The African Genomics Medicine Portal: grounding the computational environment for precision medicine in Africa

List and order of authors: to be determined after the writing of the paper and discussion in Editorial

Editorial Team Volunteers: Samar, Faisal, Kais, Nicky, Fouzia, Houcem, Jorge, Yosr, Michael, Judit, Yasmina

| **Section Volunteers** | |
| --- | --- |
| Abstract | Houcem, Yosr, Fouzia, Kais, Faisal |
| Introduction | Yosr, Fouzia, Houcem, Lyndon, Kais, Rym, Yasmina, Nicky, Lilia, Nessrine,Maroua |
| Grounding elements | Fouzia, Yosr, Lyndon, Kais, Jorge, Faisal, Anmol, Ziyaad, Melek, Samah, Samar, Chaimae, Reem, Rania |
| Portal database design (Methods) | Houcem, Kais |
| Data mining and data curation (Methods) | Houcem, Jorge, Lyndon  Maryame, Kholoud,Maroua, Nessrine, Lilia…. |
| Portal building (Methods) | Ziyaad, Alfred, Anmol, Kais, Michael, Ichrak |
| Results | Yosr, Houcem, Jorge, Lyndon, Fouzia. Kais, Yasmina, Lilia |
| Demo (section for results) | Ozlem, Bilinga, Chiamaka, Chaimae, Lyndon, Yasmina |
| Discussion | Yosr, Fouzia, Houcem, Kais, Reem, Balkiss, Ozlem, Rym, Judit, Faisal, Samar, Melek, Lilia |
| Conclusions and perspectives | Yosr, Houcem, Fouzia, Kais, Reem, Balkiss, Samar, Faisal, Judit |

**NB! Actions Items Prior to Publication**

* Portal Documentation to be completed
  + User Documentation - Background, How to Use Portal
  + Technical Documentation - Portal Design, Backend, Data Mining + Curation, Update Plan
* Final decision on Portal Name, Host Location
* URL to be changed to include “h3abionet” in url
* Portal update
* Usability testing

NB! The different sections of the paper should be added in “suggestions mode” to get a traceability record of contribution. “Add Comment” option in google docs is more suited for giving suggestions or remarks. If for any reason you have to add them to the main text, please do that in the Editing mode not the suggestions mode.

# 

# 

# 

# Abstract

Due to the reduced cost of sequencing and increased access to high throughput technologies, omics data are currently being produced at unprecedented rates, resulting in increased sizes of data being submitted to public biorepositories. Despite these advances, few genetic and genomics data on African populations are publically available for the scientific community compared to large scale data from other populations such as Europeans and Asians. Indeed, a substantial amount of African Genomics data is available in some public databases and there are many challenges related to the retrieval and availability of such data. To overcome these challenges we aimed to develop an African Genomic medicine portal (AGMP) that includes all African genomic data available in two main public databases related to genomic medicine namely PharmGKb and Disgenet. Data that has been collected in this proposed portal only includes African genomic variants that may have a clinical impact and can be used either in drug prescription or drug response…..Data were collaboratively collected and manually curated by H3ABioNet domain experts, including scientists, bioinformaticians and data scientists following a clear workflow. The African Genomics Medicine Portal is the first resource that hosts African specific genomics data and will contribute to facilitating access to these data by researchers, clinicians and patients.

**Keywords:** Africa, Clinical Genetics, Database, Genomics Medicine, Personalized Medicine, Pharmacogenomics, Portal, Precision Medicine

# Introduction

*Suggestion of the introduction content (suggestions by Rym, Judit, Reem):*

* *Precision medicine/genomic medicine background, importance*
* *Overview of the available databases, portals and other resources related to precision medicine.*
* *Genome sequencing data, bioinformatics and genome comparison*
* *Explain why it is relevant to have a Genome medicine portal specific to Africa (Some arguments: Lack of African data, data not standardized, data pulled with other populations... while on the other side Africa has several specificities (ethnic, genetic, nutritional, environmental, predisposition to diseases...)*
* *Explain purpose and need*
* *Who is the target audience*
* *Start the introduction with a paragraph highlighting the different nomenclature used in this field; such as "genomic", "precision", and "personalized" medicine; and also our decision to use the term "genomic or precision" in this article and for the portal. There are articles that discuss the subtle difference between these terms, for instance:* [*https://doi.org/10.1038/clpt.2013.101*](https://doi.org/10.1038/clpt.2013.101)*, https://doi.org/10.1038/nm0313-249*

Comment:

Many advances have been made on the pathophysiological and genetic knowledge of various diseases. This is mainly linked to the development of robust technologies to generate and understand sequence variation (Lima Marson et al 2017; DM Roden and RF Tyndale et al,2013). Several concepts have emerged since these advancs including Genomic Medicine, Precision Medicine, and Personalized Medicine. Genomic Medicine is defined as the use of the genotypic information about an individual as part of their clinical care (DM Roden and RF Tyndale et al,2013). Personalized medicine is the tailoring of a medical treatment based on genomic, epigenomic, proteomic profile and clinical information of an individual. This approach increases our capacity to predict which medical treatments will be reliable and effective for individual patients, and which ones will not be based on their genotypic data (Sunil Mathur and Joseph Sutton 2017).

Precision medicine is the concept of tailoring disease treatment and prevention to account for differences in genetics, environmental, or even lifestyle factors that are specific to groups of people or sub populations. It gathers genetic and biochemical information unique to a group of patients and uses that information to develop more specific and streamlined medications or treatments (Bresnick, Jennifer. 2018). Ultimately, precision medicine represents the significant enhancement of evidence-based medicine, rapid advances in personal, cohort, and population-scale data acquisition, through sequencing, genotyping arrays,, mass spectroscopy, biosensors, medical devices including mobile health devices, imaging e.g. X-ray, MRI, ultrasound, social network activity, electronic health records (EHR) and other applications toward new perspectives for personalized health biomedical data collection all over the world (Alexander A. Morgan, 2016).

In African populations, the huge diversity that affects disease risk, progression and response to treatment applies to both communicable and non-communicable diseases (Ref1). Nowadays, the challenges encountered by patients are compounded by the complexities associated with poly-pharmacy and comorbidity which requires careful management of medication type and dosages for the most optimal patient benefits, at the same time reducing the incidence of drug induced adverse effects (Ref2). The interaction between genomics and nutrition is also playing an increasing role in human health and wellness (Ref3). Thus performing a literature search on Precision Medicine of African populations was the first performed work which allowed us to identify the most important genomic markers to guide precision medicine or precision public health in African (Ref4, Radouani et al. 2020). Simultaneously, the idea came to make this information about these genomic markers and their phenotype accessible through an African precision database portal, aiming to provide guidance for policy makers for the implementations of precision medicine in Africa. The proposed portal will provide curated resources to guide the application of pharmacogenomics, clinical genetics and nutrigenetics/ nutrigenomics within a highly diverse population.

diagnosis therapeutic decisions.

The paradigm of precision medicine as robust techniques for characterizing the genotype-phenotype relationship become more and more in use. The first pillar in precision medicine is the ability to collect data related to disease manifestation and environmental factors. The second pillar is pharmacogenetics/pharmacogenomics, aiming to determine the relationship between the genetic marker characters of the population or an individual and the drug response. Like food, drugs or pharmaceutical agents once ingested, undergo both a pharmacokinetic (the action the body takes on the drug, i.e., metabolism and drug metabolism respectively) and pharmacodynamic processes (the action the drug has on the body) (Kumuthini et al.. 2016).

The diverse reaction to a drug i.e. favorable and unfavorable due to the genetic characteristics has been described in many instances. For example, CYP2D6\*29 results in a decreased function of the enzyme. Its frequency in Caucasian and Asian populations is very low, while it is highly prevalent in Tanzanians with a 20% frequency ( 11470994). Therefore, it is critical to establish the pharmacogenetic/pharmacogenomic landscape population-wise to account for genetic diversity. Populations in Africa have been described as the most diverse in the world with a complex human migration history and admixture (PMID 33116287).

Many collaborative initiatives were constituted worldwide to provide support for precision research implementation such as the ICPerMed consortium in Europe, the genome/phenome database in Singapore and the Personalized Medicine Coalition in North America.

There is a growing interest to optimize treatment of both rare and common diseases in the African continent where the severity of the disease and treatment response are affected by genetic variants.

An accurate investigation of actionable genetic variants is required in genomic medicine research including investigating druggable variants and variants related to drug response. Most of the variants that are available for African populations are those involved in drug response that can be found in the PharmGkB database, however very few are known on druggable and other actionable variants in Africa. Moreover, genotypic as well as phenotypic data related to African populations are scattered in different public databases in a badly annotated manner. This prompts us to collect all these data and to develop an African Genomic medicine portal that includes all relevant African specific genomic data. The development of this portal will be of keen interest to the African scientific community, physicians and all precision medicine stakeholders worldwide as it will have a direct impact on therapeutic decisions, public health care and personalized medicine practice.

**Suggestions to be included (from Fouzia, Judit)**

* **Type of genomic data needed for PM research**
  + **What are they?**

* + **What is already available?**
* **Lack of African PM databases available internationally and especially in African context.**
* **Need for African PM Resource**

**The need for an African resource is justified by the extensive evidence for the genetic diversity among populations across the continent supported by archaeological findings . Through its history, the African continent has witnessed successive invasions, colonization, admixture, several migratory flows combined with a great variety of evolutionary forces (genetic drift, natural selection) thus shaping its genetic landscape which is highly diverse. In fact, populations across Africa are highly subdivided with nearly as many ancestries within Africa as in other continents combined (Campbell et al. 2014; Tishkoff et al. 2009).**

**In the context of the African continent, the 1000 Genomes Project focused only on some sub-Saharan African populations genetically close to the Yoruba (Nature 2015). Other genome projects, namely H3Africa, Egypt Genome and African Genome Variation, have integrated samples from individuals predominantly from Niger-Congo-speaking populations. Several populations, especially North African groups, are poorly or not represented (Rotimi et al. 2017). This result calls for the inclusion of more diverse African samples in public genomic reference panels and highlights the urgent need of resources with available African genomic data.**

**Only a small fraction of the African diversity left the continent to migrate to Asia and Europe. There are four major linguistic families in Africa : the Niger-Congo, Nilo-Saharan, Afro-Asiatic and Khoesan, that have a distinct geographic distribution. It is estimated that there are more than 2000 languages, which is an indication of the complex demography. Several events of human migration have contributed to shaping the present day African genomes. The Bantu people moved from West-Central Africa to Southern and Eastern Africa and came across local populations resulting in admixtures (Tishkoff et al 2009). People had to adapt to new ecosystems, lifestyle and climate conditions. Genome wide scans have identified signatures of selection in specific loci that explain the adaptation process. Exposure to pathogens is said to be the most important type of selective pressure that led to gene evolution and regional variation. Evidence points to variations in immune genes that have resulted from selection. Fifty-seven genes of the innate immune system have variants that can explain susceptibility to infectious and non-infectious diseases (Deschamps et al 2016).**

* **What problem the portal is going to solve ?**
  + How is it going to achieve that?
* Problematic, project objectives and goal from implementing the portal (Fouzia, Judit, Gordon, Faisal, ...)

**Objectives**

The main goal of this initiative consists of the development and implementation of an African precision medicine portal or African precision medicine database which will help to: (1)Develop a road map on the implementation of precision medicine in Africa; (2) Develop a user-friendly resource specific to African data to support precision medicine in Africa; (3) Increase awareness on the role of genomic solutions in the management of human health and; (4) Provide the bases for potential inter-population comparative studies within African countries, and with other non-African countries.

# Methods

## Strategy & steps followed to develop the portal

**Suggestions of content to be included**

* + Pre-hackathon activities
    - (discussed scenarios, users, …)
    - meetings before hackathon
    - Idea of constitution of the groups (Content team, technical team)
    - Agenda to succeed the Hackathon
  + Summary of Work during Tunis hackathon (work and discussion during the hackathon, meetings)
  + Summary of Work during Cape Town hackathon
  + Post hackathons Keep working
  + **Content team**
  + Data extraction
  + Data curation
  + **Technical team**

The H3ABioNet Management Committee (MC in agreement with consultation with the Database and resources (DR) work package chairs selected active members across the H3ABionet precision medicine project to participate to a «Hackathon» in order to initiate and catalyze the development of the African Genomic Medicine portal. A total of 8 participants (content team) and 2 participants (technical team) from the Precision Medicine working group, representing 7 African countries met face to face in Tunisia for this Hackathon a week's work, from 12 to -18 April 2019 at Institut Pasteur de Tunis, and discussed launching the portal implementation (Ref hackathon paper). ~~The selected working group members and project leaders met monthly by conference calls to organize the hackathon.~~

The hackathon program was based on a series of presentations and talks highlighting the main aims of the portal, are the portal design and development breakout group sessions. In addition, the hackathon was dedicated to a detailed presentation of the use case scenarios, the different data aspects of the data that need to be integrated in the portal, how the portal will be designed and implemented, and finally how the output will be displayed for users.

+

A second hackathon on the development of the African genomic medicine portal took place in Cape Town, South Africa from 02-08 August 2019. During this hackathon, the technical team took into consideration the significant challenges that were faced during the first hackathon regarding automated retrieval of data through APIs and switched to a static relational database design of the portal that needs significant efforts in data curation. During this hackathon, the content team also decided to switch from GWAS Catalog to DisGeNET as the latter includes GWAS Catalog data in addition to many other data resources such as ClinVar, DbSNP, ClinGen and VIP. In addition, the portal interface was further designed and refined during this second hackathon (Ref hackathon paper).

After a laborious exploration of various databases, Contextual filters were set to retrieve African data from DisGeNET and PharmGkB databases; however, it was later found that a lot of African data were misclassified and annotated in larger regions and ethnic groups. Therefore, the content team opted instead to employ a manual mining approach to extract and curate African-specific data.

**Figure 1: Here, we can implement an workflow to summarize**

**work steps (Fouzia, see Kais’s workflow from the hackathon paper)**

~~After a laborious exploration of various databases from which data could be retrieved using APIs, including PharmGKB, Clinvar, GWAS Catalog, dbSNP, BIOMART, OMIM, clinGEN, VIP, Monarch Initiative, MyVariant, PharmVar and others, followed by an establishment of a protocol to evaluate the most informative databases to be used to extract African data and identify metadata to be included, two databases were selected for initial incorporation into the African Genomic Medicine Portal, namely PharmGKB and DisgeNet. These databases were selected due to the classification of data records already incorporated within them, enabling African-related records to be easily retrieved.~~

Primary data sources

We have retrieved two datasets in plain text format from PharmGKB (22992668) and DisGeNET ( 31680165) databases corresponding respectively to 'var\_drug\_ann.tsv' andall\_variant\_disease\_pmid\_associations.tsv'. These tables contain variant based information regarding known drug response associations (PharmGKB) and associations related to disease phenotypes (DisGenet). Each entry is also linked to metadata regarding the study which produced the association, including a link to its Pubmed entry (PMID). Gene names were obtained for DisGeNet variants with Ensembl’s Biomart. These linked names allow for variants to be queried by name of gene. PharmGKB data already has gene names present within its data structure.

Text mining

The PMID data from the variant-drug and the variant-disease tables were submitted to a text mining workflow (https://github.com/hothman/text\_mining/blob/master/workflow/mining\_medline.nf). This aims to filter the bibliography entries that contain data about variants identified from African ancestry. The workflow uses the 'EDirect' to fetch the PMID entries in XML format. A natural language processing treatment is then applied over the Title and Abstract fields of each entry for tokenization and filtering the relevant words of the processed text. Thereafter, we search for matches from a lexicon of 252 words that contains a list of countries, nationalities, and descriptions with African ancestry as well as a list of ethnolinguistic groups from Africa. The text mining process employs an in-house Python script to automate the search.

Moreover, all the references related to variants identified in populations with African ancestry were retrieved in XML format from PubMed. The title, publication year, and ID in the primary source (PMID) were extracted using an in-house series of python scripts to generate the table of all the bibliography resources linked to the curated entries.

Manual curation of data

The retained PMID entries were examined to remove any false positive hits that result from the text mining processing, and to extract the p-value data of the association between the variants and the phenotypes. The manual curation stage was also employed to harmonize the data about biogeographical groups and to extract the data about the country of origin of each study. All along this process, entries corresponding to studies about populations from African ancestry were labeled with one of the following nomenclature: Sub-Saharan Africa, North Africa, African American, Mixed Population. The latter label includes populations that partially contain individuals from African origins. Data about p-values were retrieved either from the PharmGKB metadata field or by checking the corresponding paper of the study.

Data mining and curation for Drug information

Drugs emerging from the PharmGKB variants-phenotypes list have been annotated by identifying the commercial drug names, the FDA approval status, the indication information, and their conventional name from the International Union of Pure and Applied Chemistry (IUPAC). All this information was extracted manually from DRUGBANK (PMC1347430) (https://go.drugbank.com/) and the corresponding drug ID was added to the annotation data.

Data mining and curation for Gene information

All the genes associated with drug response or disease phenotypes were annotated to include information about their gene product, chromosome, and biological function. The gene product is identified by mapping the gene name to the corresponding Uniprot ID using BioMart. The biological function data were obtained by manual extraction and formatting of the related field from the Uniprot database.

Data mapping, integration, and quality assessment

Data integration consists of applying the Extract, Transform, and Label (ETL) procedure to integrate data from different sources into a relational database. A series of interactive transformation codes were applied for such a purpose. Manually curated data were examined for their consistency of using unique terminology in each column. All the data found to describe populations from Non-African ancestry were discarded from the source table. Any entry with a missing p-value was assigned to an empty character, while association studies reporting p-values for mixed populations including Africans were attributed to the label 'ambiguous'. The quality assessment process consists of evaluating the fraction of missing data, the correctness of mapping to the target tables, harmonization of terminologies, and class labels, and the presence of duplicates.

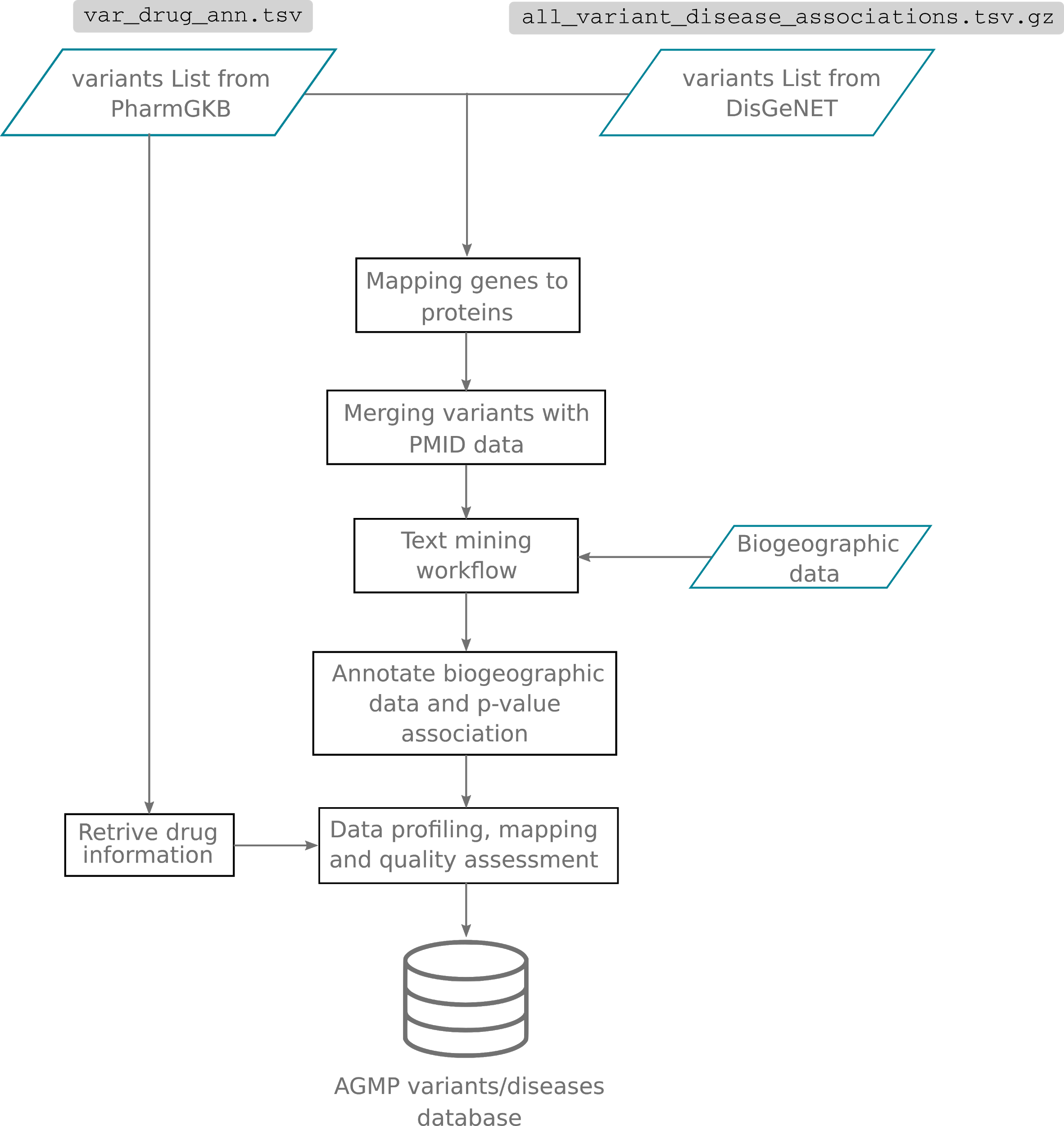


Figure2 : Workflow for data mining, curation and cleansing of the content for the AGMP portal.

## Portal building

**Suggestions of content to be included (Houcem, Michel, Anmol)**

* Algorithms and languages used to implement
* Other than Django, what libraries (python, js, css) were used??
* Deployment: Docker/NoDocker, web server used

The search menus have been set up on the basis that most of the users would need prior information about at least the gene, the variant, the disease, or the drug to easily explore the content of the portal. The portal has been designed in a gene-based centralized style which ensures that all the information relative to a gene and related to genotype-phenotype data as well as the associated metadata are harbored in one accession. With such implementation, querying, navigating, and extracting information from the portal would require only a few prior inputs from the user.

**Results**

**Our goal behind the development of the portal is to provide a reliable source of information regarding the actionable genetic variations identified in populations of African ancestry. Although the term precision medicine loosely incorporates the many disciplines to deliver a targeted therapeutic solution for a specific group, we have chosen to focus on linking genetic variants to drug response and disease manifestation. We have therefore utilized PharmGKB and DisGeNET as the main data sources of the portal mainly for their high-quality data content and for providing comprehensive and easy access to their raw data.**

**DisGeNET includes more than 357000 variants related to 66379 bibliography resources. PharmGKB includes 9151 variant-drug associations including single nucleotides and star alleles described by 3007 papers (data from August 2019). An exhaustive search for the African-related data content would be very difficult to achieve therefore we opted to apply a text mining approach to reduce the size of the bibliography resources to be curated. The text mining processing allowed us to identify 846 and 164 studies from DisGeNET and PhamGKB respectively. The literature identifiers were then merged to result in a total of 479 and 2584 variants that were manually curated from PhamGKB and DisGeNET respectively.**

## 

## Overview of the data model for AGMP

The content of the data sources belonging to PharmGKB, DisGeNET, Uniprot, DrugBank, Ensembl, and PubMed was implemented into a relational database containing 6 tables according to the Entity-Relationship Diagram (ERD) in figure XXX. The star allele variants from PharmGKB were created to account for the particularity of annotating Absorption Distribution Metabolism Excretion (ADME) genes. A star allele is defined by a set of core variants defining the phenotype associated with drug response in linkage disequilibrium with other variants whose presence or absence is not necessarily related to the phenotype outcome (PMID 26479518). The user is invited to review the content from PharmGKB, PharmVar (PMID 29134625, 30536702), or the bibliography if interested in a detailed description of the variant composition of each start allele. The table 'SNP' contains the data about single nucleotide variants from PharmGKB and DisGeNET. Along with the 'star\_allele' table, they constitute the source of information about the variants and their associated phenotypes which could be related to either the response to a drug or a disease. The population-specific component of the portal is also implemented in these two tables providing information about the country of origin of the study and the region. The other four tables were created to provide additional metadata for the variant-phenotype association. For instance, the gene table stores data about the genes encoding the variants while the drug table accommodates detailed data for the drugs described by the variant-phenotype associations including their approval status. Literature resources from which we have curated the association data are incorporated in the 'Study' table.



## Figure XXX. Entity relationship diagram of the data content for AGMP.

## Data content of the AGMP

A summary of the content of the unique entries of AGMP are given in Table XXXX. The manual curation retained 815 total studies including 143 references from PharmGKB and 652 references from DisGeNET that are associated with populations of African ancestry. We have identified a total of 556 genes of which 461 and 95 are related to 552 diseases and 71 drugs respectively. This represents only a fraction of 5.5 % and 2 % of the total genes annotated in PhamGKB and DisGenNET. The total number of variants associated with the drug response phenotype is 286 of which 234 correspond to single nucleotide variants and 52 are star allele variants.

Table xxx. Count of unique entries in the database of AGMP

| Data | Count |
| --- | --- |
| SNVs | 1186 |
| Star alleles | 52 |
| Genes | 556 |
| Drugs | 71 |
| Studies | 815 |
| Diseases | 552 |

Representativity of African researchers and data in AGMP content

Studies about genetic markers from populations of African ancestry have been conducted by different research teams from all over the world. The ‘country’ data of AGMP aims to assign the annotated variants in the database to a geographical entity. A geographical entity is an indicator of the place where the study cohort was collected such as the country and the region names. The assignment of the geographical entity is depending on the level of details given by a study many of which have used a mixed population either of total African individuals or combined with non-Africans. For example, Zimmerman et al (PMID 9132277) aimed to study the polymorphism of chemokine receptor 5 (CCR5) in a group of mixed ethnicity including Africans. However, they did not give the details about the exact origin of the African cohort, therefore their geographical entity was assigned to "Africa".

AGMP includes data from 37 African countries. With 568 variants, Tunisia is the country with the most annotated variants which represents a proportion of 23% of the entire dataset. We have also noticed a significant imbalance of representativity between the different regions of Africa. With 66% of entire shared variants, countries from North Africa are by far the most represented in the dataset. We also noticed an important gap between the central African region and the rest of the regions. With only 52 variants corresponding to a proportion of 2.4%, countries from these regions are the least represented in the AGMP dataset despite its importance in the migration history of Africa.

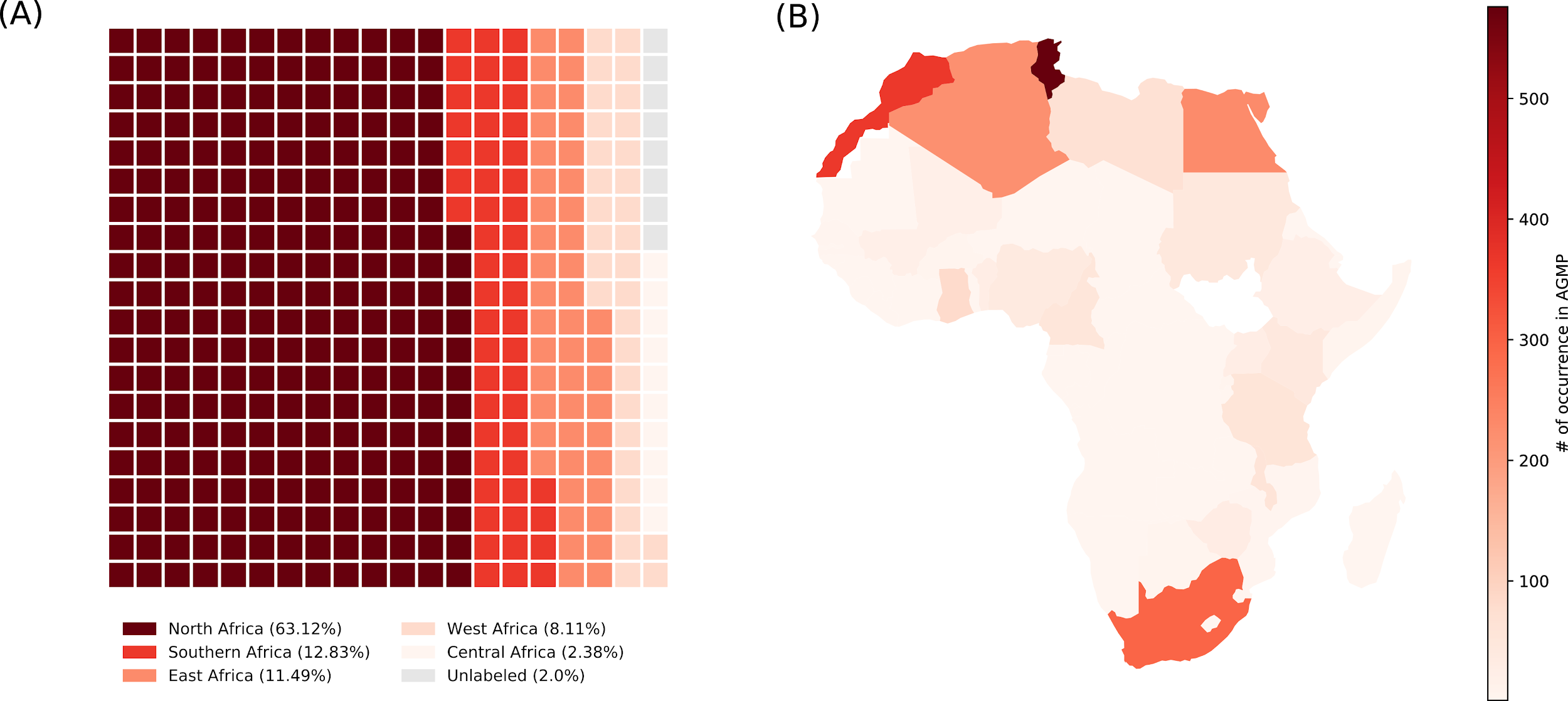


Figure XXXX. Representativity of African countries in AGMP. (A) the waffle plot shows the ratio of occurence in the geographical location annotation data by African regions (North Africa, Southern Africa, East Africa, West Africa and Central Africa). The countries were affected to their corresponding region according to the UN classification. The squares are arbitrary units of proportion. The entries that have been affected to a geographical entity in Africa but not to a country are unlabeled. (B) Representativity of countries according to their occurrence in the geographical location.

Pharmacogenetic related phenotypes

Variants identified in PharmGKG are associated with drug response phenotypes for 50 drugs approved for specific treatments. Warfarin is the drug with the highest count of associated variants with a total of 192 associations representing about 7% of the entire dataset. Together with efavirenz, they represent the only drugs in the database with more than 20 annotations. Anticoagulant drugs are the most represented therapeutic class associated with 37% of the total annotated variants.

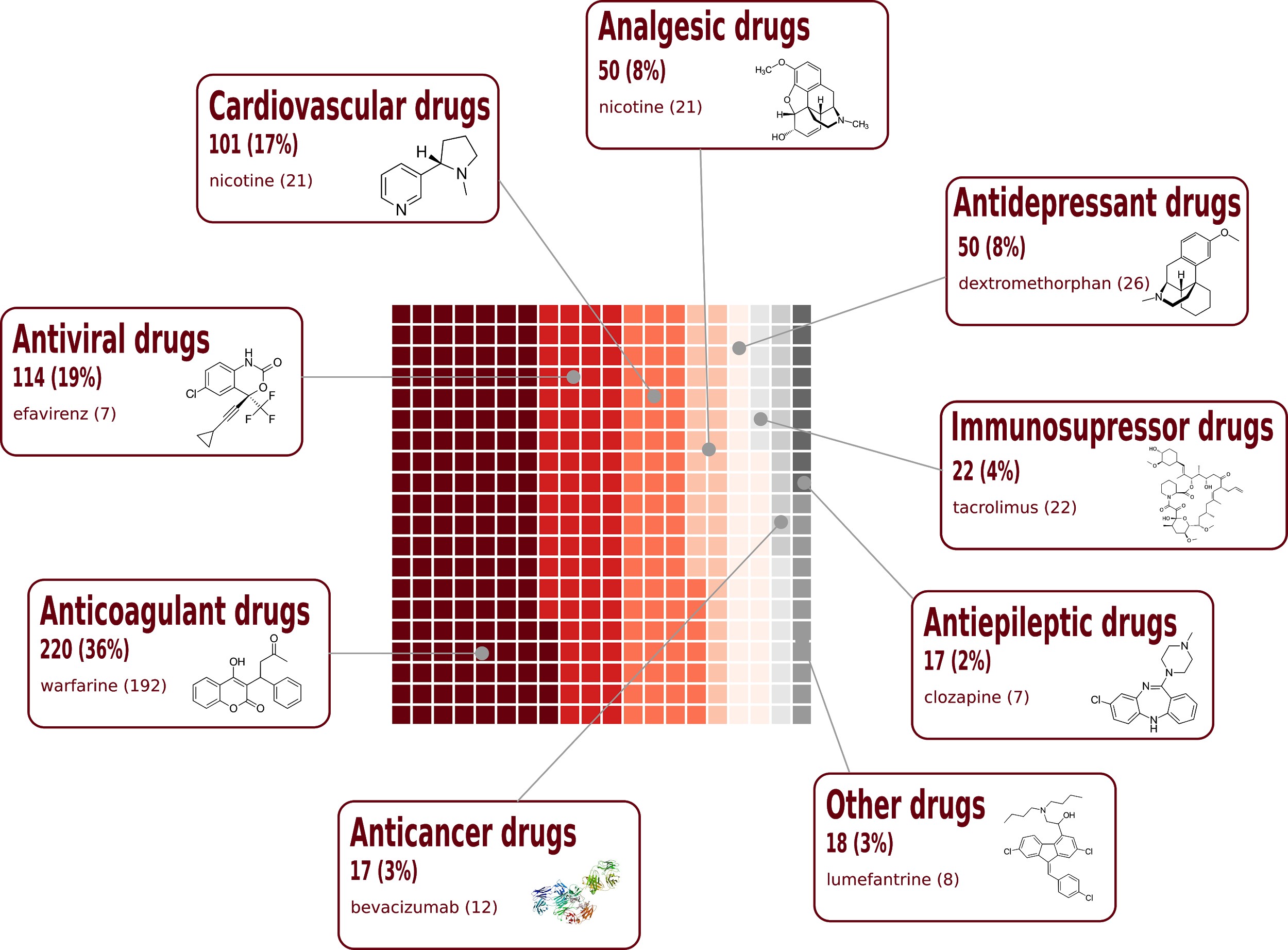


Figure XXXX. Representativity of therapeutic drug classes in AGMP. The drug with the highest count of variant-drug response phenoy association is given in the annotation boxes along with the statistics about the total count.

**Suggestions of content to be included (Fouzia, Houcem)**

* Acknowledgements of data curation team
* Alber D, Ayoub, Bilinga, Chaimae, Chiamaka, Fouzia, Haifa, Houcem, Imane, Kais, Kholoud, Lyndon,Maryame, Maroua, Melek, Mohamed Ahmed, Nessrine, Noha, Olivier, Rania, Reem Sallam, Rym Kefi, Samah, Samar, Yosr Hamdi

## **Demo (By Ozlem, Bilinga and Chiamaka, Samah, Reem, Lyndon, Chaimae, Rania)**

**Portal access by users (Subject to change with the portal update)**

The African Genomics Medicine Portal (AGMP) provides semantic information across a user-friendly navigation bar menu categorized into: Home, About, Search, Data Summary, Resources and Help sections (Fig. 1). Through the link:<https://agpm.knust.edu.gh/>, the user can access AGM web-portal homepage (Fig. 1A), where overall information pertaining to the portal is provided. Comprehensive information should be accessed by clicking the “About” section (Fig. 1B) found at the top right on the portal page. The “Search” section (Fig. 1C) provides the user with various multi-search functionality which can be navigated by disease, drug, variant or gene. More importantly, a click to “Data summary” section opens a portal page with an overview of the data contained in the AGMPThe locations from where the data is derived is also illustrated in a user-friendly map.(Fig. 1D).

# 

# Fig. xx: Illustration of the AGMP portal interface as displayed by various portal pages. A represents the “Home” page, B represents the “About” page, C represents the “Search” page, D represents the “Data Summary'' page, E and F represents the Databases and Online courses page which is found under the “Resources” icon, and G represents the “Help” page.

Considerable amount of data from several sources has been accrued in the AGMP. This webpage can be accessed by clicking “resources” on the bar menu which enables the user to choose from three categories: databases, tools & pipelines, and online courses (Fig. 1F). Notably, these categories have been interlinked to provide the user with necessary guidance to obtain data. For instance the database functionality exposes the users to various external databases such as PharmGKB, Clinvar and KEGG (Fig. 1E). Additionally, a click to the help bar menu (Fig. 1G) directs the user to the portal tutorials and frequently asked questions (FAQ).

AGMP mainly focuses on the association between diseases, drugs, genes and variants (Fig 2). Therefore, further details of each of the fore-mentioned multi-search functionality can be obtained by clicking on the specific logo of interest (Fig. 2A, B, C, D) followed by either typing in the name of disease, drug, gene or variant accession number in the search field and then clicking the search symbol.

For instance, a search for malaria entails the user to click the disease symbol “Disease” and type “malaria” in the search bar followed by a click to the directive “click here for disease association” (Fig. 2A). Similarly, a search for warfarin will involve clicking on the drug symbol, typing warfarin in the search bar (Fig. 2B) and a click to the directive “click here for drug association” as seen in Fig. 2B.1.

Subsequent searches by gene and protein sections require a click to either the variant “variant” (Fig. 2C) or gene “Gene” (Fig. 2D) symbol followed by either the variant accession number (rs28371685) or gene’s name (CFH). This will lead to Fig 2C.1 and Fig. 2D.1 respectively.

The resultant pages (Fig. 2B.1, 2C.1, and 2D.1) gives elaborate information about diseases, the diverse variants and genes in the African population associated with these diseases and their response to specified drugs. Dependently, information on the literature related to these categories, their study population, country and the level of significance can also be obtained. Other additional functionality includes: sorting information by use of drop down menu, as well as, retrieving information in form of CSV and excel which can be copied, downloaded or printed.

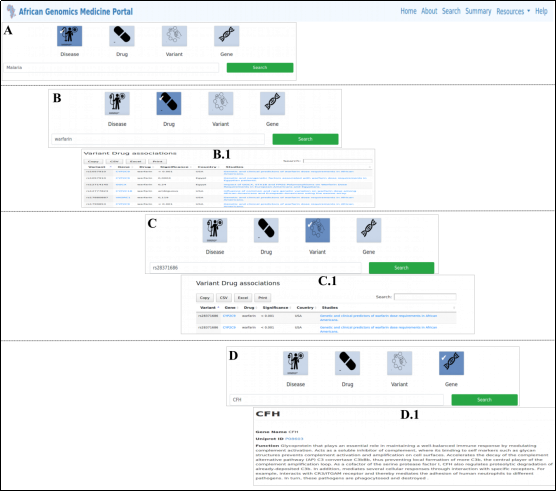


Fig 2. AGMP search by disease, drug, variant or gene results page. The results are divided into disease (A), drug (B), variant (C) and gene (D). These sections lead to their respective resultant pages (B.1, C.1, D.1).

# 

# 

# 

# Discussion

The main objective behind building the AGMP is to implement the first resource prototype that provides high quality and reliable support for information about genetic variants of populations from Africa or with African ancestry with interests in precision medicine.

The primary sources of data were fetched from PhamGKB and DisGeneNet These provide rich metadata for genetic variants in a standard format style and controlled terminologies which helped us to easily implement them according to ourdata model.. The two databases are also recognized for the high quality of their content mainly because of the rigorous process of curation. PhamGKB for example has a dedicated panel of experts that are constantly controlling the content of the literature to integrate novel variants or updating the information about the association with drug response phenotypes (PMID 17392795). On the other hand, DisGeneNET aggregates data from 8 databases including Uniprot, ClinVar, GWAS catalog, and GWASdb (PMID 27924018).

The strategy that we have adopted for filtering and curating the raw content of PhamGKB and DisGeneNET has several implications for the content of AGMP. The fact that we have selected a subset of the bibliography resources from the primary sources using text mining implies that we integrated only the variants that comply with the criteria of selection set up by the two databases. This implies that the associations described by AGMP are from literature resources that use samples of populations from Africa or with African ancestry. It does not indicate, however, that any other association described in different populations does not exist in Africa. Moreover, it also does not include studies about African samples that confirm or infirm the presence or the absence of a variant-phenotype association unless it was already integrated by the curation scheme of PhamGKB and DisGeneNET.

The data profiling and mapping processes of AGMP have highlighted some drawbacks in the current conception of DisGeneNET and PhamGKB, as well asthe practices of describing experimental protocol and presenting data in genetic and genomic studies. DisGeneNET was intended to give a broad overview of variant-phenotype associations. The data, however, are presented as a binary association without taking into consideration the ethnic composition, hence our strategy to use text mining and manual curation for extracting the genetic entity information. PharmaGKB however was more considerate in this regard by adopting specific guidelines to assign a biogeographical group to each association (PMID 18819074). According to this scheme, Africa is represented by "Saharan-African" and "African American/Afro-Caribbean'' groups while the North African region was assigned to the "Near eastern" group. For our purposes,hese annotation guidelines are inappropriate as it implies that the Sub-Saharan African region is homogenous by nature ( PMID 33116287) and inaccurate because it ignores the admixture landscape of the North African population (PMID 29895688, PMID 23733930, PMID 31039721) that is significantly different from the broad near eastern population. In addition to these issues,manual curation showed that the description of biogeographical information in scientific literature might not be adequate in some instances. Some of the variants in AGMP were assigned to Sub-Saharan and North African geographical entities because we have not found enough information to annotate based on the country of origin. While it is difficult to use a global annotation scheme for annotating biogeographical data in the genotype-phenotype association, federated activities among African researchers could be valuable to establish clear guidelines aiming to cover the genetic diversity in Africa.

The under representativity of some African regions in public databases has been extensively expressed on many occasions. It is however more compelling to have very few entries assigned to central Africa being therefore underrepresented compared to the other African regions. It is not clear however if the reason could be the lack of reliable studies originating in central Africa or the lack of data related to the region generated by consortial efforts (Ref white paper).

We expect the AGMP to be a crucial source of information for the ongoing African research activities in pharmacogenetics and precision medicine. As a sub-project of H3Africa consortium, we have the opportunity to advertise the AGMP role among a large community of researchers which will allow us to update, improve, and diffuse its content. Besides its informative added value, throughout the processes of data profiling, mapping curating and web application building, the precision medicine group of H3ABioNet was able to assemble a solid core of experts around the AGMP including, bioinformaticians, curators, computer scientists, clinical biologists, and data scientists. The core group would have a critical role in ensuring the sustainability and update of the portal and the precision medicine-related activities in Africa.

# Conclusions and Perspectives

# Extensive efforts have been dedicated during the last two years by the H3ABioNet precision medicine group to develop the AGMP. The development of this user-friendly tool will help the scientific community to access relevant data related to genomics medicine and would promote and facilitate implementing genetic testing and other genomic medicine applications in different African healthcare institutions.

Extending the capacity of AGMP has been discussed within the group and two strategies have been proposed to improve the content and features of the resource. In the first plan, a curation group is established to run the data mining, cleaning, and integration processes in a similar way to what has been aforementioned. The data curation team would also determine the feasibility and strategies to adopt in integrating new data sources. The second plan vision consists of increasing the engagement of the scientific community involved in precision medicine research in Africa. The plan would proceed by advertising and outreaching the portal to the wider community, establishing an extended network of experts and curators, and changing the portal user interface to allow researchers to submit their data. The employed network of experts would also determine the guidelines for annotating genetic variants related to precision medicine and establish a comprehensible classification for the levels of curation presented by the portal.

# 

# References

* Bresnick, Jennifer. 2018. “What Are Precision Medicine and Personalized Medicine?” HealthITAnalytics. HealthITAnalytics. January 11, 2018. <https://healthitanalytics.com/features/what-are-precision-medicine-and-personalized-medicine>
* Alexander A. Morgan. PRECISION MEDICINE: DATA AND DISCOVERY FOR IMPROVED HEALTH AND THERAPY. Pac Symp Biocomput. 2016 ; 21: 243–248.
* Ref1
* Ref2
* Ref3
* Ref4
* Radouani F, Zass L, Hamdi Y, Rocha JD, Sallam R, Abdelhak S, Ahmed S, Azzouzi M, Benamri I, Benkahla A, Bouhaouala-Zahar B, Chaouch M, Jmel H, Kefi R, Ksouri A, Kumuthini J, Masilela P, Masimirembwa C, Othman H, Panji S, Romdhane L, Samtal C, Sibira R, Ghedira K, Fadlelmola F, Kassim SK, Mulder N. [A review of clinical pharmacogenetics Studies in African populations.](https://pubmed.ncbi.nlm.nih.gov/32125935/) Per Med. 2020 Mar;17(2):155-170. doi: 10.2217/pme-2019-0110. Epub 2020 Mar 3

1. Kumuthini, J. et. al. Minimum Information required for a DMET experiment reporting. 2016. Pharmacogenomics:17(14):1533-45. doi: 10.2217/pgs-2016-0015