LARGE DATA AND BIOMEDICAL COMPUTATIONAL PIPELINES FOR COMPLEX DISEASES

Ezekiel Adebiyi, PhD
Professor and Head, Covenant University Bioinformatics Research and CU NIH H3AbioNet node
Covenant University, Ota, Nigeria

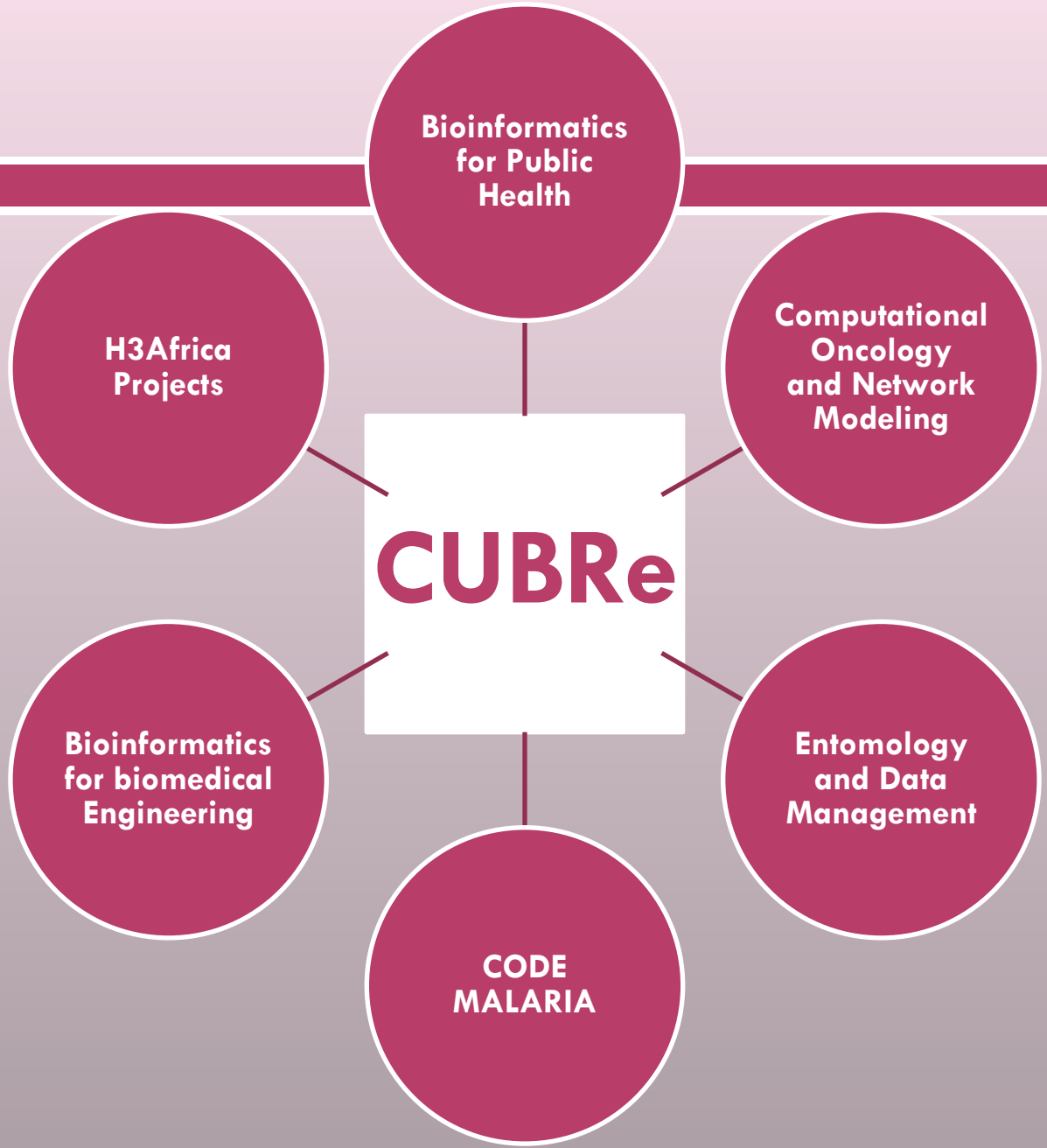
11th March 2016

A talk given at the joint workshop on promoting open science in Africa (15 March 2016, Dakar, Senegal)

Outline

- Overview of research area
- Impact of research on Africa and beyond
- Challenges in our research area
- Technologies in biomedical research
- Existing systems
- Recent project: CUBRe HPC facility accreditation for Genome Wide Association Studies (GWAS)
- Related new one (to commence!): A Federated Genomes analysis based in Memory Database Computing Platform (FEDGEN)

Overview of research area



Impact of our research to Africa & beyond

- Support for established Bio-medical institutes and companies.
- Personalized medicine based on the robust biomedical databases at CU.
- Production of high tech products for the control and final eradication of malaria starting with Nigeria.
- Support for other tropical health issues and other important health issues in the West.

Challenges in our research area

- Large data transfer and sharing
- Data accessibility
- Data security: Lack of adoption of encryption to secure patients' data on the cloud.
- Limited communication networks among research institutes, centres and Universities. (We need to connect all nodes)
- Lack of sufficient High Performance Computing machines and web services
- Lack of sufficient trained/skilled personnel

Technologies in Biomedical Research

- **Services**
  - Galaxy
- **Data transfer**
  - Globus
- **Cloud services**
  - Amazon Web Services (AWS)
  - Genomics virtual library (GVL)
  - Big data in personalized medicine

Galaxy

GENOME RESEARCH
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2005 Oct; 15(10): 1451–1455.
doi: 10.1101/gr.4086505

PMCID: PMC1240089

Galaxy: A platform for interactive large-scale genome analysis

Belinda Giardine,[1] Cathy Riemer,[1] Ross C. Hardison,[1] Richard Burhans,[1] Laura Elnitski,[2] Prachi Shah,[1,2] Yi Zhang,[1] Daniel Blankenberg,[1] Istvan Albert,[1] James Taylor,[1] Webb Miller,[1] W. James Kent,[3] and Anton Nekrutenko[1,4]

Galaxy is an open, web-based platform for data intensive biomedical research.

It is used for genomics, gene expression, genome assembly, proteomics, epigenomics, transcriptomics.

Galaxy

Globus

ELSEVIER

Computational and Structural Biotechnology Journal

Volume 13, 2015, Pages 64–74

Open Access

A case study for cloud based high throughput analysis of NGS data using the globus genomics system

Krithika Bhuvaneshwar[a], Dinanath Sulakhe[b, c], Robinder Gauba[a], Alex Rodriguez[b], Ravi Madduri[b, c], Utpal Dave[b, c], Lukasz Lacinski[b, c], Ian Foster[b, c], Yuriy Gusev[a], Subha Madhavan[a]

- **Globus Genomics:** is designed for researchers; bioinformatics core, genomics center, medical centers and health delivery providers to perform high volume genomics analysis

- **Globus Connect Server:**
  Delivers advanced file transfer and sharing capabilities to researchers on your campus no matter where their data lives.

- It makes it easy to add your lab cluster, campus research computing system or other multi-user HPC facility as a Globus endpoint

Amazon Web Services (AWS)

Journal List › PLoS Comput Biol › v.7(8); 2011 Aug › PMC3161908

PLOS | COMPUTATIONAL BIOLOGY A Peer-Reviewed, Open Access Journal

View this Article | Submit to PLOS | Get E-Mail Alerts | Contact Us

PLoS Comput Biol. 2011 Aug; 7(8): e1002147. PMCID: PMC3161908

Published online 2011 Aug 25. doi: 10.1371/journal.pcbi.1002147

Biomedical Cloud Computing With Amazon Web Services

Vincent A. Fusaro,[1,*] Prasad Patil,[1] Erik Gafni,[1] Dennis P. Wall,[1,2] and Peter J. Tonellato[1,2]

Fran Lewitter, Editor

Case Study: Creating a Whole Genome Mapping Computational Framework

- Analysis of a large amount of NGS data with the AWS
- process an entire human genome's worth of NGS reads using a short read mapping algorithm. We use the ~4 billion paired 35-base reads sequenced from a Yoruba African male.
- The African genome read set is 370 GB with individual files containing nearly 7 million reads each.
- Computation time for just one of the 303 read file pairs typically ranges from 4 to 12 hours.
- The cloud is an ideal platform for processing this dataset because the computational resources required to run these intensive mapping steps.

Genomics virtual library (GVL)

- A middleware layer of machine images, cloud management tools, and online services.
- It enables researchers to build arbitrarily sized compute clusters on demand.
- These clusters are pre-populated with fully configured bioinformatics tools, reference datasets and workflow and visualization options.
- Users can conduct analyses through web-based (Galaxy, RStudio, IPython Notebook) or command-line interfaces, and add/remove compute nodes and data resources as required.

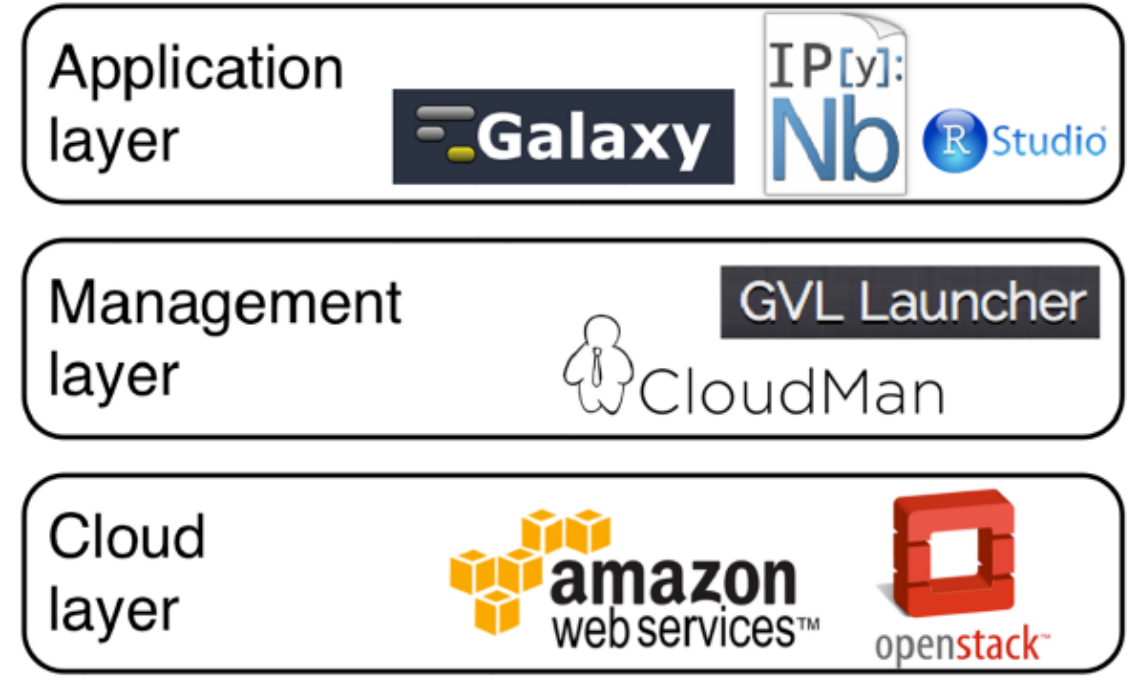
GVL

PLOS ONE

RESEARCH ARTICLE

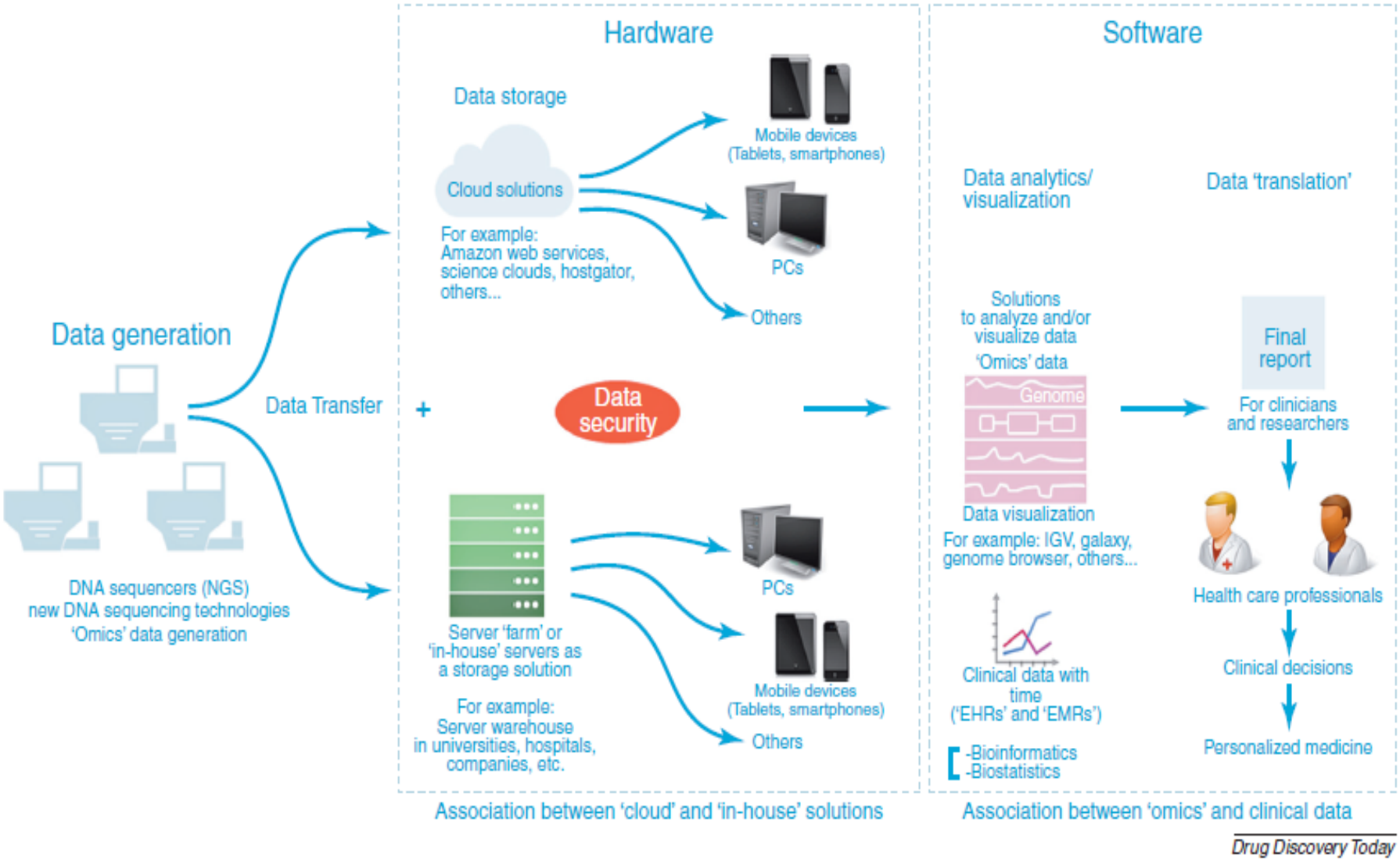
Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud

Enis Afgan[1,2,3], Clare Sloggett[1], Nuwan Goonasekera[1], Igor Makunin[4], Derek Benson[4], Mark Crowe[5], Simon Gladman[1], Yousef Kowsar[1], Michael Pheasant[4], Ron Horst[4], Andrew Lonie[1]*



Basic architecture for GVL workbench. (Afgan et al., 2015)

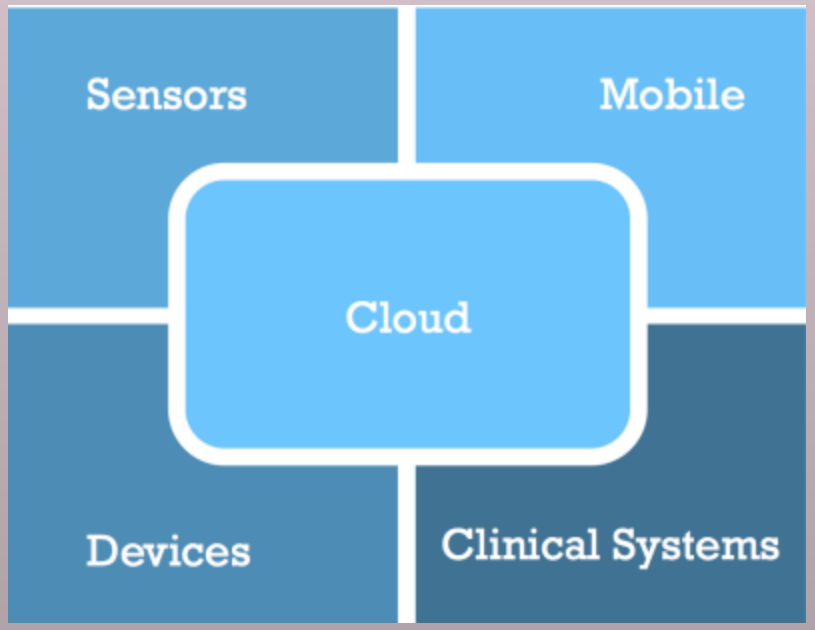
Big data in personalized medicine



Sample pipeline for personalized medicine. (Costa, 2013)

Companies with big data solutions to personalized medicine

- **Pathfinder:** They design and build connected care systems that integrate medical devices, sensors and **diagnostics with mobile applications, cloud computing and clinical systems.**



Companies with big data solutions to personalized medicine

- **NextBio:**
- A technology owned by Illumina which enables users to integrate and interpret molecular data and clinical information.
- Users can import their private experimental molecular data.
- Correlate their data with continuously curated signatures from public studies.
- Discover genomic signatures for tissues and diseases.
- Identify genes and pathways that contributes to drug resistance.

Existing systems

CHPC
1. Lease out their facility to Universities, Research Institutes and Scientific Centres to work.
2. TSESSEBE cluster (Sun).
3. Lengau Cluster (peta-scale system consisting of Dell Servers, powered Intel.
4. Galaxy for automating bioinformatics workflow

CHPC
1. The CHPC enables scientific and engineering progress in SA by providing world-class high performance computing facilities and resources.
2. Train personnels
3. Support research & human capital development.

The UCT Computational Biology Group hosts a number of bioinformatics tools, in-house and external, and services for researchers at UCT. Data analysis support can be provided for:

1. Proteomics data
2. Genotyping data
3. Next generation sequencing data
4. Genome or EST annotation
5. Microarray data

CBIO has a Galaxy installation for developing and running bioinformatics workflows and can provide support for creating custom pipelines or packaging new modules into Galaxy.

CBIO Galaxy tools

- The UCT CBIO main tool suite

UCT CBIO Tools

GET DATA

- Upload File from your computer (Tool is specifically for UCT CBIO data)

SEQUENCE PROCESSING

- Call bases and get quality values from chromatograph sequences.
- Trim fasta sequences from vector, ecoli, poly(A) and poly(T) sequences.
- EMBOSS modified vectorstrip Strips out DNA between a pair of adapter sequences
- Cluster sequences .
- Assemble sequences .
- Predict peptide sequences .
- Predict peptide sequences from metagenomic data .

BLAST UTILITIES

WITS BIOINFORMATICS

- **Tools:** Wits has a number of on-line tools available for bioinformatics. Their wEMBOSS server is used for training as well as by researchers who need to use bioinformatics tools.
- **High-Performance Computing:** Wits run a research computer cluster which is available to members of the bioinformatics community. The cluster contains 150 cores and roughly 70TB of data storage. They have some large memory machines (128-256GB of RAM). This is also a node on the SA National Compute Grid.
- **Databases:** Wits mirror some of the key databases including Genbank and PDB and they can mirror or host other data bases.

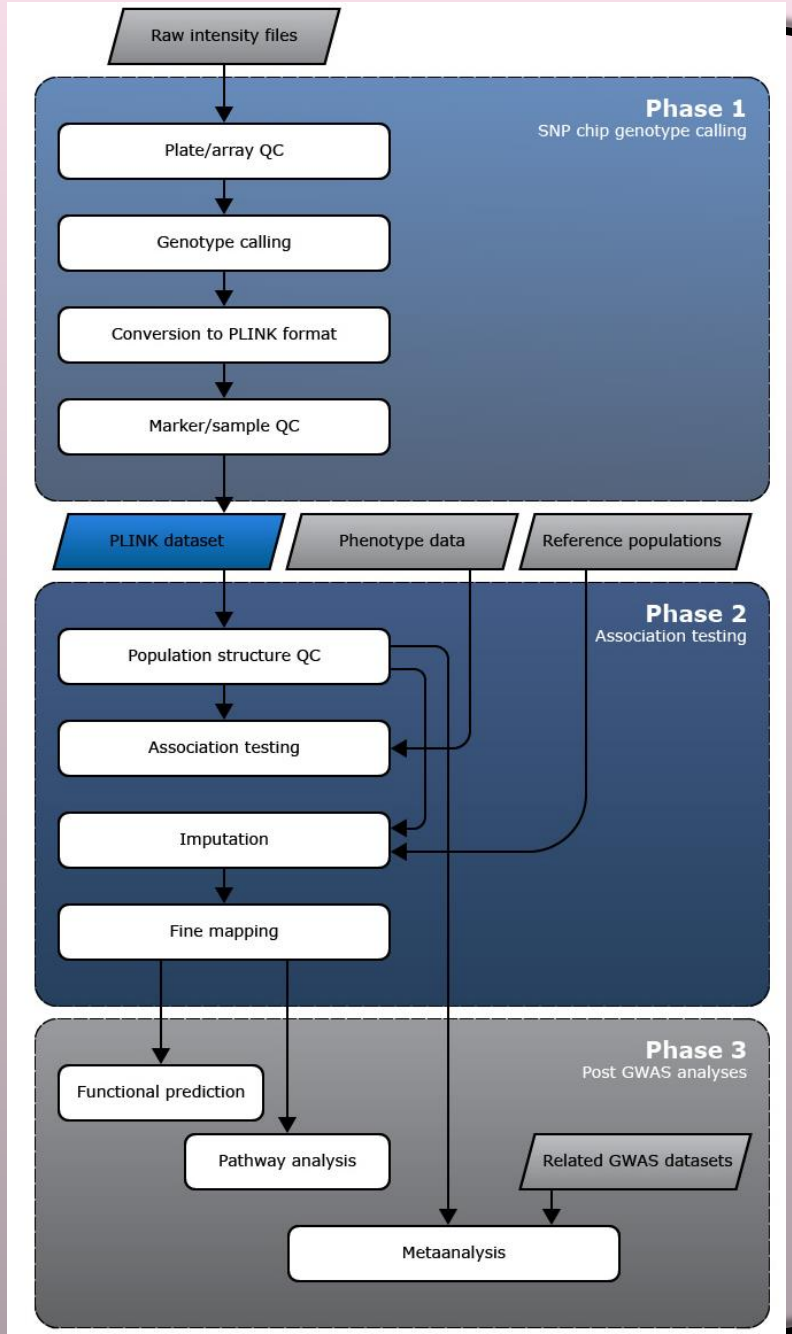
Recent project: CUBRe HPC facility ACCREDITATION for GWAS analysis

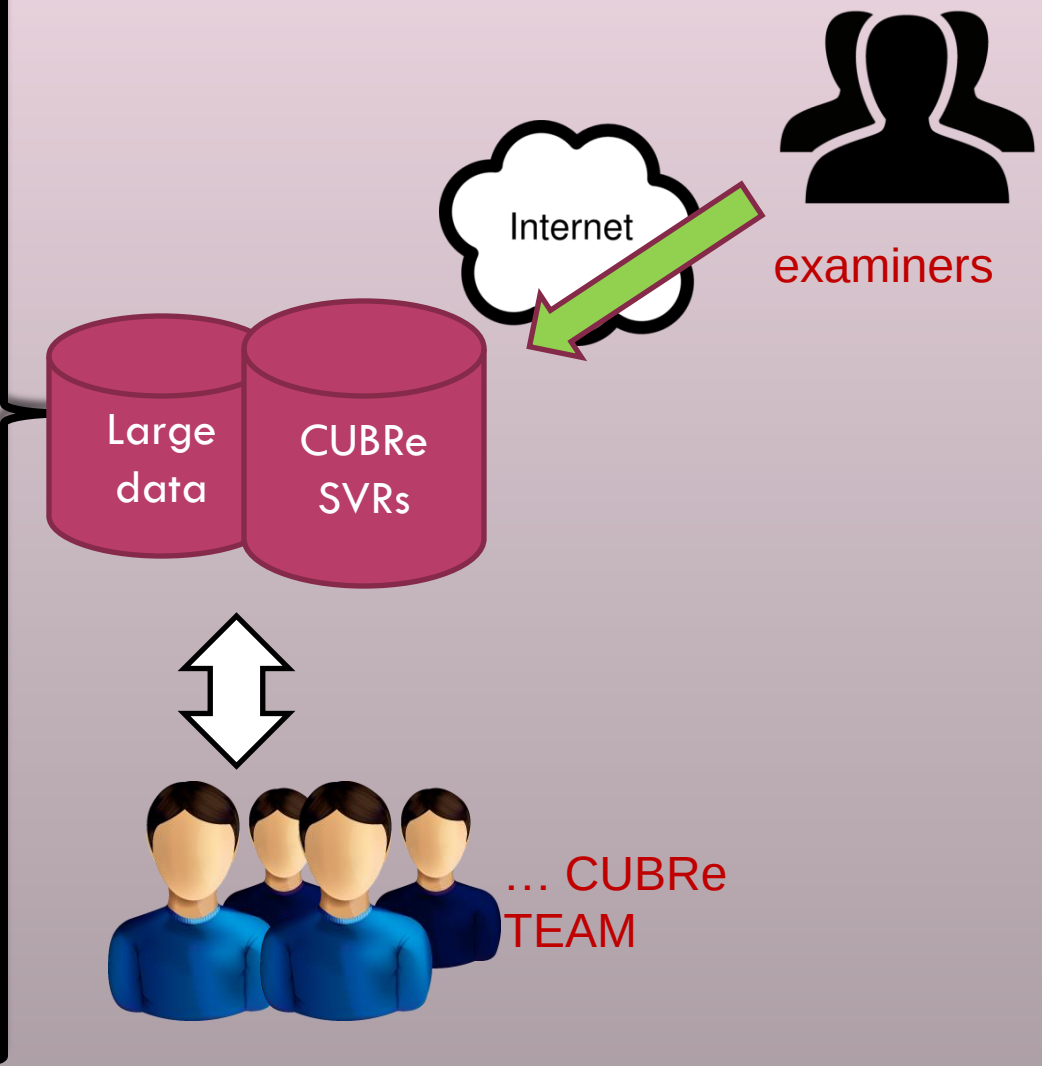
- The CUBRe accreditation for GWAS analysis included the use of pipelines, workflows, protocols, and HPC facilities to analyze GWA datasets.
- GWAS is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease.
- Genetic associations found can help researchers develop better strategies to detect, treat and prevent the disease.

CUBRe HPC facility Accreditation for GWAS analysis

- CUBRe HPC facilities used for the accreditation include 52 CPU cores, 5TB, and 230GB ram.
- The analysis included 3 phases: SNP chip genotype calling, Association testing and Post GWAS analysis.
- Data included 384 cels files which was about 8GB for phase 1.
- Phase 2 dataset included 716 people (203 males, 512 females, 1 ambiguous) and 194432 variants from Massai tribe in Kenya.

Pipeline for GWAS analysis





RESULTS

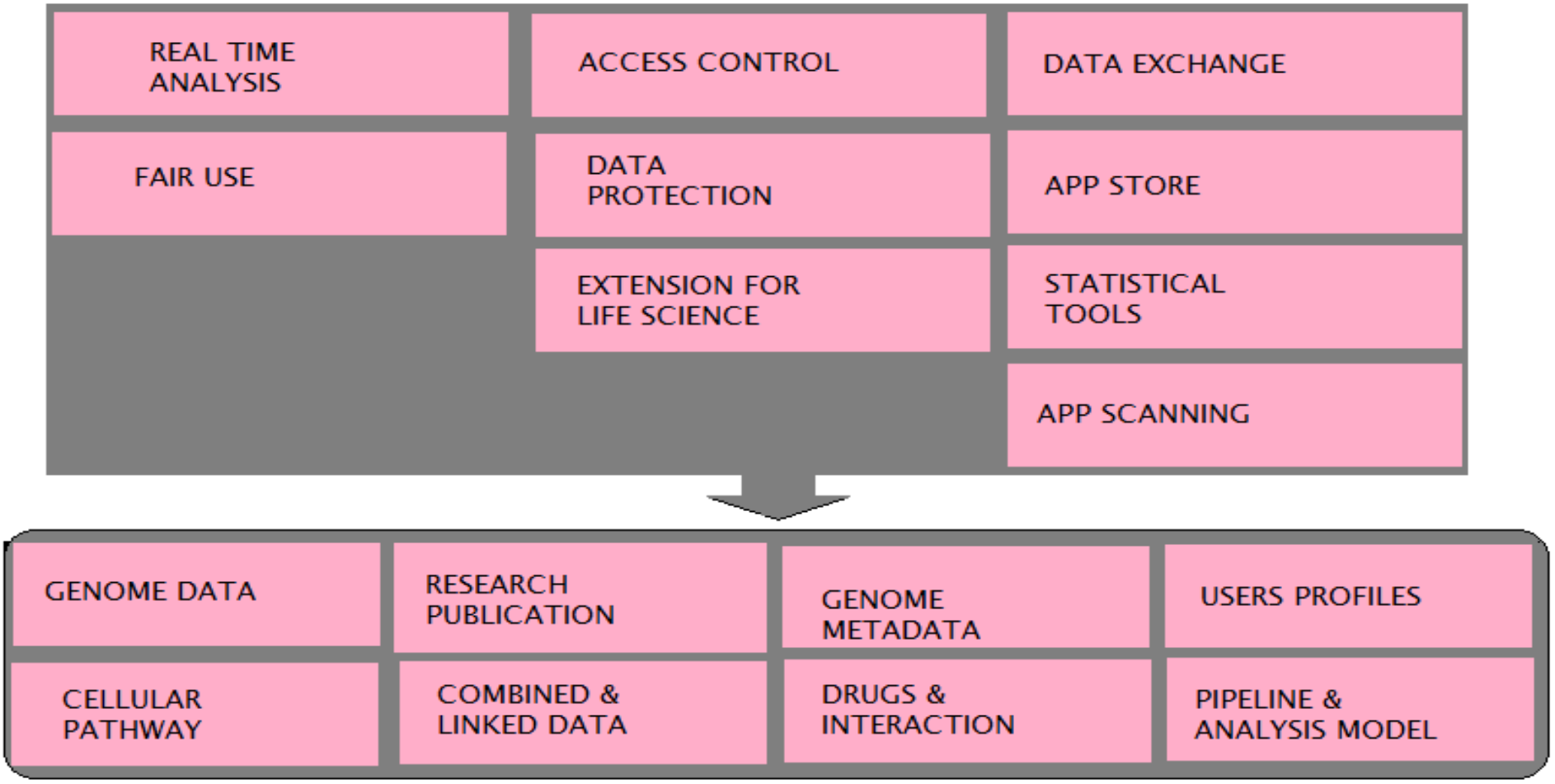
- We identified 24 biologically significant SNPs that have been associated with 5 pathways which have been ranked and mapped.

- A pathway that was highly implicated was leukocyte transendothelial migration in rheumatoid and osteoarthritis.
- Finalizing a manuscript on this for publication.

Related new one (to commence!): A Federated Genomes analysis based in Memory Database Computing Platform (FEDGEN)

- Distributed Heterogeneous Data Sources: Human genome and proteome, Hospital Inf. Sys, Patient records, Prescription data, Clinical trials, Medical sensor data (for example, scan of a single organ in 1s creates 10GB of raw data) and PubMed Database.
- Target providing in the 1st instance in WA, improve Health Care free services on mobile devices, by delivering a) Health Education, b) Medication efficiency and c) Enhanced early disease diagnosis.
- The intention is to "improve the health of our people".

A Federated Genomes analysis based in Memory Database Computing Platform (FEDGEN) - workflow



Acknowledgements

- Covenant University, Ota, Nigeria
- H3ABioNet supported by NHGRI grant number U41HG006941
- Covenant University Bioinformatics Research (CUBRe) group members (please see cubre.covenantuniversity.edu.ng)

THANK YOU FOR YOUR ATTENTION

DANKESCHOEN

ESEO