H3ABioNet

Pan African Bioinformatics Network for H3Africa

Progress relevant to consortium
May 2015

Nicky Mulder: nicola.mulder@uct.ac.za
Summary of initial goals

• Short/medium term:
  – Store H3Africa data & enable submission to EGA
  – Build human and computing infrastructure
  – Train in bioinformatics theory and techniques
  – Apply techniques to data, with skills transfer

• Long term:
  – Train bioinformatics academics and support staff in H3ABioNet and H3Africa projects
  – Build network of bioinformaticians and data analysers
What’s new?

• User support and infrastructure update
  – Developing computing infrastructure
  – Helpdesk

• H3Africa archive progress

• Training activities

• Research activities

• H3Africa consortium projects
Developing computing infrastructure

- Several nodes now have computing facilities
  - Use H3ABioNet helpdesk to get info and access
- Developed documentation for sys admins
  - Useful for anyone investing in computing equipment
- Doing survey on HPC and Galaxy
- Globus for data transfers
- NetMap for bandwidth testing
H3ABioNet help desk

- http://www.h3abionet.org/support/help-desk
- Managed by 15 volunteers with experience and expertise in Bioinformatics and Genomics.

Stats 11/11/14-05/05/15
H3Africa Archive

• Submission of H3Africa data 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
    – Can assist with additional analyses, but only by agreement with PI
  • Archive area
    – Purely for storage, no processing
    – encrypted files are mirrored to a separate physical location
H3Africa archive submission process

• Register project with the H3A Archive 2 months prior to submission and provide:
  – Submission timeline
  – Blank consent form, ethics clearance numbers
  – Estimated sample count
  – What phenotype data
    • Will be collected by the project
    • Will be submitted for H3A storage
    • Submitted to EGA

• EGA accessions will be provided
  – Study accession on H3A registration
  – Sample accessions on EGA submission
First submission: AWIGEN project

- Pilot data
- 973 samples
- Illumina MetaboChip (~200k SNPS after QC)
- Phenotypes captured:
  - Sex
  - Age
  - Height
  - Weight
  - Hip circumference
  - Waist circumference
  - Ethnicity
H3ABioNet training activities (1)

- Co-organised metagenomics workshop in Mauritius in Dec 2014
  - Participants from 11 African countries
  - Trainers from H3ABioNet, Greece, Sweden, Mauritius
- Provided travel fellowships to CafGEN GWAS course in Botswana
- Ran Advanced Systems Administration workshop in Pretoria in Feb 2015
  - Course in 2 parts, beginner and advanced
  - 24 participants from H3Africa projects and H3ABioNet
  - Trainers local and/or H3ABioNet
H3ABioNet training activities (2)

- Organised 2 workshops at end of ISCB Africa ASBCB Bioinformatics conference in Tanzania –Exploring variation data (EBI) and Data visualization (Amel Ghouila)
- Ran Introduction to biostatistics workshop in Tunis in March 2015
  - ~20 participants from H3Africa projects and H3ABioNet
  - Local and H3ABioNet trainers
- Ran Medical population genetics and GWAS for complex diseases workshop and symposium in Cape Town in April 2015
  - 32 participants from H3Africa, AIMS and other locals
  - Divided into 2 streams –data analysis, advanced modelling
  - 3 trainers from abroad, 2 H3ABioNet
  - Followed by symposium and discussions
H3ABioNet training activities (3)

- Trained >450 people to date
- Placed 8 interns – internships available* 
- Ongoing evaluation of online courses
- Wordpress sites available for all courses – many have recorded lectures. Links can be found at: http://training.h3abionet.org
- Progress on African Bioinformatics Education Committee activities:
  - University of Bamako, Mali started their first Masters course based on our curriculum with several H3ABioNet trainers
Plans for future training

- UVRI NGS course in July
- Introduction to bioinformatics course – live broadcasting to multiple sites
- Regional GWAS workshop for Tanzania

• Awaiting your needs!!
Research and outreach

• Several projects within the network to build capacity and skills, as well as tools
• Started special interest groups in e.g. GWAS, NGS, etc. –all welcome!
• Organised webinar series, testing platforms
• Work presented at ISCB Africa ASBCB Bioinformatics conference –March, Tanzania
• Increased presence on Facebook and Twitter
ISCB Africa ASBCB conference

H3ABioNet and Wellcome Trust provided ~30 travel fellowships
H3Africa consortium projects

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

• To track participant recruitment and data/sample workflow
• Aim to do as much pre-population of data as possible
• Need to customise it for each project
Recruitment database mock-up

H3A BioNet
Pan African Bioinformatics Network for H3Africa
Ontologies & controlled vocabularies TF

- The group aims to harmonize H3Africa data using standardized vocabularies and ontologies.
- This harmonization will make it easier to query H3Africa metadata.
Progress and Challenges (OCVT)

• Access to CRFs only - phenotype data would be better.
  – We recommend standardization of CRF design (use PhenX)
  – Ambiguous questions in CRFs difficult to harmonize

• Previously no access to experimental information

• Need to work with Biobanks/Biorepository for their metadata
Trainer/trainee DB – HTrainDB (1)

Welcome to the Trainee Trainer DB for H3Africa.

H3Africa aims to strengthen research and training capacity in Africa particularly among young African scientists. Over the next couple of years, various training efforts and activities will be supported by the H3Africa consortium. Given the large and diverse number of projects and potential trainees in this consortium, it is imperative that training efforts are captured and stored in a systematic and consistent manner. The proposed project aims to develop a database which:

- captures H3Africa training efforts and trainees’ progress on a regular basis
- provides a web-portal for the H3Africa funders and managers to track progress and adapt the H3Africa efforts based on evidence.
- in future this resource will provide the international scientific community an invaluable resource to learn lessons and successes of the H3Africa project in building research and training capacity in Africa via various training activities.

Zahra Mungloo- Dilmohamud
Shakuntala Baichoo – University of Mauritius
Trainer/trainee DB – HTrainDB (2)

H3Africa Projects

<table>
<thead>
<tr>
<th>ProjectID</th>
<th>ProjectEndDate</th>
<th>Start Date</th>
<th>Project Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2015-12-30</td>
<td>2014-01-30</td>
<td>African Collaborative Center for Microbiome and Genomics Research (ACCME)</td>
</tr>
<tr>
<td>2</td>
<td>2016-04-08</td>
<td>2014-02-19</td>
<td>Burden, Spectrum and Aetiology of Type 2 Diabetes in sub-Saharan Africa</td>
</tr>
<tr>
<td>3</td>
<td>2015-02-26</td>
<td>2015-02-02</td>
<td>Clinical and Genetic Studies of Hereditary Neurological Disorders in Mali</td>
</tr>
<tr>
<td>4</td>
<td>2015-06-13</td>
<td>2014-10-06</td>
<td>Collaborative African Genomics Network (CAfGEN)</td>
</tr>
<tr>
<td>5</td>
<td>2015-11-13</td>
<td>2014-08-11</td>
<td>Contribution of Genetic Variation to Pharmacokinetic Variability and Toxicity in Patients Undergoing</td>
</tr>
<tr>
<td>6</td>
<td>2015-09-18</td>
<td>2014-05-05</td>
<td>Development of H3 Africa Biorepositories to Facilitate Studies on Biodiversity, Disease &amp; Pharmacoge</td>
</tr>
<tr>
<td>7</td>
<td>2015-09-24</td>
<td>2014-07-07</td>
<td>Establishment of an H3Africa Biorepository at Clinical Laboratory Services</td>
</tr>
<tr>
<td>8</td>
<td>2015-11-07</td>
<td>2014-03-02</td>
<td>Ethical, Legal and Social Implications (ELSI) Addis Ababa University, Ethiopia</td>
</tr>
<tr>
<td>9</td>
<td>2016-03-11</td>
<td>2014-07-13</td>
<td>Ethical, Legal and Social Implications (ELSI) Institute of Human Virology, Nigeria</td>
</tr>
<tr>
<td>10</td>
<td>2015-08-14</td>
<td>2015-02-01</td>
<td>Exploring Perspectives on Genomics and Sickle Cell Public Health Interventions</td>
</tr>
</tbody>
</table>

Zahra Mungloo- Dilmohamud
Shakuntala Baichoo –University of Mauritius
H3Africa custom chip design

- Chip design working group through Genome Analysis WG, in collaboration with Wits, Sanger and many other experts
- Computing: CHPC, NCSA
- Data ~ 3400 samples
- Timelines:
  - Depends on Baylor sequencing (when samples arrive)
  - Aim for list of SNPs in Aug/Sept
Pre-processing → QC → Pre-processing

Raw Reads → Lanelet BAM QC
Alignment/Mapping to Reference (BWA) → Mark Duplicates (Picard)
Indel Realignment (GATK) → Base Recalibration (GATK)

Joint variant calling (GATK UG) → PLs in VCFs 
Filtering (GATK VariantRecalibrator) across the genome → Filtered VCFs 
Cohort VCF QC → Union set of variants (GATK CombineVariants – minimalVCF – filteredrecordsmmergeType KEEP_UNCONDITIONAL)

Genotype likelihood refinement with BEAGLE4 across samples within regions → Refined GPs in VCFs 
Biallelic GPs in VCFs → Phasing with SHAPEIT2 onto haplotype scaffold across chromosomes 
Phased haplotypes → Calculate LD in windows including buffer regions (PLINK or calculateLD) for each population 
Biallelic SNPs in LD → Selection of biallelic tag SNPs (RagTagger and IMPUTE2) across the genome in a cyclical/hybrid approach 
Biallelic tag SNPs → Validation of tag SNPs by Affymetrix or Illumina 
Validation of improvement in capture over other chip designs AND validation of improvement in imputation accuracy using new reference panel

H3ABioNet: Pan African Bioinformatics Network for H3Africa
H3ABioNet service offerings

- Participant cohorts
- Extract DNA
- Data analysis and storage

Collect samples → Biorepository → Generate data → Data repository

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

H3ABioNet
Pan African Bioinformatics Network for H3Africa
H3ABioNet service offerings

Participant cohorts
Collect samples

Extract DNA

Generate data

Data analysis and storage

Biorepository

Computing Infrastructure
Research/tool development
User support
Training

Participant databases,
LIMS, data integration

Participant recruitment
database

H3ABioNet
Pan African Bioinformatics Network for H3Africa
H3ABioNet service offerings

- Participant cohorts
- Collect samples
- Extract DNA
- Generate data
- Data analysis and storage

Biorepository

Compared Infrastructure | Research/tool development | User support | Training

- Chip design
- Data management
- Data transfer
- Data analysis
- Pipelines and SOPs
- Galaxy

H3ABioNet
Pan African Bioinformatics Network for H3Africa
H3ABioNet service offerings

- Participant cohorts
- Collect samples
- Extract DNA
- Biorepository
- Generate data
- Data analysis and storage
- Data archive
- EGA submission
- Ontologies
- Data integration

Computing Infrastructure | Research/tool development | User support | Training
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 Alzohairy</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 Jaufeerally 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 Ibrahim</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 Seghouchni</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 Sadki</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 Ouve 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 Patterton</td>
<td>University of the Free State</td>
<td>South Africa</td>
</tr>
<tr>
<td>Faisal Fadilemola</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>

Funding: NIH Common Fund, NGHRI grant: U41HG006941

Project Manager: Sumir Panji