H3ABioNet

Pan African Bioinformatics Network for H3Africa

Preparing for the data phase

PI: 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
Progress

• H3Africa consortium projects
  – Participant recruitment database
  – Biobank catalogue
  – Ontologies
  – eGenomics catalogue
  – Chip design
• H3Africa archive
• Tools & resources for data analysis
• Training activities for data analysis
H3Africa Participant Recruitment database

- To track participant recruitment and data/sample workflow
- RedCAP is being used as the data capture system
- Project specific forms have been designed in RedCAP and emailed to H3Africa project PIs to complete
- The reporting dashboard is accessible by login only and provides a variety of graphs
- Second round of submissions about to begin
- If unsure about submission, please ask an H3ABioNet member
H3APRD funders’ reports

Dashboard

- 11 Projects
- 2 Collections
- 53,175 Participants Expected
- 43,705 Participants Recruited
- 82% Completed

Project title: All NIH projects

Choose a project ...
H3APRD funders’ reports

Note, each new entry is aggregate data!
H3APRD query tool
H3Africa BioBank (data) Catalogue

• A centralized web portal catalogue is being developed to act as a single point of entry for anyone to query H3Africa bio-repositories to determine:
  – What H3Africa biospecimens and quantities are available
  – Which H3Africa biorepository the biospecimen resides in
  – What phenotypic attributes the biospecimen has been annotated with
  – How to apply to gain access to H3Africa biospecimens

• Minimum Information About Biobank data Sharing (MIABIS) terms have been curated by the Ontology task force
H3Africa BioBank Catalogue

Catalogue will be extended to genomic data
Harmonizing H3Africa data

- To enable queries of phenotype information across different H3Africa projects which used different CRFs
- To query biobanks for H3Africa samples from different studies
- For EGA submission – standardised metadata description
- Ontology and Standardized Vocabulary Task Force:
  - Promote and oversee the use of ontologies and standardized vocabularies for H3Africa project data
  - Use existing ontologies where possible, create new ontologies only when necessary

H3ABioNet
Pan African Bioinformatics Network for H3Africa
Ontologies work

Mapping CRFs to ontologies, e.g. phenotype or disease ontology

Adapting OMIABIS ontology to H3Africa data

Mapping genomics data to Experimental Factor ontology

Developing Sickle Cell Disease Ontology
eGenomics catalogue

- Started out as a list of online courses for bioinformatics, extended to all genomics
- Provides a searchable catalogue of genomics resources
- Contributions from anyone welcome, fellows and E&CTWG have contributed
- Review facility
- [http://egenomics.h3abionet.org/](http://egenomics.h3abionet.org/)
DISCLAIMER: The eGenomics Catalogue is a community effort, developing out of the H3ABioNet consortium. The resources listed were added and reviewed by volunteers. Reviews are personal opinions and a listing or review should not be construed as an endorsement of any particular resource by the consortium. The funders do not independently verify information added to the catalogue, we rely on submitters to provide information that is accurate and not misleading.

What is the eGenomics Catalogue?

The eGenomics catalogue was initiated by H3ABioNet and is maintained by volunteers from across the globe, in particular H3Africa Fellows. This catalogue is a community effort and resources are added and reviewed by volunteers. Reviews are personal opinions and a listing or review should not be construed as an endorsement of any particular resource by the consortium. The funders do not independently verify information added to the catalogue, the website relies on submitters to provide information that is accurate and not misleading.
eGenomics catalogue

What is the eGenomics Catalogue?

The eGenomics catalogue was initiated by H3ABioNet and is maintained by volunteers from across the globe, in particular H3Africa Fellows. This catalogue maintains free online Genomics educational material. This material is rated based on community-based reviews/evaluations. The material collected on this website includes: Books, Journals, Courses, MOOCs, OpenCourseware Databases. The material is categorized based on subject categories proposed by the H3Africa Education and Coordinated Training Working Group, in addition the EDAM ontology has been used to categorize Bioinformatics relevant courses.

Latest Listings
7. GenePhony
   Category: Software
8. Open-access journal - Journal of Applied Mathematics and Bioinformatics
   Category: Journals
9. Open-access journal - International Scholarly Research Notices (ISRN) Bioinformatics
   Category: Journals

Latest Reviews
   Reviewed by Yicky Nembare
   4.3
5. Computational Molecular Biologics
   Stanford University
H3Africa archive

- First dataset is in from AWIGEN
- Plan to include Baylor sequence data
- Anyone else ready to submit?
- Register project with the 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 and submitted to EGA
- Will be assigned EGA accession numbers for project and samples
- Dashboard developed for interface with submitter
Pre-register a submission

Your name:
Ayton Meintjes

Project:
AWIGEN

Data types:
- [x] Genotyping array data

Estimated date of submission:
2016/06/01

Save

Pan African Bioinformatics Network for H3Africa
Pre-register a submission

Your name: Ayton Meintjes
Project: AWIGEN

Data types:
- Human whole genome sequence data
- Human exome sequence data
- Genotyping array data
- Metagenomics sequence data
- Pathogen sequence data

Estimated date of submission:
2016/06/01

Save
Pre-register a submission

Your name:
Ayton Meintjes

Project:
AWI-GEN

Data types:
- [ ] Human whole genome sequence data
- [ ] Human exome sequence data
- [x] Genotyping array data
- [ ] Metagenomics sequence data
- [ ] Pathogen sequence data

Estimated date of submission:
2016/06/01
test submission

Save

Projects

- TGEN - TrypanoGEN: An integrated approach to the identification of genetic determinants of susceptibility to trypanosomiasis
- KDRN - H3Africa Kidney Disease Research Network
- AWI-GEN - Genomic and Environmental Risk Factors for Cardiometabolic Disease in Africans
- RH-GEN - The RH-GEN Network: Genetics of rheumatic heart disease and molecular epidemiology of Streptococcus pyogenes pharyngitis
- T2DSSA - Burden, spectrum and etiology of type 2 diabetes in sub-Saharan Africa
- RAFAGENE - RAFaGene: Contribution of genetic variation to pharmacokinetic variability and toxicity in patients undergoing multi-drug tuberculosis treatment in Sub-Saharan Africa
- OSSAX - Identification of genes important for schizophrenia based on genomic analysis of the South African Xhosa population
- ACCME - African Collaborative Center for Microbiome and Genomics Research
- CAI-GEN - Collaborative African Genomics Network
- NMRDQAC - The Nasopharyngeal Microbiome and Respiratory Disease in African Children
- HMGWA - Host and Microbial Genetic Determinants of febrile illness in West Africa
- SIREN - Stroke Investigative Research & Educational Network

AWI-GEN
Genomic and Environmental Risk Factors for Cardiometabolic Disease in Africans

Owner
Michele Ramsay

Submitters
Ayton Meintjes

Submissions

<table>
<thead>
<tr>
<th>Date started</th>
<th>Submitter</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 11, 2016, 7:10 p.m.</td>
<td>ayton.meintjes</td>
<td>New</td>
</tr>
</tbody>
</table>
Data transfers

- Need for **transferring large amounts of genomic data** around Africa & abroad
- Globus Online is a free, fast, secure, reliable and easy to use solution for transferring data via the internet
- Deployed across 21 nodes within the H3ABioNet consortium
- However, still challenges, transfer can take weeks to months, so also have backup hard drives

<table>
<thead>
<tr>
<th>GO endpoints</th>
<th>Transfer speeds (Mbps) (min, max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baylor &lt;-&gt; Blue Waters</td>
<td>340, 1900</td>
</tr>
<tr>
<td>Blue Waters -&gt; UCT</td>
<td>204, 322</td>
</tr>
<tr>
<td>CHPC &lt;-&gt; Blue Waters</td>
<td>81, 243</td>
</tr>
<tr>
<td>UCT &lt;-&gt; CHPC</td>
<td>34, 406</td>
</tr>
<tr>
<td>Sanger &lt;-&gt; UCT</td>
<td>38, 76</td>
</tr>
</tbody>
</table>
Tools and resources for data analysis

- SOPs available on website, including compute resources required
Tools and resources for data analysis

• Developing Docker containers for these pipelines so they can be run on Cloud infrastructures
• Cloud-based imputation server using African reference panel –NCSA (Victor Jongeneel):
  – From chip data, impute additional markers
• GWAS pipeline –Wits (Scott Hazelhurst, Lerato Magosi, Robert Clucas):
  – Pipeline for analysis of Affymetrix chip data (will be adapted to Illumina)
• 16SrRNA analysis pipeline
WitsGWAS pipeline for the Cloud

- Affymetrix chip data QC
- Genotype calling
- Sample QC (missingness, heterozygosity etc.)
- SNP QC (MAF, Hardy Weinberg, etc.)
- Association testing – different options

<table>
<thead>
<tr>
<th>Overview</th>
<th>Setting up witsGWAS</th>
<th>Running pipelines</th>
<th>Extending pipelines</th>
<th>FAQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>About witsGWAS</td>
<td>Getting started</td>
<td>Running pipelines</td>
<td>Extending pipelines</td>
<td>Getting help</td>
</tr>
</tbody>
</table>
Training activities update

• Ran Bioinformatics Train-the-Trainer workshop with GOBLET in Cape Town in November 2015
• Sudan node ran an NGS course in March
• UCT ran a GWAS course in March taught by a visitor we hosted
• Ran a Bioinformatics analysis and annotation of variants in NGS data workshop in Cape Town as satellite to Pharmacogenomics & Precision Medicine conference
• Provided support for Microbiome course run by Nicol project at UCT
• Teaming up with Institute Pasteur INDA for NGS course
HTrainDB

• Still progressing
• Mark Nicol's group piloted and used the database to:
  – Advertise their workshop
  – Use the centralized application form (60 participants registered successfully via HtrainDB)
New course - Intro to Bioinformatics

Aim
• **Basic** bioinformatics **training** for interested H3Africa members (**bioinformatics users**)  
• bioinformatics **tools, algorithms** and **resources** and how to **use them**

Course logistics
• ~4\(^{th}\) July 2016: **3 months, 2 days** contact time **per week**  
  (**3 hours** per contact session)  
• Distance learning model – physical classrooms connected to virtual classroom  
• Mconf – video conferencing  
• Vula – course management
Intro to Bioinformatics course

Curriculum modules

- Resources and databases
- Linux
- Sequence alignment theory & applications
- Multiple sequence alignment
- Molecular evolution and phylogenetics
- Genomics
- Protein Structural Bioinformatics

Lecturers & teaching assistants from Africa

Proposed classrooms (20 and counting!)
Training for data analysis

• No additional formal courses planned unless there is a need. Focussing rather on 2 formats of workshops:
  • Hackathons:
    – Tool, database or resource needed – bring expertise together to do it as a collaborative effort
    – Recently attended biobanking hackathon from B3Africa project
    – Planning cloud hackathon – ”Dockerizing” pipelines
    – Working on a DREAM challenge “hackathon”
  • Bring your data hands on workshops:
    – Projects with complete or pilot datasets
    – We bring trainers and infrastructure to you
    – Mix of training and analysis
• Benefits: get some work done and develop/exchange skills
• Masters degrees - regional courses for shared modules
Other activities & plans

• Special interest groups still going – NGS, health informatics – all welcome!
• Monthly webinar series – many interesting talks on variety of subjects
• Prepare for new application
• Links with international initiatives:
  – ELIXIR
  – BD2K
  – Global Alliance for Genomics and Health (G4GH)
Global Alliance’s mission and role

- **Mission:** To accelerate progress in human health by helping to establish a common framework of harmonized approaches to enable effective and responsible sharing of genomic and clinical data, and by catalyzing data sharing projects that drive and demonstrate the value of data sharing

- Convene stakeholders
- Catalyze sharing of data
- Create harmonized approaches
- Foster innovation
- Commit to responsible data sharing
- Have working groups and task forces:
  - Clinical WG
  - Security WG
  - Data WG
  - Regulatory & ethics WG

400+ Organizational Members
690+ Individual Members
69 Countries
Relevant G4GH outputs

- Framework for responsible sharing of data – in 12 languages
- Beacon project
- Matchmaker exchange
- BRCA challenge
- Actionable cancer genome initiative
- DWG:
  - Standards for reads
  - File formats
  - Reference genomes
  - Variant analysis

H3Africa data should take advantage of the standards, tools and APIs and follow international trends
Acknowledgements

The H3ABioNet Consortium, Project Manager Sumir Panji

Funding: NIH Common Fund, NGHRI grant: U41HG006941