Wellcome Trust, EMBO, etc.
- Work with GOBLET on training objectives, quality etc.
- Set up internship program
- Set up co-supervision program
- Work towards degree programs at African institutions
Communication

• Meetings
  • Kick-off meeting Nov 2012
  • Meeting in Casablanca March 2013
  • WG meetings fortnightly, MC monthly

• Web portal:
  • Members info and expertise
  • Documents
  • Tools and resources, etc.

• Mailing lists, biweekly bulletin
• Internal database – info on nodes
Kick-off meeting
H3ABioNet is the Pan African Bioinformatics network for H3Africa (Human Heredity and Health in Africa). H3ABioNet consists of partners from over 30 institutions in many African countries and abroad. For a complete list of partners see [Consortium Members](#).

H3ABioNet has four main components:

- User support and computational infrastructure development
- Training and capacity development
- Research and tool development
- Outreach and communication
# Terms of Reference

## Members

<table>
<thead>
<tr>
<th>Name</th>
<th>Institute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niyi Oyeleke</td>
<td>NAABDA</td>
</tr>
<tr>
<td>Shaun Aron</td>
<td>Wits</td>
</tr>
<tr>
<td>James Brandful</td>
<td>NMIMR</td>
</tr>
<tr>
<td>Segun Fatumo</td>
<td>NAABDA</td>
</tr>
<tr>
<td>Judit Kumuthini</td>
<td>CPGR</td>
</tr>
<tr>
<td>Ahmed Moussa</td>
<td>AEU</td>
</tr>
<tr>
<td>Shakesh Nalbho</td>
<td>UoM</td>
</tr>
<tr>
<td>Nicky Mulder</td>
<td>UCT</td>
</tr>
<tr>
<td>Jonathan Kayondo</td>
<td>UVRI</td>
</tr>
<tr>
<td>Faisal M. Fatileinola</td>
<td>FUSU</td>
</tr>
<tr>
<td>Victor Osamer</td>
<td>CU</td>
</tr>
<tr>
<td>Jean-Baka Demelevo Entellah</td>
<td>SANBI</td>
</tr>
<tr>
<td>Sylvester Leonid Lyantagaye</td>
<td>UDSM</td>
</tr>
<tr>
<td>Odile Ouwe Mosi Oukom</td>
<td>CERMES</td>
</tr>
<tr>
<td>Owusu-Dabo Ellis</td>
<td>KCCR</td>
</tr>
<tr>
<td>Samson Salifu Pandum</td>
<td>KCCR</td>
</tr>
<tr>
<td>PostDoc C / O Dean Everett</td>
<td>MLWCTR</td>
</tr>
<tr>
<td>Winston Hide</td>
<td>HSPH</td>
</tr>
<tr>
<td>Shanman Ho Sui</td>
<td>HSPH</td>
</tr>
<tr>
<td>Ahmed Mansour Mohamed Mansour Elzohair</td>
<td>ZU</td>
</tr>
<tr>
<td>Samar Kassim</td>
<td>ZU</td>
</tr>
<tr>
<td>Anne Fischer</td>
<td>ICIPE</td>
</tr>
<tr>
<td>Sumir Panji</td>
<td>UCT</td>
</tr>
</tbody>
</table>

## Minutes

- Meeting #2 (2013-02-12)
- Meeting #3 (2013-02-26)
- Meeting #4 (2013-03-13)
- Meeting #5 (2013-04-09)
- Meeting #6 (2013-04-23)
Training page

Online application system for courses
H3ABioNet funded personnel

Composition of H3ABioNet Members (N=153)

- Bioinformaticians
- Computer Scientists
- Developers
- Domain Experts
- Post-Graduate Students
- Post-Doctoral Fellows
- Statisticians
- Scientists
- Systems Administrators
- Trainers
Internal database
Outreach

• Casablanca conference exhibition
• Represented at SASA meeting
• Coordination with other training programs
• Work with GOBLET – attended their TGAC workshop
• External mailing lists – interest groups
• Work more closely with H3Africa projects
# Acknowledgements

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simani Gaseitsiwe</td>
<td>Botswana Harvard AIDS Institute Partnership</td>
<td>Botswana</td>
</tr>
<tr>
<td>Ahmed Mansour Alzohairiy</td>
<td>Zagazig University</td>
<td>Egypt</td>
</tr>
<tr>
<td>James Brandful</td>
<td>NMIMR</td>
<td>Ghana</td>
</tr>
<tr>
<td>Ellis Owusu-Dabo</td>
<td>KNUST</td>
<td>Ghana</td>
</tr>
<tr>
<td>Daniel Masiga</td>
<td>ICIPE</td>
<td>Kenya</td>
</tr>
<tr>
<td>Dean Everett</td>
<td>Malawi-Liverpool Wellcome Trust Clinical research Programme</td>
<td>Malawi</td>
</tr>
<tr>
<td>Seydou Doumbia</td>
<td>University of Bamako</td>
<td>Mali</td>
</tr>
<tr>
<td>Yasmina Jaferally Fakim</td>
<td>SANBio</td>
<td>Mauritius</td>
</tr>
<tr>
<td>Hassan Ghazal</td>
<td>University Mohammed First</td>
<td>Morocco</td>
</tr>
<tr>
<td>Azedine Ibrahimi</td>
<td>Faculte de Medecine et de Pharmacie de Rabat</td>
<td>Morocco</td>
</tr>
<tr>
<td>Ahmed Moussa</td>
<td>Abdelamlek Essaadi University, Tangier</td>
<td>Morocco</td>
</tr>
<tr>
<td>Fouzia Radouani</td>
<td>Pasteur Institute Casablanca</td>
<td>Morocco</td>
</tr>
<tr>
<td>Fouad Seighrouchni</td>
<td>Institut National d'Hygiène, Rabat</td>
<td>Morocco</td>
</tr>
<tr>
<td>Fatima Gaboun</td>
<td>National Institute of Agronomic Research, Rabat</td>
<td>Morocco</td>
</tr>
<tr>
<td>Khalid Saddiki</td>
<td>Mohammed V University, Rabat</td>
<td>Morocco</td>
</tr>
<tr>
<td>Alami Raouf</td>
<td>Centre National de Transfusion Sanguine, Rabat</td>
<td>Morocco</td>
</tr>
<tr>
<td>Odile Ouwe Missi</td>
<td>CERMES</td>
<td>Niger</td>
</tr>
<tr>
<td>Ezekiel Adebibi</td>
<td>Covenant University Bioinformatics Research</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Oyekanmi Nash</td>
<td>NADBA</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Nicky Mulder</td>
<td>University of Cape Town</td>
<td>South Africa</td>
</tr>
<tr>
<td>Judit Kumuthini</td>
<td>CPGR</td>
<td>South Africa</td>
</tr>
<tr>
<td>Nicki Tiffin</td>
<td>SANBI, University of the Western Cape</td>
<td>South Africa</td>
</tr>
<tr>
<td>Ozlem Tastan Bishop</td>
<td>Rhodes University</td>
<td>South Africa</td>
</tr>
<tr>
<td>Scott Hazelhurst</td>
<td>Wits University</td>
<td>South Africa</td>
</tr>
<tr>
<td>Fourie Joubert</td>
<td>University of Pretoria</td>
<td>South Africa</td>
</tr>
<tr>
<td>Hugh Patterson</td>
<td>University of the Free State</td>
<td>South Africa</td>
</tr>
<tr>
<td>Faisal Fadlelmola</td>
<td>Future University</td>
<td>Sudan</td>
</tr>
<tr>
<td>Sylvester Lyantagaye</td>
<td>University of Dar es Salaam (UDSM)</td>
<td>Tanzania</td>
</tr>
<tr>
<td>Nzovu Ulenga</td>
<td>MDH</td>
<td>Tanzania</td>
</tr>
<tr>
<td>Julie Makani</td>
<td>MUHAS</td>
<td>Tanzania</td>
</tr>
<tr>
<td>Alia Benkahla</td>
<td>Institute Pasteur of Tunis</td>
<td>Tunisia</td>
</tr>
<tr>
<td>Jonathan Kayondo</td>
<td>UVRI</td>
<td>Uganda</td>
</tr>
<tr>
<td>Victor Jongeneel</td>
<td>NCSA</td>
<td>USA</td>
</tr>
<tr>
<td>Win Hide</td>
<td>Harvard School of Public Health</td>
<td>USA</td>
</tr>
</tbody>
</table>
Thanks: Sumir Panji – Project Manager
Questions to H3Africa

- How will you get your data from the generator to your lab?
- Are there any specific journal ToCs you want access to?
- What else do you want from us?
- How best can we engage with you?