FGAI4H-O-046

Berlin, 31 May – 2 June 2022

Source:	H3ABioNet (South Africa)
Title:	Workshop: WG-CO - Development of genomics data infrastructure and standards within Africa
Purpose:	Discussion

Contact: Sumir Panji E-mail: sumir.panji@uct.ac.za
H3ABioNet
South Africa

Abstract: This PPT contains a presentation from the WG-CO workshop on "Equitable data infrastructures to support equitable and effective pandemic intelligence".

H3ABioNet

Pan African Bioinformatics Network for H3Africa

DEVELOPMENT OF GENOMICS DATA INFRASTRUCTURE AND STANDARDS WITHIN AFRICA

Sumir Panji

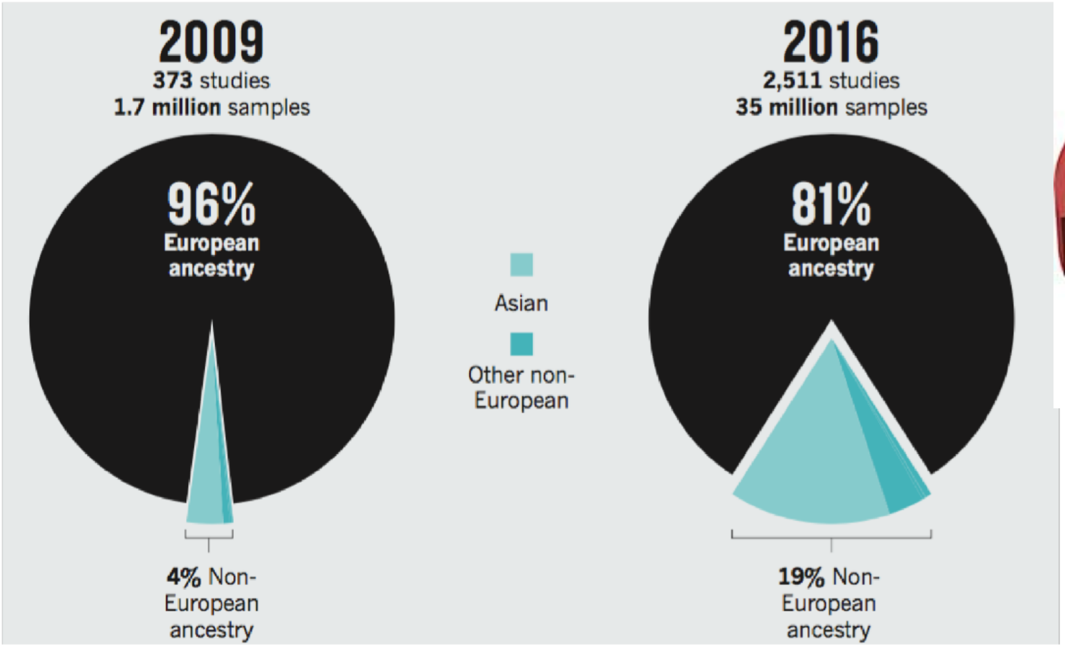
ITU-WHO Focus Group on Artificial Intelligence for Health
30th May 2022

www.h3abionet.org #h3abionet

Presentation Overview

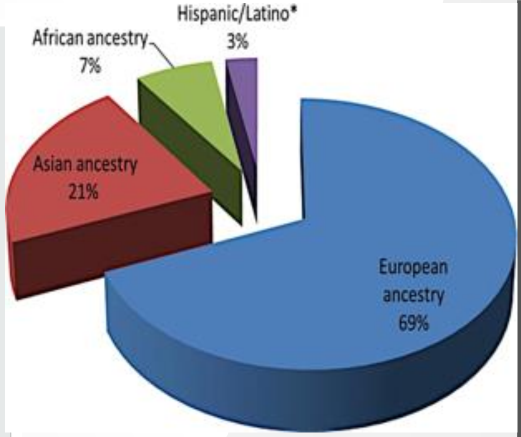
- Introduction
- H3Africa and H3ABioNet
- Overview of H3ABioNet
- Development of data standards
- Development of an H3Africa data archive and data catalogue
- Computational tool development
- Findable, Accessible, Interoperable and Resusable (FAIR) and H3Africa Data and meta-data
- Examples of SARS works and early release of data
- Data Science Initiative Africa (DS-A Africa)
- eLwazi Open Data Science Platform
- Conlcusions

The Genomics Revolution and Africa



0.57% 3%

African Ancestry
(Majority African-American)



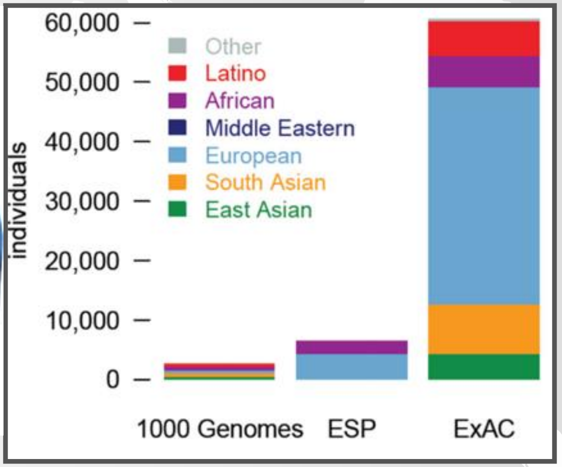
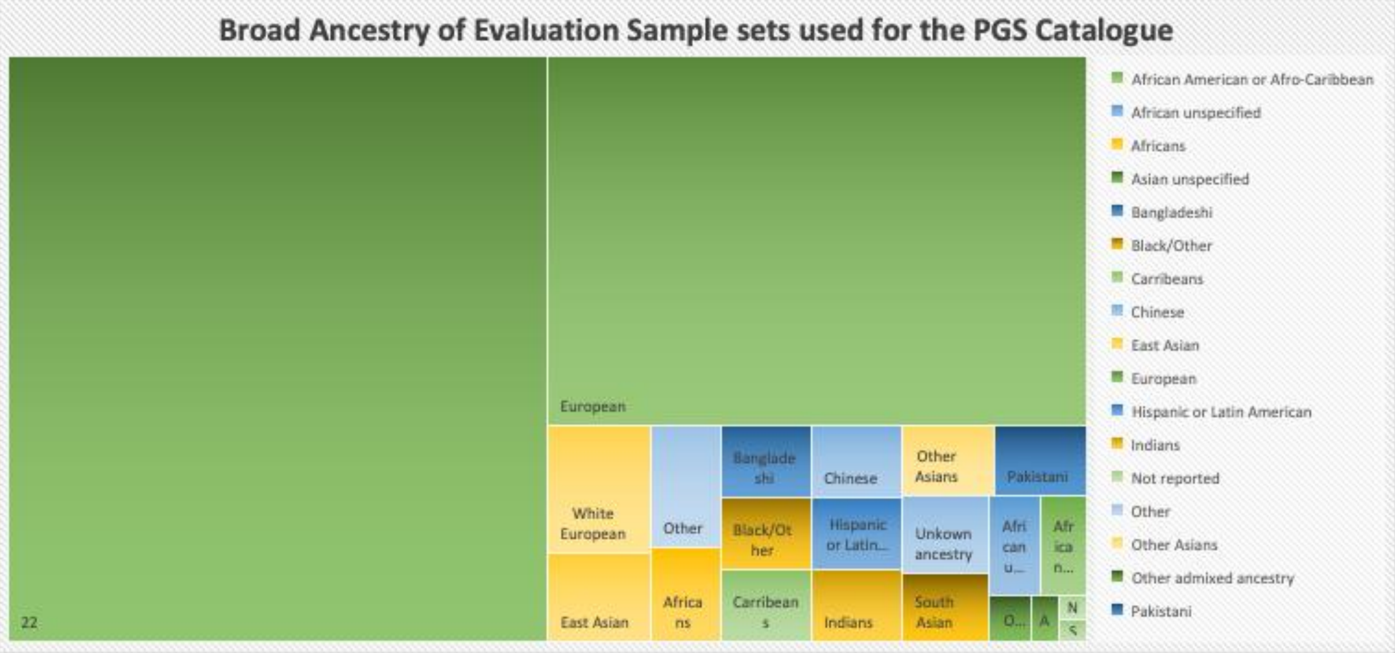


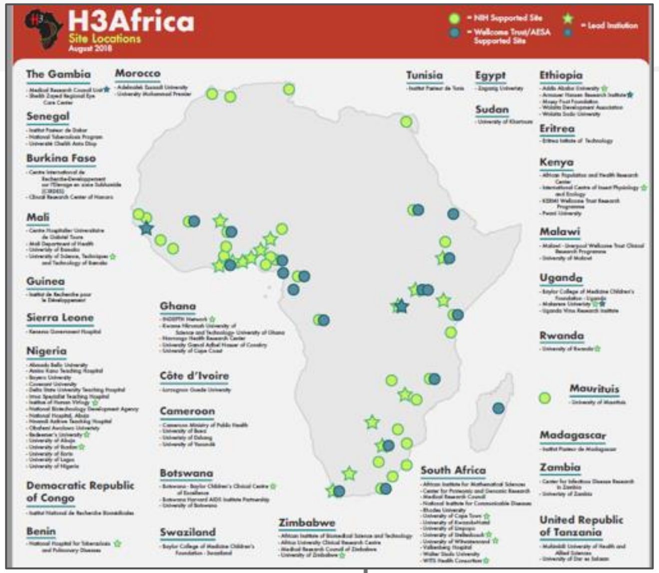
Image credits and reference: Popejoy AB, Fullerton SM. Nature. 2016 Oct 13;538(7624):161-164 https://doi.org/10.1038/538161a

Slide credits - Prof. Michèle Ramsay, SBIMB, Wits and Prof. Nicky Mulder UCT CBIO



Data summary from the PGS Catalogue accessed in August 2020: https://www.pgscatalog.org/

The H3Africa Consortium



The Human Heredity and Health in Africa (H3Africa) Initiative aims to facilitate a contemporary research approach to the study of genomics and environmental determinants of common diseases with the goal of improving the health of African populations." (http://h3africa.org/)

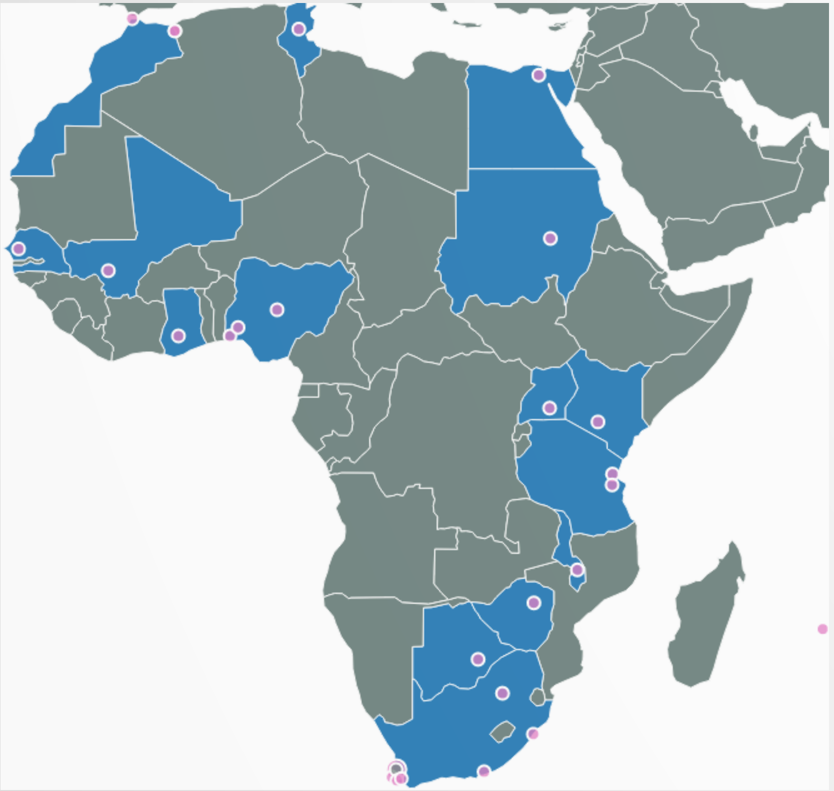
- 7 Ethical, Legal and Social Issues (ELSI) Projects
- 15 Collaborative Centers
- 3 Biorepository Projects
- 16 Research Projects
- H3ABioNet
- 4 Global Health Bioinformatics Degree Programs

- 48 H3Africa projects funded
- 30 African countries involved in phase 2
- 500+ consortium members
- >160+ publications so far
- 170 million USD invested

http://h3africa.org/consortium/projects

Pan African Bioinformatics Network for H3Africa (H3ABioNet)

- Pan African Bioinformatics Network to develop bioinformatics capacity in Africa and support the H3Africa research projects

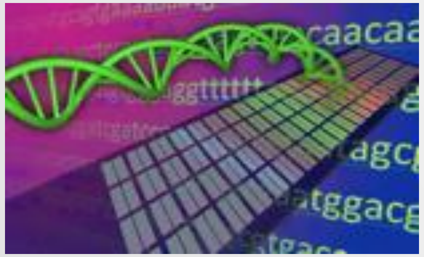


- 28 Institutions
- 16 African Countries

Overview of H3Africa Data

Phenotype data (associated with genotype data)
- Demographic information
- Anthropometric data
- Disease and health related phenotype data

Genetic Variation data human and pathogen
- Genomic Sequence data
- Whole genome exome and targeted sequencing

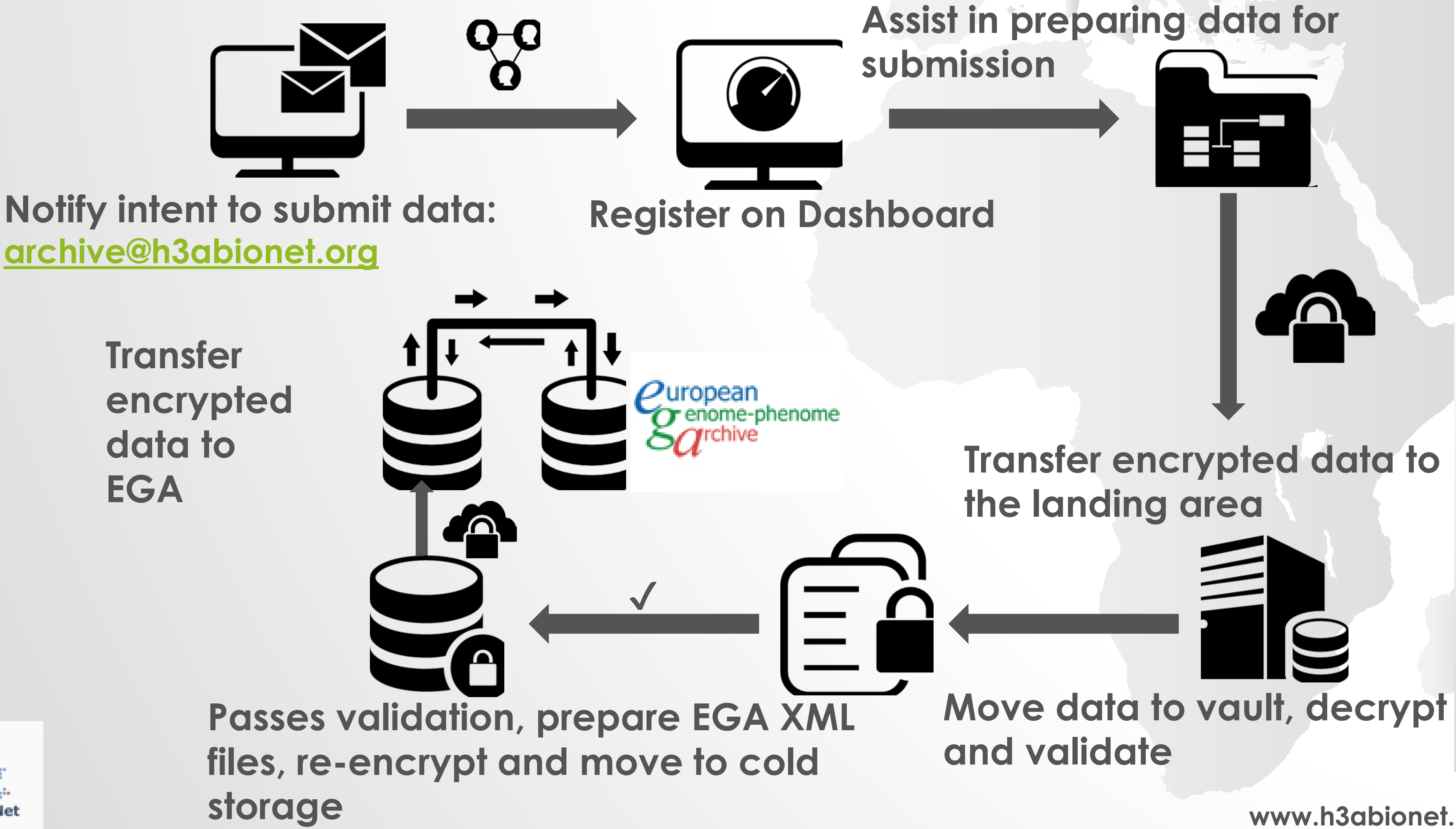


Genotyping chip array data
- ~55,000 samples to be run on an H3Africa African custom chip
- Raw intensity files linked

Microbiome sequence data
- Patient/sample phenotypes
- Non-human 16S rRNA sequence data for microbiome
- Non-human full genome sequence data for microbiome
- Possible human sequence contamination

Image credits: National Human Genome Research Institute (https://www.genome.gov/imagegallery/)

H3Africa Data Archive overview



Notify intent to submit data: archive@h3abionet.org

Register on Dashboard

Assist in preparing data for submission

Transfer encrypted data to the landing area

Move data to vault, decrypt and validate

Passes validation, prepare EGA XML files, re-encrypt and move to cold storage

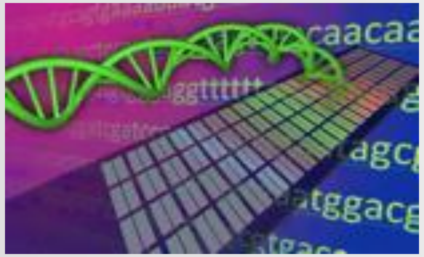
Transfer encrypted data to EGA

- 14 African Genomic Datasets in the archive
- 134.9 Terabytes of African Genomics data
- 8 Datasets QC-ed, validated and submitted to EGA
- 77.3 Terabytes transferred for storage at the EGA

www.h3abionet.org #h3abionet

Specific Data submission files and checks

- ✓ FastQ files linked with de-identified participant ID (minus adapters, linkers, barcodes)
- ✓ Binary Alignment files (BAMs, de-multiplexed) – linked with participant de-identified ID
- ✓ Variant calling files (VCFs)
- ✓ Phenotypic data
- ✓ Study type e.g. case / control, trio, longitudinal
- ✓ Sequencing platform / technology
- ✓ Mapping file indicating relationship between files



- ✓ Chip platform and model / name
- ✓ Raw intensity files linked with de-identified participant IDs (IDATs, CELs)
- ✓ Manifest file describing SNP or probe content on chip
- ✓ Software used for calling the genotypes
- ✓ Final reports and analyses files generated
- ✓ Phenotypic data
- ✓ Study type e.g. case / control, trio, longitudinal
- ✓ Mapping file indicating relationship between files

- ✓ FastQ files linked with de-identified participant ID (minus adapters, linkers, barcodes)
- ✓ Sequencing platform
- ✓ Final analyses BIOM files (at minimum contain OTUs)
- ✓ Phenotypic data
- ✓ Study type e.g. case / control, longitudinal
- ✓ Mapping file indicating relationship between files

Move data to the vault, decrypt and validate:

- ? Do the checksums match
- ? Are there mapping files present
- ? Do the number of samples match what is expected
- ? Are all the files present for each de-identified participant ID
- ? Is there a mismatch between participant IDs and files
- ? Do all the files have phenotypic data present
- ? Is there a dataset summary description and study abstract present

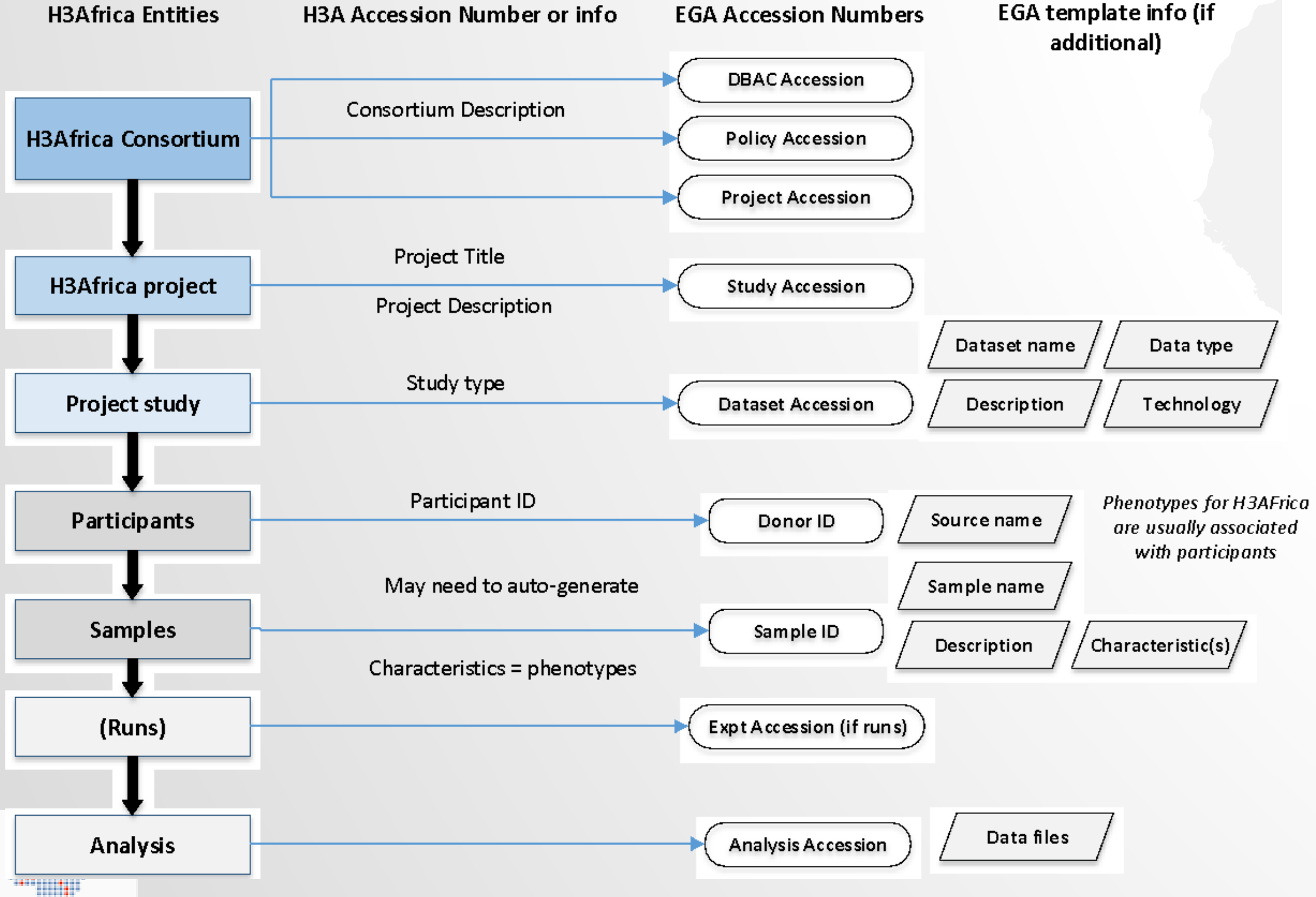
Passes validation:

- Map data to EGA XML schemas
- Re-encrypt the data and move to cold storage
- Submit data to EGA at pre-determined date

Image credits: National Human Genome Research Institute (https://www.genome.gov/imagegallery/)

Mapping and submission of H3Africa data

Passes validation:
- Map data to EGA XML schemas
- Re-encrypt the data and move to cold storage
- Submit data to EGA at pre-determined date



H3Africa Entities | H3A Accession Number or info | EGA Accession Numbers | EGA template info (if additional)

H3Africa Consortium → Consortium Description → DBAC Accession, Policy Accession, Project Accession

H3Africa project → Project Title, Project Description → Study Accession

Project study → Study type → Dataset Accession → Dataset name, Data type, Description, Technology

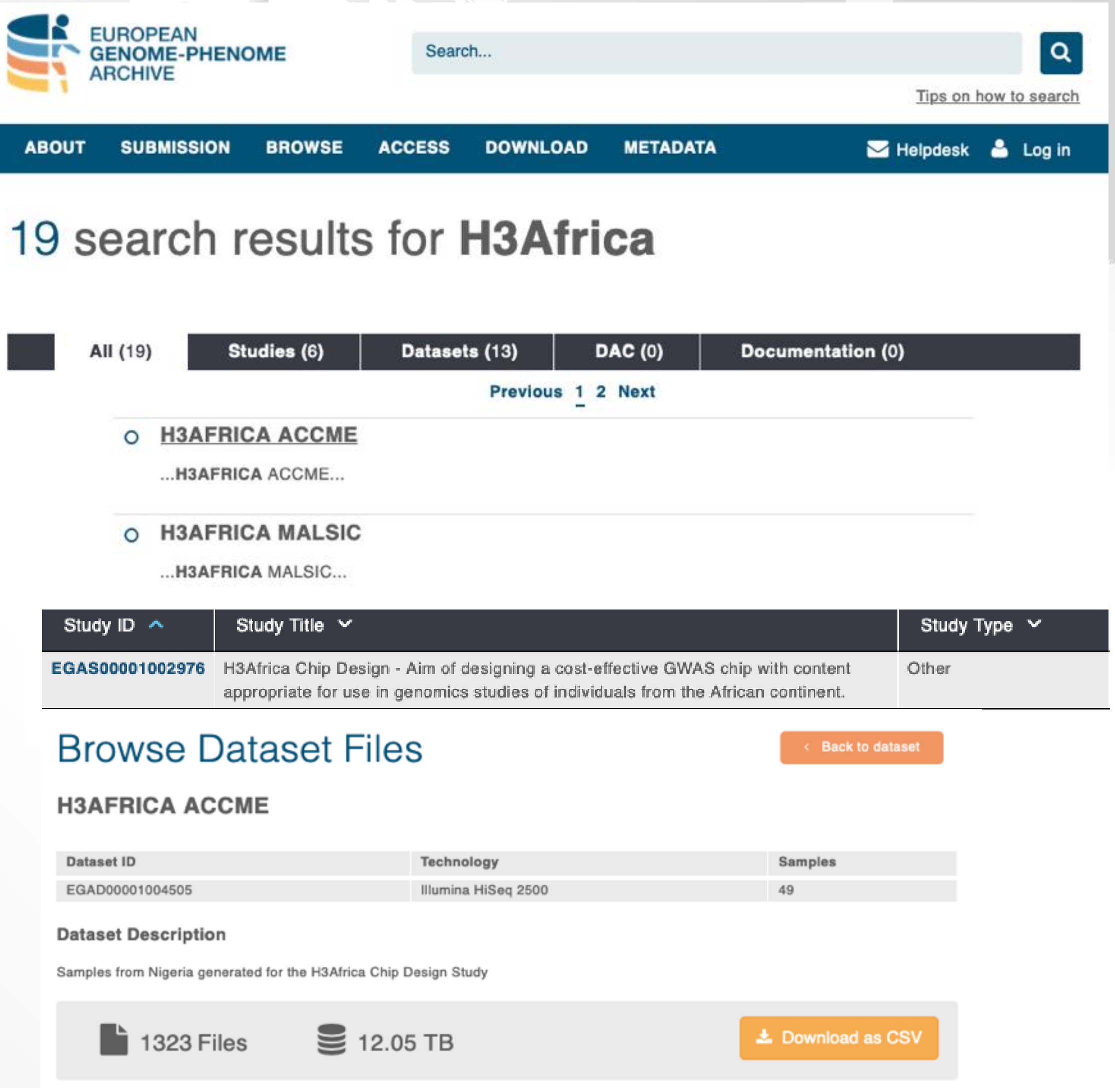
Participants → Participant ID → Donor ID → Source name

Phenotypes for H3AFrica are usually associated with participants

Samples → May need to auto-generate, Characteristics = phenotypes → Sample ID → Sample name, Description, Characteristic(s)

(Runs) → Expt Accession (if runs)

Analysis → Analysis Accession → Data files



EUROPEAN GENOME-PHENOME ARCHIVE

19 search results for **H3Africa**

- H3AFRICA ACCME
  ...H3AFRICA ACCME...
- H3AFRICA MALSIC
  ...H3AFRICA MALSIC...

Study ID	Study Title	Study Type
EGAS00001002976	H3Africa Chip Design - Aim of designing a cost-effective GWAS chip with content appropriate for use in genomics studies of individuals from the African continent.	Other

Browse Dataset Files

H3AFRICA ACCME

Dataset ID	Technology	Samples
EGAD00001004505	Illumina HiSeq 2500	49

Dataset Description

Samples from Nigeria generated for the H3Africa Chip Design Study

1323 Files 12.05 TB

https://ega-archive.org/search-results.php?query=H3Africa

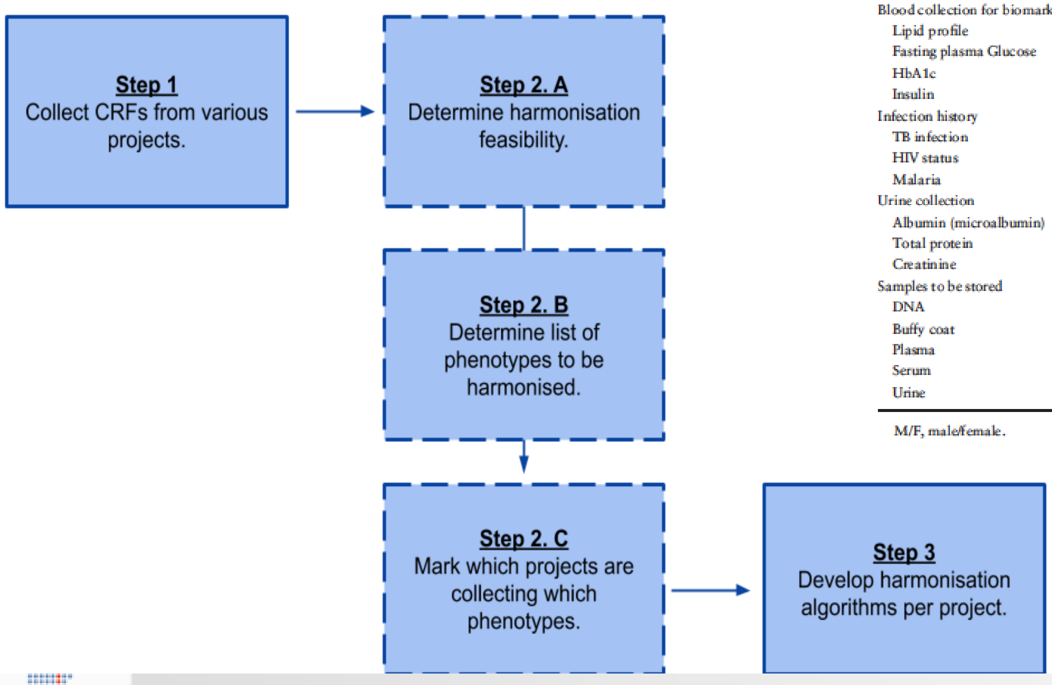
Phenotype harmonization

Meta-data (phenotype) data is collected via case report forms (CRFs)

(1) Age & (2) Sex
(3) Country of birth
(4) Current residence
(5) Native language
(6) Ethno-linguistic/tribal affiliation
(7) Country of birth of father and mother
(8) Native language of father and mother
(9)Ethno-linguistic/tribal affiliation of mother and father
(10) Height
(11) Weight
(12) Current medications
(13) Smoking history
(14) Alcohol history

	Same question – data coded in different ways	Same variable – collected in different ways
Project 1 CRF	Female	Daily units
Project 2 CRF	Woman	Weekly units
Project 3 CRF	1	User defined time period

Phenotype harmonization



Step 1
Collect CRFs from various projects.

Step 2. A
Determine harmonisation feasibility.

Step 2. B
Determine list of phenotypes to be harmonised.

Step 2. C
Mark which projects are collecting which phenotypes.

Step 3
Develop harmonisation algorithms per project.

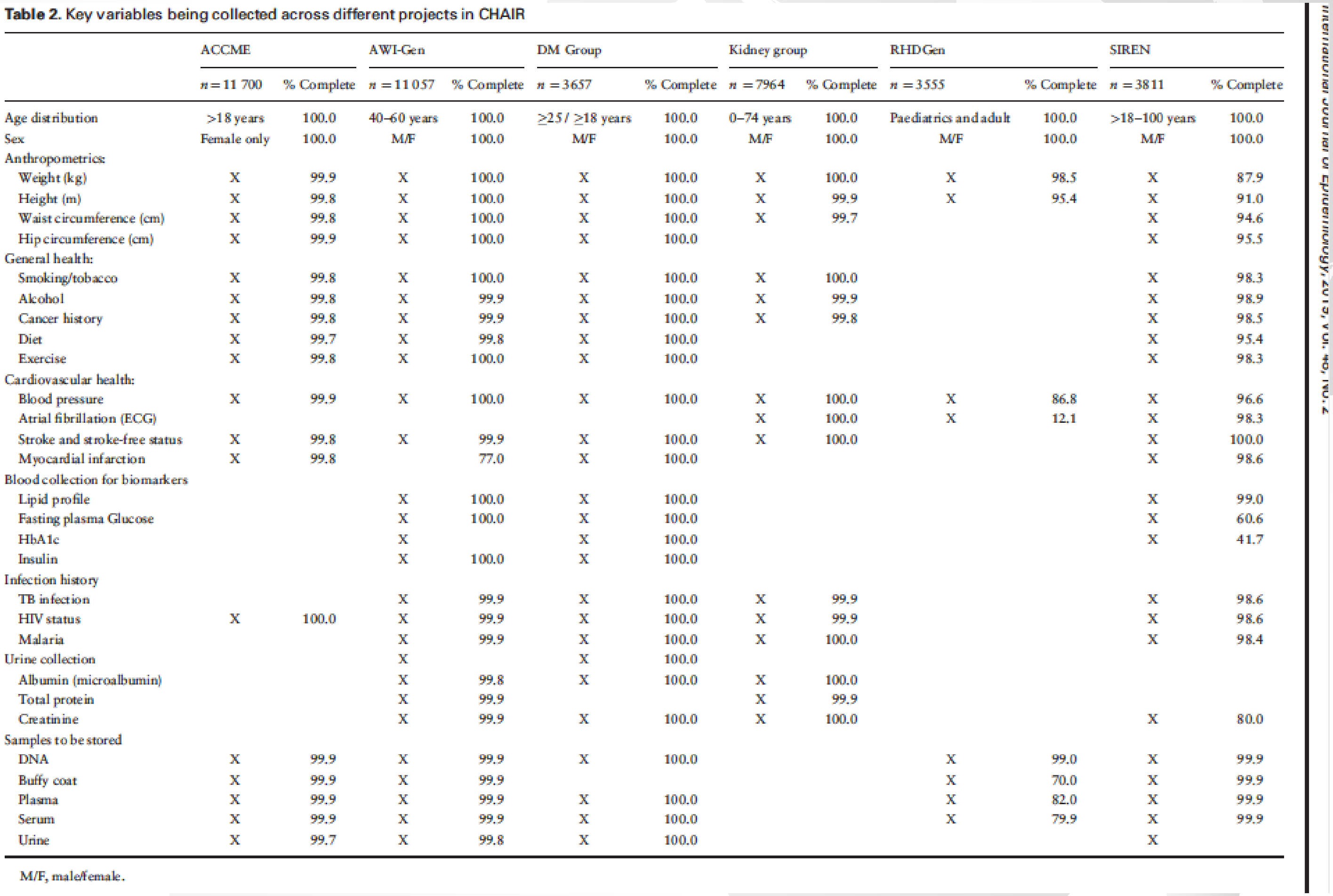


Table 2. Key variables being collected across different projects in CHAIR

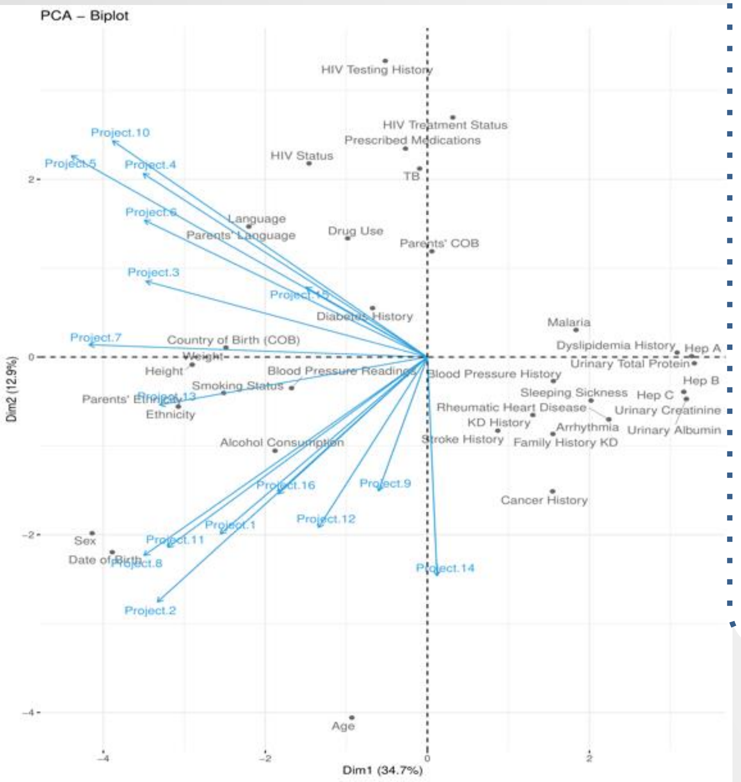
	ACCME		AWI-Gen		DM Group		Kidney group		RHDGen		SIREN	
	n = 11 700	% Complete	*n* = 11 057	% Complete	*n* = 3657	% Complete	*n* = 7964	% Complete	*n* = 3555	% Complete	*n* = 3811	% Complete
Age distribution	>18 years	100.0	40–60 years	100.0	≥25 / ≥18 years	100.0	0–74 years	100.0	Paediatrics and adult	100.0	>18–100 years	100.0
Sex	Female only	100.0	M/F	100.0	M/F	100.0	M/F	100.0	M/F	100.0	M/F	100.0
Anthropometrics:												
Weight (kg)	X	99.9	X	100.0	X	100.0	X	100.0	X	98.5	X	87.9
Height (m)	X	99.8	X	100.0	X	100.0	X	99.9	X	95.4	X	91.0
Waist circumference (cm)	X	99.8	X	100.0	X	100.0	X	99.7			X	94.6
Hip circumference (cm)	X	99.9	X	100.0	X	100.0					X	95.5
General health:												
Smoking/tobacco	X	99.8	X	100.0	X	100.0	X	100.0			X	98.3
Alcohol	X	99.8	X	99.9	X	100.0	X	99.9			X	98.9
Cancer history	X	99.8	X	99.9	X	100.0	X	99.8			X	98.5
Diet	X	99.7	X	99.8	X	100.0					X	95.4
Exercise	X	99.8	X	100.0	X	100.0					X	98.3
Cardiovascular health:												
Blood pressure	X	99.9	X	100.0	X	100.0	X	100.0	X	86.8	X	96.6
Atrial fibrillation (ECG)							X	100.0	X	12.1	X	98.3
Stroke and stroke-free status	X	99.8	X	99.9	X	100.0	X	100.0			X	100.0
Myocardial infarction	X	99.8		77.0	X	100.0					X	98.6
Blood collection for biomarkers												
Lipid profile			X	100.0	X	100.0					X	99.0
Fasting plasma Glucose			X	100.0	X	100.0					X	60.6
HbA1c			X		X	100.0					X	41.7
Insulin			X	100.0	X	100.0						
Infection history												
TB infection			X	99.9	X	100.0	X	99.9			X	98.6
HIV status	X	100.0	X	99.9	X	100.0	X	99.9			X	98.6
Malaria			X	99.9	X	100.0	X	100.0			X	98.4
Urine collection			X		X	100.0						
Albumin (microalbumin)			X	99.8	X	100.0	X	100.0				
Total protein			X	99.9			X	99.9				
Creatinine			X	99.9	X	100.0	X	100.0			X	80.0
Samples to be stored												
DNA	X	99.9	X	99.9	X	100.0			X	99.0	X	99.9
Buffy coat	X	99.9	X	99.9					X	70.0	X	99.9
Plasma	X	99.9	X	99.9	X	100.0			X	82.0	X	99.9
Serum	X	99.9	X	99.9	X	100.0			X	79.9	X	99.9
Urine	X	99.7	X	99.8	X	100.0					X	

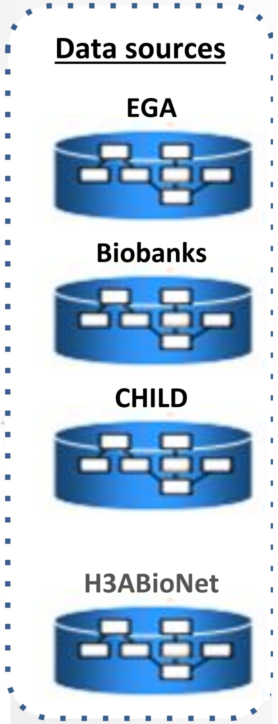
M/F, male/female.

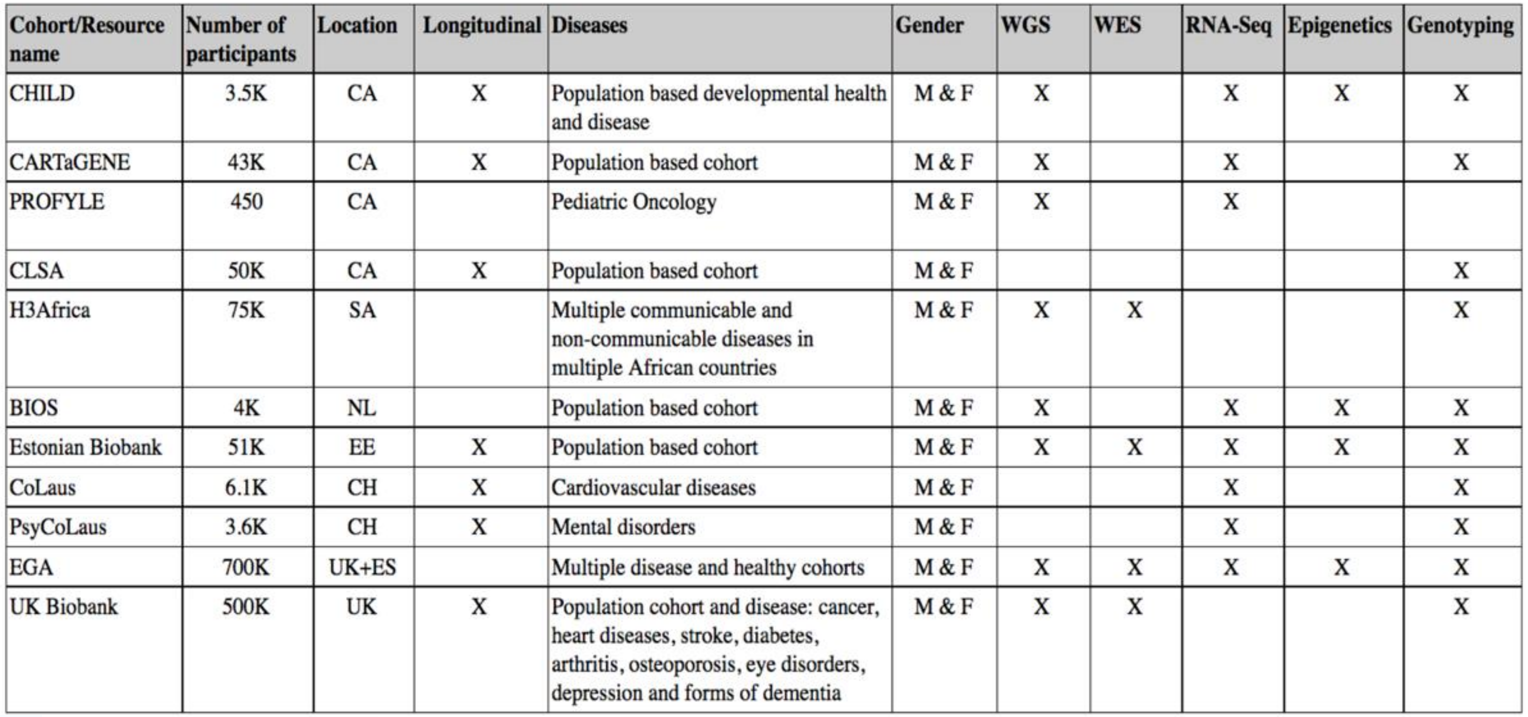
International Journal of Epidemiology, 2019, Vol. 48, No. 2

Table reference: Owolabi MO, Akpa OM, Made F, Adebamowo SN, Ojo A, Adu D, Motala AA, Mayosi BM, Ovbiagele B, Adebamowo C, Tayo B, Rotimi C, Akinyemi R, Gebregziabher M, Sarfo F, Wahab KW, Parekh RS, Engel ME, Chisala C, Peprah E, Mensah G, Wiley K, Troyer J, Ramsay M; as members of the CVD Working Group of the H3Africa Consortium. Data Resource Profile: Cardiovascular H3Africa Innovation Resource (CHAIR). Int J Epidemiol. 2019 Apr 1;48(2):366-367g. doi: 10.1093/ije/dyy261. PMID: 30535409; PMCID: PMC6469307.

Phenotype harmonization







Cohort/Resource name	Number of participants	Location	Longitudinal	Diseases	Gender	WGS	WES	RNA-Seq	Epigenetics	Genotyping
CHILD	3.5K	CA	X	Population based developmental health and disease	M & F	X		X	X	X
CARTaGENE	43K	CA	X	Population based cohort	M & F	X		X		X
PROFYLE	450	CA		Pediatric Oncology	M & F	X		X		
CLSA	50K	CA	X	Population based cohort	M & F					X
H3Africa	75K	SA		Multiple communicable and non-communicable diseases in multiple African countries	M & F	X	X			X
BIOS	4K	NL		Population based cohort	M & F	X		X	X	X
Estonian Biobank	51K	EE	X	Population based cohort	M & F	X	X	X	X	X
CoLaus	6.1K	CH	X	Cardiovascular diseases	M & F			X		X
PsyCoLaus	3.6K	CH	X	Mental disorders	M & F			X		X
EGA	700K	UK+ES		Multiple disease and healthy cohorts	M & F	X	X	X	X	X
UK Biobank	500K	UK	X	Population cohort and disease: cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia	M & F	X	X			X

CINECA

HPO

MeSH

SICKLE CELL DISEASE ONTOLOGY SCDO

DISEASE ONTOLOGY

efo Experimental Factor Ontology

Human Genomic Data Privacy

H3Africa rich source of meta-data (phenotypes)

(1) Age & (2) Sex
(3) Country of birth
(4) Current residence
(5) Native language
(6) Ethno-linguistic/tribal affiliation
(7) Country of birth of father and mother
(8) Native language of father and mother
(9)Ethno-linguistic/tribal affiliation of mother and father
(10) Height
(11) Weight
(12) Current medications
(13) Smoking history
(14) Alcohol history

- Combination of phenotype and genetic data makes it possible to identify different populations and individuals – controlled access

Image credits: National Human Genome Research Institute (https://www.genome.gov/imagegallery/)

H3Africa data sharing and access policy

Balance between ensuring adequate safeguards to protect participants, while not being a barrier for scientists to advance research:

- Maximizing the availability of research data, in a timely and responsible manner.
- Protecting the rights and privacy of human subjects who participated in research studies.
- Recognizing the scientific contribution of researchers who generated the data.
- Considering the nature and ethics of the research proposed in establishing the timely release of data, and mechanisms of data sharing.
- Promoting deposition of genomic data in existing community data repositories whenever possible

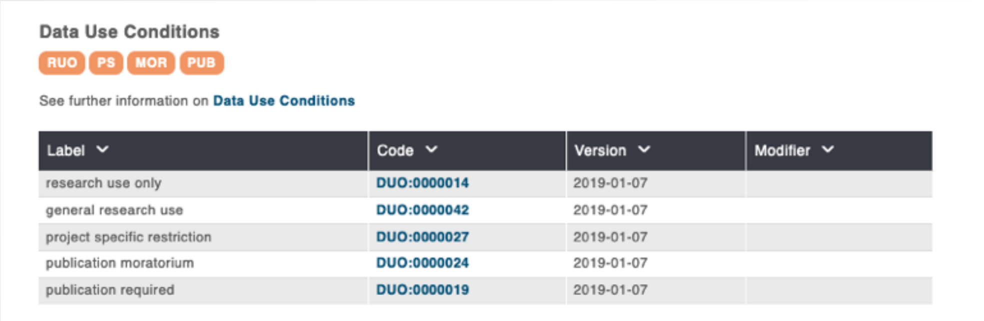
http://h3africa.org/images/DataSARWG_folders/FinalDocsDSAR/H3Africa%20Consortium%20Data%20Access%20%20Release%20Policy%20Aug%202014.pdf

H3Africa Biospecimen and Data Catalogue

https://catalog.h3africa.org/

H3Africa Biospecimen and Data Catalogue

https://catalog.h3africa.org/

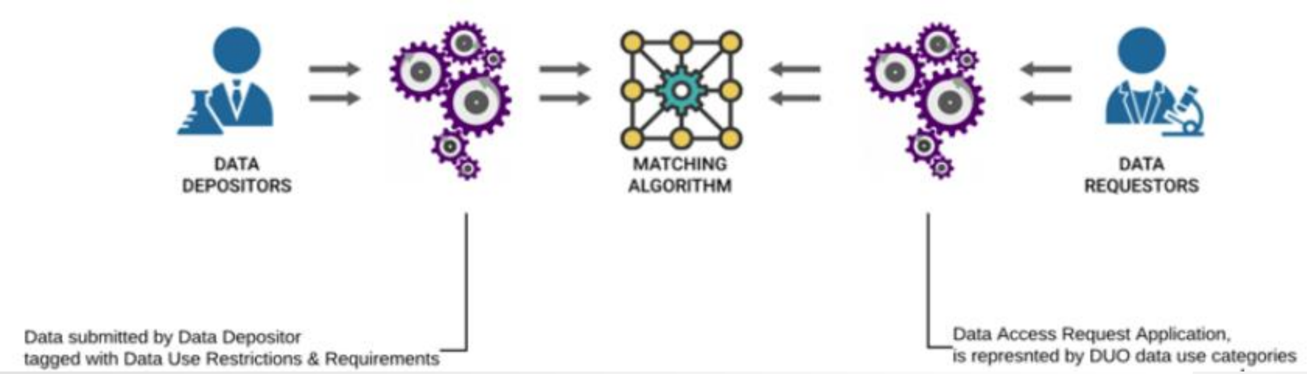


Data Use Conditions

See further information on **Data Use Conditions**

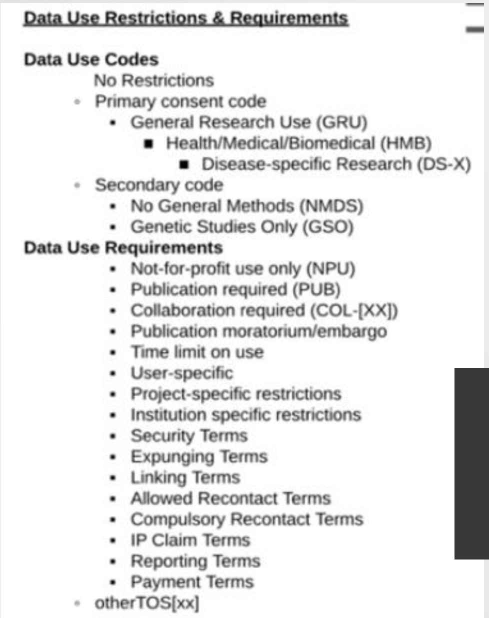
Label	Code	Version	Modifier
research use only	DUO:0000014	2019-01-07	
general research use	DUO:0000042	2019-01-07	
project specific restriction	DUO:0000027	2019-01-07	
publication moratorium	DUO:0000024	2019-01-07	
publication required	DUO:0000019	2019-01-07	

Data Use Ontology



Data submitted by Data Depositor tagged with Data Use Restrictions & Requirements

Data Access Request Application, is represnted by DUO data use categories



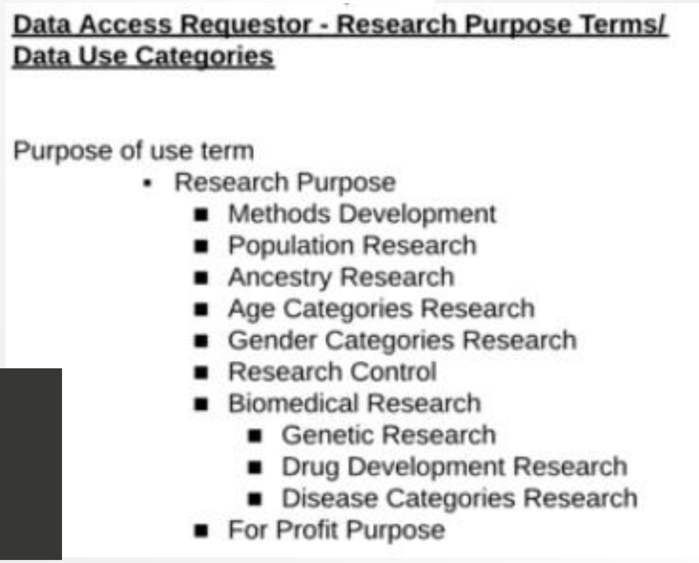
Data Use Restrictions & Requirements

Data Use Codes

No Restrictions
- Primary consent code
  - General Research Use (GRU)
    - Health/Medical/Biomedical (HMB)
      - Disease-specific Research (DS-X)
- Secondary code
  - No General Methods (NMDS)
  - Genetic Studies Only (GSO)

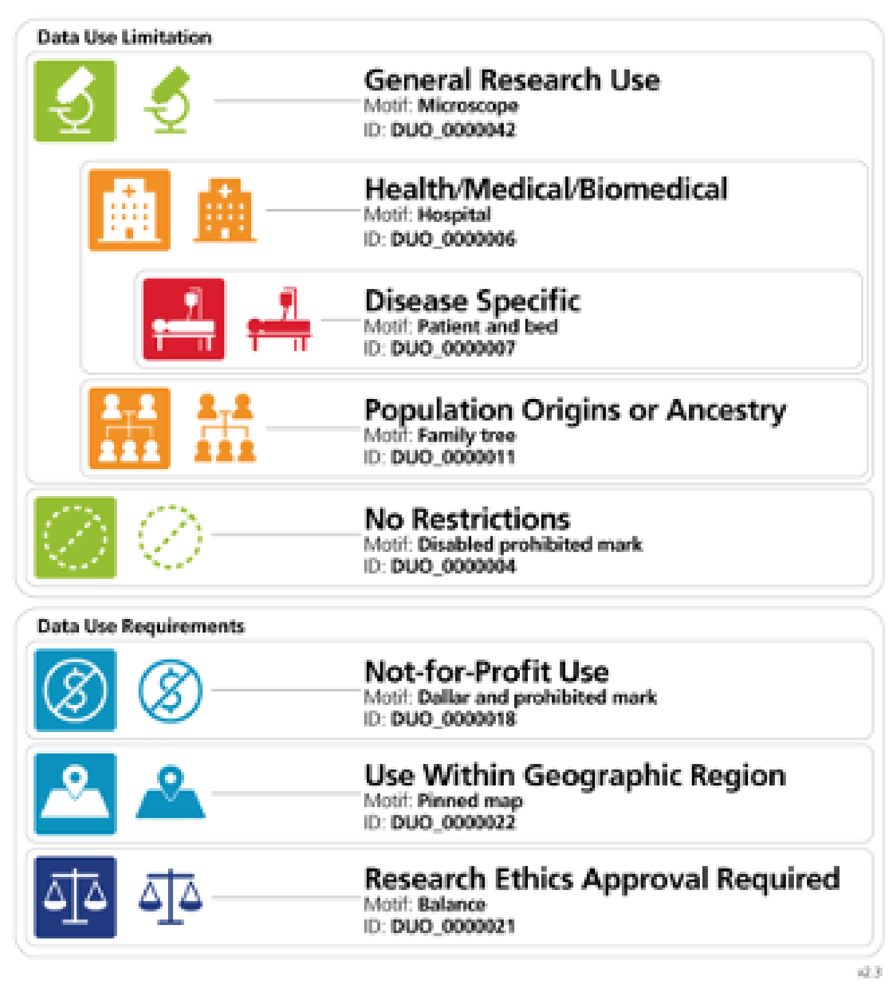
Data Use Requirements
  - Not-for-profit use only (NPU)
  - Publication required (PUB)
  - Collaboration required (COL-[XX])
  - Publication moratorium/embargo
  - Time limit on use
  - User-specific
  - Project-specific restrictions
  - Institution specific restrictions
  - Security Terms
  - Expunging Terms
  - Linking Terms
  - Allowed Recontact Terms
  - Compulsory Recontact Terms
  - IP Claim Terms
  - Reporting Terms
  - Payment Terms
- otherTOS[xx]

Data Use Ontology



Data Access Requestor - Research Purpose Terms/ Data Use Categories

Purpose of use term
- Research Purpose
  - Methods Development
  - Population Research
  - Ancestry Research
  - Age Categories Research
  - Gender Categories Research
  - Research Control
  - Biomedical Research
    - Genetic Research
    - Drug Development Research
    - Disease Categories Research
  - For Profit Purpose



Data Use Limitation

General Research Use
Motif: Microscope
ID: DUO_0000042

Health/Medical/Biomedical
Motif: Hospital
ID: DUO_0000006

Disease Specific
Motif: Patient and bed
ID: DUO_0000007

Population Origins or Ancestry
Motif: Family tree
ID: DUO_0000011

No Restrictions
Motif: Disabled prohibited mark
ID: DUO_0000004

Data Use Requirements

Not-for-Profit Use
Motif: Dallar and prohibited mark
ID: DUO_0000018

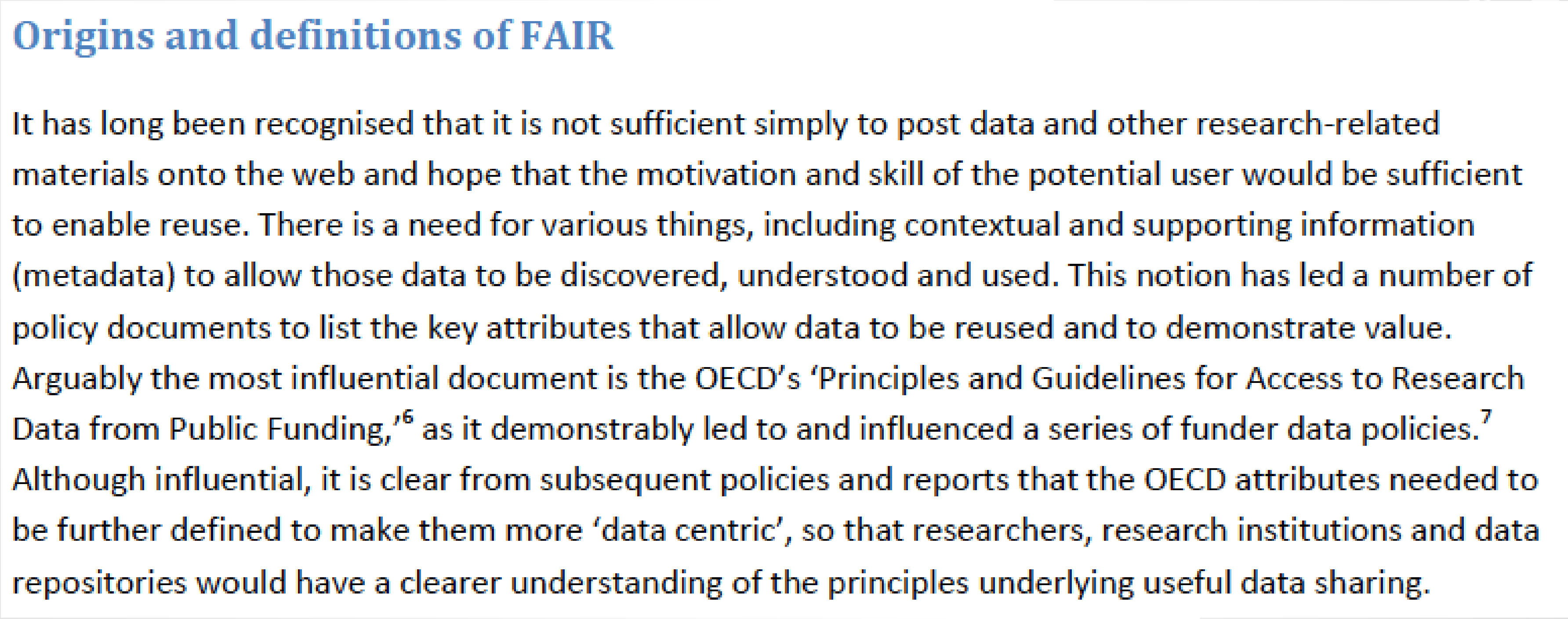
Use Within Geographic Region
Motif: Pinned map
ID: DUO_0000022

Research Ethics Approval Required
Motif: Balance
ID: DUO_0000021

v2.3

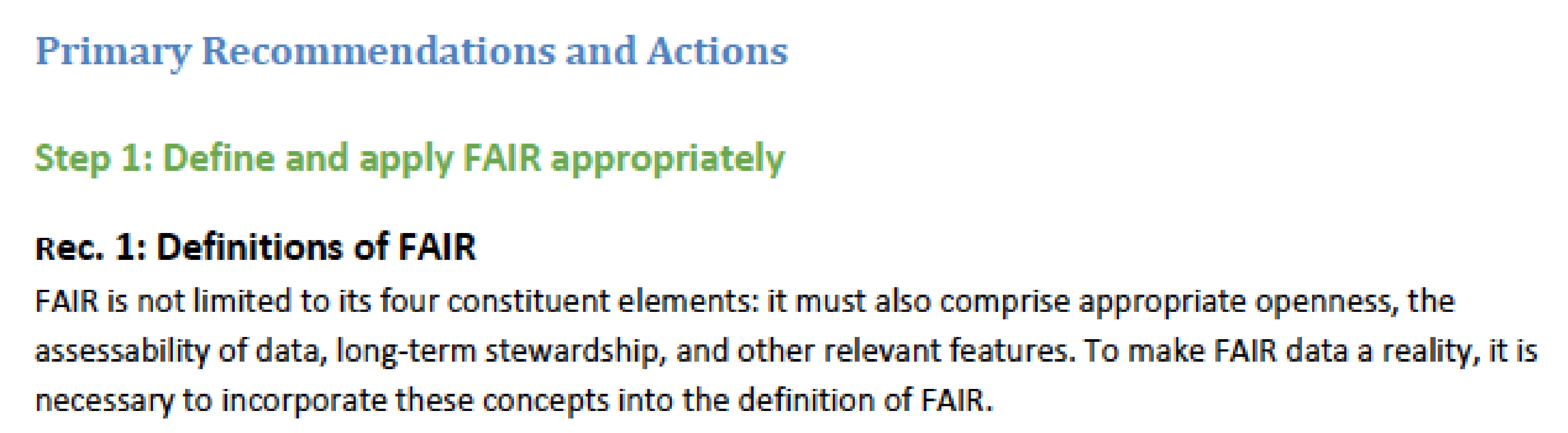
Image credits: GA4GH DUO Stream: https://github.com/EBISPOT/DUO

FAIR?



Origins and definitions of FAIR

It has long been recognised that it is not sufficient simply to post data and other research-related materials onto the web and hope that the motivation and skill of the potential user would be sufficient to enable reuse. There is a need for various things, including contextual and supporting information (metadata) to allow those data to be discovered, understood and used. This notion has led a number of policy documents to list the key attributes that allow data to be reused and to demonstrate value. Arguably the most influential document is the OECD's 'Principles and Guidelines for Access to Research Data from Public Funding,'[6] as it demonstrably led to and influenced a series of funder data policies.[7] Although influential, it is clear from subsequent policies and reports that the OECD attributes needed to be further defined to make them more 'data centric', so that researchers, research institutions and data repositories would have a clearer understanding of the principles underlying useful data sharing.



Primary Recommendations and Actions

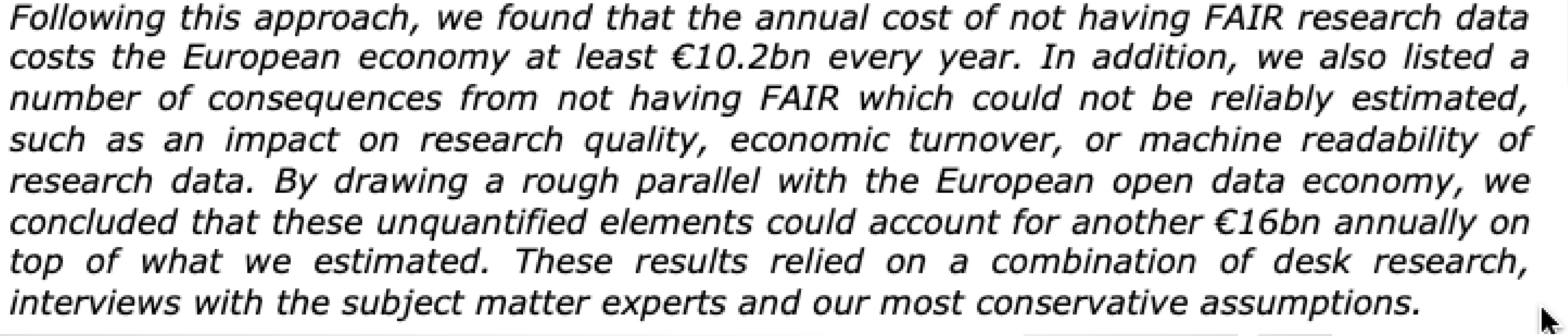
Step 1: Define and apply FAIR appropriately

Rec. 1: Definitions of FAIR

FAIR is not limited to its four constituent elements: it must also comprise appropriate openness, the assessability of data, long-term stewardship, and other relevant features. To make FAIR data a reality, it is necessary to incorporate these concepts into the definition of FAIR.

https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf

Why is FAIR important?



Following this approach, we found that the annual cost of not having FAIR research data costs the European economy at least €10.2bn every year. In addition, we also listed a number of consequences from not having FAIR which could not be reliably estimated, such as an impact on research quality, economic turnover, or machine readability of research data. By drawing a rough parallel with the European open data economy, we concluded that these unquantified elements could account for another €16bn annually on top of what we estimated. These results relied on a combination of desk research, interviews with the subject matter experts and our most conservative assumptions.

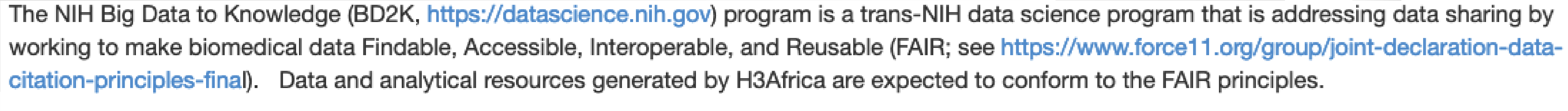
https://op.europa.eu/s/sniv

Who does FAIR apply to?

Stakeholder groups assigned Actions

1. **Research communities:** practitioners from all fields of humanities and science, clustered in groups around disciplinary interests, data types or cross-cutting grand challenges.
2. **Data services:** domain repositories, Research Infrastructures (ESFRIs) and E-Infrastructures, institutional provision, community and commercial tools and services.
3. **Data stewards:** support staff from research communities and research libraries, and those managing data repositories.
4. **Standards bodies:** formal organisations and consortia coordinating data standards and governing procedures relevant to FAIR, e.g. repository certification, curriculum accreditation.
5. **Global coordination fora:** the Research Data Alliance, CODATA, WDS Communities of Excellence, FORCE11, GO FAIR and other similar initiatives.
6. **Policymakers:** governments, international entities like OECD, research funders, institutions, publishers and others defining data policy.
7. **Research funders:** the European Commission, national research funders, charitable organisations and foundations, and other funders of research activity.
8. **Institutions:** universities and research performing organisations
9. **Publishers:** commercial and not-for-profit, paywall and Open Access publishers of research papers and data.

https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf



The NIH Big Data to Knowledge (BD2K, https://datascience.nih.gov) program is a trans-NIH data science program that is addressing data sharing by working to make biomedical data Findable, Accessible, Interoperable, and Reusable (FAIR; see https://www.force11.org/group/joint-declaration-data-citation-principles-final). Data and analytical resources generated by H3Africa are expected to conform to the FAIR principles.

https://grants.nih.gov/grants/guide/rfa-files/RFA-rm-16-011.html

FAIR – H3Africa Data

- RDA FAIR Data Maturity Model (https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and%20guidelines_v1.00.pdf)
  - AKA FAIR Indicators
  - Indicators of 3 levels
    - Essential
    - Important
    - Useful
  - 40 Indicators
    - 7 Findability Indicators
    - 12 Accessibility Indicators
    - 12 Interoperability Indicators
    - 9 Reusability Indicators

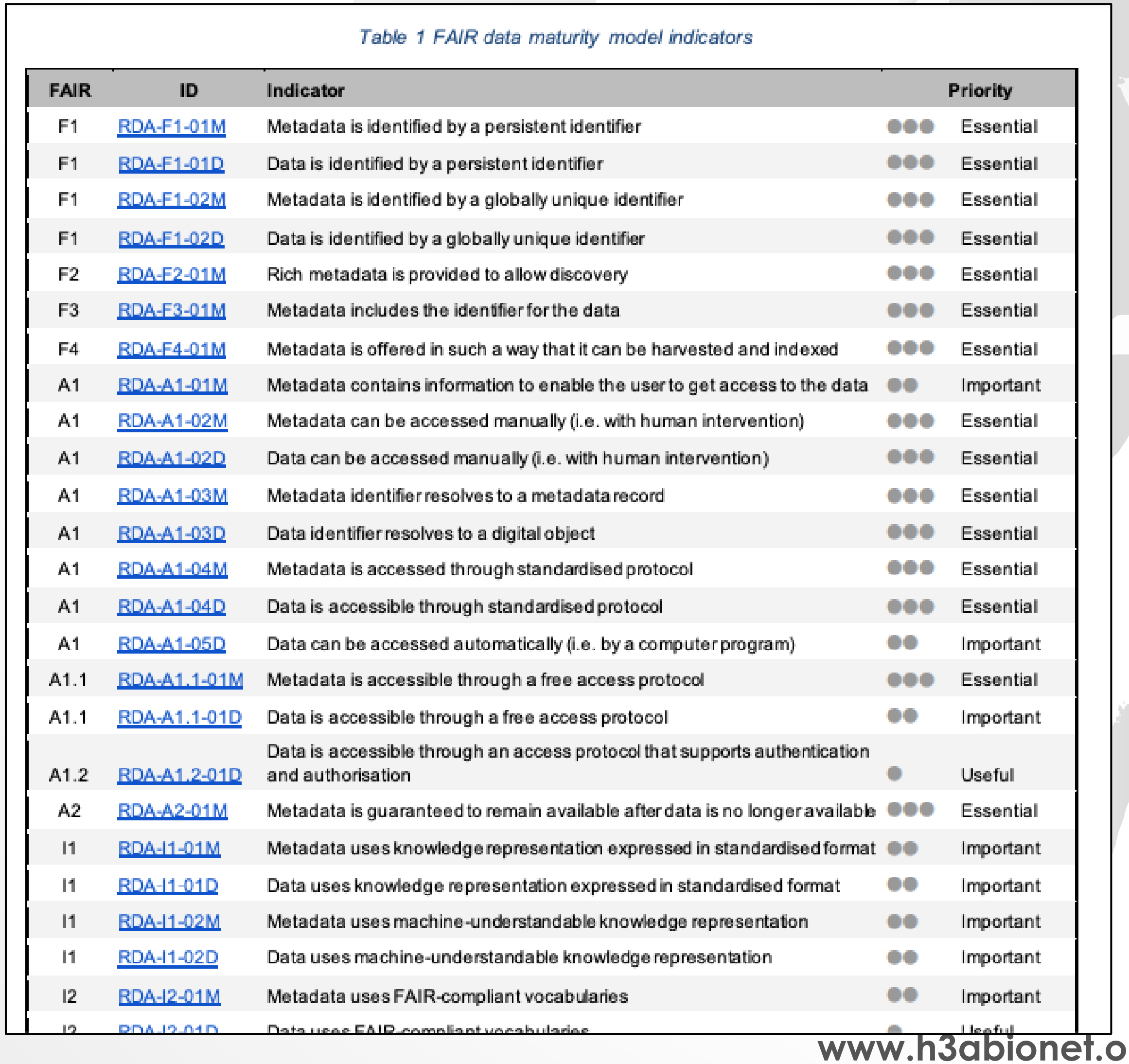
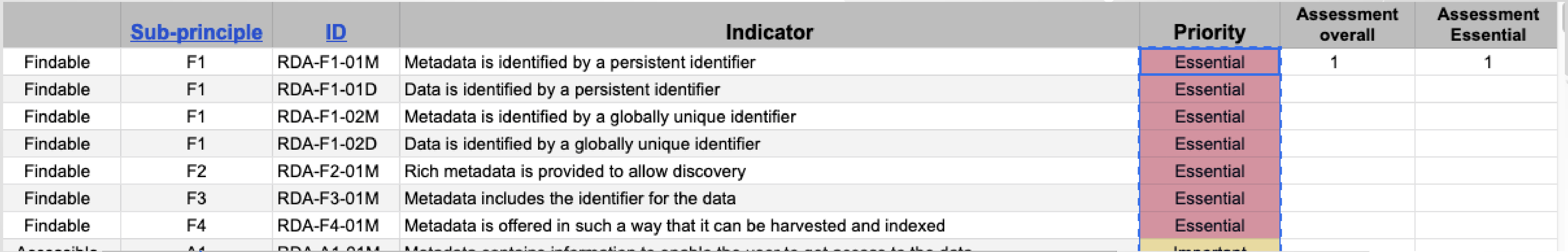


Table 1 FAIR data maturity model indicators

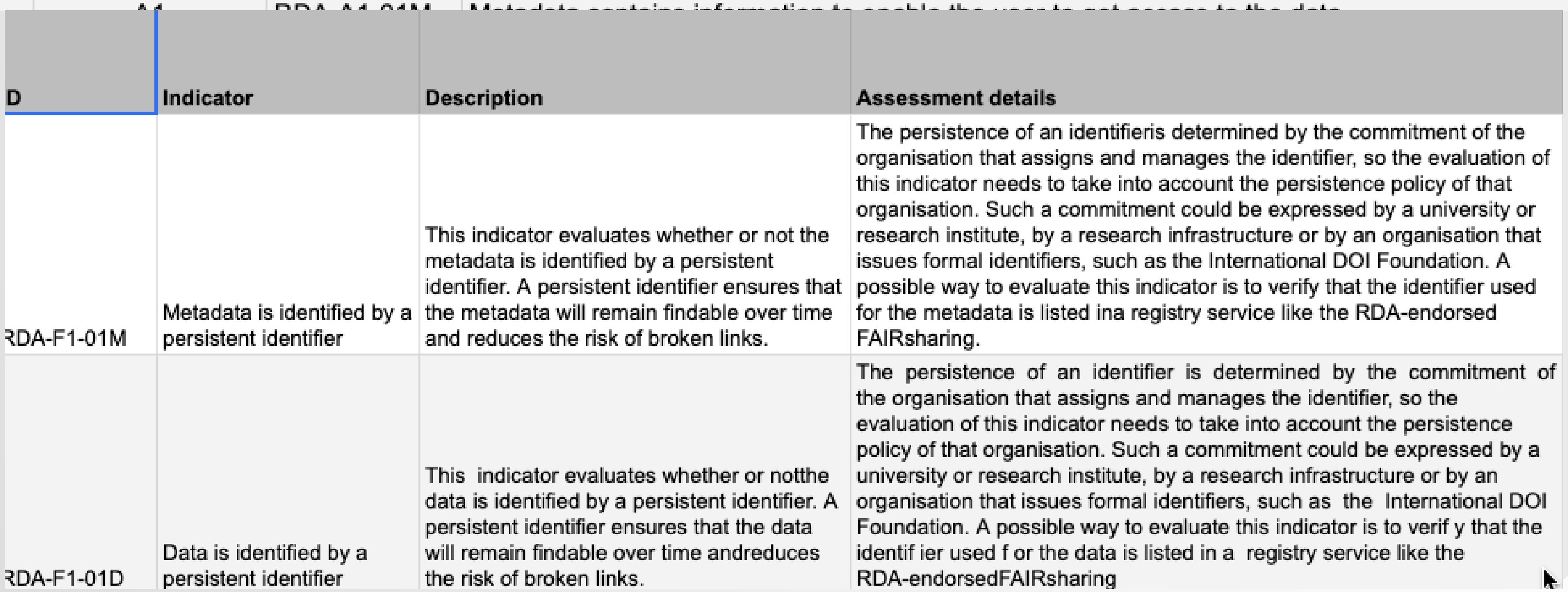
FAIR	ID	Indicator	Priority
F1	RDA-F1-01M	Metadata is identified by a persistent identifier	●●● Essential
F1	RDA-F1-01D	Data is identified by a persistent identifier	●●● Essential
F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	●●● Essential
F1	RDA-F1-02D	Data is identified by a globally unique identifier	●●● Essential
F2	RDA-F2-01M	Rich metadata is provided to allow discovery	●●● Essential
F3	RDA-F3-01M	Metadata includes the identifier for the data	●●● Essential
F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	●●● Essential
A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	●● Important
A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	●●● Essential
A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	●●● Essential
A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	●●● Essential
A1	RDA-A1-03D	Data identifier resolves to a digital object	●●● Essential
A1	RDA-A1-04M	Metadata is accessed through standardised protocol	●●● Essential
A1	RDA-A1-04D	Data is accessible through standardised protocol	●●● Essential
A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	●● Important
A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	●●● Essential
A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	●● Important
A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	● Useful
A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	●●● Essential
I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	●● Important
I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	●● Important
I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	●● Important
I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	●● Important
I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	●● Important
I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	● Useful

FAIR – H3Africa Data

- RDA FAIR indicators assessment (https://docs.google.com/spreadsheets/d/1vaRf4PXJ7FvxLZplKXRlkiEYDSZP5ajsb3KnETz7Em0/edit#gid=1375297260)



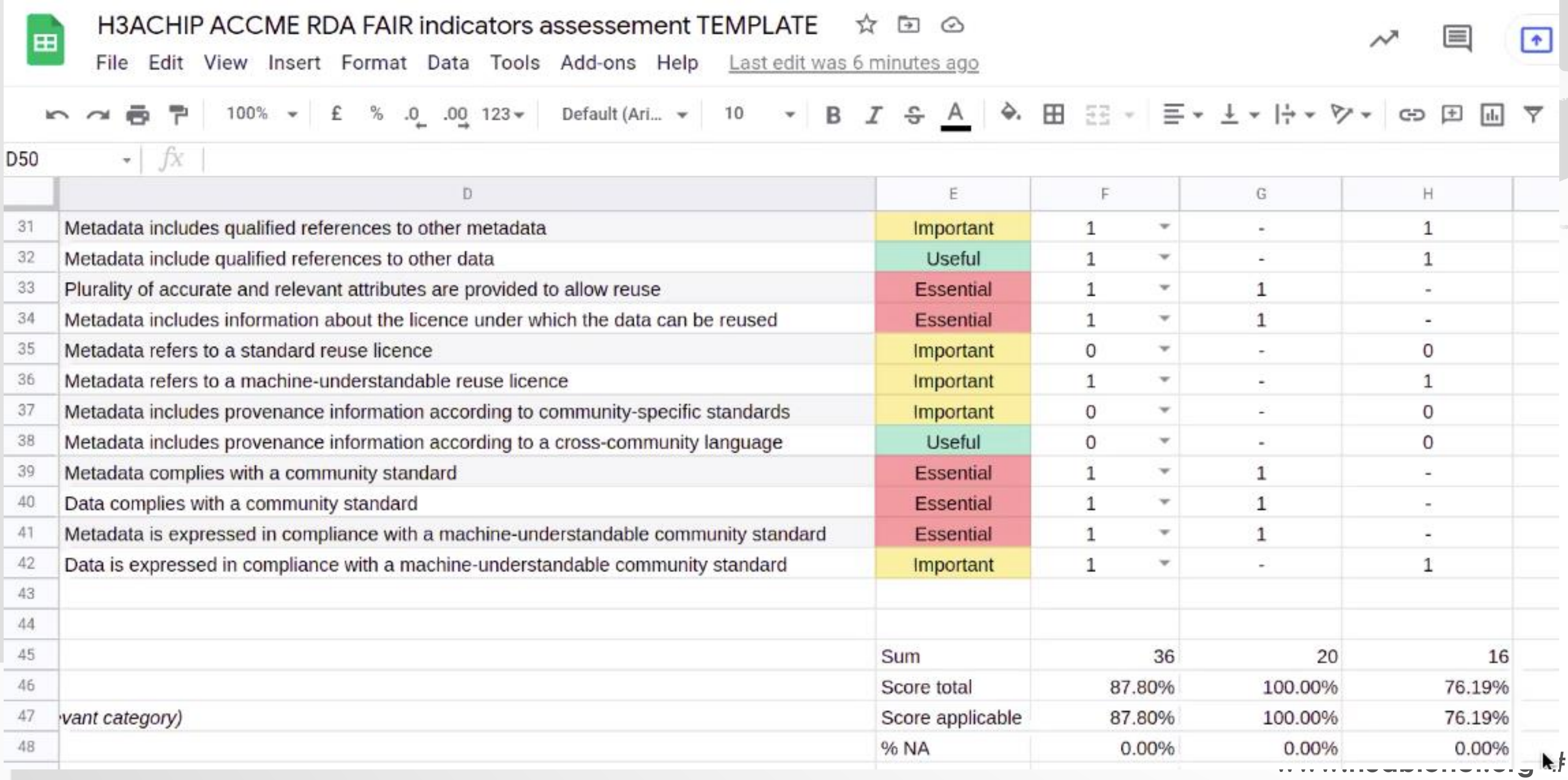
	Sub-principle	ID	Indicator	Priority	Assessment overall	Assessment Essential
Findable	F1	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential	1	1
Findable	F1	RDA-F1-01D	Data is identified by a persistent identifier	Essential		
Findable	F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential		
Findable	F1	RDA-F1-02D	Data is identified by a globally unique identifier	Essential		
Findable	F2	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential		
Findable	F3	RDA-F3-01M	Metadata includes the identifier for the data	Essential		
Findable	F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential		
Accessible	A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	Important		



D	Indicator	Description	Assessment details
RDA-F1-01M	Metadata is identified by a persistent identifier	This indicator evaluates whether or not the metadata is identified by a persistent identifier. A persistent identifier ensures that the metadata will remain findable over time and reduces the risk of broken links.	The persistence of an identifieris determined by the commitment of the organisation that assigns and manages the identifier, so the evaluation of this indicator needs to take into account the persistence policy of that organisation. Such a commitment could be expressed by a university or research institute, by a research infrastructure or by an organisation that issues formal identifiers, such as the International DOI Foundation. A possible way to evaluate this indicator is to verify that the identifier used for the metadata is listed ina registry service like the RDA-endorsed FAIRsharing.
RDA-F1-01D	Data is identified by a persistent identifier	This indicator evaluates whether or notthe data is identified by a persistent identifier. A persistent identifier ensures that the data will remain findable over time andreduces the risk of broken links.	The persistence of an identifier is determined by the commitment of the organisation that assigns and manages the identifier, so the evaluation of this indicator needs to take into account the persistence policy of that organisation. Such a commitment could be expressed by a university or research institute, by a research infrastructure or by an organisation that issues formal identifiers, such as the International DOI Foundation. A possible way to evaluate this indicator is to verif y that the identif ier used f or the data is listed in a registry service like the RDA-endorsedFAIRsharing

FAIR – H3Africa Data

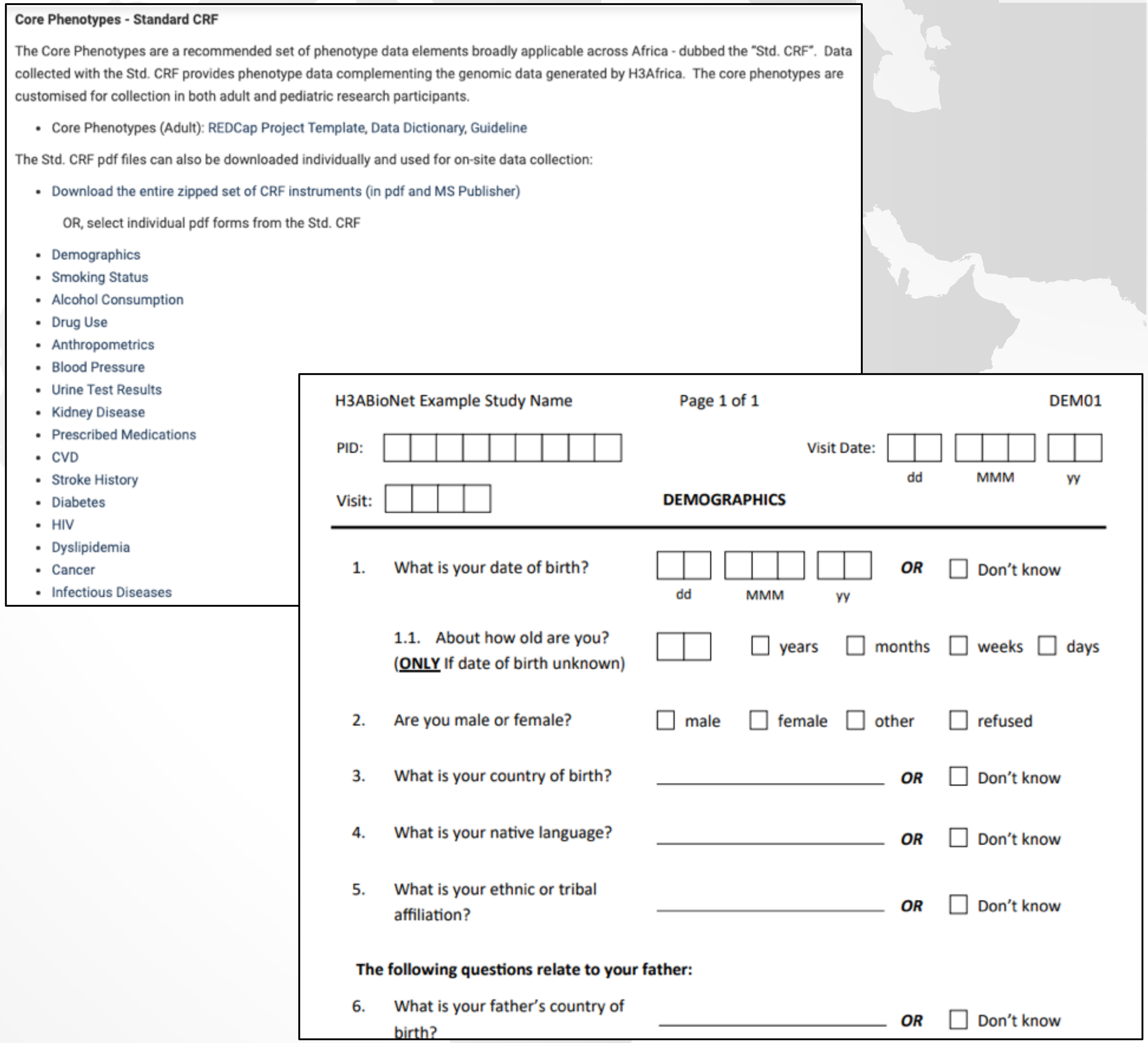
- Assessed the H3Africa Chip project data according to the indicators: - 100% on the Assessment for the level of Essential, 76.19% on Assessment Non-Essential with an overall Assessment score of 87.8%


H3ACHIP ACCME RDA FAIR indicators assessement TEMPLATE
File Edit View Insert Format Data Tools Add-ons Help
Last edit was 6 minutes ago
100%
£ % .0 .00 123
Default (Ari...
10
B I S A
D50
fx

	D	E	F	G	H
31	Metadata includes qualified references to other metadata	Important	1	-	1
32	Metadata include qualified references to other data	Useful	1	-	1
33	Plurality of accurate and relevant attributes are provided to allow reuse	Essential	1	1	-
34	Metadata includes information about the licence under which the data can be reused	Essential	1	1	-
35	Metadata refers to a standard reuse licence	Important	0	-	0
36	Metadata refers to a machine-understandable reuse licence	Important	1	-	1
37	Metadata includes provenance information according to community-specific standards	Important	0	-	0
38	Metadata includes provenance information according to a cross-community language	Useful	0	-	0
39	Metadata complies with a community standard	Essential	1	1	-
40	Data complies with a community standard	Essential	1	1	-
41	Metadata is expressed in compliance with a machine-understandable community standard	Essential	1	1	-
42	Data is expressed in compliance with a machine-understandable community standard	Important	1	-	1
43					
44					
45		Sum	36	20	16
46		Score total	87.80%	100.00%	76.19%
47	*vant category)*	Score applicable	87.80%	100.00%	76.19%
48		% NA	0.00%	0.00%	0.00%

FAIR - H3Africa Phenotype Data

- Developed the H3Africa STD CRF
  - Form is submitted when H3A projects are ready to submit to the H3Africa Data Archive
  - Standardized response options were included for majority of the requested fields
  - Metadata is mapped to Genomics Cohorts Knowledge Ontology (GECKO: http://www.obofoundry.org/ontology/gecko.html)
  - Consent mapped to Data Use Ontology (DUO: https://github.com/EBISPOT/DUO/blob/master/README.md



Core Phenotypes - Standard CRF

The Core Phenotypes are a recommended set of phenotype data elements broadly applicable across Africa - dubbed the "Std. CRF". Data collected with the Std. CRF provides phenotype data complementing the genomic data generated by H3Africa. The core phenotypes are customised for collection in both adult and pediatric research participants.

- Core Phenotypes (Adult): REDCap Project Template, Data Dictionary, Guideline

The Std. CRF pdf files can also be downloaded individually and used for on-site data collection:

- Download the entire zipped set of CRF instruments (in pdf and MS Publisher)

OR, select individual pdf forms from the Std. CRF

- Demographics
- Smoking Status
- Alcohol Consumption
- Drug Use
- Anthropometrics
- Blood Pressure
- Urine Test Results
- Kidney Disease
- Prescribed Medications
- CVD
- Stroke History
- Diabetes
- HIV
- Dyslipidemia
- Cancer
- Infectious Diseases

H3ABioNet Example Study Name   Page 1 of 1   DEM01

PID: □□□□□□□□□   Visit Date: □□ □□□ □□ dd MMM yy

Visit: □□□□   **DEMOGRAPHICS**

1. What is your date of birth? □□ □□□ □□ dd MMM yy ***OR*** □ Don't know

1.1. About how old are you? (**ONLY** If date of birth unknown) □□ □ years □ months □ weeks □ days

2. Are you male or female? □ male □ female □ other □ refused

3. What is your country of birth? ____________ ***OR*** □ Don't know

4. What is your native language? ____________ ***OR*** □ Don't know

5. What is your ethnic or tribal affiliation? ____________ ***OR*** □ Don't know

The following questions relate to your father:

6. What is your father's country of birth? ____________ ***OR*** □ Don't know

NIH DS-I Africa Initiative

- Harnessing Data Science for Health Discovery and Innovation in Africa
  - **Research Hubs:** Advance and demonstrate feasibility of data science research and innovation to improve health in Africa
  - **Training:** Increase capacity for data science research in Africa
  - **ELSI Research:** Explore Ethical, Legal, and Social Implications of data science research from an African perspective and contribute to policy discussion on the continent
  - **Open Data Science Platform & Coordination Center:** Facilitate the development of a trans-African network of data scientists

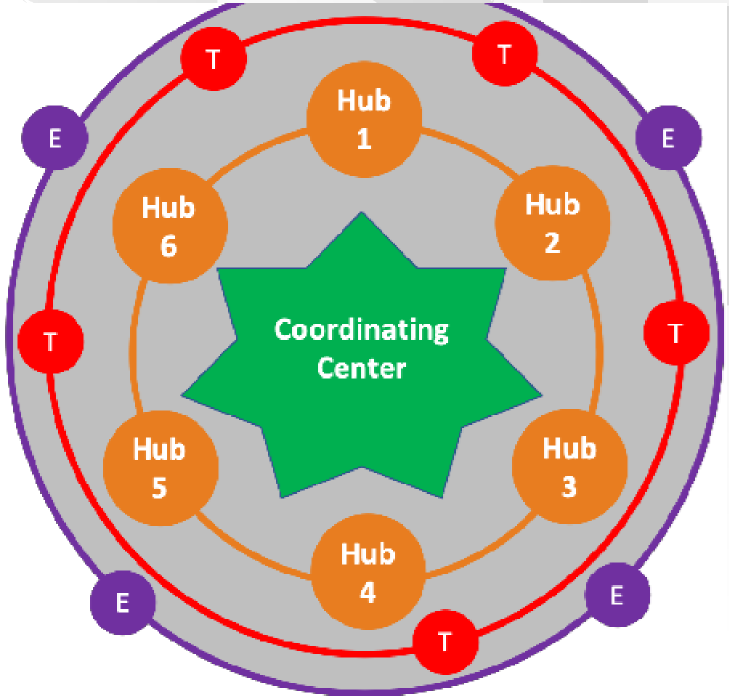


Image and text extracted from: https://commonfund.nih.gov/africadata

eLwazi Open Data Science Platform

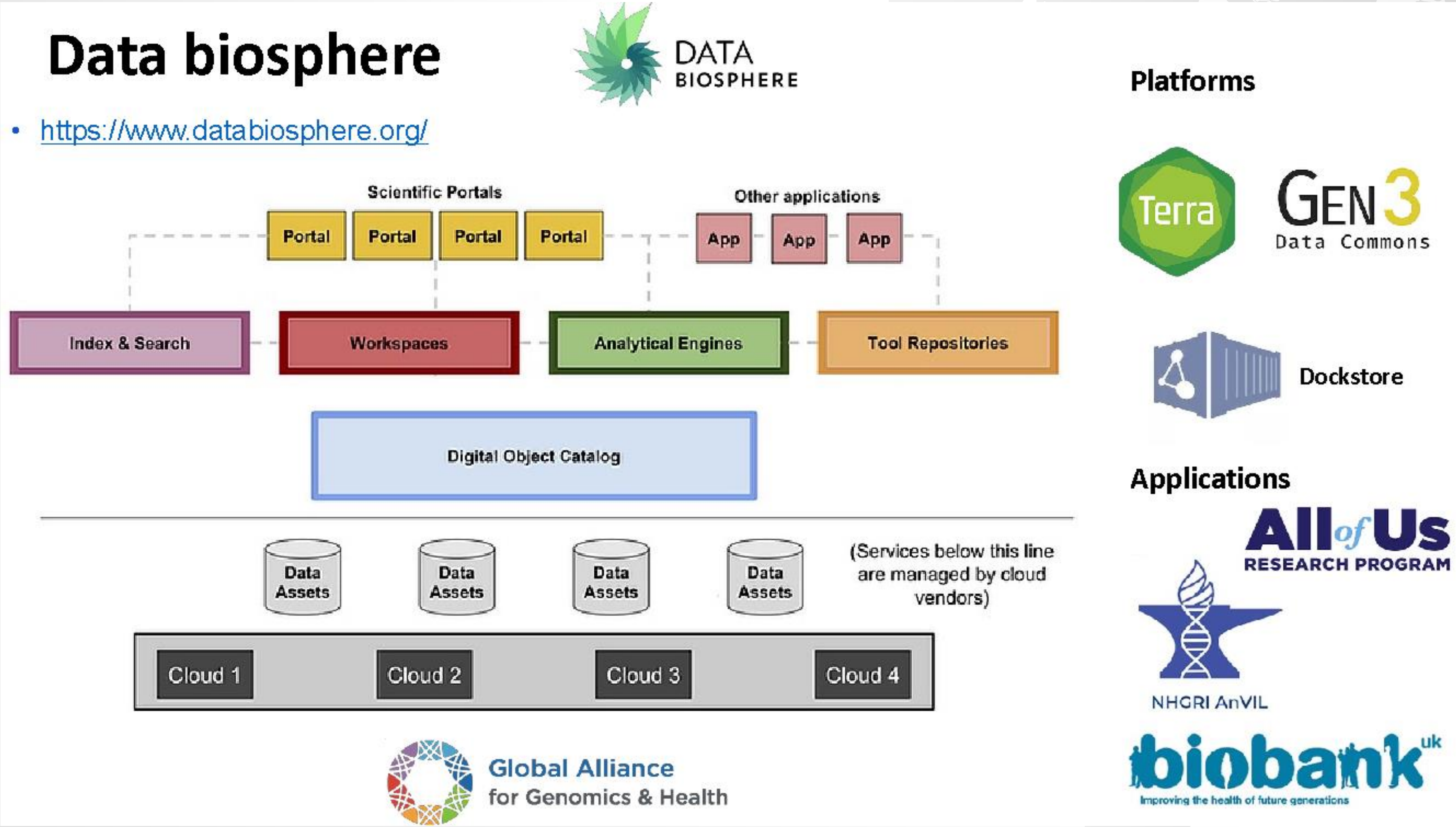
- To develop an African Open Data Science Gateway, including the eLwazi platform and associated resources, to support the Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa) consortium and beyond. eLwazi, will be a flexible, scalable Open Data Science Platform enabling the implementation of data science for health, that is relevant to the African context.
- Facilitate the development of a trans-African network of data scientists

Ulwazi is the Xhosa word meaning “knowledge” or “information”,
and Olwazi means big rock in Luganda

⬇

eLwazi

eLwazi Open Data Science Platform



Data biosphere

DATA BIOSPHERE

- https://www.databiosphere.org/

Scientific Portals

Portal Portal Portal Portal

Other applications

App App App

Index & Search

Workspaces

Analytical Engines

Tool Repositories

Digital Object Catalog

Data Assets Data Assets Data Assets Data Assets

(Services below this line are managed by cloud vendors)

Cloud 1 Cloud 2 Cloud 3 Cloud 4

Global Alliance for Genomics & Health

Platforms

Terra

GEN3 Data Commons

Dockstore

Applications

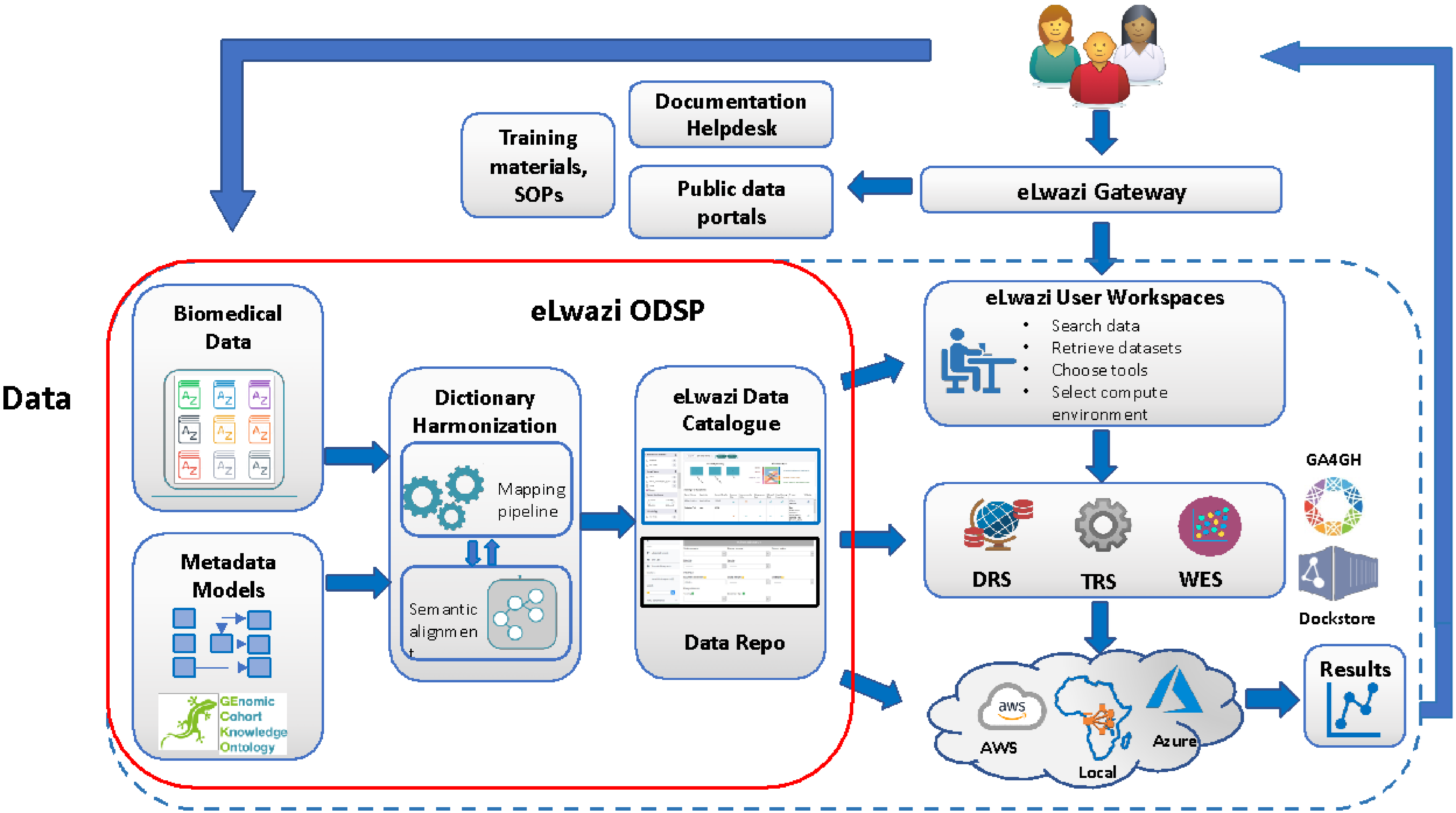
All of Us RESEARCH PROGRAM

NHGRI AnVIL

biobank uk

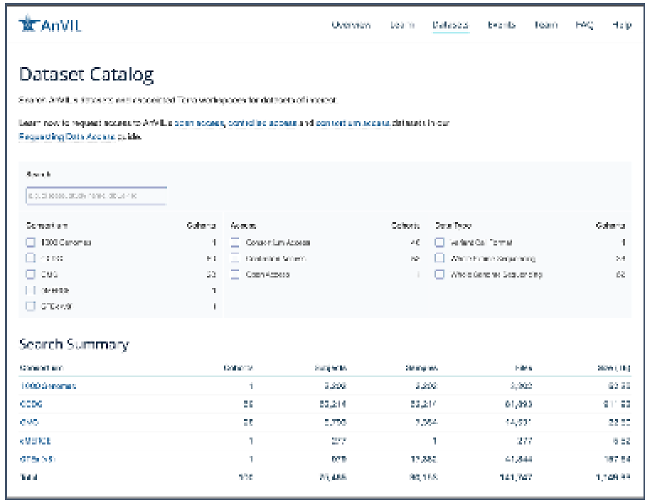
Improving the health of future generations

Overview of the ODSP



Data and Meta-data store

Data registry



AnVIL

Overview Learn Datasets Events Team FAQ Help

Dataset Catalog

Search AnVIL's [illegible] and associated Terra workspaces for datasets of interest.

Learn how to request access to AnVIL's open access, controlled access and consortium access datasets in our Requesting Data Access guide.

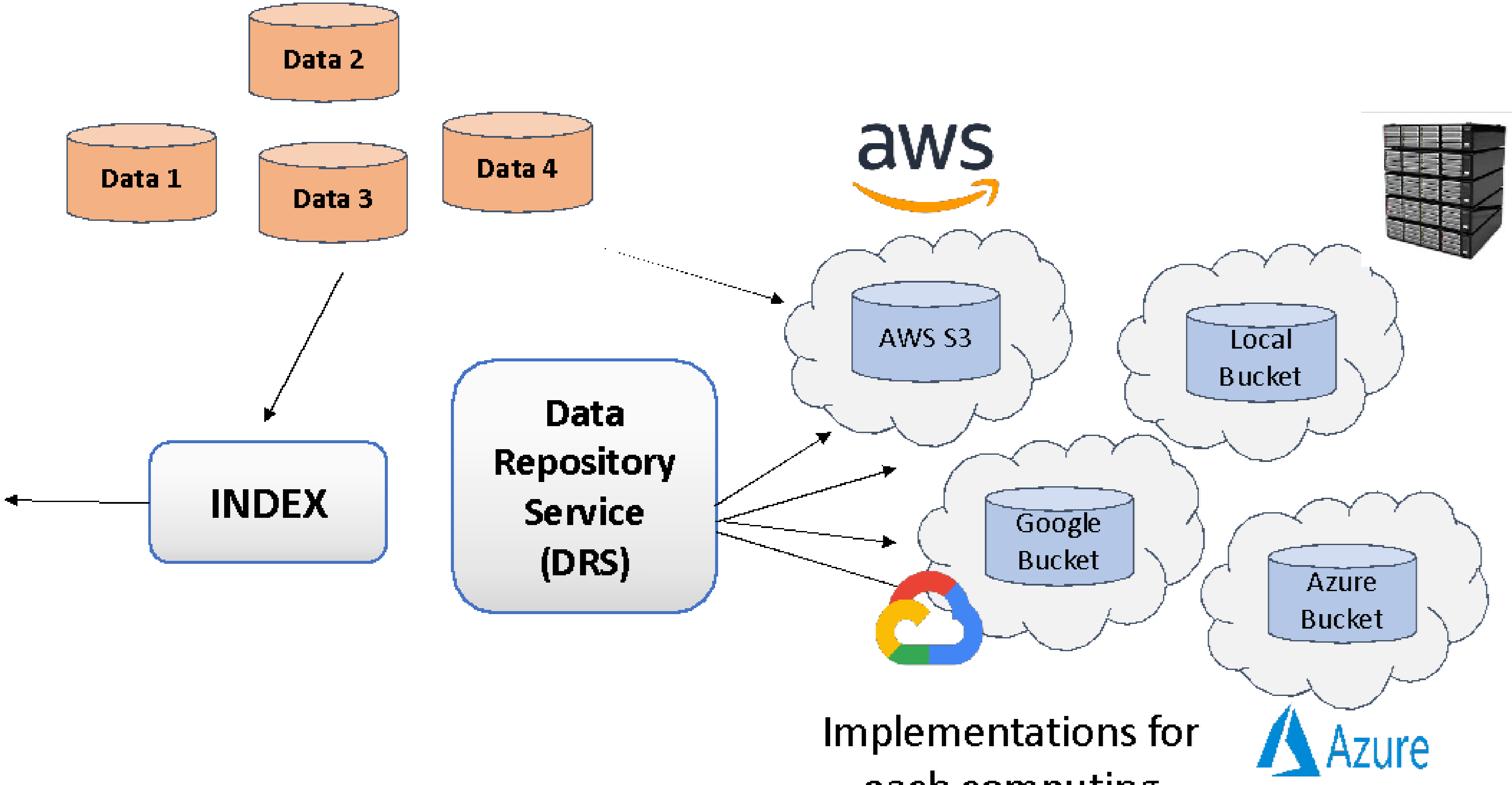
Search

Consortium	Cohorts	Access	Cohorts	Data Type	Cohorts
□ 1000 Genomes	1	□ Consortium Access	40	□ Variant Call Format	1
□ CCDG	51	□ Controlled Access	62	□ Whole Exome Sequencing	24
□ CMG	23	□ Open Access	1	□ Whole Genome Sequencing	62
□ eMERGE	1				
□ GTEx (v8)	1				

Search Summary

Consortium	Cohorts	Subjects	Samples	Files	Size (TB)
1000 Genomes	1	3,202	3,202	3,202	62.33
CCDG	55	83,214	83,217	91,893	911.83
CMG	25	5,790	7,354	14,531	22.52
eMERGE	1	277	1	277	6.62
GTEx (v8)	1	979	17,382	41,644	187.84
Total	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Data Browser
query interface



Data 2

Data 1

Data 3

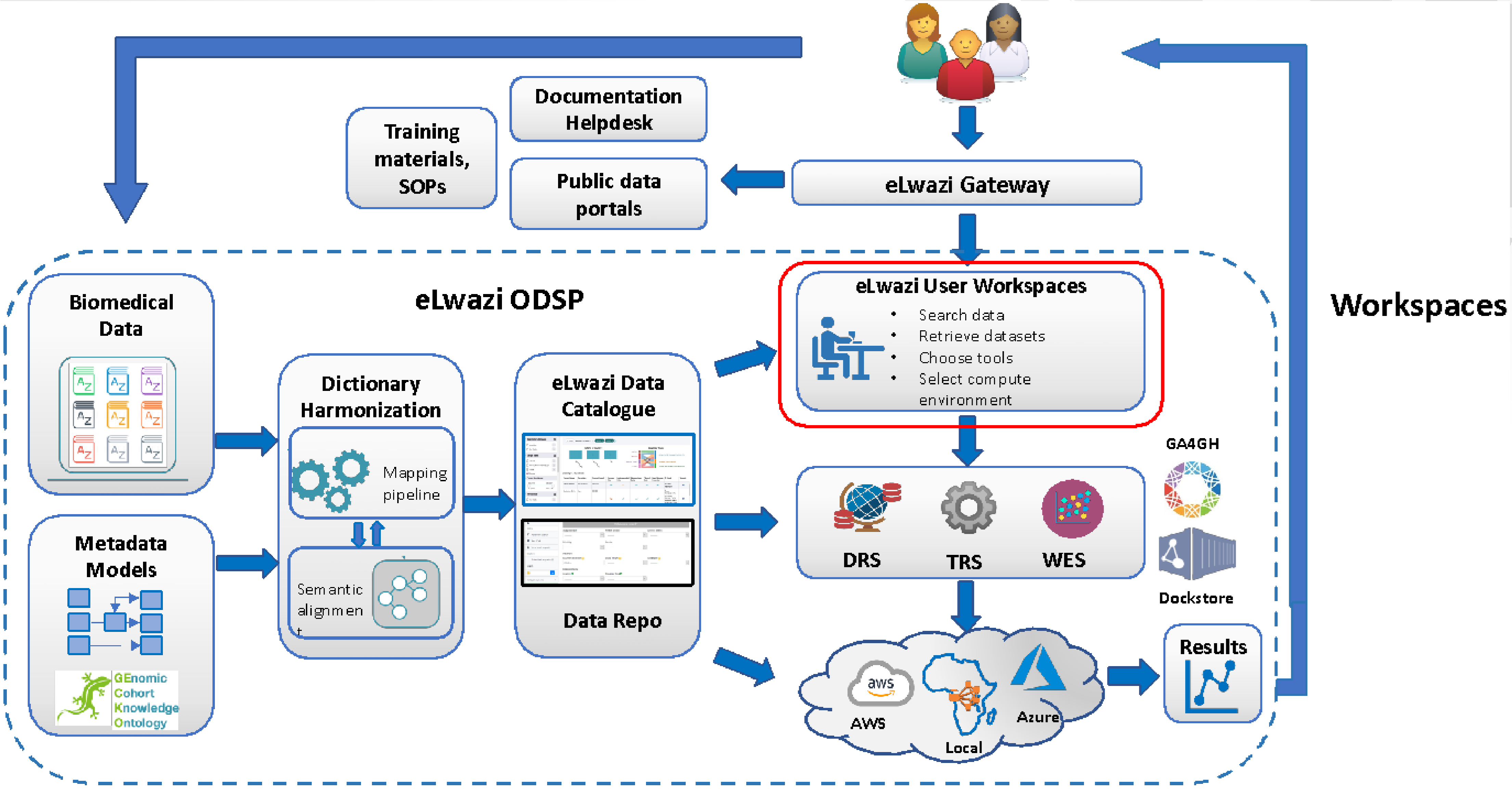
Data 4

INDEX

Data Repository Service (DRS)

Implementations for each computing environment

Data Analysis



Documentation Helpdesk

Training materials, SOPs

Public data portals

eLwazi Gateway

eLwazi ODSP

Biomedical Data

Dictionary Harmonization

Mapping pipeline

Semantic alignment

eLwazi Data Catalogue

Data Repo

Metadata Models

GEnomic Cohort Knowledge Ontology

eLwazi User Workspaces

- Search data
- Retrieve datasets
- Choose tools
- Select compute environment

DRS

TRS

WES

GA4GH

Dockstore

AWS

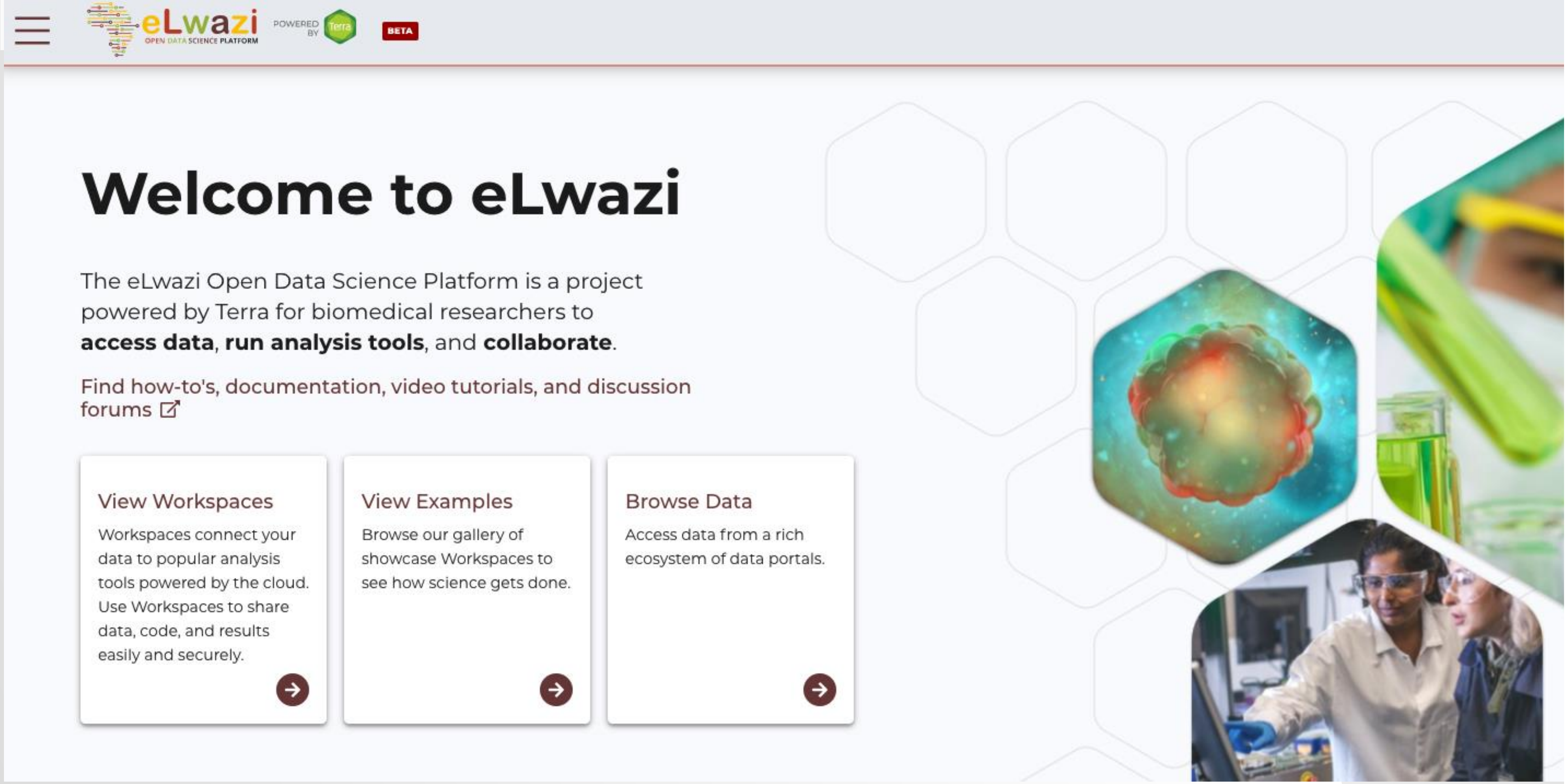
Local

Azure

Results

Workspaces

Data Analysis



Welcome to eLwazi

The eLwazi Open Data Science Platform is a project powered by Terra for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**.

Find how-to's, documentation, video tutorials, and discussion forums

View Workspaces

Workspaces connect your data to popular analysis tools powered by the cloud. Use Workspaces to share data, code, and results easily and securely.

View Examples

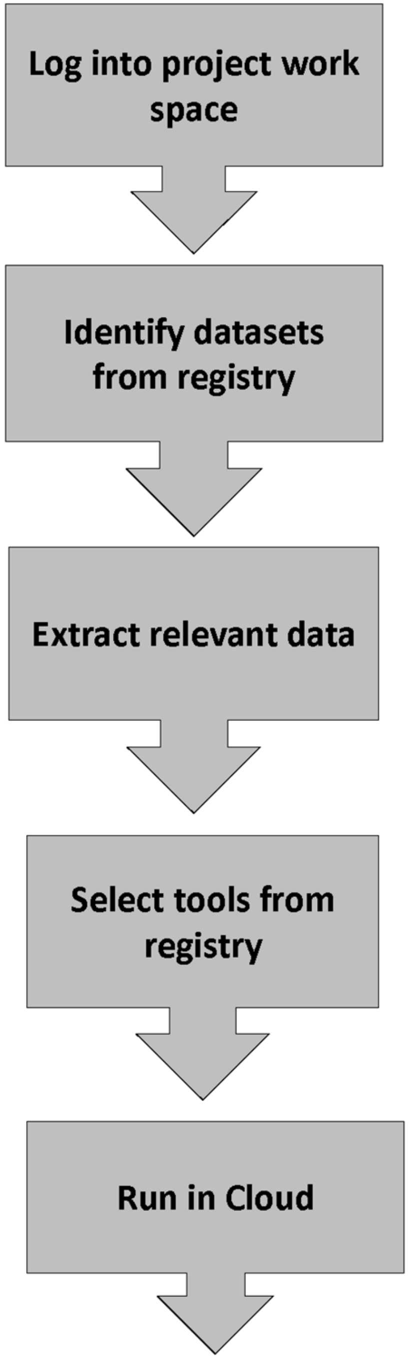
Browse our gallery of showcase Workspaces to see how science gets done.

Browse Data

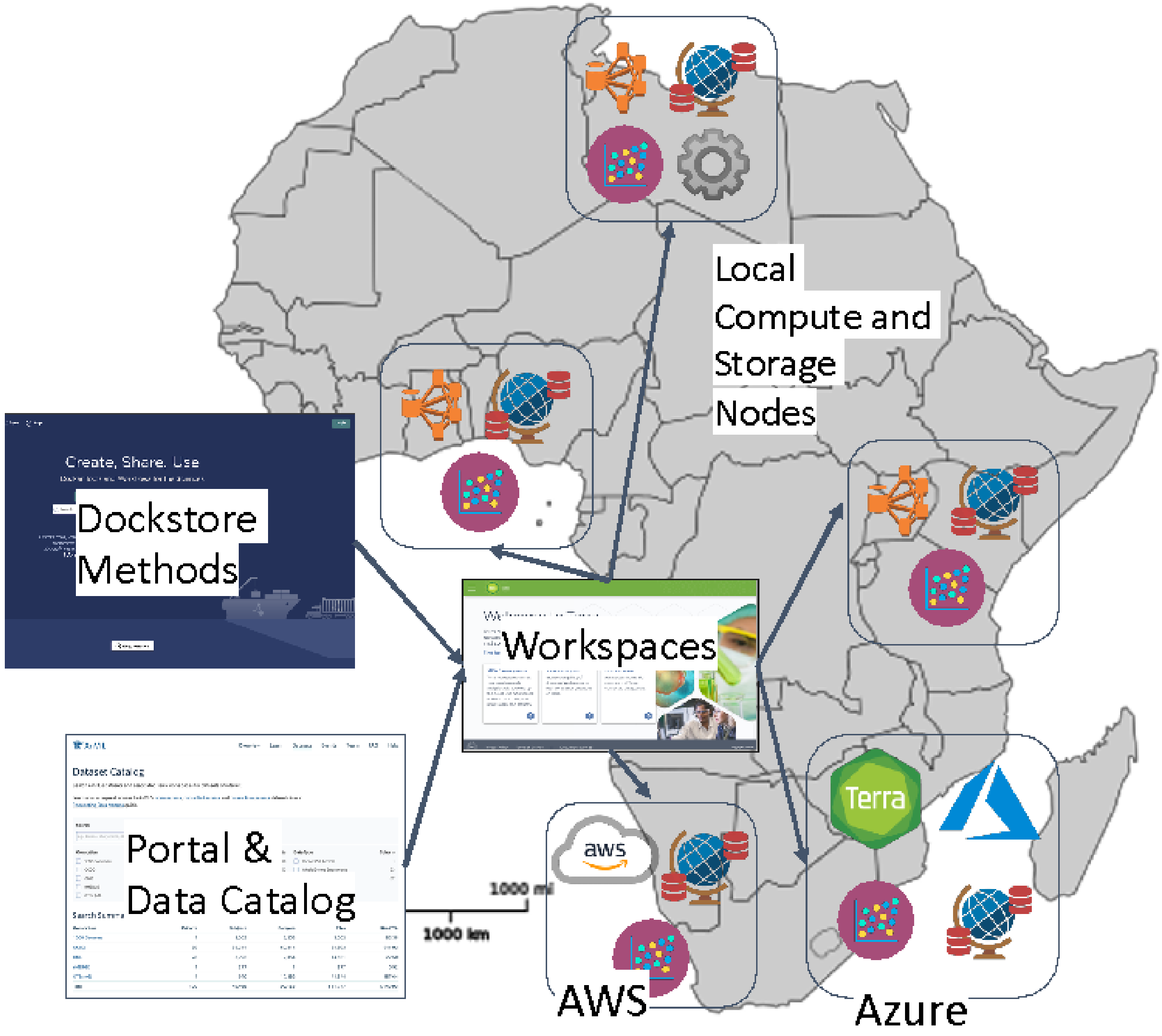
Access data from a rich ecosystem of data portals.

https://elwazi.terra.bio/

Workspaces bring together metadata, data, and compute environments through a web-based user interface. Access for beginner and advanced users


Log into project work space
Identify datasets from registry
Extract relevant data
Select tools from registry
Run in Cloud

eLwazi ODSP hosting sites



Cloud & HPC Environments

- AWS, Azure, and Local

DRS for data access

WES for compute

TRS for workflow sharing

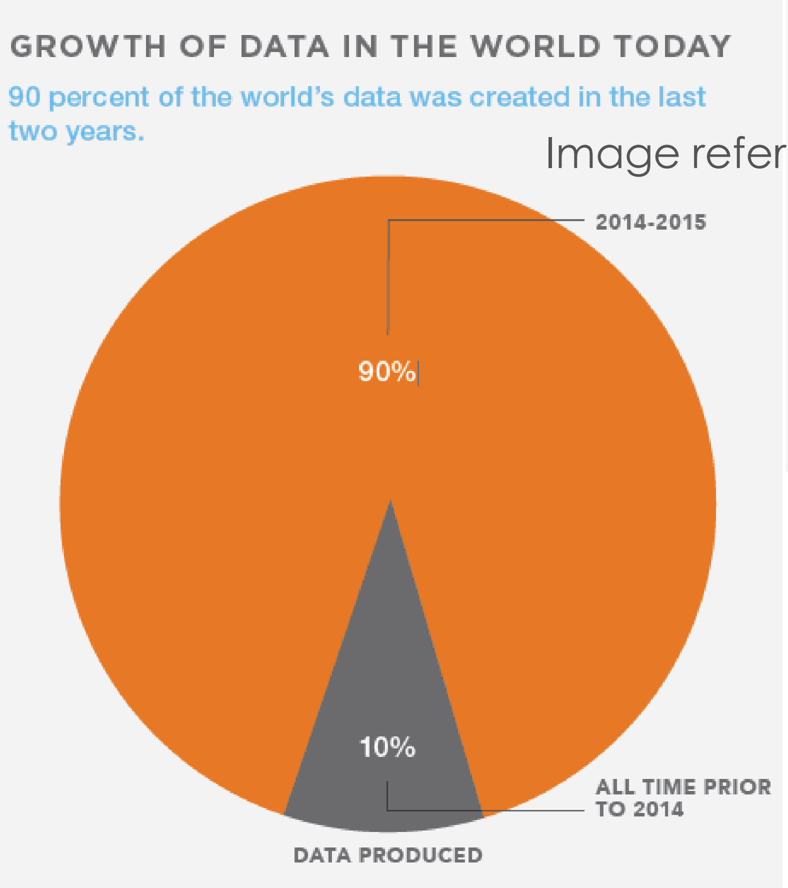
Workspaces for:

Data modeling and access

- Compute on data in various locations
- Interactively analyze via Notebooks
- Sharing and Collaborating

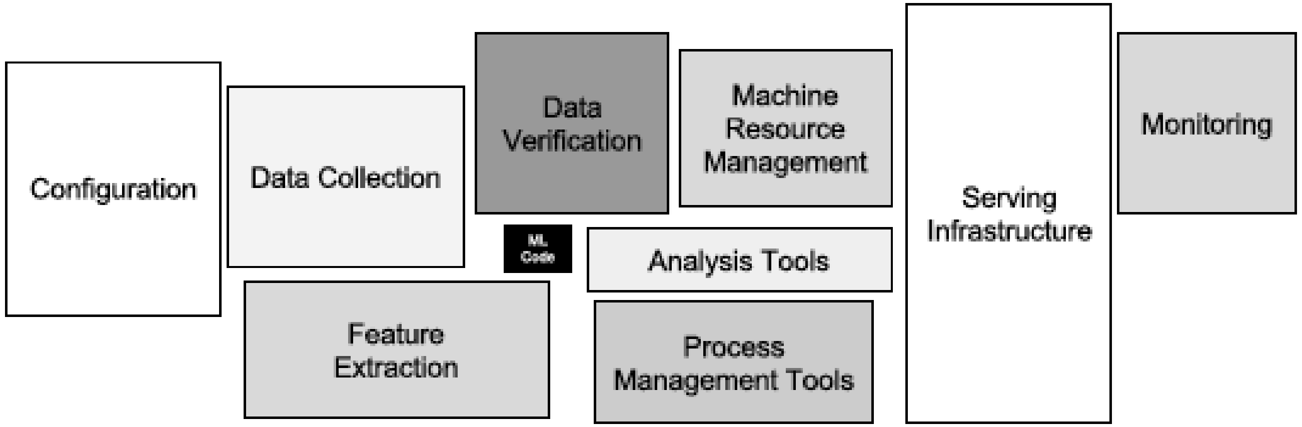
Conclusions

Image reference: https://data.bsa.org/wp-content/uploads/2015/12/bsadatastudy_en.pdf



GROWTH OF DATA IN THE WORLD TODAY

90 percent of the world's data was created in the last two years.



Configuration
Data Collection
Data Verification
Machine Resource Management
Serving Infrastructure
Monitoring
ML Code
Analysis Tools
Feature Extraction
Process Management Tools

Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Hidden Technical Debt in Machine Learning Systems: https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

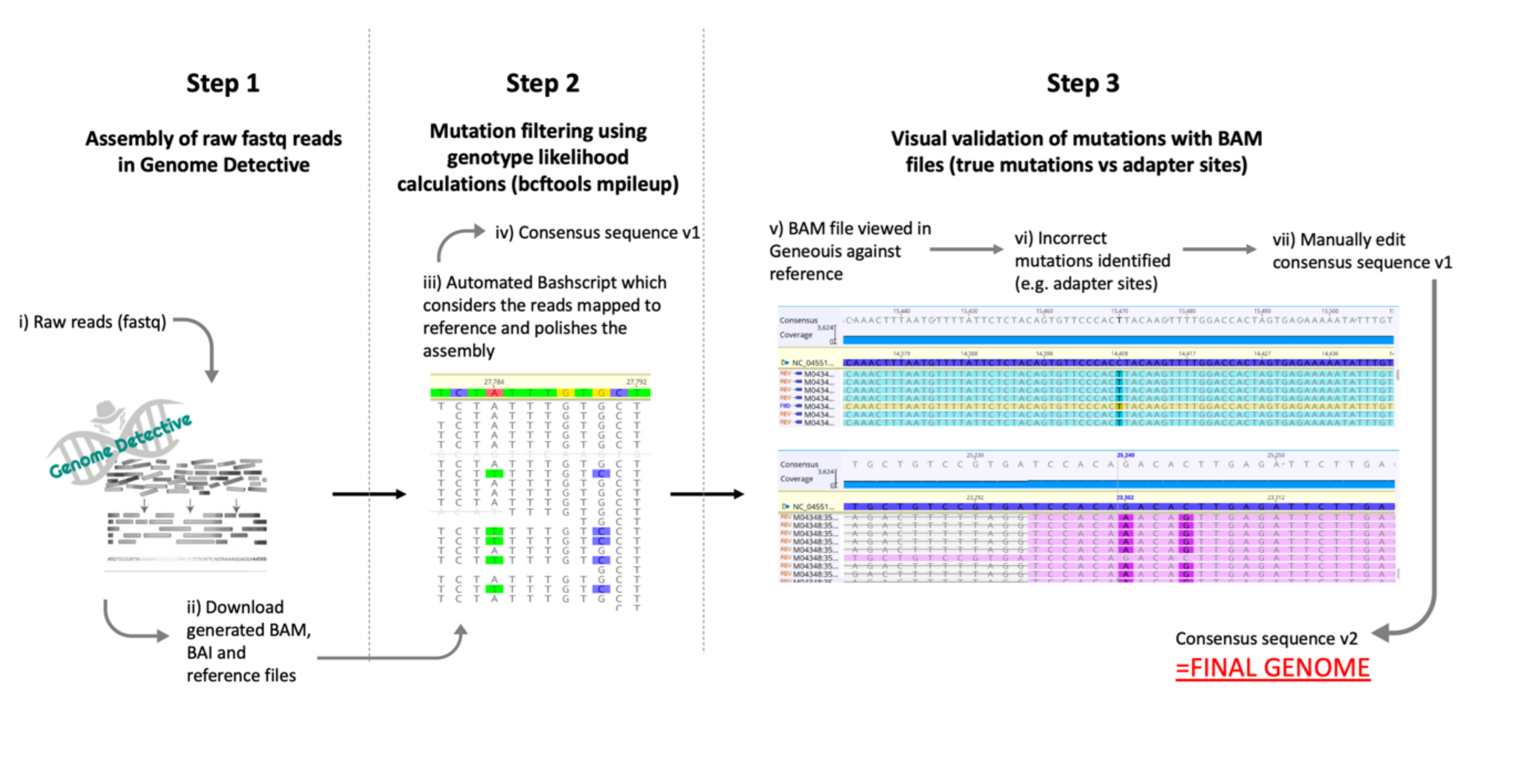
Acknowledgements

H3ABioNet Consortium

National Human Genome Research Institute: U24HG006941

Assembly of SARS-CoV-2

Objective: To quickly and accurately generate high-quality whole genomes of SARS-CoV-2 that are well annotated and available in public databases.



Step 1

Assembly of raw fastq reads in Genome Detective

i) Raw reads (fastq)

ii) Download generated BAM, BAI and reference files

Step 2

Mutation filtering using genotype likelihood calculations (bcftools mpileup)

iii) Automated Bashscript which considers the reads mapped to reference and polishes the assembly

iv) Consensus sequence v1

Step 3

Visual validation of mutations with BAM files (true mutations vs adapter sites)

v) BAM file viewed in Geneouis against reference

vi) Incorrect mutations identified (e.g. adapter sites)

vii) Manually edit consensus sequence v1

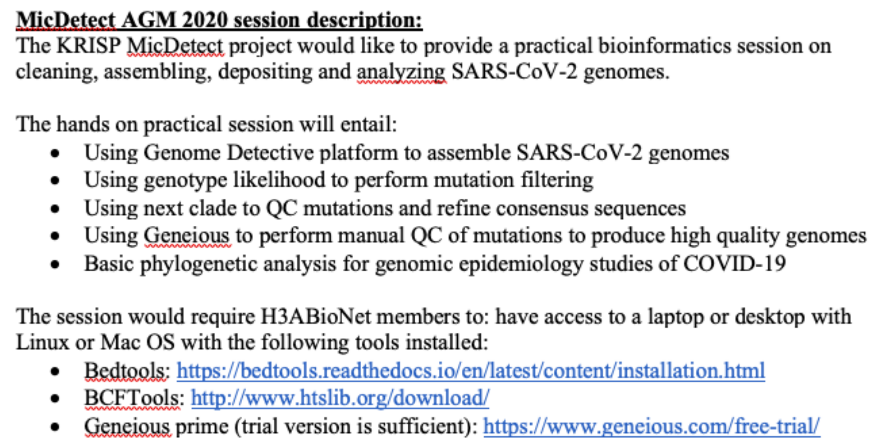
Consensus sequence v2

=FINAL GENOME

Slides courtesy of: Houriiyah Tegally, San Emmanuel James and Prof. Tulio De Olveira, KRISP, UKZN
https://www.genomedetective.com/app/typingtool/virus/

https://www.science.org/doi/10.1126/science.abj4336

Data curation for SARS-CoV-2



MicDetect AGM 2020 session description:
The KRISP MicDetect project would like to provide a practical bioinformatics session on cleaning, assembling, depositing and analyzing SARS-CoV-2 genomes.

The hands on practical session will entail:

- Using Genome Detective platform to assemble SARS-CoV-2 genomes
- Using genotype likelihood to perform mutation filtering
- Using next clade to QC mutations and refine consensus sequences
- Using Geneious to perform manual QC of mutations to produce high quality genomes
- Basic phylogenetic analysis for genomic epidemiology studies of COVID-19

The session would require H3ABioNet members to: have access to a laptop or desktop with Linux or Mac OS with the following tools installed:

- Bedtools: https://bedtools.readthedocs.io/en/latest/content/installation.html
- BCFTools: http://www.htslib.org/download/
- Geneious prime (trial version is sufficient): https://www.geneious.com/free-trial/

Consensus v1 — Manual correction of wrong mutations — Consensus v2 = FINAL GENOME

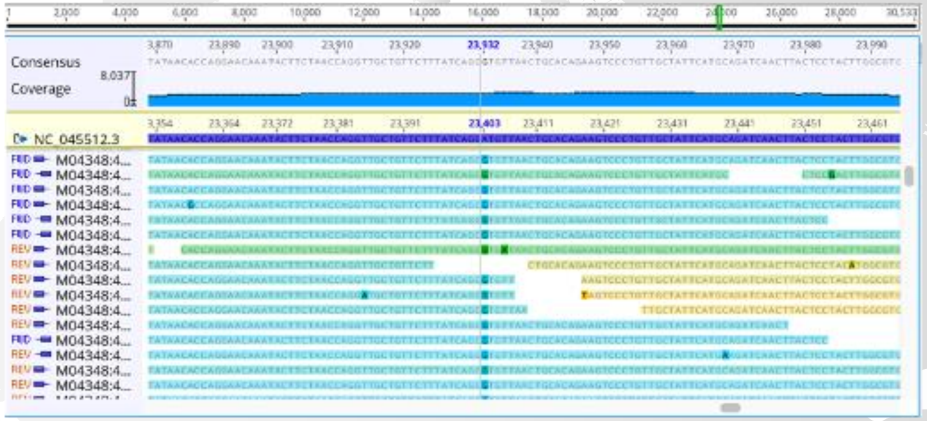
How to spot wrong mutations:

- Indels
- Mutations that cause stop codons in the middle of a genes
- Apparent adapter sites not being trimmed out properly
- “Expected” mutations

How to correct genomes:

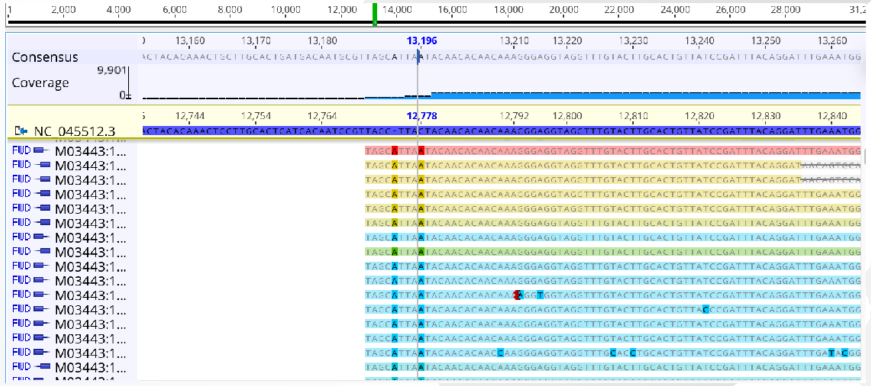
- Upload the FASTA sequencing into Genome Detective
- Find location to be edited by inspecting the mutation on Genome Detective and copying a short length of sequence before or after the mutation
- Search for that sequence in the FASTA text file
- Manually edit sequence in FASTA text file (back to the reference base if mutation is wrong)
- Be very careful at this step to not mistakenly insert other false mutations

Correct mutation



Here, we are looking at a correctly called mutation, indicated by consistent base change (23403A>G), mostly in the middle of long reads.

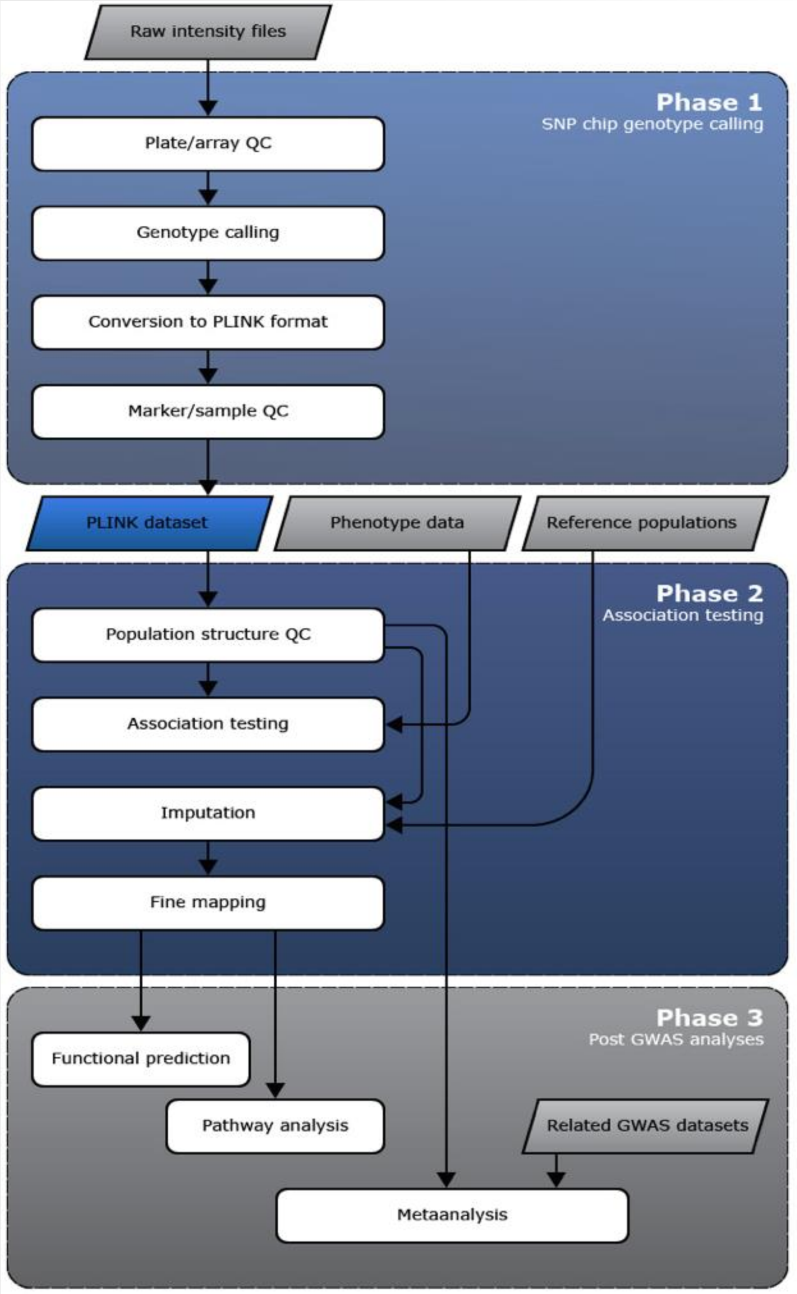
Untrimmed adapter sites



Here, we are looking at an incorrectly called mutation (12778C>A), which appears to be the result of an adapter site that was not trimmed. These sites will often appear at the end of reads.

Slides courtesy of: Houriiyah Tegally, San Emmanuel James and Prof. Tulio De Olveira, KRISP, UKZN
https://www.genomedetective.com/app/typingtool/virus/

Reproducible Science - Pipelines



Raw intensity files

Phase 1
SNP chip genotype calling

Plate/array QC

Genotype calling

Conversion to PLINK format

Marker/sample QC

PLINK dataset

Phenotype data

Reference populations

Phase 2
Association testing

Population structure QC

Association testing

Imputation

Fine mapping

Phase 3
Post GWAS analyses

Functional prediction

Pathway analysis

Related GWAS datasets

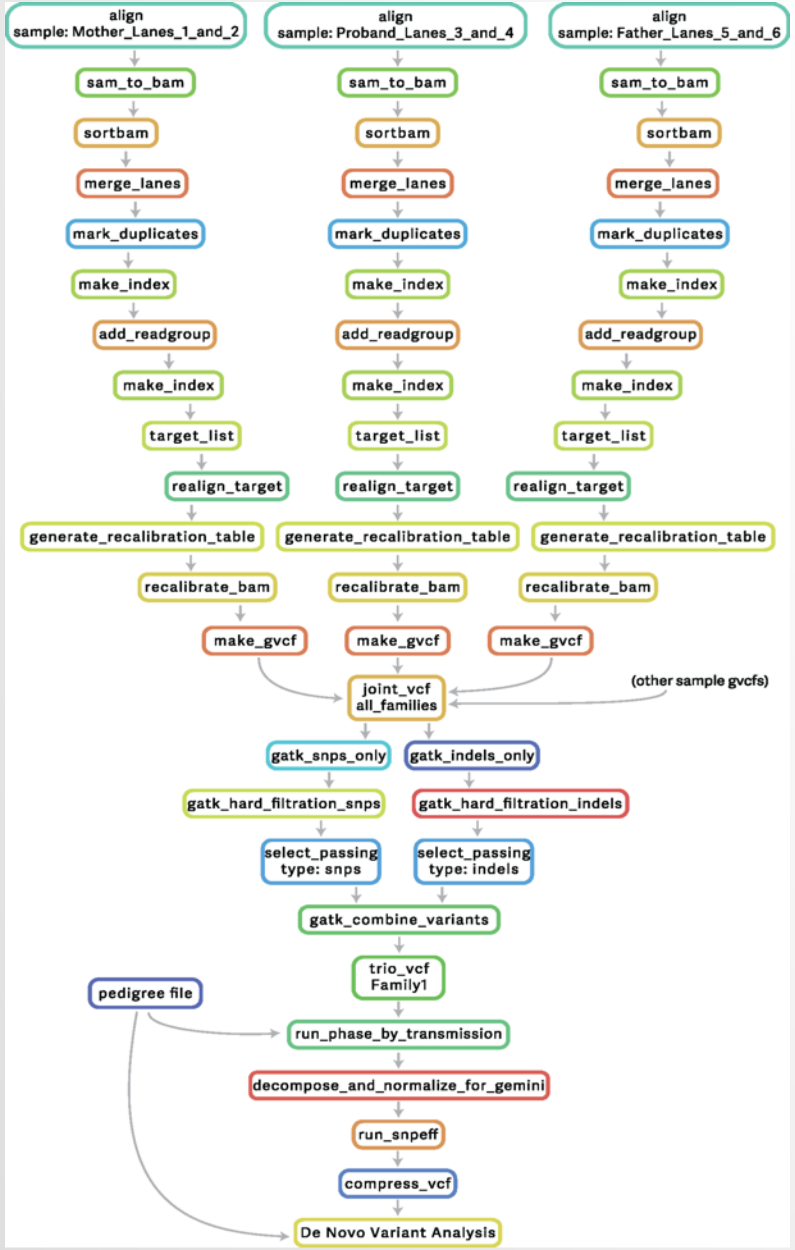
Metaanalysis

- Bioinformatics analyses □ directing files through a series of transformations and programs to a final output i.e. a computational pipeline

- Transformations typically done by third-party executable command line software written for Unix-compatible operating systems

- Manually started when previous transformation step completed e.g. qsub command run

Reproducible Science - Workflows



align
sample: Mother_Lanes_1_and_2

align
sample: Proband_Lanes_3_and_4

align
sample: Father_Lanes_5_and_6

sam_to_bam

sortbam

merge_lanes

mark_duplicates

make_index

add_readgroup

make_index

target_list

realign_target

generate_recalibration_table

recalibrate_bam

make_gvcf

joint_vcf
all_families

(other sample gvcfs)

gatk_snps_only

gatk_indels_only

gatk_hard_filtration_snps

gatk_hard_filtration_indels

select_passing
type: snps

select_passing
type: indels

gatk_combine_variants

trio_vcf
Family1

pedigree file

run_phase_by_transmission

decompose_and_normalize_for_gemini

run_snpeff

compress_vcf

De Novo Variant Analysis

- A workflow is a description of a process (pipeline) that consists of a series of tasks connected in the form a directed graph

- Tasks can be defined as single units of work e.g. split files

- A workflow comprises of an initial unique task and ends with a unique terminal task

- Completion of a task can initiate one or more tasks

- Enables automation e.g. run a pipeline from start to finish without manual input (pipeline)

Image credit: Leipzig J. Brief Bioinform. 2016;18(3):530-536. https://doi.org/10.1093/bib/bbw020

H3ABioNet Workflows project

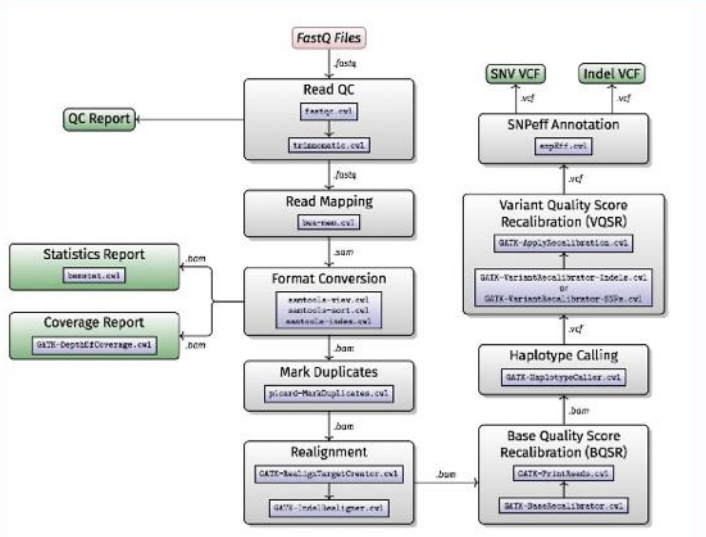


Figure 1 - Workflow A: Whole Genome/Exome NGS Data Analysis

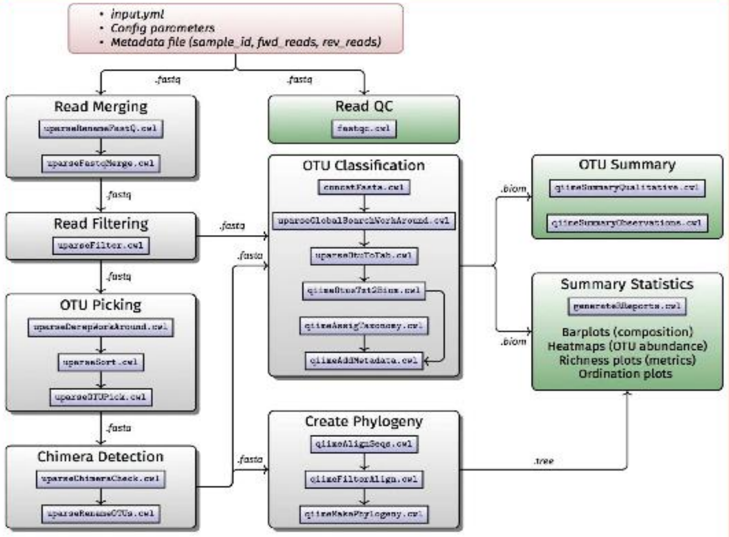


Figure 2 - Workflow B: 16S rDNA Diversity Analysis

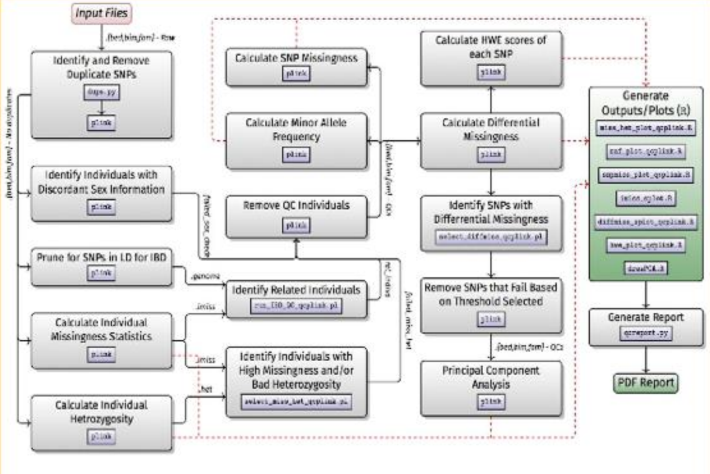

Input Files
.{bed,bim,fam} - Raw
Identify and Remove Duplicate SNPs
dups.py
plink
Identify Individuals with Discordant Sex Information
plink
Prune for SNPs in LD for IBD
plink
Calculate Individual Missingness Statistics
plink
Calculate Individual Hetrozygosity
plink
.genome
.imiss
.imiss
.het
failed_sex_check
Calculate SNP Missingness
plink
Calculate Minor Allele Frequency
plink
Remove QC Individuals
plink
Identify Related Individuals
run_IBD_QC_qcplink.pl
Identify Individuals with High Missingness and/or Bad Heterozygosity
select_miss_het_qcplink.pl
.{bed,bim,fam} - QCs
rel_indivs
failed_miss_het
Calculate HWE scores of each SNP
plink
Calculate Differential Missingness
plink
Identify SNPs with Differrential Missingness
select_diffmiss_qcplink.pl
Remove SNPs that Fail Based on Threshold Selected
plink
.{bed,bim,fam} - QCs
Principal Component Analysis
plink
Generate Outputs/Plots (R)
miss_het_plot_qcplink.R
maf_plot_qcplink.R
snpmiss_plot_qcplink.R
imiss_qplot.R
diffmiss_splot_qcplink.R
hwe_plot_qcplink.R
drawPCA.R
Generate Report
qcreport.py
PDF Report

Figure 3 - Workflow C: Genome Wide-association studies

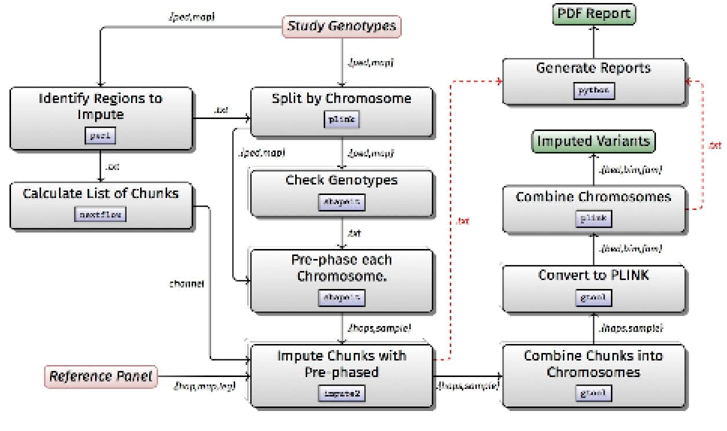


Figure 4 - Workflow D: SNPs Imputation: Boxed subgraphs indicate pathways which are executed in parallel (per chromosome and per region within each chromosome) as computational resources permit.

Image credits: Phelelani Mpangase, SBIMB, Wits
Availability: https://quay.io/search?q=h3abionet
https://github.com/h3abionet