H3ABioNet Project Overview

H3ABioNet Project Goal

To build H3ABioNet - a sustainable African Bioinformatics Network

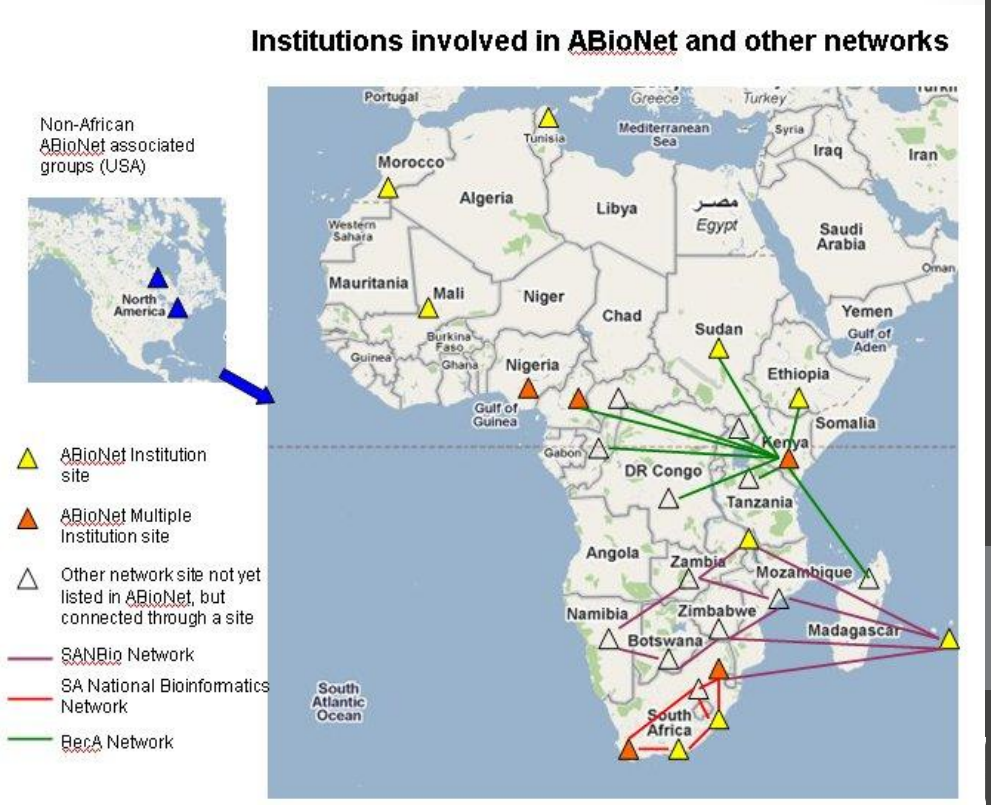
to provide bioinformatics infrastructure and support for the H3Africa consortium

Specific Aims

- Enable the H3Africa consortium to **exploit genomic and environmental data** generated within H3Africa
- Provide **core infrastructure** to aid research in genomic medicine, high throughput biology, systems biology, molecular biology, genetics, and medicine
- **Enable development and use of bioinformatics tools** and resources for H3Africa projects
- Provide secure, high-fidelity **storage of data** and facilitate **submission to public databases**
- Ensure relevant **data is accessible** to all African researchers
- Rapidly **develop sustainable bioinformatics capacity** of African scientists

Motivation: current landscape

- Individual Bioinformatics groups exist with varying levels of capacity
- These provide local support and expertise
- No overarching coordination of efforts
- Existing networks:
  - ASBCB
  - ABioNet
  - BecA
  - SA NBN
  - SANBio



H3ABioNet: A Pan-African Bioinformatic

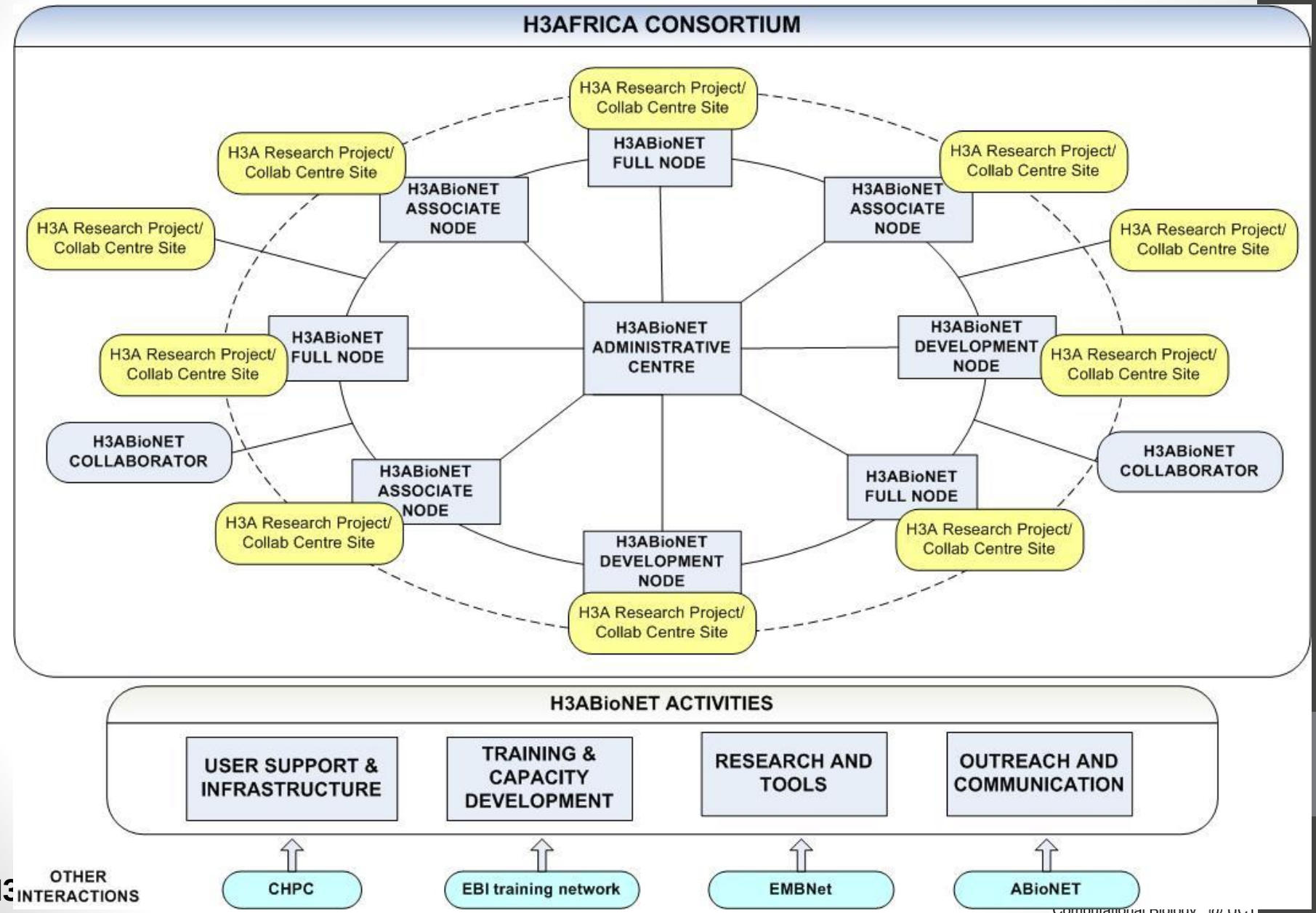
Motivation: H3Africa's needs

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

Approach

- Formalize existing networks and connections between bioinformatics institutions
- Determine exact capacity at participating nodes, identify gaps
- Determine which projects are funded and identify their bioinformatics needs
- Formulate plan to fill gaps and address needs through:
  - Training
  - Computing infrastructure development
  - Research into new tools
  - Effective communication
  - Interaction/collaboration with foreign institutions

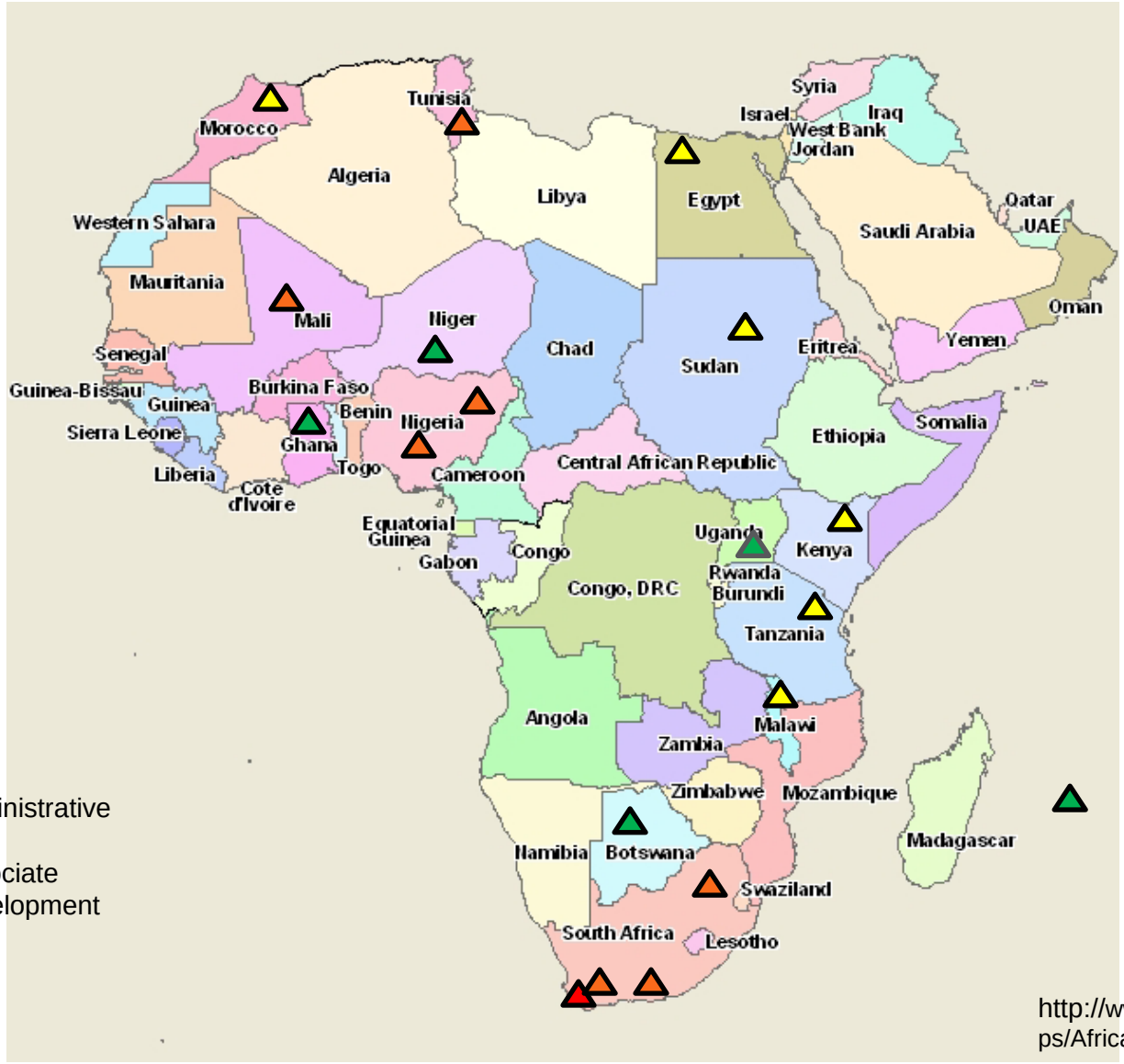
Network structure



H3ABioNet Partners

- Core Administrative Hub: UCT
- Nodes x 20 (node may be 1 or more institutions):
  - Full Nodes x 9
  - Associate nodes x 6
  - Development nodes x 5
- Associated groups:
  - 2 USA partners
  - Several associated groups (CHPC, EBI, EMBNet, JCVI)

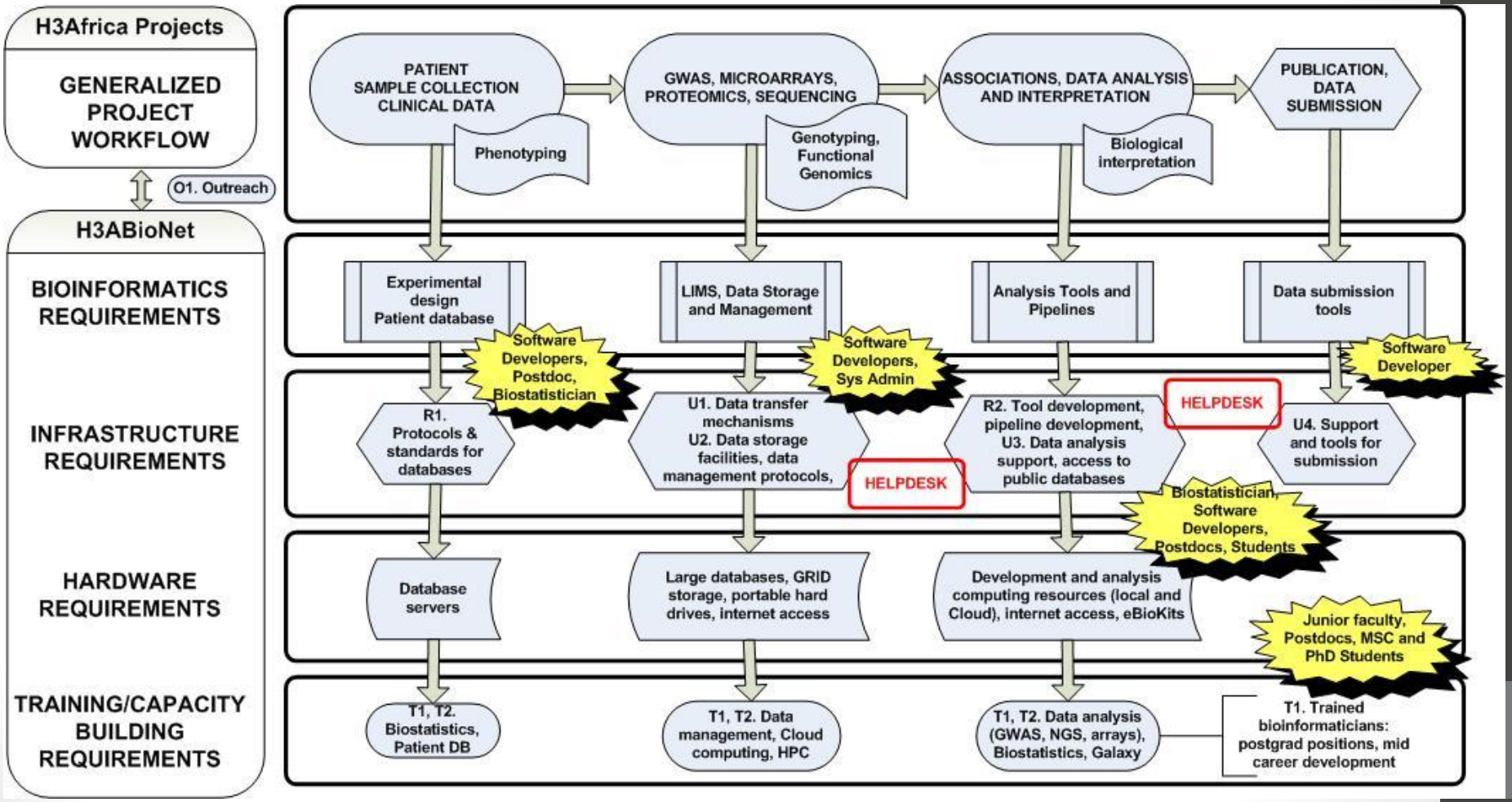
Partner institutions



- Administrative
- Full
- Associate
- Development

http://www.mapshop.com/Travel_Maps/Africa/Africa_Maps.htm

Network activities



Network activities

- Outreach and communication
- User support
- Research and tool development
- Training and capacity development

Communication

- Web portal:
  - Members info and expertise
  - Documents
  - Tools and resources, etc.
- Social networking media
- Mailing lists
- Meetings
- Internal database –info on nodes

Outreach (1)

- Develop associate nodes and development nodes
- Use nodes to identify research partners involved in H3Africa projects
- Identify other potential bioinformatics nodes involved in H3Africa
- Maintain close ties with other ABioNet nodes that are not directly involved in H3Africa activities (with potential for Affiliate status)
- Associate the network with meetings of other groups with complementary objectives e.g. ASBCB, EMBNet, etc

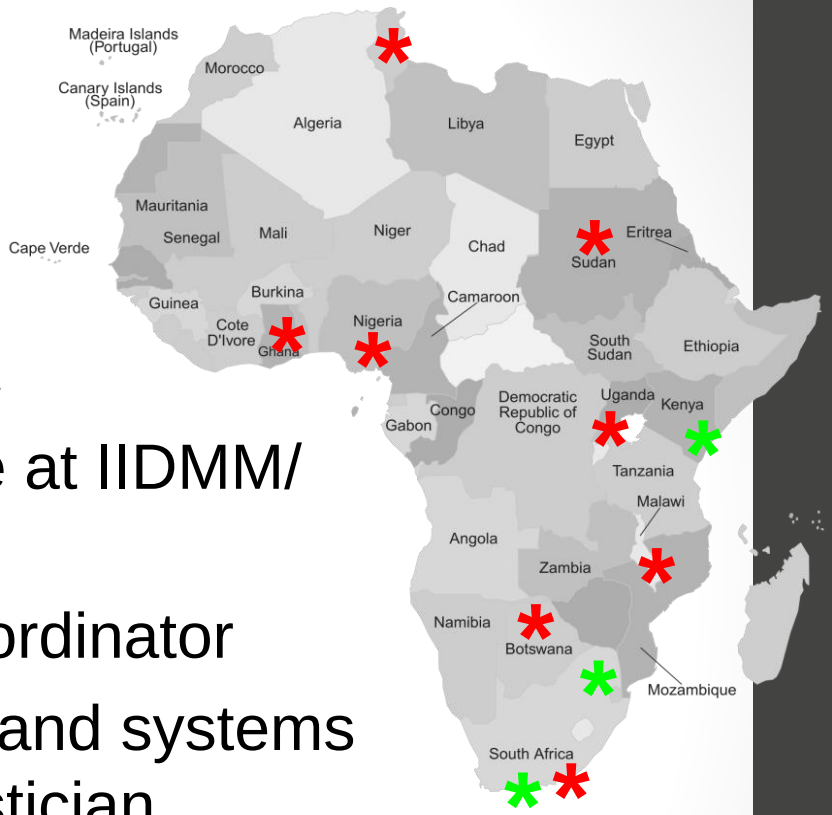
Outreach (2)

- **Bioinformatics stakeholders meetings** to communicate with H3Africa Consortium
- Development of **Special Interest Groups (SIGS)** to promote skills sharing and development
- Dedicated web portal – with outreach to African research diaspora
- Marketing materials: fliers, conference posters at international meetings
- Webinars, as an educational tool
- Negotiation by H3ABioNet for reduced publication costs at major publication houses

User support (1)

- **Existing skills and infrastructure**
  - identify and consolidate existing skills and infrastructure -document
  - identify gaps in skills for H3Africa projects
- **Developing capacity**
  - facilitate skills sharing and knowledge transfer
  - coordinate training in bioinformatics at different levels
  - ensure skills retention

User support (2)



- **Bioinformatics Help Desk**
  - Based at administrative centre at IIDMM/ University of Cape Town
  - Manned by Bioinformatics Coordinator
  - Support: software developers and systems administrators onsite, biostatistician
  - Managed with standard tracking system to monitor requests, response times

User support: technical support

- Investigate appropriate technologies for high performance computing in Africa, including:
  - High performance computing centres
  - Cloud computing
  - Grid computing
  - Addressing internet bandwidth/access limitations
  - Training to implement these technologies
- Build core hardware centres and infrastructure
- Build infrastructure in under-resourced nodes, including:
  - Servers
  - eBioKits, where required
  - Technical support from network Systems Administrators

User support: data management

- Find alternatives to internet for data transfer and analysis –hard drives
- Hardware infrastructure to the consortium for processing and storage
- Standard operating procedures and guidelines for data management
- Software developer and systems administrator availability via help desk
- Updated central mirror site for public databases
- Tools/support for data analysis
- Tools/support for data submission

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
  - Functional & structural SNP analysis tools
  - Admixture mapping and network analysis tools
  - Data visualization tools
  - Recombination tools

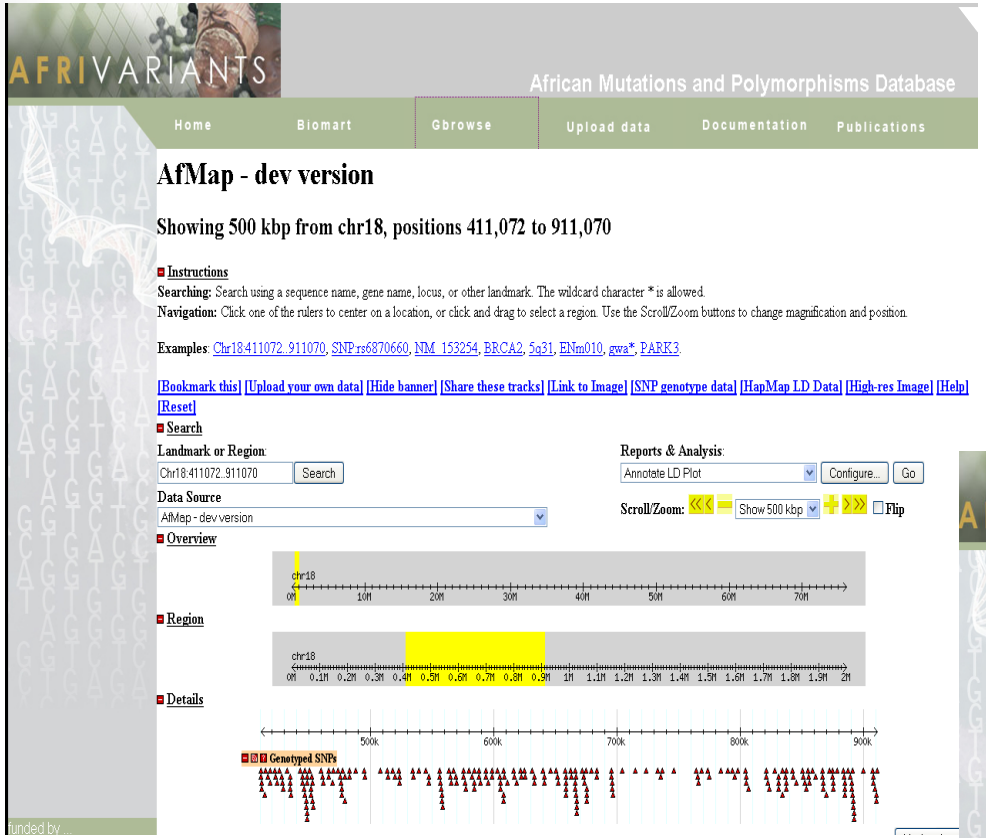
Research: Patient Databasing

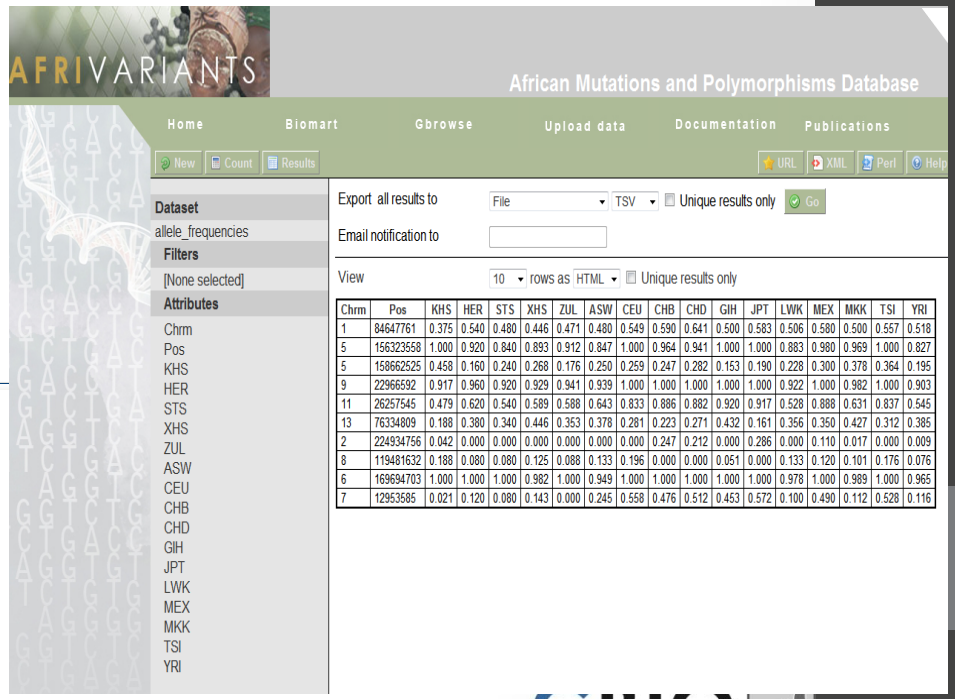
- Develop a generic framework for patient databases to include:
  - Ontologies and controlled vocabularies
  - User friendly front-end
- Develop protocols and standards for storing and submitting data
- Support/advice on protocols for effective data security, patient anonymity and data backup

Research: BioMart for genotype 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
- Will develop BioMart for all public data
- Can query across BioMarts
- Can protect data in BioMart

Example of BioMart interface





Chrm	Pos	KHS	HER	STS	XHS	ZUL	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
1	84647761	0.375	0.540	0.480	0.446	0.471	0.480	0.549	0.590	0.641	0.500	0.583	0.506	0.580	0.500	0.557	0.518
5	156323558	1.000	0.920	0.840	0.893	0.912	0.847	1.000	0.964	0.941	1.000	1.000	0.883	0.980	0.969	1.000	0.827
5	158662525	0.458	0.160	0.240	0.268	0.176	0.250	0.259	0.247	0.282	0.153	0.190	0.228	0.300	0.378	0.364	0.195
9	22966592	0.917	0.960	0.920	0.929	0.941	0.939	1.000	1.000	1.000	1.000	1.000	0.922	1.000	0.982	1.000	0.903
11	26257545	0.479	0.620	0.540	0.589	0.588	0.643	0.833	0.886	0.882	0.920	0.917	0.528	0.888	0.631	0.837	0.545
13	76334809	0.188	0.380	0.340	0.446	0.353	0.378	0.281	0.223	0.271	0.432	0.161	0.356	0.350	0.427	0.312	0.385
2	224934756	0.042	0.000	0.000	0.000	0.000	0.000	0.000	0.247	0.212	0.000	0.286	0.000	0.110	0.017	0.000	0.009
8	119481632	0.188	0.080	0.080	0.125	0.088	0.133	0.196	0.000	0.000	0.051	0.000	0.133	0.120	0.101	0.176	0.076
6	169694703	1.000	1.000	1.000	0.982	1.000	0.949	1.000	1.000	1.000	1.000	1.000	0.978	1.000	0.989	1.000	0.965
7	12953585	0.021	0.120	0.080	0.143	0.000	0.245	0.558	0.476	0.512	0.453	0.572	0.100	0.490	0.112	0.528	0.116

Note: also exploring other storage options

Research: Grid-based data storage

- Storing data on the Grid
- Enables storage, retrieval and annotation of data
- Need to get a grid certificate
- Get different levels of data protection
- Can share data with collaborators
- User friendly web front-end
- Temporary solution for pre-publication phase

Research: 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

Research: Functional & structural SNP analysis

- Develop functional SNP detection pipeline
  - Incorporate the Polyphen, SIFT, SNP-3D
  - Pipeline can be precomputed for known SNPs (dbSNP)
  - Develop user friendly interface
  - Develop user friendly outputs
- Develop a structural SNP tool
  - Tool to analyse genetic variations in coding regions and relate them to structures
  - Develop homology modelling tool with interactive interface

Research: Admixture mapping & network tools

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

Research: Data visualization tools

- Existing tools include Ensembl, Gbrowse, Galaxy Trackster, GATK
- We will use DAS (|Distributed Annotation System) technology based on karyoDAS
- Enables quick visualization of locally generated data –remains private
- Has annotation tool available

Research: Recombination tools

- Recombination events are important for linkage disequilibrium
- Existing tools for measuring recombination rates vary in capacity and effectiveness
- Will develop a heuristic approach to characterise recombination events across the 10-100 Mb range

Training

- Bioinformatics graduate training
- Researcher training
- Computing courses
- Other training

Courses will be open to H3Africa consortium and others (where possible), recorded and video streamed

Bioinformatics Graduate Training

- MSc/PhD
  - Course curriculum
  - Co-supervision across nodes
  - Internships with external partners
  - Project approval by advisory board
- Postdocs
  - Fellowships
- Early career start-up funds
  - University-secured posts

Researcher Training

- Train-the-trainer –training faculty
- Short courses on specific topics relevant to H3Africa projects (NGS, genotyping data analysis, pop genetics, metagenomics)
- Internships for H3Africa researchers to visit bioinformatics nodes

Computing Training

- Basic computing
  - scripting/programming
- High performance
  - supported by CHPC and NCSA
- Week long summer/winter school
- Other courses
    - grant writing/management workshops

Project outcomes

- Development of a sustainable bioinformatics infrastructure (human and hardware) that can be extended beyond H3Africa
- Improved coordination and resource sharing in bioinformatics
- Training to bring wet and dry lab scientists closer together
- Increased awareness of bioinformatics applications in human health research
- Empowerment of African scientists and clinicians to analyze and exploit their own data
- Increased scientific output thus international competitiveness of African scientists
- Increased buy-in and support for bioinformatics from African institutions and funding bodies

Keeping African science on the African continent

Consortium Partners

Name	Institution	Country
Nicky Mulder	University of Cape Town	South Africa
Victor Jongeneel	NCSA	USA
Sylvester Lyantagaye	University of Dar es Salaam (UDSM)	Tanzania
Dean Everett	Malawi-Liverpool Wellcome Trust Clinical Research Programme	Malawi
Judit Kumuthini	CPGR	South Africa
Nicki Tiffin	SANBI	South Africa
Appolinaire Djikeng	ILRI	Kenya
Alia Benkahla	Institute Pasteur of Tunis	Tunisia
Ezekiel Adebiyi	Covenant University Bioinformatics Research	Nigeria
Odile Ouwe Missi	CERMES	Niger
Ozlem Tastan Bishop	Rhodes University	South Africa
Seydou Doumbia	University of Bamako	Mali
Faisal Fadlelmola	Future University	Sudan
Dan Masiga	ICIPE	Kenya
James Brandful	NMIMR	Ghana
Nzovu Ulenga	MDH	Tanzania
Ellis Owusu-Dabo	KNUST	Ghana
Oyekanmi Nash	NADBA	Nigeria
Win Hide	Harvard	USA
Hassan Ghazal	University Mohammed First	Morocco
Simani Gaseitsiwe	Botswana Harvard AIDS Institute Partnership	Botswana
Julie Makani	MUHAS	Tanzania
Yasmina Jaufeerally Fakim	SANBio	Mauritius
Scott Hazelhurst	Wits	South Africa
Jonathan Kayondo	UVRI	Uganda
Fourie Joubert	University of Pretoria	South Africa
Hugh Patterton	University of the Free State	South Africa
Ahmed Mansour Alzohairy	Zagazig University	Egypt

Funding: NIH/NGHRI H3Africa Initiative

Contact: nicola.mulder@uct.ac.za

H3ABioNet: A Pan-African Bioinformatics Network for H3Africa

CBIO Computational Biology @ UCT