An African Genome Variation Database and its applications in human diversity and health

Davis Todt

Student number: TDTDAV001

Supervisor: Dr Nicola Mulder

Division of Computational Biology, Department of Integrative Biomedical Sciences.

Submitted in fulfilment of the requirements for the degree MSc (Med) in Bioinformatics.

Faculty of Health Sciences

University of Cape Town

15 March 2021

Declaration

I, Davis Marshall Todt (student number TDTDAV001), hereby acknowledge that all of the work presented in this dissertation is my own, except where external work has been explicitly cited.

Any ideas, results, or quotations represented in this dissertation which have been derived from external sources have been referenced appropriately, using the Harvard style.

Additionally, I acknowledge that none of the work presented herein has been used in fulfilment of another degree, either in the past, as well as intended to be used as such in the future.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 2021-03-12

Acknowledgements

I would like to extend my gratitude to the following individuals for supporting me through this trying journey: Prof. Nicky Mulder, Ayton Meintjes, Junaid Gamieldien and Stacey Page.

To Prof. Nicola Mulder: thank you for your helpful feedback – from your contributions to the ideation of the AGVD and the desired feature set all the way through to the final moments of review prior to hand-in. Thank you also for the financial support, for which I am immensely grateful. I have thoroughly enjoyed working with you and your colleagues and have learnt a great deal over the last wo years.

I would also like to extend a special thank you to Ayton Meintjes and Mohammed Elsiddieg for your help with various aspects of platform development. Mohammed, your assistance in helping me get to grips with OpenCGA were paramount in my early development efforts. Ayton, you have extended a helping hand to me from the first day I came through to meet everyone at CBIO and have never hesitated to reach out with useful advice on the direction of my project and to help with various troubleshooting efforts – thank you for everything.

To Junaid Gamieldien: words really cannot express my gratitude for all the support and guidance you have provided me over the last two years. I really mean it when I say that I would not have reached this point if it weren't for your continued support – both academically and emotionally. That lockdown-induced depression and two-year long imposter syndrome was something of a nightmare, but you helped me navigate it and somehow, I managed to get through to the other side. You are one of the most generous and knowledgeable individuals I have ever had the pleasure to get to know, and I am forever here if there is anything I can do for you in return.

Finally, to Stacey, my dodo: I'm thinking back to two and a half years ago when we were staying at that eco-resort in Clanwilliam. I distinctly remember feeling arrested with decision anxiety about whether I should commit to a two-year MSc or not. It didn't help that there were absolutely no distractions where we were staying (oh, the irony). However, you pushed me to challenge myself, and here we are today. Thank you for all your encouragement and understanding. Thank you for putting up with my strained psyche and subpar emotional state at the best of times. You mean the world to me and so, I'd like to dedicate this work to you. I hope you find it a nice reprieve from the dreary world of accounting. If I can offer one piece of advice though: consider reading it over a glass of red wine – it might go down a little easier.

Abstract

African genomes exhibit the highest levels of sequence and haplotype diversity of all extant human populations. A combination of historical as well as geographical factors have contributed toward the high level of genetic diversity in Ancestral populations in Africa. Additionally, a series of concomitant migration events out of Africa, with founder populations harbouring only a subset of this genetic variation, have contributed to the relatively lower genetic diversity observed in non-Africans.

Population genetic studies have refined our understanding of human evolutionary history and clinical genomic studies have resulted in improved patient outcomes. However, despite the increased throughput and decreased cost afforded from next-generation sequencing (NGS) and despite the relatively higher genetic variation in Africans, relatively little of the genomic data currently available is representative of diverse African populations. This may result in adverse outcomes in the context of minority populations with little representation in clinical databases.

Given the under-representation of African genetic variation and the importance of highlighting and further characterizing it, the objectives of this project were to design, develop and deploy a proof of concept database and web application for the storage, analysis and visualization of African genetic variant data – the African Genome Variation Database (AGVD). The AGVD was developed according to software industry design standards. The project also explored available genomic tools and databases in order to leverage existing software solutions where suitable. Additionally, relevant data sets were identified for use during testing and validation of the pilot phase of the project. To this end, the open access 1000 Genomes Project phase 3 dataset was selected and the genotypes for several chromosomes were loaded into the AGVD.

The AGVD leverages the scalable, performant, and open source genomics engine OpenCGA for data storage and analysis. A custom front-end web application was developed by applying a novel approach to render and serve static Vue JS assets from the Python Flask microframework. The web application supports rich data search and filtering operations of loaded variants and allows end-users to visualize annotations of genomic loci and allele change, variant type, associated gene and transcript consequences, clinical significance, and allele frequency information for all annotated cohorts in a highly interactive manner. A bespoke REST API also supports future analytical functionality. The AGVD has demonstrated proof of concept in the secure and scalable storage and visualization of African genomic data, providing

a viable solution for H3ABioNet to further extend in future iterations of the project and a valuable resource for researchers to explore African genetic variation.

Table of Contents

List of abbreviations

AF: allele frequency

AAF: alternate allele frequency

AMH: anatomically modern humans

bp: base pair

CNV: copy number variant

dbGaP: The database of Genotypes and Phenotypes

DGV: The Database of Genomic Variants

EVS: The Exome Variant Server

ExAC: Exome Aggregation Consortium

FR: functional requirement

gnomAD: The Genome Aggregation Database

GoNL: Genome of the Netherlands

GUI: graphical user interface

GWAS: genome-wide association study

HGDP: Human Genome Diversity Project

HGMD: Human Gene Mutation Database

IGSR: The International Genome Sample Resource

kya: thousand years ago

MAF: minor allele frequency

MGP: Human Metabolome Gene/Protein Database

NHGRI-EBI: National Human Genome Research Institute, European Bioinformatics Institute

NCBI: National Center for Biotechnology Information

OMIM: Online Mendelian Inheritance in Man

OR: odds ratio

PCA: Principal Component Analysis

SO: The Sequence Ontology

SNP: single nucleotide polymorphism

TCGA: The Cancer Genome Atlas

UI: user interface

UR: user requirement

US: user scenario

UX: user experience

WES: whole exome sequencing

WGS: whole genome sequencing

YRI: Yoruba tribe from Nigeria, Africa

1kGP: 1000 Genomes Project

Glossary of terms

Alternate Allele Frequency: The frequency at which the alternative (i.e. non-reference) allele of interest occurs in a given population.

Anatomically modern humans: Members of *H. sapiens* that share a high degree of anatomical similarity with that of extant humans, and can therefore be broadly distinguished from archaic humans.

Archaic humans: A broad group under which early Homo species are classified and which do not present with modern anatomical features such as a reduced brow ridge and prominent chin.

Chibanian: a geological epoch dating to approximately 770–126 kya; previously known as the Middle Pleistocene before formal ratification in 2020.

Fst: A measure of the total genetic variance within a population that can be explained by population structure

Functional Requirement: A plain text description of a specific software feature that is required to be implemented as part of a software application in order to fulfil one or more User Requirements.

Haplotype: A set of genetic polymorphisms that tend to be inherited together as a result of their close genetic linkage.

MoSCoW method: A framework for task prioritisation typically used for the prioritisation of software requirements or features during the software development process and which leverages the categories “Must have”, “Should have”, “Could have” and “Would have”.

REST API: A Representational State Transfer Application Programming Interface (API) is a methodology and architectural style for exchanging information between a client and server, using HTTP protocol for transfer of (typically JSON formatted) data.

The Old World: Regions of the world that have long been inhabited by humans, relative to those more recently explored in the West; i.e. Africa, Asia and Europe.

User Requirement: A plain text description of a particular, high-level solution to a problem that a user would require to be included as part of a software application.

User Scenario: A contextualised description of a specific intention of an imaginary user which can be used to generate one or more User Requirements.

1. Introduction and Background

1.1 The origin of modern human populations

1.1.1 Temporal evolution of anatomically modern humans

Homo sapiens, much like other hominins, have observed an interesting and complex evolutionary history. With a wealth of paleontological-, archaeological- and more recently, genetics-based tools at our disposal, the story of our origins as a species has become an increasingly well studied area of research (Stringer, 2002; Willoughby, 2007). Although an exact timeline of human evolution remains elusive, most studies conducted using fossil records and genetic evidence suggest that 'anatomically modern' humans appeared within at least the last 200 000 years (Nielsen *et al.*, 2017).

Studies applying radiocarbon dating methods to fossilized human artefacts have been instrumental in resolving a timeline of human origins. One such study, by White et al. in 2003, radioisotopically dated fossilized crania in Herto, Ethiopia to between 160,000–154,000 years ago (160–154 kya). However, this timeline is constantly shifting, owing in large part to the discovery of new fossil evidence as well as improvements in the techniques applied to study them. For example, more contemporary evidence from fossilized pelvic remains in Ethiopia have suggested the presence of anatomically modern humans circa 200 kya (Hammond, Royer and Fleagle, 2017). Even more divergent are the results from thermoluminescence dating of the recently discovered Jebel Irhoud artefacts which have pushed back the estimated timeline for the emergence of modern humans to circa 315 kya (Hublin *et al.*, 2017; Richter *et al.*, 2017).

Certain genetic studies have supplanted the timelines suggested by fossil evidence for the origin of modern humans. For example, the recent sequencing of seven ancient genomes from Southern Africa by Schlebusch *et al.*, in 2017 hints at an earlier divergence of modern humans closer to the mid Chibanian (Cohen *et al.*, 2020). Their results, generated using a Bayesian coalescent model to estimate divergence times from ancient whole genome sequence data support a more ancient timeline of 260-350 kya. Studies using mtDNA analysis have largely agreed with these timelines (Endicott, Ho and Stringer, 2010; Stringer, 2016).

The term 'anatomically modern humans' (AMH) is not associated with a single, widely agreed upon metric by which human taxonomy can be assessed. Indeed, there is great debate within the scientific community as to the minimum feature set required to constitute anatomical modernity in humans (Lieberman, McBratney and Krovitz, 2002). Generally speaking,

however, the term is used to distinguish between later Pleistocene *Homo* species with a close physical resemblance to present day humans, and earlier, archaic *Homo* species from the early Chibanian (previously Middle Pleistocene; dating approximately 770–126 kya) which share a somewhat reduced feature set (Rightmire, 2009; Manzi, 2011; Cohen *et al.*, 2020). Classical examples of anatomically modern human features include a somewhat reduced brow ridge and vertically sloped forehead corresponding to an increase in brain size (particularly of the forebrain); decreased tooth size, jawbone and chin, associated with differences in diet and tool usage; as well as a less robust body skeletal structure (Lieberman, McBratney and Krovitz, 2002; Holt, 2015; Scerri *et al.*, 2018). It may be useful to consult Stringer et al., 2012, as well as Relethford, 2008 for a more detailed discussion on anatomical modernity in humans and the ongoing contention surrounding this topic within the scientific domain.

Regardless of the inherent ambiguity in the classification of 'archaic' versus 'modern' humans, it is generally accepted that archaic humans represented a diverse group of early *Homo* species that likely descended from *Homo erectus* (or *Homo ergaster*, according to paleoanthropologists who classify this as a distinct, African lineage of *H. erectus*) and later begat *H. sapiens* (Relethford, 2008; Tuttle, 2020). They also exhibited a wide range of morphological, cultural and geographical heterogeneity. In fact, prior to *H. sapiens*, archaic humans from the Chibanian likely occupied the widest geographical range of any hominin group (Bae, 2013). Belonging to the group of archaic humans are at least *Homo neanderthalensis*, Denisovans and *Homo heidelbergensis,* although several additional species have also been described by anthropologists (Relethford, 2008).

The first *H. heidelbergensis* fossil was discovered in 1907, near Heidelberg, Germany (Schoetensack, 1908). Fossils have since been discovered across a range of excavation sites around the Old World. Collectively, they suggest that *H. heidelbergensis* occupied a wide geographical as well as temporal region, having lived between 800-200 kya (Harvati, 2007; Relethford, 2008; Stringer, 2012b). Taxonomy of *H. heidelbergensis* within the context of other archaic *Homo* species has been a challenge since its discovery (Manzi, 2011). Due to similarities between fossilised crania of archaic humans, some paleoanthropologists have suggested that many fossils from the Chibanian should collectively be referred to as *H. heidelbergensis* (Tattersall, 1986; Rightmire, 1998). In contrast, some palaeontologists believe that *H. heidelbergensis* represents a strictly Eurasian chronospecies, while a distinct lineage, *Homo rhodesiensis*, evolved independently in Africa (Hublin, 2009; Stringer, 2012b). Nevertheless, it is generally accepted that *H. heidelbergensis* is likely to have given way to the

evolution of *H. sapiens* (Buck and Stringer, 2014; Tuttle, 2020). The same is probably true for other recent hominins, such as *Homo neanderthalensis* and the Denisovans which co-existed with *H. sapiens* during their tenure on Earth (circa 400–25 kya) (Stringer, 2002, 2012b; Hublin, 2009). The evolution of *H. sapiens*, *H. neanderthalensis*, and the Denisovans may have occurred gradually through a series of intermediate speciation events, or more rapidly in an example of "punctuated equilibrium" (Eldredge and Gould, 1972; Tuttle, 2020). Due to inherent weaknesses in empirical evidence and gaps in the fossil record, it is still not known for certain whether *H. heidelbergensis* represents a shared ancestor of extant *H. sapiens, H. neanderthalensis* and the Denisovans (Tuttle, 2020).

1.1.2 Geographical origins and migration patterns of anatomically modern humans

There exist a diverse range of hypotheses which attempt to explain the temporal and geographical origin(s) of AMH as well as their relationship with archaic humans. Many of these models vary only slightly in their interpretation, while others offer vastly different suppositions. Arguably, all of them may be oversimplified to some extent (Scerri *et al.*, 2018). The majority of contemporary models agree that archaic humans originated in Africa and later spread to Eurasia and Oceania in one or several early migration events predating the existence of modern humans (Stringer, 2016). These models mostly differ in their interpretation of the subsequent events which played a significant role the evolutionary history of extant *Homo sapiens* (Jin and Su, 2000; Scerri *et al.*, 2018). More specifically, the central questions they attempt to answer are the following: did modern humans arise in a single location and from a single source population, or from multiple geographic locations and antecedent populations? And secondly, to what extent did admixture occur between modern and archaic humans?

The most commonly accepted model of early human migration is known as "Out of Africa" (OOA). The OOA model has received strong support from studies of both archaeological as well as genetic evidence; the latter predominantly by way of mtDNA and Y chromosome studies, historically (Cann, Stoneking and Wilson, 1987; Stringer and Andrews, 1988; Hammer, 1995). The central tenet of the OOA model is that anatomically modern humans first appeared in Africa and later migrated across the Old World in one or multiple migration events, eventually replacing local populations. Founding populations may have arisen from a single location or from a range of locations and ancestral populations spread throughout Africa, as will be discussed in the coming text (Relethford, 2008; Scerri *et al.*, 2018). As a result, founder

populations would have given rise to all extant non-African populations to date – meaning that all humans share a recent common ancestor in Africa (Stringer and Andrews, 1988).

Multiple studies have supported an East African origin for modern humans. This hypothesis was predominantly based on inferences from available fossil data but has also gained support from genetic approaches such as mtDNA and microsatellite analysis (Ramachandran *et al.*, 2005; Gonder *et al.*, 2007; Brown, McDougall and Fleagle, 2012). However, several fossil- and genetics studies have hypothesized alternative origins for modern humans, such as a Southern African origin (Lema *et al.*, 2009; Henn *et al.*, 2011; Mounier and Mirazón Lahr, 2019). Additionally, certain studies have proposed a Central African (Ramachandran *et al.*, 2005) or even West African origin (Cruciani *et al.*, 2011; Scozzari *et al.*, 2012; Mendez *et al.*, 2013). For example, in 2011, Henn *et al.* conducted large scale micro-array studies on a cohort of 27 different African populations. They applied regression to Linkage Disequilibrium (LD) as measured by r^2, as well as Fst, versus geographical distance from various origin points in Africa. Their results indicate a 300-1000 fold increase in the likelihood of a Southwest African, rather than East African origin (Henn *et al.*, 2011).

In another more recent example, Chan *et al.*, conducted a cross-disciplinary study in an attempt to reconstruct modern human origins and migration patterns in Africa. They sequenced mitochondrial DNA (mtDNA) from a cohort of 198 Southern African individuals and combined them with existing mtDNA data in order to study the deep-rooted L0 haplotype. They then clustered groups of southern African populations using ethnolinguistic data and constructed a phylogenetic tree. Their results trace the ancient L0 maternal lineage to around 260-140 kya in a region of Botswana known as the Makgadikgadi–Okavango (Chan *et al.*, 2019). However, this study has drawn criticism for drawing far-reaching conclusions based off only a small subset of empirical evidence (Conniff, 2019).

Perhaps a more likely origin story for modern humans within Africa is one that includes multiple founding populations across a range of geographic regions (Scerri *et al.*, 2018; Mounier and Mirazón Lahr, 2019). Indeed, the diverse range of conclusions drawn from fossil and genetic studies would indicate that a more complex model is needed to fully explain our origins. What's more, the fossil record, and by extension, a large amount of the available genetic data from ancient specimens, is only partly descriptive. This kind of empirical evidence is wholly dependent on their discovery. It is also likely that the genetic material derived from such specimens and/or regions may be more degraded in some samples than others. For instance, hot, tropical conditions in Africa would have likely contributed to the degradation of

artefacts and their genetic material over the last 200 kya (Gallego Llorente, Eriksson and Siska, 2015; Loosdrecht *et al.*, 2018).

1.2 Characteristics of African genomic diversity

Of all modern human populations, Africans are generally considered to exhibit the highest levels of ethnolinguistic, phenotypic as well as genetic diversity (Atkinson, 2011; Rotimi *et al.*, 2017). There are an estimated 2000 distinct ethnolinguistic groups in Africa, which represents around one third of known languages in the world (Eberhard, Simons and Fennig, 2020). Africans also exhibit extensive population substructure – most notably between hunter-gatherer and pastoralist groups – as a result of human evolutionary history within Africa (Campbell and Tishkoff, 2008; Li *et al.*, 2008). Additionally, African populations also make use of a wide range of dietary and subsistence patterns, including a variety of hunter-gatherer and agricultural practices (Lema *et al.*, 2009). African populations also appear to exhibit the highest levels of genetic diversity across several key measures, with diversity reportedly decreasing with increasing geographical distance from Africa (Ramachandran *et al.*, 2005). Several of these metrics, as well as some of the possible historical causes shaping their evolution will be explored below.

1.2.1 Factors contributing toward high genetic diversity in Africans

A variety of language and cultural barriers, in combination with a range of environmental and geographical factors, have contributed toward the high levels of genomic diversity and population substructure observed in Africans to date (Campbell and Tishkoff, 2008). Geographical factors include prolonged exposure to highly divergent habitable zones including arid landscapes, tropical rainforest, swamps and mountainous regions in founder populations (Reed and Tishkoff, 2006). Major climatic events also played a role in the migration patterns as well as subsistence patterns of modern humans. For instance, sudden shifts in climate from periods of warming to cooling and vice versa incurred certain dietary, cultural and geographical restrictions (Kuper and Kröpalin, 2006). As a recent example in the literature, Chan et al. hypothesize that the ancient maternal lineage, L0, was in fact originally geographically isolated to a wetland region in Botswana until a period of warming opened up favourable migration routes across Africa for founding populations (Chan *et al.*, 2019). Drastic differences in exposure to pathogens between sub-populations also contributed toward a high level of genetic

differentiation between African populations as well as the selection of both protective and deleterious alleles in extant African populations. Several such examples will be discussed in later sections.

Many ancestral African populations also maintained relatively large effective population sizes during the course of their evolutionary history in Africa, helping to retain accumulated variation by disrupting the effects of genetic drift. Finally, all modern humans evolved from ancestral populations in Africa as a result of the OOA migrations. Such migration events resulted in founder effects, whereby founding populations outside of Africa carried only a subset of the available genetic diversity, thereby leading to concomitant loss of diversity in the subsequent generations. Collectively, these circumstances played a major role in shaping the high levels of genetic diversity observed in African populations at present and contributed toward the relatively lower levels of genetic diversity observed in non-African populations.

1.2.2 African population substructure

African populations have consistently demonstrated a high degree of substructure across various methods of assessment (Campbell and Tishkoff, 2008). Historically, human population structure has commonly been assessed by way of mtDNA and Y chromosome haplogroup analysis as well as low resolution microsatellite analysis across few informative markers (Bowcock *et al.*, 1994; Cavalli-Sforza and Feldman, 2003; Zhivotovsky, Rosenberg and Feldman, 2003; Cruciani *et al.*, 2011). Such studies have, for instance, illuminated population structure between major continental groups (Rosenberg *et al.*, 2002). However, more contemporary genomics methods have since been applied to study population structure at a finer scale, such as higher resolution genotyping of autosomes across 1000's of markers simultaneously (Belmont *et al.*, 2005; Pemberton, DeGiorgio and Rosenberg, 2013; Sudmant *et al.*, 2015). Such approaches increasingly paint a picture of deep divergence of ancestral African populations which were able to maintain large effective population sizes and extensive substructure between populations, with comparatively less structure observed within non-African populations (Bergström *et al.*, 2020).

Mitochondrial DNA (mtDNA) phylogenetic trees describe relationships between known mtDNA haplogroups and have been used extensively to trace human origins and substructure (Gonder *et al.*, 2007). Typically, these are divided into haplogroups L0-L6, with L3 itself giving rise to haplogroups M, N and R. Of these, the L lineages are amongst the deepest rooted

and belong only to humans of African descent, while non-Africans represent descendants of M, N and R. The L0 mtDNA haplogroup commonly denotes the most divergent maternal lineage which arose from the maternal most recent common ancestor (MRCA) of *Homo sapiens* (Gonder *et al.*, 2007; van Oven and Kayser, 2009). The Khoe-san have been demonstrated to harbour this haplogroup at a frequency of around 73% – the highest of all remaining populations – lending support to the notion of a Southern African origin for modern humans (Rosa *et al.*, 2004).

Population genetics studies which leverage high-throughput sequencing of mostly autosomal markers have helped to examine African population substructure in finer detail. Studies have, for instance, genotyped large arrays of SNPs as well as microsatellite loci in order to perform unsupervised clustering of populations by ancestry (Rosenberg *et al.*, 2002; Pemberton, DeGiorgio and Rosenberg, 2013). An example of this approach is to apply the commonly used population analysis tool STRUCTURE, which applies a Bayesian approach to genetic variants within diverse samples in order to create and assign K different populations (Porras-Hurtado *et al.*, 2013). Such studies have demonstrated that extensive population substructure exists between African populations; for example a prominent split exists between hunter-gatherer and the remaining African populations (Jakobsson *et al.*, 2008). Principal Component Analysis (PCA) has also been frequently used as a means to cluster populations based on variant or haplotype information and patterns of genetic linkage, with similar results (Lema *et al.*, 2009; Pemberton, DeGiorgio and Rosenberg, 2013).

Another commonly used metric for assessing population structure is Wright's fixation index, F_{st}, as well as additional estimators of F_{st} (Wright, 1949; Weir and Cockerham, 1984; Nei, 1986). The fixation index is a measure of the total genetic variation within a population that can be ascribed to population structure, based on allele frequency distributions, and has been used to assess population genetic diversity (Holsinger and Weir, 2009).

One of the earliest studies comparing F_{st} between global populations as calculated from a large panel of autosomal SNPs was the HapMap project. Results from this study reported the highest levels of population differentiation between the Yoruba of Nigeria and East Asian populations included in the study (YRI and CHB/JPT; $F_{st} = 0.12$), whereas differentiation was lowest between European and East Asian populations (CEU and CHB/JPT; $F_{st} = 0.0.07$) (Belmont *et al.*, 2005). Similar results were obtained for these populations when compared in a separate study (D. L. Altshuler *et al.*, 2010). Additionally, studies of Fst have suggested that African

hunter-gatherer populations display the highest levels of population diversity in the world, in alignment with PCA, STRUCTURE and mtDNA analyses (Henn *et al.*, 2011).

1.2.3 Patterns of Linkage Disequilibrium

One of the most well characterised features of African populations in the context of genomic diversity is that African populations tend to exhibit lower levels of linkage disequilibrium (LD) among loci compared non-Africans (Lonjou *et al.*, 2003a). Linkage disequilibrium is a term used to describe two or more alleles from different loci that tend to segregate together more often than expected by chance (Lewontin and Kojima, 1960). Contrary to the name, LD does not necessarily imply genetic linkage between alleles nor lack of equilibrium of allele frequencies within a population (Slatkin, 2008). Characterising LD can be useful for studying evolutionary processes and genetic diversity, as well as mapping clinically important loci in disease correlation studies (Jorde, 2000).

Linkage disequilibrium is commonly quantified in terms of the measure D. Various standardisation measures are also commonly applied to D and certain of these may be more robust under different conditions; for a review of these measures consult Devlin and Risch, 1995 (Devlin and Risch, 1995) . The D measure can be calculated according to the following formula, in the simple case of two loci being compared:

$$D_{AB} = P_{AB} - P_A P_B$$

Here, the extent of LD (D_{AB}) between two alleles, A and B, is calculated as the difference between the frequency of occurrence of the haplotype AB (P_{AB}) and the product of the individual frequencies of allele A (P_A) and allele B (P_B) (Lewontin and Kojima, 1960). The D measure takes on a value between –0.25 and 0.25 and little known association between alleles results in a D value tending toward zero, while no association (i.e. linkage equilibrium) exists when D=0 (Ramakrishnan, 2013).

LD is affected by population parameters such as migration events, the effective population size (N_e) of a population, which describes the size of an ideal population able to contribute offspring toward subsequent generations, as well as the substructure within a population. Such factors can also contribute to genetic bottlenecking and admixture events, which may generally increase LD within a given population. Additionally, LD is dependent on several locus-specific

effects such as selection, mutation, genetic recombination, genetic linkage and genetic drift (Lewontin and Kojima, 1960; Pritchard and Przeworski, 2001; Ramakrishnan, 2013).

When studying LD in the context of African genomes, it is generally evident that African populations demonstrate greater genomic diversity both in terms of allelic frequencies and structure of haplotype blocks, consistent with the OOA model of modern human evolution (Campbell and Tishkoff, 2008; Auton *et al.*, 2015). This is particularly evident when comparing sub-Saharan African populations with Europeans (Lonjou *et al.*, 2003b). African populations also tend to harbour shorter haplotype blocks on average than non-Africans (Edwards *et al.*, 2013).

These characteristic LD patterns observed in African genomes arose as a result of the evolutionary history of Ancestral African populations. African populations were able to sustain relatively larger effective population sizes over a longer timeframe than non-Africans, thereby reducing the effects of genetic bottlenecks as well as genetic drift. This, in combination with the greater timeframe over which recombination has had a chance to disrupt the accumulation of associated alleles, has had a major impact on the patterns of LD observed in Africans to date (Campbell and Tishkoff, 2008). Inversely, non-African populations tend to display increased LD between loci, as well as more homogenous LD patterns as a result of the loss of genetic diversity due to founder effects as a result of the Out of Africa migrations (McEvoy *et al.*, 2011).

1.2.4 Patterns of short nucleotide diversity

Multiple large scale population genetics studies have demonstrated that African populations exhibit a higher level of sequence and haplotype diversity compared to non-Africans (Altshuler, 2010; Auton *et al.*, 2015). Additionally, it has been shown that non-African populations possess largely only a subset of the sequence variation found in African populations and that genetic diversity generally decreases as a function of geographic distance from Africa (Lema *et al.*, 2009; Bergström *et al.*, 2020).

As part of the HapMap, 60 individuals were genotyped and re-sequenced from each of 7 different populations in an effort to assess the informativeness of one population toward SNP discovery in the remaining populations. This was assessed in a pairwise manner between groups of individuals from each pair of populations, and reported as the fraction of SNPs found in both populations across 1000 resampling events. For both of the African populations

considered (YRI, Yoruba from Nigeria; LWK, Luhya in Webuye, Kenya), polymorphic SNPs were far more informative of SNPs found in the non-African populations than for the converse (Altshuler, 2010).

In 2007, Guthery *et al.* performed targeted resequencing on four different American populations, including one of African descent. A total of 3,873 genes were sequenced across 152 chromosomes, for a total of 76 individuals. Their results indicated that African Americans harboured the greatest proportion (83%) of all non-singleton SNPs observed across all populations, with the greatest difference detected between African Americans and Asian Americans (83% versus 50%, respectively). A similar trend was true for rare SNPs (AAF <= 5%), with African Americans accounting for 64%. Additionally, they found that African American samples contributed the highest number of common SNPs (i.e. those with a MAF of >= 5% or MAF >= 10%) of the four populations, but consistently shared the lowest percentage of common SNPs with remaining populations, during pairwise comparisons between the four populations. Finally, they observed that around 44% of all SNPs observed in the African American population were private to that population, and 40% of all private SNPs were common (i.e. MAF >= 5%) (Guthery *et al.*, 2007).

The significance of these results can be emphasised in the context of the common disease/common variant (CD/CV) hypothesis, which states that the genetic contribution toward most common complex diseases is underpinned by variants occurring relatively often in the population (Hemminki, Försti and Bermejo, 2008). They help to demonstrate the importance of implementing highly stratified population sampling strategies in population genomics studies, especially in the case of variant-disease association studies. This is particularly true for African populations, which have historically been under-represented by such studies (Popejoy and Fullerton, 2016; Sirugo, Williams and Tishkoff, 2019).

Another, more recent study was published by Bergström *et al.* in 2020. In this large-scale experiment, deep sequencing (mean coverage of 35X) was conducted on a total of 929 individuals across 54 populations, using the Human Genome Diversity Project (HGDP)–Centre d'Etude du Polymorphisme Humain (CEPH) panel. The results of their study agree with previous findings that there exists a great deal of substructure within African populations. Additionally, they also discovered a large number of SNPs that were private to as well as common in African populations, whereas this number was much lower for European, central & south Asia, Middle East, and East Asia populations. For example, they observed several thousand private alleles at a frequency of 50% or greater in African, Oceanian and American

samples, whereas the highest frequency of alleles private to any of the remaining populations was less than 30%. Similar trends were also observed for short insertions and deletions (INDELs) as well as copy number variants (CNVs), although in the latter case the number of high frequency CNVs was significantly higher for Oceanian populations than all remaining populations (Bergström *et al.*, 2020).

1.2.5 Patterns of structural variation

Structural variants, such as large insertions or deletions (i.e. typically greater than 1kb in size), copy number variants (CNVs), translocations and inversions make up the majority of sequence specific differences between genomes (Weischenfeldt *et al.*, 2013). Their role in health and disease has been continually established following large scale genome sequencing projects such as the 1000 Genomes Project (Auton *et al.*, 2015). Briefly, examples of structural variant/phenotype associations include: neurological disorders such as Parkinson's and Alzheimer's disease, cardiovascular diseases such as Atherosclerosis, as well as impacts on drug metabolism and cancer (Zhang *et al.*, 2009; Almal and Padh, 2012; Torres, Barbosa and Maciel, 2015). Their clinical significance is therefore well established, albeit not fully characterised. On a population genetics scale, African genomes have been shown to exhibit both greater amounts, as well as more variability in structural variants than non-African genomes (Sudmant *et al.*, 2015; Levy-Sakin *et al.*, 2019). Several of these characteristics will be discussed below, alongside key studies that have played a role in their understanding to date.

The Structural Variant Analysis Group of the 1000 Genomes Project was set up to better characterise classes of structural variants across 26 different populations. They employed a combination of techniques including short-read sequence data from a total of 2,504 individuals from the 1000 genomes project dataset as well as long-read sequencing in order to generate high confidence structural variant calls. Their results, published in 2015, indicate that individuals of African ancestry exhibit a high diversity of structural sequence variation. For instance, they observed an average of 27% more heterozygous deletions greater than or equal to 50 bp in African populations versus non-African populations. Additionally, the inverse pattern was apparent for homozygous deletions, indicating a higher level of genetic diversity in African populations. This was true both at the resolution of the African continental population as well as the individual sub-populations examined. They also observed a higher degree of within-population variability in African populations relative to the remaining 4

continental populations, when both deletions and SNPs were considered, assessed by pairwise comparisons of the number of nucleotide differences between and within populations. However, the ratio of deletion bp difference to SNP bp difference was lowest in African populations, possibly a consequence of the relatively high amount of SNP variation in Africans (Sudmant *et al.*, 2015).

Genomes of African descent have been shown to harbour a far greater number of low frequency (i.e. MAF < 5%) copy number variants (CNVs) than non-African genomes, according to the HapMap3 dataset. This trend was consistent across all four African populations (namely, Yoruba from Nigeria, African ancestry in the south-western USA, Luhya from Kenya and Maasai from Kenya) of the eleven different populations studied. African genomes were also significantly more diverse with regard to the amount of alleles that differed in copy number between two individuals (Altshuler, 2010).

1.2.6 Microsatellite diversity

Microsatellites (also known as short tandem repeats) are short, repetitive sequences which are highly abundant in eukaryotic genomes and account for a large proportion of genetic diversity between individuals. Stretches of repeated sequence typically consist of motifs 1-6 bp in length which are repeated a variable number of times. Microsatellites are highly polymorphic in terms of the number of motifs repeated but generally display well conserved sequence between motifs. They usually occur in non-coding regions of the genome, and as such, may be less subject to either positive or negative selection pressures than other classes of sequence variation (Ellegren, 2004; Kinney *et al.*, 2019). Considering these properties, microsatellites are an ideal marker for assessing human genetic variation. An abundance of studies have shown that African individuals generally exhibit greater diversity than non-Africans when various different microsatellite loci are examined (Bowcock *et al.*, 1994; Jorde *et al.*, 1997; Zhivotovsky, Rosenberg and Feldman, 2003) .

As an example, microsatellite diversity was compared between different populations in a study by Lema *et al.* in 2009. In this study, individuals were genotyped from a wide range of populations spanning a host of geographical locations, for a total of 2,551 individuals of African, African American and Yemeni ancestry. These data were then integrated with additional data from the CEPH-HGDP panel. As a proxy for genetic diversity, they calculated the measure theta using two different methods. Firstly, theta was calculated as twice the

variance of the microsatellite allele repeat length unit ($\theta = 2\sigma^2$) for each different population. Secondly, theta was calculated as a function of the expected heterozygosity of microsatellite loci (Kimura and Ohta, 1973). Across both measures of theta it was found that African and African American populations exhibited greater levels of genetic diversity than the remaining (non-African) populations examined (*fig. 1*). It was also observed that African individuals possessed greater numbers of private alleles of the 848 microsatellite loci examined, compared to the remaining continental groups (Lema *et al.*, 2009).

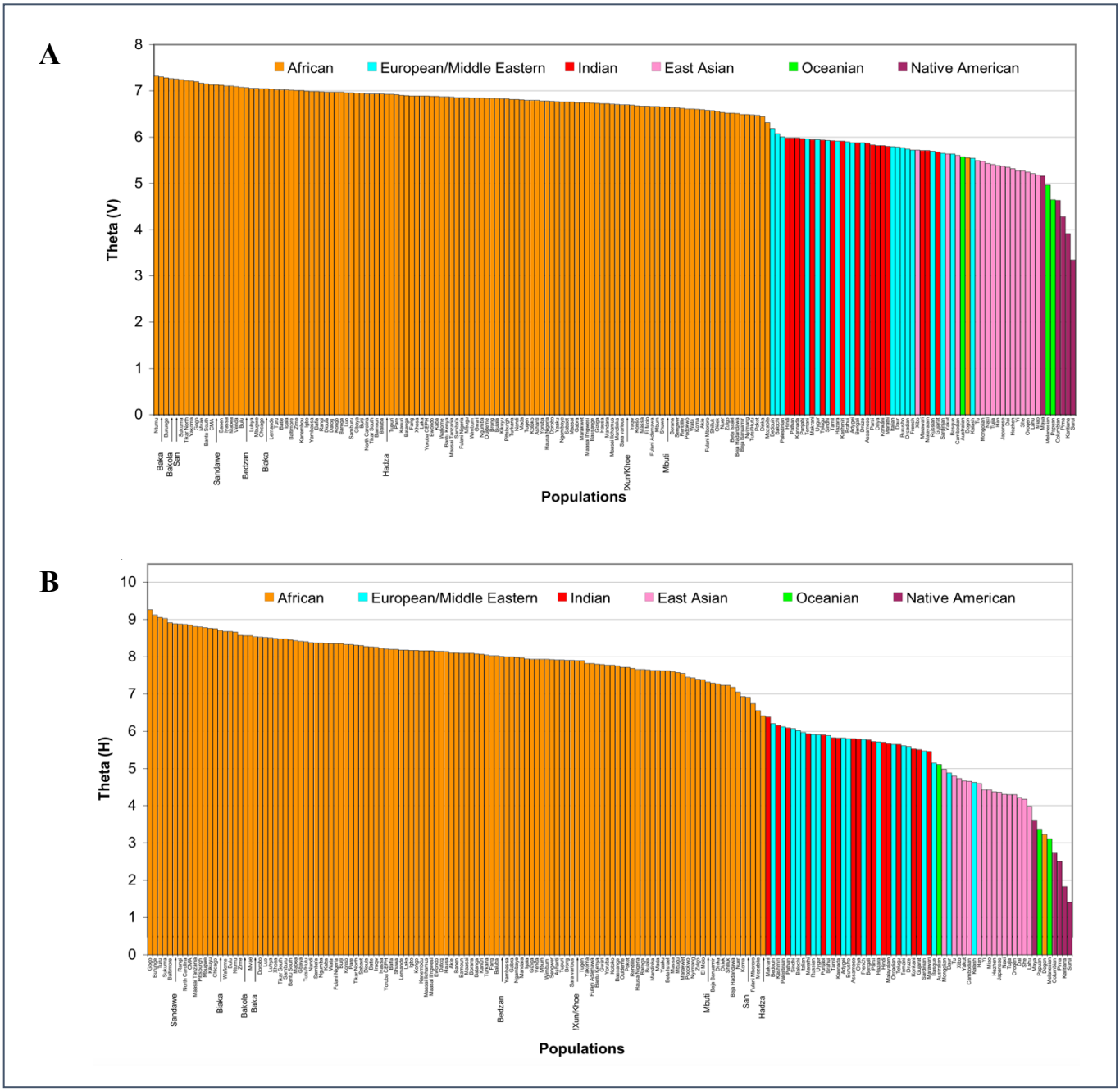

Figure 1: Population genetic diversity estimates inferred from two different calculations of theta using microsatellite loci. ***A****, theta calculated as twice the variance of the microsatellite allele repeat length unit ($\theta = 2\sigma^2$).* ***B****, theta calculated as a function of the expected heterozygosity of microsatellite alleles ($H = 1 - \frac{1}{\sqrt{1+2\theta}}$), as derived from Kimura and Ohta, 1973.*

Source: Lema et al., 2009, supplementary materials

A similar trend was observed by Pemberton *et al.* in 2013. In one of the largest studies of human microsatellite diversity at the time, they combined eight different microsatellite datasets in order to generate a single dataset containing 645 shared microsatellite loci across 5,795 individuals, representing 267 different populations. They used this dataset to calculate the expected heterozygosity for different populations and report this as a function of geographic distance from Addis Ababa, Ethiopia. Consistent with the findings by Lema *et al*., they observed the highest levels of microsatellite loci heterozygosity amongst individuals of African ancestry. Also consistent was the correlation between a decrease in expected heterozygosity and an increased distance from Ethiopia ($R^2 = 0.841$) (Pemberton, DeGiorgio and Rosenberg, 2013).

1.3 African genetic variation within a health research and clinical setting

Studying African genetic variation can be useful in the discovery of risk alleles for disease, ethnically linked genotype-environment interactions, such as the effect of dietary or lifestyle choices on phenotype, as well as drug responses in different populations (Campbell and Tishkoff, 2008; Manrai *et al.*, 2016; Crawford *et al.*, 2017). Malaria and Human African Trypanosomiasis (HAT) are two examples of diseases which have resulted in the strong positive selection of several resistance loci in African ancestries (Gomez, Hirbo and Tishkoff, 2014; Pays *et al.*, 2014). Each of these diseases will be discussed and several such resistance loci will be examined below.

1.3.1 Malaria

Malaria is an infectious disease caused by several species of the *Plasmodium* genus and spread through the bite of infected female Anopheles mosquitoes. Symptoms of malaria include fever, chills, headache, fatigue, respiratory complications and – if left untreated – may result in organ failure or death (Trampuz *et al.*, 2003; WHO, 2020a). According to the World Health Organisation, there were 228 million positive cases of malaria in 2018 of which an estimated 405,000 were fatal. Malaria has a particularly large disease burden in Africa. Indeed, around 93% of malaria cases and 94% of deaths in 2018 occurred in only a subset of African countries. Additionally, more than half of the global malaria burden was carried by only six African countries during this time (WHO, 2020a). Consequently, malaria continues to be an important area of research, particularly within the African context.

Malaria represents an interesting case study of an infectious agent as a major driving force for recent human evolution. In response to strong selective pressures incurred by malaria, African populations have accumulated a range of beneficial mutations that confer varying levels of resistance to the disease (Gomez, Hirbo and Tishkoff, 2014). Genomic changes in affected populations have been swift and robust, several of which demonstrate near fixation in affected populations. For instance, research suggests that a variant of the *Duffy Antigen/Chemokine Receptor (DARC),* which confers resistance against malaria from *P. vivax,* is present in 99% of contemporary sub-Saharan Africans and may have arisen over a span of only 8,000 years (McManus *et al.*, 2017). In fact, there are a wealth of examples of selection in humans for alleles conferring resistance to malaria in the last 5,000-10,000 years (Hedrick, 2012). This timeline coincides with the shift toward more agricultural based subsistence patterns in Africans and subsequent population expansion, which is thought to have been a causative factor for the rapid spread of the parasite (Coluzzi, 1999; Carter and Mendis, 2002). Resistance to *P. falciparum* may have even evolved more recently, according to at least one study which dated the wide-spread emergence of this species to within the last 3,000-4,000 years (Otto *et al.*, 2018).

Perhaps the most well characterised of the malaria resistance mutations within African populations is the sickle-cell mutation, HbS. The HbS mutation refers to a "loss-of-function" (LOF) variant of the beta-globin gene, HBB, which occurs at locus 11p15.5 on chromosome 11 and is important for healthy oxygenation of the blood. The mutation results in the substitution of glutamic acid with valine in the beta haemoglobin subunit which interferes with its tertiary structure and function (Ashley-Koch, Yang and Olney, 2000; Piel *et al.*, 2010). Consequently, this affects the shape and oxygen-binding affinity of the adult haemoglobin (HbA) tetramer and the red blood cells that they comprise. Affected red blood cells present with a characteristic, "sickle" shape (Maakaron and Taher, 2020). The HbS allele results in sickle-cell disease (SCD) in homozygous individuals, but confers partial carrier-resistance to *P. falciparum* malaria in heterozygotes. Individuals that are heterozygous for HbAS are carriers for the sickle-cell trait (SCT) and are generally asymptomatic, thus offsetting the effects of negative selection (Hedrick, 2012; Ashorobi and Bhatt, 2019). Sickle cell trait may confer up to 90% protection against severe forms of malaria (Williams *et al.*, 2005). HbS variants may be associated with one of at least five known haplotypes, four of which are African in origin. These region-linked haplotypes are believed to have evolved independently in an example of convergent evolution, although this supposition has also been challenged by recent research

(Ngo Bitoungui *et al.*, 2015; Pule *et al.*, 2017). A strong association between the sickle-cell trait and resistance to malaria in malaria-endemic regions was first demonstrated in 1954 and is widely accepted as the first formal evidence for recent selection against an infectious agent in humans (Allison, 1954). Conclusions drawn from more contemporary studies performed on African populations support these initial observations (Flint *et al.*, 1998; Piel *et al.*, 2010; Damena *et al.*, 2019).

Sub-Saharan Africa has the highest occurrence of SCT carrier individuals globally, with an incidence rate of approximately 20-30% in most affected countries in this region, and up to 45% in heavily affected parts of Uganda (Regional Committee for Africa, 2011). Additionally, one out of every three West Africans carries a variant HbS allele (Assembly, 2006). This value is much lower for populations outside of malaria-endemic countries, such as white Americans, for whom the prevalence is an estimated 0.05%. A lower incidence rate is also apparent in certain African-American populations living in the United States (a non-malaria-endemic region), with an estimated prevalence of 8% (Tsaras *et al.*, 2009).

One of the largest studies conducted on HbS allele frequency data across global populations was conducted in 2010. This study generated a high-quality aggregate dataset to explore the relationship between the geographical distribution of pre-intervention malaria (circa 1900) and the HbS allele within indigenous global populations. The results of this study estimate an HbS allele frequency of greater than 0.5% across the majority of the African continent, with higher frequencies prevalent in most of SSA and regions of the Middle East and India. The highest estimated allele frequency reported was in a region of northern Angola (18.18%), while HbS was virtually non-existent in Southern Africa and across a large expanse of East Africa. Central African populations including regions of Gabon, The Republic of the Congo, Central African Republic and the Democratic Republic of the Congo (the D.R.C.) shared high allele frequencies, in the range of approximately 9-14% (*fig. 2*). Furthermore, this study confirmed the hypothesis that the SCT arose within SSA as a consequence of positive selection against *P. falciparum* malaria (Piel *et al.*, 2010).

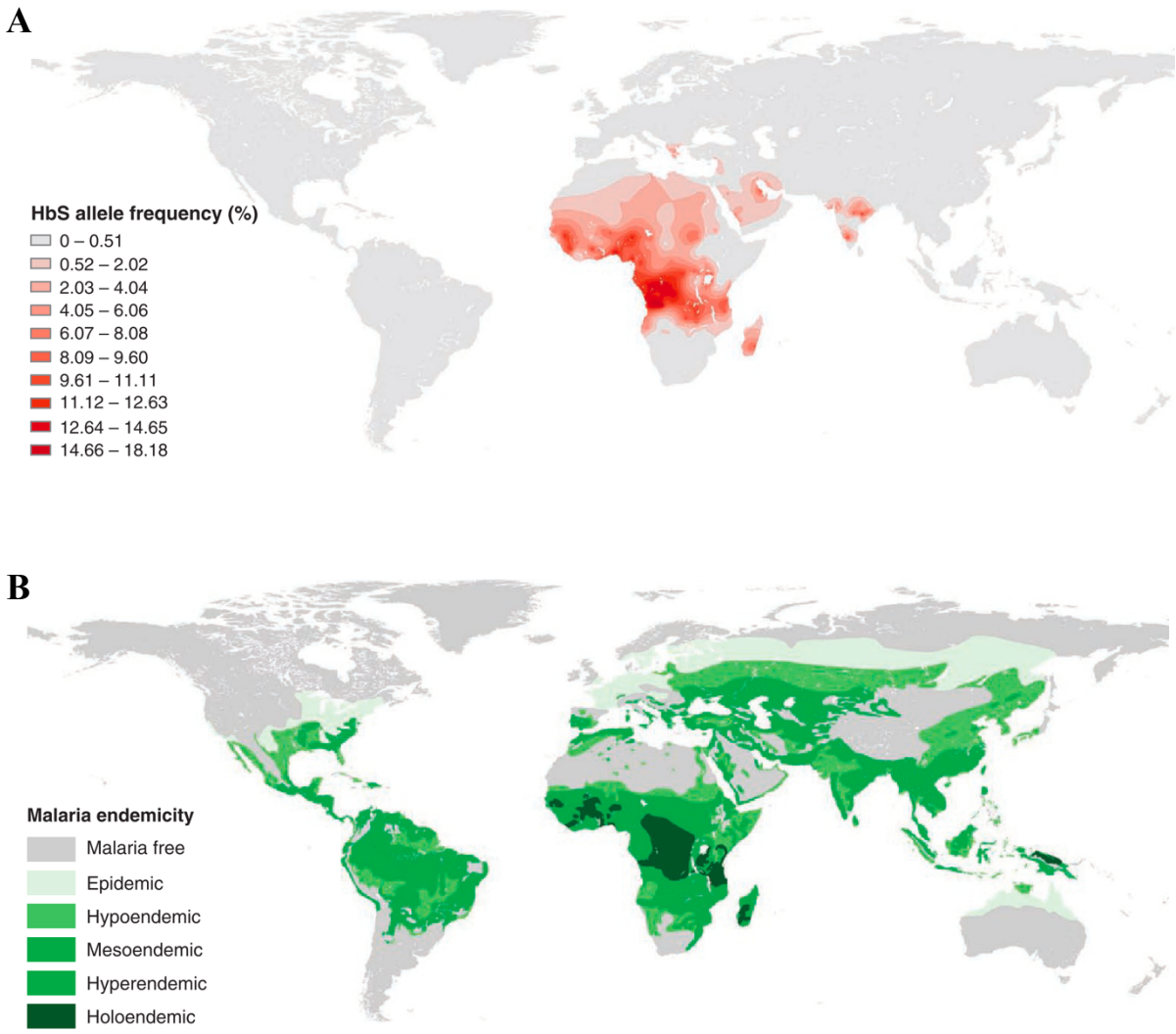

Figure 2: Geographical distribution of the HbS allele and malaria endemicity. ***A,*** *the estimated geographical distribution of HbS allele frequency across indigenous global populations.* ***B,*** *the global distribution of pre-intervention malaria endemicity,*

Source: Piel et al.*, 2010*

A large proportion of the malaria resistance mutations which have been well characterised to date are relatively simple, LOF variants, such as those described above. Many of these manifest as SNPs and short indels, or larger deletions in some cases (Hedrick, 2011). However, there also exist interesting examples of more complex structural variants that have been associated with malaria resistance in Africans. A recent example was identified by Leffler et al. in a study which set out to further resolve novel variation loci near the glycophorin genes (*GYPA*, *GYPB* and *GYPE*) which were previously associated with malaria resistance in Africans (Band *et al.*, 2015). Their study constructed a large reference panel using the 1000 Genomes Project Phase

3 data in combination with additional deeply sequenced African individuals across various different ethnolinguistic groups. They used this panel to impute a larger GWAS dataset representing approximately 10 000 malaria case/control samples and were able to infer significant association from structural variants of glycophorin genes. Specifically, they identified a novel structural variant, termed *DUP4*, with a strong association signal for malaria resistance. This is a complex CNV which consists of a complete duplication of *GYPE* and partial duplications and deletions of *GYPA* and *GYPB* to form *GYPA/GYPB* hybrids. They estimated that DUP4 conferred 40% resistance against severe malaria (odds ratio of 0.60; 95% confidence interval 0.50-0.72; $P = 9.9 \times 10^{-8}$). This effect was limited to the East African populations studied (Malawi and Kenya), as DUP4 was altogether absent from remaining populations (Leffler *et al.*, 2017).

1.3.2 African Trypanosomiasis

African Trypanosomiasis, commonly referred to as African sleeping sickness, is a tropical disease most prevalent in rural parts of sub-Saharan Africa. It is a vector-borne illness carried by the tsetse fly and is caused by infection from one of two parasites: *Trypanosoma brucei gambiense* (accounting for 98% of cases) and *Trypanosoma brucei rhodesiense*, which are more prevalent in west and central Africa, and east and southern Africa, respectively (Simarro *et al.*, 2014; WHO, 2020). Symptoms of the disease include fever, joint pain, and swollen lymph nodes in the early stages of the disease, followed by severe neurological symptoms in later stages. The prognosis of infected individuals is generally considered fatal if left untreated (CDC, 2020). Contemporary medicine as well as control programmes implemented by governments and NGOs have made significant progress in reducing the burden of disease over the last decade (from less than 10 000 new cases to less than 1000 at time of writing) (Franco *et al.*, 2020). Despite this, Trypanosomiasis remains a very serious threat to millions of people across 36 African countries. Importantly, those at the highest risk of infection are least likely to be able to find and afford appropriate treatments (Simarro *et al.*, 2014; Franco *et al.*, 2020; WHO, 2020).

Nucleotide sequences conferring resistance against African Trypanosomes have been positively selected in human genomes during the course of their co-evolution in Africa (Pays *et al.*, 2014). Perhaps the most well characterised resistance-associated gene is that of apolipoprotein L1 or APOL1, located on the long arm of chromosome 22, haplotypes of which

confer resistance against parasitaemia by certain subspecies of *Trypanosoma brucei* (Thomson *et al.*, 2014). While the wild type APOL1 does not offer a protective advantage against *T.b. rhodesiense* or *T.b. gambiense*, two haplotypes of APOL1 – referred to as G1 and G2 – are associated with resistance to *T.b. rhodesiense* (Genovese *et al.*, 2010; Limou *et al.*, 2015). Importantly, these variants have also been strongly associated with chronic kidney disease (CKD) in populations of African ancestry, such as African Americans (Genovese *et al.*, 2010; Parsa *et al.*, 2013; Sumaili *et al.*, 2018). In light of their protective effect, G1 and G2 have undergone strong positive selection in Trypanosomiasis-endemic areas within the last 10 000 years and, as a result, occur at high frequencies in populations with recent African descent, but are virtually absent elsewhere (Genovese *et al.*, 2010; Simarro *et al.*, 2014).

The G1 haplotype includes two nonsynonymous SNPs, rs73885319 (resulting in a serine to glycine amino acid substitution at position 342) and rs60910145 (resulting in an isoleucine to methionine amino acid substitution at position 384) (Genovese *et al.*, 2010). These two mutations are in perfect LD, likely as a result of their close genetic linkage (Genovese *et al.*, 2010; Ko *et al.*, 2013). The G2 haplotype is less common than G1 and comprises a small, 6 bp deletion of the sequence TTATAA (rs71785313), which results in the deletion of 2 amino acids (asparagine and tyrosine at position 388 and 389, respectively). Functionally, these variants lead to disruption of the interacting domain of APOL1 for the Trypanosoma virulence factor serum resistance-associated protein (SRA), thereby restoring APOL1-mediated lysis of *T.b. rhodesiense* (Genovese *et al.*, 2010; Thomson *et al.*, 2014).

A recent candidate gene association study conducted by Kamoto *et al.* demonstrated a significant protective effect of G2 against *T.b. rhodesiense* in northern Malawians ($p = 0.0000105$, OR = 0.14, CI95 = [0.05–0.41], BONF = 0.00068). In the control samples studied, this variant occurred at a frequency of almost 20%, versus approximately 3% in cases (Kamoto *et al.*, 2019). Their study also highlighted significant heterogeneity in disease susceptibility and prognosis between different African populations, reiterating the need for genomics studies that are more inclusive of different African populations. This may be further illustrated by the results from a separate association study by Cooper *et al*. Their results indicate a similar, five-fold protective effect of G2 heterozygotes against *T.b. rhodesiense* susceptibility within Ugandan individuals and no protective effect for the G1 haplotype. No significant protective effect against *T.b. gambiense* was observed for either variant, as expected (Genovese *et al.*, 2010; Cooper *et al.*, 2017). Interestingly, however, they observed an increased risk of clinical (rather than latent) stage Trypanosomiasis from *T.b. gambiense* for Guinean individuals with

the G2 haplotype. Furthermore, they observed an association between G1 and latent phase *T.b. gambiense* Trypanosomiasis in Guinean individuals (Cooper *et al.*, 2017).

The demographic distribution of the G1 and G2 haplotypes is highly heterogenous, but is mostly restricted to populations of sub-Saharan African ancestry. Generally, G1 is far more common than G2 in West African populations, and occurs at much lower frequencies in East Africa (Genovese *et al.*, 2010; Thomson *et al.*, 2014; Cooper *et al.*, 2017). For instance, G1 is present in almost 50% of the Ibo and Esan and approximately 38-40% of Yoruba individuals from Nigeria (from the HapMap3 dataset), whereas the highest frequency observed in East Africa is approximately 12% for the Chewa of Malawi and the Sena of Mozambique (Genovese *et al.*, 2010; Altshuler *et al.*, 2012; Thomson *et al.*, 2014). African Americans, who are of recent West African ancestry, have also been shown to exhibit G1 allele frequencies of between 20-30% (Limou *et al.*, 2014, 2015). The G2 haplotype is less common overall than G1, and frequencies are more uniform across most of sub-Saharan Africa when assessed by pairwise Fst calculation (Thomson *et al.*, 2014; Limou *et al.*, 2015). In comparison with G1, G2 occurs in only 7-8% of Yoruba individuals from the HapMap3 dataset (Genovese *et al.*, 2010; Thomson *et al.*, 2014). However, higher frequencies (17%) have been observed for Yoruba by studies using the HGDP dataset. Additionally, neighbouring West African populations such as the Mandenka from Senegal have observed minor allele frequencies of up to 20% (Thomson *et al.*, 2014). To date, the highest frequencies of the G2 allele outside of West Africa have been observed in the Shangaan of Mozambique (AF = 21.7%) (Pinto *et al.*, 2016).

1.4 Project Rationale

1.4.1 Population genetics in the post-genomic era

With the advent of massively parallel, high-throughput sequencing technologies (collectively referred to as next-generation sequencing, or NGS) in the early twenty-first century, there has been a marked increase in the volume of genomic data generated (Cook *et al.*, 2016). This effort has been in a large way catalysed by the completion of the Human Genome Project in 2003 - an international collaboration that realized the first ever map of the human genome, including structural and functional annotation that has been instrumental in human health and genomics studies to date (Cook *et al.*, 2016). Rapid improvements in sequencing technologies have facilitated a sharp decrease in the cost per sequenced Megabase (MB) of DNA, at a rate which has long since surpassed prediction by Moore's Law.

Additionally, improvements in bioinformatic analysis methods have sought to keep up with the pace and volume of data generated. Particularly within the course of the last decade, both the efficiency and accuracy of alignment algorithms, as well as downstream applications such as variant calling have been greatly optimized (Kulski, 2016; Wetterstrand, 2019). It is no wonder that such an abundance of genomic data exists to date.

1.4.2 Lack of diversity in human genetics studies

Despite the increase in the volume of genetic data generated over the last twenty years, there exists a large over-representation of individuals of European descent in genomic databases. This systemic bias toward Europeans has accumulated as a result of African and other minority populations being largely neglected from genomics studies until fairly recently (Popejoy and Fullerton, 2016). This has occurred despite the high level of genetic diversity observed in African genomes and the demonstrable impact of this heterogeneity on human genetics and health (Sirugo, Williams and Tishkoff, 2019). As is discussed in the current text, these historically neglected populations in genomic studies are an essential component of the "genomics puzzle" researchers are hard at work to decipher, and more attention in the form of funding, resources, and research focus should be directed toward the design and implementation of diverse, ethnically inclusive sampling strategies in genomic studies. The following section will highlight both the current state, causes and consequences of this under-representation within the genomic domain, as well as discuss recent progress being made to address this deficit and the future trajectory of African genomic studies.

To illustrate the depth of this issue, consider that around 78% of the array data stored in the NHGRI-EBI GWAS Catalog in 2018 belonged to European individuals, with the remaining portion representing almost exclusively individuals of Asian ancestry. Of the total available data, only 2.4% derived from African individuals. This is despite the fact that 7% of the total genotype-phenotype associations were contributed from these African samples (Morales *et al.*, 2018; Lee, 2020). Africans are also under-represented by the landmark 1000 Genomes Project (1kGP), in which only seven African populations are represented out of an estimated 2000 different ethnolinguistic groups in Africa (Auton *et al.*, 2015; Eberhard, Simons and Fennig, 2020). Similar trends can be observed in other prominent databases, such as the Genome Aggregation Database (gnomAD) – a successor to the ExAC database – and the database of Genotypes and Phenotypes (dbGaP) (Landry *et al.*, 2018; Karczewski *et al.*, 2020). In

particular, gnomAD represents a useful proxy for assessing the contemporary state of population representation in genomics studies as it consists of aggregated data from over 15,700 genomes across more than 100 independent studies. A total of 12,487 African and African American individuals are represented in the gnomAD v2 release; this equates to ~8.8% of the total number of genome and exome sequences represented, or approximately one-fifth of the amount of non-Finnish Europeans represented in the database (Karczewski *et al.*, 2020).

This significant under-representation of Africans in genomic studies may be attributed to both historical and contemporary causes. Historically, the early pioneering era of NGS brought with it prohibitive sequencing costs which could generally only be pursued by wealthier nations on local study participants; this inadvertently instilled a depth-vs-breadth research focus, which in many ways is still prevalent today (Wu, 2019). Unfortunately, the unique challenges of conducting research in the developing world have hindered progress in this regard. Mulder *et al.* recently published a detailed review article which describes the wide array of challenges facing researchers in an African setting, and can be consulted for more a more in-depth discussion (Mulder *et al.*, 2017). Briefly however, limited funding for large-scale genomics studies, which can be costly even when conducted in developed countries, compounded by a lack of resources can make it difficult to implement such studies in Africa. Examples of this include infrastructure limitations, such as a lack of basic services including roads, public transport, internet and telecoms, can make it difficult to access study participants and research facilities or store and share large volumes of data. In many cases, a lack of adequately trained researchers and efficient organizational structures can also impede research. Unfortunately, these limitations may also impact the willingness of African researchers to share their data, due to the initial challenges in obtaining it. Additionally, there are less obvious concerns such as language and cultural barriers, which may get in the way of obtaining consent from study participants or effectively communicating experimental results. Therefore, careful consideration of the ethical implications also needs to be made when conducting such studies (Mulder *et al.*, 2017).

Lack of diversity of African samples in genome databases has had a dramatic impact on human health and medicine. Sampling bias has knock-on effects as research results are translated to clinical relevance: rare disease associations may evade discovery, clinical interpretation can have little relevance when generalised across diverse groups and replication of results on minority groups can be challenging (Landry *et al.*, 2018). For instance, the statistical power of GWA studies to detect rare variants associated with disease in a given population is contingent

upon imputed variants determined from diverse reference panels and is not limited to the population of study (Pulit, Voight and de Bakker, 2010). Furthermore, individuals of non-European ancestry observe a greater proportion of variants of unknown significance (VUS), which impedes the ability of clinicians to fully elucidate disease risk in these populations (Caswell-Jin *et al.*, 2018). As a consequence of these factors, many of the variants which have historically been considered causative or otherwise associated with phenotypic outcomes in European ancestries have not been replicated in African populations, or in some instances have even been associated with the reverse outcome. Several such cases have already been discussed in detail in the text. However, thus far little attention has been given to the historical implications such instances have had on the diagnosis and treatment of African individuals.

There exist several examples of misdiagnosis of African individuals for certain diseases. A useful example is hypertrophic cardiomyopathy (HCM) – a monogenic disease which affects an estimated one in five-hundred individuals globally (Liew *et al.*, 2017; Tuohy *et al.*, 2020). HCM is often diagnosed in conjunction with genetic screening for variants MYBPC3 G278E, TNNT2 K247R and TNNI3 P82S, amongst others, which can be used to preclude disease phenocopies or determine patterns of familial inheritance of the disease (Liew *et al.*, 2017; Owens and Reza, 2020). Recently, a study conducted using historical patient data revealed that multiple patients of African ancestry were falsely diagnosed with HCM based on the presence of variants classified as pathogenic, but later reclassified as benign. They also identified that the top five HCM associated variants were present at significantly higher MAFs ($p<0.001$) in black Americans than white Americans (for example, TNNT2 K247R had a MAF of 27.1% in black Americans versus 2.9% in white Americans). In addition, they reported that the MAFs observed in black Americans were substantially higher than expected and prevalence in the control population was not consistent with disease penetrance – factors which would have provided strong evidence for lack of pathogenicity if these populations were initially included as part of the reference database (Richards *et al.*, 2015; Manrai *et al.*, 2016). This has the potential to affect not only the lives of the misdiagnosed individuals – who may be prescribed unnecessary lifestyle changes and medical interventions as a result – but can also have adverse effects on the lives of relatives (who, for example, may receive false negative diagnoses based off incorrect assumptions about causative variants).

Type 2 Diabetes (T2D) represents another example of a disease that has been misdiagnosed in minority groups due to incomplete reference databases. T2D is commonly diagnosed by assessing changes in glycated haemoglobin (HbA1c) over time and T2D risk may be influenced

by at least 18 variants with an additive effect (Sherwani *et al.*, 2016; Wheeler *et al.*, 2017)(diff ref). Recently, it was observed that many so called "erythrocytic variants" – variants associated with changes to RBC structure or function – did not impact T2D risk despite reducing HbA1c blood concentration and may result in false negative diagnoses of T2D. Additionally, one such variant common in Africans (MAF = 11%), G6PD G202A, was found to have a substantial effect on Hb1Ac blood concentration but was virtually absent from the remaining ethnicities included in the study. This indicates that Africans may be particularly susceptible to misdiagnosis using standard screening protocols and may benefit from combinatorial genotyping that is inclusive of G6PD deficiency markers. As a result, an estimated 2% of African Americans with T2D would remain undiagnosed if the entire African American population were to be screened for Hb1Ac (Wheeler *et al.*, 2017). Such examples of false negative diagnoses can have potentially dire consequences for affected individuals, who may have otherwise benefitted from lifestyle interventions and treatment to improve disease prognosis and outcomes.

1.4.3 An African-centric human variant database

Fortunately, as a scientific community, we are beginning to move in the direction of genomic study designs that are more inclusive of diverse African populations (Mills and Rahal, 2019). Shortly after the completion of the 1000 Genomes Project in 2013, the African Genome Variation Project (AGVP) sought to further scientific interest in, as well as data generation and results interpretation of, African genomic studies. To this end, the AGVP genotyped 1,481 SSA individuals from 18 different ethno-linguistic groups and a further 320 individuals were sequenced using a WGS approach. Their results were used to further inform upon SSA population differentiation, admixture and historical migration patterns and, additionally, elucidated millions of novel rare variants that were not previously annotated in prior studies (Gurdasani *et al.*, 2015).

In keeping with the trajectory of increasing African genomic data generation, the AGVP has since been surpassed by larger scale studies of African populations in terms of both the number of individuals as well as the number of different ethno-linguistic groups included in such studies. For instance, the Simons Genome Diversity Project (SGDP) generated 300 deeply sequenced genomes (mean depth of 43X) from 142 global populations, of which 22 populations were of African Ancestry (Mallick *et al.*, 2016). More recently, the most inclusive human

genetic diversity sequencing experiment was conducted on the Human Genome Diversity Project Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) panel. In 2020, Bergstrom *et al.*, deeply sequenced 929 genomes from 54 distinct human populations – including seven of African ancestry – and published these data in a freely accessible manner (Bergström *et al.*, 2020). Similarly, a study from the Human Heredity and Health in Africa (H3Africa) Initiative was recently published in which 426 genomes from 50 different ethnolinguistic groups in Africa were sequenced at medium to high depth of coverage (314 individuals sequenced at a mean depth of 30X). The results of the study revealed novel insights into human migration patterns and reiterated the shortcomings of variant pathogenicity classification in populations that have been historically underrepresented in clinical and population databases (Choudhury *et al.*, 2020).

Evidently, over the last decade, there has been a significant shift in the focus of genomic funding and research efforts toward more deeply stratified population sampling strategies which are more representative of the gamut of African genetic diversity. Broad collaborations such as H3ABioNet – an H3Africa initiative – have been established to address historical disparities in African genomic studies and have made significant scientific contributions toward this objective (Mulder *et al.*, 2016, 2018). Since its inception, H3ABioNet has cultivated a collaborative network of African bioinformatics researchers with representative nodes in 16 African countries and 1 in the USA (H3ABioNet, 2020). Furthermore, H3ABioNet has facilitated skills development of African researchers, developed standardised protocols for conducting research effectively and ethically within the African context, as well as contributed infrastructure such as open source tools and workflows for bioinformatic data analysis (Mulder *et al.*, 2016, 2017; Aron *et al.*, 2017).

Given the trend of increasing genomic data generation for African genomes in particular, it is therefore surprising that no single, central database and data exploration platform exists for managing and exploring the array of African genetic variation data generated from such studies (currently, African data that do exist are mostly presented at a high level in international databases, with little granularity). Such a database would likely demonstrate particular value to large consortiums such as the H3Africa consortium, and H3ABioNet in particular, which facilitate genomic studies across the African continent and have also acknowledged the need for an African variant database (Mulder *et al.*, 2016). Ideally, such a database would also be African owned and operated. This would help to ensure that data generated from African

researchers are collected ethically and that data use and accreditation are conducted fairly for researchers and study participants alike.

With this in mind, the aim of this project was to develop a proof of concept solution to address the need for long term storage and management of African genetic data. Specifically, the objectives of this project were to design, build and deploy a prototype database for storing and managing African genetic variant data generated from NGS studies and to make this data accessible to researchers by way of a user-friendly web platform.

Part of these objectives involved identifying one or more example data sets containing variants which are representative of diverse African populations. Such a data set will serve as a validation data set and help demonstrate utility of the application for engagement with potential stakeholders. In addition to storing and managing this data effectively and in a scalable manner, the application also sought to annotate variants and to report on the allele frequencies of each variant within various African populations. Finally, annotation and allele frequency information will be made available to end-users by way of a client facing front-end application. This will necessitate the development of a custom developed web application, leveraging existing bioinformatics tools where possible, and will be designed in close alignment with software industry standards.

Finally, the project then sought to demonstrate utility of this data base and accompanying web application within an African context. This will be achieved by examining variants which confer protection against malaria and HAT and will be applied to examples of exploratory data analysis and data validation. This process will be conducted with the aim of meeting specific user and functional requirements set out during the project's design phase. Additionally, proof of concept will be partly assessed by way of demonstration to internal H3Africa stakeholders in order to assess the viability of the solution and garner additional feedback. The remaining chapters therefore detail the process that was followed to design, develop and implement the African Genome Variation Database (AGVD) pilot release and to demonstrate its potential utility within the context of health research.

2. Investigating Existing African Genomic Resources

The purpose of this chapter is to outline currently available resources with potential utility in the design, development and testing of the AGVD. The scope of this resource curation process includes datasets, bioinformatic software tools, as well as databases within a population genetics context. The intention is not to build a comprehensive list of all available resources at present. Rather, the resource list gathered will serve as a product development aid during subsequent development efforts.

2.1 Existing African variant datasets

In order to determine the relevance and use cases of an exclusively African genomic database, it is pertinent to first explore existing studies that include African samples in a genomic context. The purpose of conducting such an assessment was threefold. Firstly, the purpose was to assess the volume of African genotype or variant data that has been generated thus far. This has implications for both the appropriateness and feasibility of such a database at present. Data volume is also likely to impact on design and prototyping considerations, such as scalability of the database and the accompanying visualisation platform. Secondly, it is important to assess candidate data sets available for inclusion in the database in order to demonstrate utility of the application and to perform validation activities, as well as to serve as a test dataset during code development. Importantly, valid data sets are ultimately limited to those which are "open access" or otherwise available with restrictions that do not limit their utility in the AGVD. An example of the latter scenario may be restricted genotype data for which population allele frequencies may be publicly shared. Finally, is useful to evaluate the manner in which existing data are being utilized at present. Assessing the various ways in which the data are being stored, analysed and visualised at present is useful for the development of appropriate use cases, concomitant feature sets, and ultimately the value proposition of the AGVD.

2.1.1 Literature search

With the above aims in mind, a thorough literature search was conducted to ascertain a list of relevant studies and accompanying data sets. For a study to initially be considered in downstream classification steps, it was required to: include either whole-genome, whole-

exome, or targeted sequencing data as part of the dataset analysed; include data from African samples (preferably across a diverse set of cohorts); as well as perform bioinformatics analysis on the data within the context of variant discovery or annotation, or within a population genetics context.

To address these criteria, a PubMed search (https://pubmed.ncbi.nlm.nih.gov/) was conducted on 2020-07-20 using several search parameters (*Box 1*, below). The initial search yielded 681 results. Additional filters were then applied to further narrow the search space via the available filtering options. Only articles published within the last ten years were included at time of writing, and the species was restricted to "Humans". Additionally, the text availability was restricted to "Full text". The final set of search results included a total of 333 journal articles (see *Appendix A*).

Box 1: Initial search parameters supplied to PubMed literature search for African genomic studies

```
(((((("African"[Title/Abstract] AND "population"[Title/Abstract]) OR
("African"[Title/Abstract] AND "ancestry"[Title/Abstract])) OR ("African"[Title/Abstract]
AND "genome"[Title/Abstract])) OR ("genomics"[Title/Abstract] AND
"Africa"[Title/Abstract])) AND ((((("variant"[Title/Abstract] OR
"mutation"[Title/Abstract]) OR ("allele"[Title/Abstract] AND
"frequency"[Title/Abstract])) OR ("linkage"[Title/Abstract] AND
"disequilibrium"[Title/Abstract])) OR "Fst"[Title/Abstract])) AND
((((("WGS"[Title/Abstract] OR "WES"[Title/Abstract]) OR "TE"[Title/Abstract]) OR
```

A manual curation effort was then applied to search results with an increasing level of stringency observed between filtering stages. During the first phase of curation, article titles were scanned in order to discern between the following three categories within the context of the AGVD: "High relevance" (HR), "Possible relevance" (PR) and "Low relevance" (LR). In addition, duplicate entries were searched for and removed. Relevancy of articles was determined as a function of the aims outlined above. In other words, those which demonstrated possible utility in terms of either the ideation or development of the AVGD itself or had the potential to be used as part of the demo dataset in the database were assigned a greater weight. As such, articles which hinted at an African-centric subject matter and/or larger scale population genetics studies were assigned the highest priority. Alternatively, articles which hinted at a focus on admixed or non-continental African populations, or those with small cohort sizes (for example, a single individual or population) were relatively less important. Furthermore, titles which suggested smaller panel sizes, alternative sequencing chemistries (such as non-HTS approaches), or an analytical focus that does not closely align with the

intended use cases of the AGVD were assigned the lowest importance. This first pass manual curation effort resulted in a total of 78 HR, 127 PR and 128 LR articles.

The surviving HR results went on to the next phase of curation, while the PR articles were retained for later use, should additional studies be required in future. During this phase of curation, HR article abstracts were parsed manually in order to garner additional insights into each underlying study. Articles were removed that were deemed less relevant than initially assumed in the previous step. This approach followed the same guiding principles as outlined previously but ultimately sought to create a single list of studies with high potential value in downstream database curation or platform development. In comparison to previous steps, additional filtering stringency was also applied to prioritize larger panel sizes (i.e. WGS, WES and larger genotyping arrays). Following this step, a total of 14 studies were included in the final list.

The final phase of manual curation involved reading each remaining article in its entirety in order to make further distinctions about the study design and relevant datasets used as part of each study. This classification process included several additional considerations that help to position the usefulness of each study within an appropriate context for downstream use. Specifically, the origin and type of data used in the study was identified, as well as any applicable access restrictions. Additionally, metadata were captured which described, for example, the study design and the samples included in the study. This included the number of participants and different populations included in the study as well as the sequencing strategy employed. The complete resource list, along with the data from intermediate steps, can be found in *Appendix A*. Figure 3 provides an overview of the steps followed to generate the final resource list.

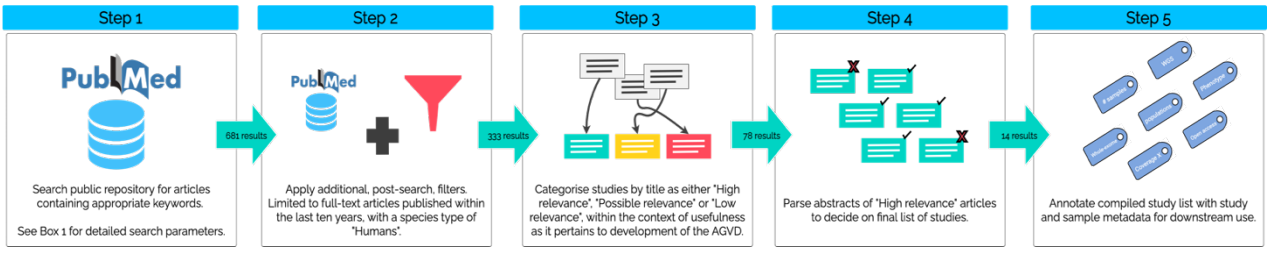

Figure 3: Process outline for literature review of African genomic studies.

2.1.2 Study selection

Once the resource list had been compiled, one or more candidate data sets were chosen for inclusion in the concept build of the AGVD. After consideration of the available data sets from the studies in the resource list, the 1000 Genomes Project (1kGP) Phase 3 dataset was selected (Auton *et al.*, 2015). This dataset was amongst the most widely used and cited across the available studies, which is advantageous from an analysis and validation perspective. It is also freely available for use and dissemination and therefore served as an ideal candidate for inclusion within the pilot phase of the AGVD. Also important within the context of the AGVD, the 1kGP dataset represents a large cohort of samples across diverse ethnolinguistic groups even though the geographical distribution across Africa is limited. With the above in mind, it is worth noting a short-coming of the literature review process described above: although the 1000 Genomes Project was frequently cited in the study result set, the 1kGP study was itself not present as part of the search results, which reveals limitations in the current methodology. The discussion that follows provides an overview of the history and aims of the 1000 Genomes Project, as well as the types of data and samples included in the study.

The 1000 Genomes Project was a large-scale population sequencing effort which concluded in 2015 – almost a decade after its inception in 2007. At the time, it was the largest whole genome sequencing project ever conducted and continues to give rise to novel insights into human genetic variation, disease susceptibility and demography. It is worth noting that some overlap exists between the 1kGP and the International HapMap Project – a large genotyping effort which ran for several years prior (Frazer *et al.*, 2007). The 1kGP aimed to build on part of the work of the HapMap Project, but to expand this work to a larger scale by taking advantage of more contemporary, high-throughput sequencing (HTS) solutions. Ultimately, the goal of the 1kGP was to provide a catalogue of most common human genetic variation and to publish this resource in a freely accessible manner, for posterity (Auton *et al.*, 2015).

The project was divided into a pilot phase, which concluded in 2010, followed by three main project phases. The pilot phase applied a combination of low coverage (2-4 X) whole-genome sequencing (WGS), high coverage (20-60X) WGS family trio analysis and high coverage (50X) targeted re-sequencing approaches to approximately 1000 individuals across a limited range of ancestries (D. L. Altshuler *et al.*, 2010). The results of the pilot phase were then used to inform the sequencing strategy for the following three phases of the main project. This was followed soon after by the phase 1 release of the main project in 2012, in which a total of 1092 genotypes across 14 different populations were reported (Altshuler *et al.*, 2012).

The main project concluded in the final release – the 1000 Genomes Project phase 3 dataset – in 2013. This phase realised a mean depth of approximately 7X and 65X for low coverage WGS and high coverage targeted exome sequencing, respectively, across samples. The NGS sequence data, alongside array genotyping data, was jointly analysed to yield greater than 88 million variants across the entire dataset. This represents approximately 80% of variants available in dbSNP at the time of publication (Sherry *et al.*, 2001). The 1kGP phase 3 release represents a total of 2504 individuals across 5 continental super populations – African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). These cohorts include a total of 26 different global populations, 7 of which are of African Ancestry, for a total of 661 African samples (*Table 1*, below) (Auton *et al.*, 2015). The final release is freely accessible to the public and is available via the International Genome Sample Resource repository (IGSR), at: https://www.internationalgenome.org/ (Clarke *et al.*, 2017).

Table 1: Overview of the African samples included in the final release (phase 3) of the 1000 Genomes Project dataset.

Population Description	Population Code	Number of samples
African Caribbean in Barbados	ACB	96
African Ancestry in Southwest US	ASW	61
Esan in Nigeria	ESN	99
Gambian in Western Division, The Gambia	GWD	113
Luhya in Webuye, Kenya	LWK	99
Mende in Sierra Leone	MSL	85
Yoruba in Ibadan, Nigeria	YRI	108
Total		**661**

2.1.3 Data extraction

The complete set of 1kGP phase 3 genotype data for the GRCh37 reference assembly was downloaded via FTP from the IGSR repository. This was achieved by using the Unix tool "wget" and supplying the URL: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ with the regular expression: "*genotypes.vcf.*". The compressed VCF files contain the genotypes for each of the 2504 samples included in the study as individual identifier columns, while each unique allele is represented as an independent row in the file. Additional information regarding each allele is also included as part of the "INFO" column, for example

the pre-calculated allele frequencies for each site as observed within each super population. Several of these VCF files were later selected in downstream steps, loaded into the AGVD and further annotated, according to section *4.2 Data ingestion and annotation*.

Additionally, phenotype data describing each individual in the VCF files was downloaded according to the same methodology as outlined above but by instead supplying the regular expression "*integrated_call_samples_v3.20130502.ALL.panel". This tab separated file describes each sample included in the VCF file as a row, with subsequent columns mapping to the population, super population, and gender of each individual. Metadata from the panel file was extracted and combined with additional metadata in subsequent steps, according to the processes outlined in section *4.2 Data ingestion and annotation*.

2.2 Existing genomics tools and databases

Particularly within the last decade, resources have increasingly become available for facilitating the storage, analysis, and dissemination of large-scale genomic data. Each tool or database is typically designed to address a particular set of research or clinical inquiries or may focus on a limited cohort of individuals or populations. In order to design a uniquely African resource for this purpose, it was necessary to first review currently available genomic databases, analysis tools and visualization platforms. A brief overview of several popular genomic analysis tools will be discussed below, followed by an overview of some of the most prominent genomic databases.

The primary focus of this investigation was to identify the main feature set offered by each resource, and how they may potentially be leveraged during development of the AGVD, if required. In the ideal case, a single resource (or set of resources) may already provide much of the functionality required for the AGVD, requiring only a limited set of changes. In this case, using such a resource would necessitate that the code base be open sourced, so that it may be used and modified without restriction. Alternatively, tools that are not open sourced but that still provide a convenient application programming interface (API) for accessing data or analytical methods may also be useful in the design of the AGVD. Tools that do not meet these criteria may still be indirectly useful if they can be used to inform on the design of the AGVD and its resulting feature set. Software requirements of the AGVD are discussed in more detail in *Chapter 3 – Developing Software Requirements*.

2.2.1 GLOW

GLOW (https://projectglow.io/) is a sophisticated genomic toolkit designed for working with large-scale data. It was developed from a collaboration between Databricks and the Genetics Center at Regeneron as an extension to the Databricks data analytics platform (Nothaft *et al.*, 2019). The GLOW toolkit facilitates a wide array of bioinformatic inquiries across various different data formats, including both array and sequence data. It can be used alongside existing tools to perform sequence alignment, variant calling, filtering and normalization, regression, statistical testing and machine learning operations on population-scale data. Additionally, metadata can easily be integrated into GLOW analyses by way of an intuitive API. Many of these functions are exposed as high-level operations available from the command line or via various APIs, including Python, R and Scala APIs. This allows genomic analyses using GLOW to be carried out in a largely language agnostic manner. In particular, native integration with the Python Pandas and Jupyter libraries make GLOW an accessible option for many Python developers and its use of Spark SQL DataFrames can help to simplify integration across different sources of data. As an added benefit of this design, GLOW is also highly amenable to existing workflows.

GLOW was optimized for large data sets and is highly scalable owing to its underlying implementation of Apache Spark (https://spark.apache.org/) – a well-regarded standard for the scalable analysis of generic data types – and leveraging cloud-based infrastructure such as Amazon Web Services (AWS) or Azure. It is thus particularly well suited for tasks pertaining to variant annotation and phenotype data aggregation across various data sources. This makes it a good candidate for performing calculations on aggregated, stratified data (for instance across predefined cohorts), or for inferring large-scale genotype-phenotype associations between populations. GLOW also simplifies the storage of large data by leveraging Delta Lake for distributed storage. The GLOW toolkit itself is open-source but forms part of a larger ecosystem of data analytics solutions which collectively form the Databricks suite. GLOW does not currently offer extensive user authentication and authorization capabilities.

2.2.2 Hail

Similar to GLOW, Hail (https://hail.is/index.html) is a genomic analysis tool suite designed to scale well to large data sets, such as population genotype data and genetic biobanks. Hail was developed by the Broad Institute in 2016 and has already been included in a number of large

projects since its inception, such as gnomAD and the Neale lab UK Biobank mega-GWAS (Neale, 2018; Karczewski *et al.*, 2020; Krissaane *et al.*, 2020). Hail includes a wide array of data import, quality control and analytical functions. Data import functionality includes native support for various next-generation sequencing file formats, distributed file storage – thus reducing memory requirements for large data sets – and conversion to Hail Matrix Table format for efficient querying and analysis. Importantly, the use of Hail Matrix Table object over conventional plain text file types such as VCF offer significant performance advantages (Krissaane *et al.*, 2020). Native support for data manipulation and analytical methods ranges from simple operations such as variant normalization, filtering and annotation to computing rare variant associations across large and complex user-defined cohorts (Ganna *et al.*, 2016).

Hail is written in Scala and is made accessible via a rich Python API. Hail workflows can also be scaled up for large data sets on cloud-based infrastructure, such as AWS or the Google Cloud Platform (GPC), by leveraging Apache Hadoop (https://hadoop.apache.org/) and Apache Spark for distributed storage and compute. Like GLOW, Hail's emphasis on scalability make it well suited for performing data aggregations and calculating statistics across large cohorts of individuals. This is an important requirement of the AGVD, as will be discussed in future sections. Hail is open source and does not depend on a larger platform of proprietary software solutions. This makes it a good candidate for inclusion as part of the backend data processing operations required for the AGVD. Finally, although Hail supports extensive data joining operations and boasts a performant file format for data storage, it does not offer a comprehensive database solution for storing and managing data, nor user authentication and authorization functionality.

2.2.3 OpenCGA

OpenCGA (https://github.com/opencb/opencga) is a high-performance, scalable, genomic analysis and storage platform. OpenCGA forms part of the open-source for Computational Biology (OpenCB) genomics tool suite, created in 2012 (Medina and OpenCB, 2019). It has been adopted by several large-scale genomics projects since its inception, such as the Human Genome Variation Archive (HGVA) and BiERApp (Alemán *et al.*, 2014; Lopez *et al.*, 2017). OpenCGA natively supports the ingestion, storage and manipulation of various common bioinformatics file formats, including raw sequence data (e.g. FASTQ and FASTA format), secondary and tertiary analysis outputs (e.g. BAM and VCF format) as well as support for

annotation and panel files (e.g. GFF3, BED and PED format). First class support for such a variety of biological formats simplifies the loading and manipulation of data, and is indeed one of the most prominent features of OpenCGA. All available data loading, querying, and analysis methods are made accessible to users via an extensive CLI, as well as REST API. These include a set of commonly used primary and secondary analysis tools for micro-array, WGS, WES, and RNA-Seq experiments (for instance, BWA, GATK and PLINK are currently supported). Custom analysis tools that are not supported by a standard OpenCGA installation may also be integrated by writing custom JAVA plugins or wrappers around existing binaries. Although the CLI and REST API represent the primary interface for end-users, Python, R, Java and JavaScript APIs offering a reduced feature set are also available and are under active development.

OpenCGA is written in Java and is scalable to large datasets in the order of Petabytes. The process of storing data on the file system is unique from GLOW and Hail in that data is persisted in NoSQL databases, with multiple NoSQL database options available for use on the backend, such as MongoDB (https://www.mongodb.com/) or Apache HBase (https://hbase.apache.org/) for distributed systems. Once persisted, data can also be indexed for efficient querying by means of the Apache Solr indexing tool (https://lucene.apache.org/solr/), or similar. By extension, OpenCGA also includes an extensive metadata catalogue system for managing stratified data types and allowing tiered user access. For instance, a well-defined "User" catalogue can be used to store and authorise authenticated users within the database; the "Study" catalogue is a collection of multiple files (potentially shared across multiple data sets), individuals (consisting of one or more unique samples), samples (for instance, FFPE tissue samples from treatment versus controls) and cohorts (comprising a many-to-many relationship with samples). Together, this data catalogue can be leveraged to build comprehensive genomic pipelines from data creation and analysis to investigation of results in a tiered, access-controlled manner. The OpenCGA code-base is also open source and actively maintained. Therefore, OpenCGA provides a good overall candidate for a genomics storage and analysis backend for the AGVD.

2.2.4 Prominent human genetic databases

Genetic databases aim to store, share and facilitate the exploration of genomic data. Data stored in genetic databases may originate from multiple studies and research areas. For example, the

NCBI Genome Database (https://www.ncbi.nlm.nih.gov/genome/), Ensembl (https://www.ensembl.org/index.html) and Ensembl Genomes (https://ensemblgenomes.org/) serve as centralized repositories of genome reference data and accompanying annotations spanning across multiple eukaryotic, prokaryotic and viral species. Genome databases may also be specialized for a particular organism or application. Specialised databases of model organisms are particularly common. Such databases include: Mouse Genome Informatics (MGI) for the mouse genome (*Mus musculus; http://www.informatics.jax.org/)*, FlyBase for *Drosophila melanogaster* (https://flybase.org/), and WormBase for *Caenorhabditis elegans* (https://wormbase.org/) (Southwood and Ranganathan, 2019).

Increasingly, specialized genome databases exist which focus exclusively on humans but within a particular research domain – prominently within a clinical setting (Richards *et al.*, 2015). For example, The Cancer Genome Atlas (TCGA; https://portal.gdc.cancer.gov/) consists of over 2.5 Petabytes of -omics data for the diagnosis, treatment and prevention of cancer, while ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) maintains a broader focus on the relationship between human disease and health. Furthermore, certain databases consist of population genetic data from healthy or healthy and disease-matched individuals across either a specific population or broad spectrum of populations. For instance, the UK10K Project (https://www.uk10k.org/) focuses on generating and analysing sequence data from 10,000 individuals of healthy and extreme phenotypes from the United Kingdom. In contrast, gnomAD (https://gnomad.broadinstitute.org/) is a population genetic database consisting of aggregated data from various independent studies and represents individuals from multiple global populations, which can be useful for comparing allele frequencies between different cohorts. Table 2 provides an overview of several prominent human genetic databases available at present.

Table 2: Examples of prominent, publicly available human genetic databases.

Database	Description	Available	Source
The Genome Aggregation Database (gnomAD)	Database of high-quality variants called from a total of 125,748 exomes and 15,708 genomes from unrelated individuals, predominantly from case-control studies of common disease. Data are aggregated across various independent sequencing projects and have been made accessible via a rich data querying and visualization interface. A replacement of the Exome Aggregation Consortium (ExAC).	https://gnomad.broadinstitute.org/	Karczewski et al., 2020
dbSNP	Public archive serving as a central catalogue of all known variants of low numbers of base pair. SNPs form the predominant variant class represented by the database but other class such as microsatellites and short indels are also included.	https://www.ncbi.nlm.nih.gov/snp/	Sherry *et al.*, 2001
dbVar	Database of structural variation greater than 50bp (e.g. inversions and CNVs) from submitted studies on WGS, WES and micro-array experiments. Variants derive from both healthy control samples as well as clinically relevant variants from ClinVar.	https://www.ncbi.nlm.nih.gov/dbvar/	Lappalainen *et al.*, 2013
ClinVar	Public archive of variants and supporting evidence of associated clinically relevant phenotypes.	https://www.ncbi.nlm.nih.gov/clinvar/	Landrum *et al.*, 2018
The database of Genotypes and Phenotypes (dbGaP)	Database containing a diverse set of data types from studies investigating interactions between genotype and phenotype; for example, GWAS and WGS sequencing assays. Included phenotypes include both disease and non-disease traits.	https://www.ncbi.nlm.nih.gov/gap/	Tryka *et al.*, 2014

The NHLBI Exome Sequencing Project (ESP) Exome Variant Server (EVS)	Database of exome capture variant data and allele frequencies compiled from 2,203 African American and 4,300 European-American individuals from a subset of well-phenotyped dbGaP cohorts, including healthy controls as well as heart, lung and blood disease phenotypes.	https://evs.gs.washington.edu/EVS/	EVS, 2014
The International Genome Sample Resource (IGSR)	Database for storing, updating and disseminating open access genome and exome sequence and variant data from various studies. This includes, most prominently, data from all phases of the 1000 Genomes Project but also includes follow up studies on samples and populations not originally included in the initial study.	https://www.internationalgenome.org/	Clarke et al., 2017
The Database of Genomic Variants (DGV)	Database of structural variation (e.g. inversions and CNVs) derived from greater than 22,300 genomes of healthy individuals, with the primary purpose of serving as control datasets in external disease-phenotype studies.	http://dgv.tcag.ca/	MacDonald *et al.*, 2014
Ensembl	Database and genome browser of vertebrate genome sequences, variation and annotations. Ensembl offers a rich data query and download interface as well as various ancillary tools for performing sequence alignment, analyzing gene function and the prediction of variant consequence on genes, transcripts and proteins.	https://www.ensembl.org/index.html	Yates et al., 2020

3. Developing Software Requirements

Following on from the outcomes of the research phases of the project, a structured software design methodology was followed in order to generate a list of requirements during the design phase of the AGVD. At a high-level, the design phase of the AGVD began by first assembling one or more "user scenarios" – each being informed in some way by earlier research outcomes. Together, the user scenarios were then used to compile an overlapping list of "user requirements", which were later distilled into a set of "functional requirements" and prioritized for inclusion in the concept build of the AGVD. Each of these steps will be discussed below, along with selected examples for further clarity. This section will also conclude by sharing the final list of features earmarked for development.

3.1 Defining User Scenarios

The user scenario (US) is a concept widely adopted by the software engineering industry for the development of meaningful software requirements and feature ideation. Typically, each US consists of one or more paragraphs which detail the intent of an *imaginary* user of an application by describing a particular problem that should be solved in the hypothesized case. By defining the goals of the software application from the user's perspective, this process can help to ensure the usefulness and appropriateness of the end product and streamline the software development life cycle (Arms, 2014; Affairs, 2020a). For the purposes of the AGVD, each scenario was also designed to include additional background information about the hypothesised user and their intentions. These descriptions provide additional, nuanced context into each hypothesized case and can help to inform on the considerations that need to be made in order to address these problems during the downstream build phase.

Throughout this process it was important to maintain a single area of focus for each US in order to minimise overlap between downstream requirements. Each US was also written in such a way that – although the desired solution may be alluded to – the specific actions that ought to be taken by the user in order to achieve the desired outcomes are not explicitly defined upfront. These high-level descriptions provide a good framework to use for generating more specific functional requirements in later steps, while still allowing for flexibility in the way the solutions ought to be engineered in subsequent build phases. For example, a solutions-oriented development focus – rather than one concerned primarily with interface design early on – can

make it more likely that any particular aspect of the application will be accessible for multiple end-users across different user-scenarios.

A total of nine initial USs were generated as part of the design phase. *Box 2*, below, describes one example scenario generated during this process. It is important to note that the US list was not intended to be comprehensive in terms of the use cases that the AGVD may eventually be applied to. Rather, the intention was to generate a small, focused subset of scenarios that may be addressed during the concept build of the AGVD in order to assess the feasibility of the application and to showcase utility to potential stakeholders. Additional scenarios generated in later stages of the project will become the focus of future iterations of the application.

Box 2: Example of a user scenario generated as part of the design phase of the AGVD concept build

> *A geneticist has received the genotyping results for their patient, a 12-old **Malawian** child exhibiting **developmental delay** and intellectual disability. The results of the report indicate that a mutation, **c.892C>T** on the **NACC1** gene, may be clinically relevant, based on in silico prediction. The geneticist would like to try and identify whether this variant may be pathogenic (and causative) using, in part, the **allele frequencies** observed for this variant within the broader population.*

3.2 Developing User Requirements

The completed US list was then applied to the downstream process of identifying a set of key requirements from the user's perspective. These types of requirements are referred to as user requirements (URs). This step of the requirements generation process was useful for refining the essential needs and intentions of the user, such that their particular situation could be accounted for, in the hypothetical case (Affairs, 2020b). This process involved representing each US as one or more UR. The results were compiled as a table to allow for easy traceability between requirements (*Appendix B*).

Each UR was written in prose to describe only one aspect of the user's intentions at a time. User requirements were also generated to document the requisite functionality needed to support such intentions (for example, the need for a log in and authentication system would be required for accessing private patient information by clinicians). It is worth emphasizing that, similar to that of the US, the focus of each UR was on the problem or task to be addressed, rather than the implementation thereof. Part of this process also included searching for overlap

between different requirements, such that a common feature set may be derived in later steps. This means that the specific context of the user's intentions as derived from the US was primarily used to inform the UR rather than being included as part of the requirement itself. In some cases, however, it was helpful to retain part of this detail in the form of one or more examples to accompany the requirement (for example, referencing a gene's common name can help to set up test cases for acceptance criteria). To further illustrate these concepts, at least four URs were derived from the example US in *Box 2,* above, and are outlined in *Table 3*.

Table 3: Example of the user requirements (URs) and derived functional requirements (FRs) from the user scenario (US) in Box 2. Functional requirements were later prioritised for inclusion in the concept build of the AGVD using the MoSCoW model and a subset of USs were retained.

User Scenario	User Requirement	Functional requirement	Priority for AGVD concept (MoSCoW model)
1	User shall be able to **find** information related to a **specific mutation** (e.g. "c.892C>T on the NACC1 gene").	Search bar (by standard variant nomenclature).	Must have
		Results filter box (by standard variant nomenclature).	Should have
	User shall be able to easily **identify** the gene name to which a variant belongs.	Results table that includes the gene common name that a variant belongs to.	Must have
	User shall be able to obtain the allele frequency of a variant as observed in the African population.	Results table that includes the allele frequency of the variant aggregated across all African populations.	Must have
		Results table that includes the allele frequency of the variant aggregated across a specific African subpopulation (e.g. ethnolinguistic group, nationality).	Must have
	User shall be able to obtain the allele frequency for a variant as observed in the general population (not limited to African).	Results table that includes the allele frequency of the variant aggregated across all populations (may be obtained from an external source).	Could have

3.3 Generating Functional Requirements

The next stage of the software requirements generation process involved producing one or more functional requirement (FR; sometimes also referred to as a "system requirement") for

each UR (Parker, 2020). In this context, a functional requirement refers to an *actionable* feature that can be included as part of the application. These are specific requirements that, given their implementation, should satisfy one particular aspect of an associated UR. Contrary to the way URs are composed, FRs are typically devised from the perspective of the software engineer, rather than the end-user. (US Department of Defense Systems Management College, 2001; Wu and Buyya, 2015). Throughout this process, care was also taken to not introduce additional implementation detail as part of each FR. For example, the particular programming languages, paradigms, supporting infrastructure, or functional optimisations that ought to be utilised in order to realise the associated aims should be disregarded at this step of the design process.

As an example of the process followed, the ability to "find information related to a specific mutation" – an excerpt of an UR from *Table 3* – can be partially satisfied by implementing a search bar as a feature of the application. It is worth noting from the preceding example that this particular UR is not entirely satisfied by the proposed FR (namely, the search bar). As in most cases, this UR in fact necessitates the addition of several accompanying features in order for its objectives to be fulfilled. This may include, at minimum, a means of displaying the results of the search. As far as possible however, each additional feature was captured as its own independent FR throughout this process. This approach aligns closely with the principle of "separation of concerns" – a design pattern in which the atomicity of software functions is central to the design and implementation of the software. The eventual benefit of structuring features in this manner is that of streamlining development efforts and reducing redundancy and complexity within the source code and the user interface (Hürsch and Lopes, 1995).

After the preliminary list of FRs had been generated, the requirements document underwent a refactoring process by way of manual inspection. The refactoring process sought to identify overlap between different requirements in order to reduce redundancy, and was performed in a manner akin to the user requirements curation process. Once available, the complete list of FRs could then be prioritized for inclusion in the pilot build of the AGVD.

3.4 Prioritising requirements

The software requirements generation process culminated in the action of assigning a priority to each FR. Priorities sought to denote the relative importance of including each feature as part of the AGVD concept build. Prioritisation followed the MoSCoW method, a model commonly used as part of the Dynamic Systems Development Methodology (DSDM) agile software

development framework (Clegg and Barker, 1994; Agile Business Consortium, 2014). Under this model, the priority of each FR or feature was designated to one of the following categories: "must have", "should have", "could have", or "won't have" and colour coded to simplify grouping (*Table 3*). Although priorities were allocated in a subjective manner, careful consideration was given to align as closely as possible with aims of the AGVD; that is to say, to establish the importance and utility of an African focused genetic variant data base within the context of health and disease. Prioritisation thus served the purpose of relegating FRs of lower importance to future iterations of the application build and placing more focus on those areas which would be immediately influential to the success of the pilot project in achieving these aims.

Functional Requirements from the complete set were then grouped based on priority. This step was performed in order to identify a set of user requirements and scenarios which could be jointly addressed by the implementation of one or more features across the gamut of FRs, considering the high degree of overlap of FRs between different URs. Ultimately, the final step involved selecting a representative set of user scenarios that could be used to demonstrate utility of the AGVD in its pilot phase, and to narrow the project scope of the initial concept build. Overall, a total of three User Scenarios were selected for inclusion in the subsequent development phase of the project (*Box 3*).

Box 3: Final list of User Scenarios (US) selected for inclusion in the AGVD pilot build.

US #4: *A PhD student is performing research on infectious diseases. He is interested in learning more about the influence of specific variants of the HBB gene on susceptibility to malaria, particularly within an African setting. As part of his research, he would like to:*

(i) Identify the genomic location and allele change information of his list of variants
(ii) Compare the allele frequencies of these variants between different African populations

US #8: *A researcher of Human African Trypanosomiasis (HAT) and its link to Chronic Kidney Disease in African ancestries would like to explore known variants which confer protection against HAT. She would like to know more about variants of the G1 haplotype that she read about in a scientific journal. Specifically, she would like to obtain a graphical representation of the allele frequencies of these variants within different African ancestries and compare these results with values from external studies.*

US #9*: An MSc student is performing research on the Duffy null phenotype. He is interested in learning more about the influence of variants within the DARC locus on susceptibility to malaria. As part of his research, he would like to:*

(i) Identify all known SNPs (as RefSeq IDs) related to his genes of interest
(ii) Generate a list of variants with potential clinical relevance, using available annotations

4. Design and Implementation of the AGVD

4.1 Application architecture

The AGVD is composed of four primary components: a genomic data storage and analysis engine, a RESTful interface for two-way communication between the genomics engine and the user-interface (UI), an additional layer of middle ware for servicing custom features of the application which are not satisfied by the existing genomics engine, and a custom web application to serve as the ultimate UI for end-users. Together, these components are loosely coupled in a way that maintains separation of concerns and allows specialization of each module for the particular set of tasks required. An overview of the application architecture is given in *Figure 4* and will be discussed in more detail below.

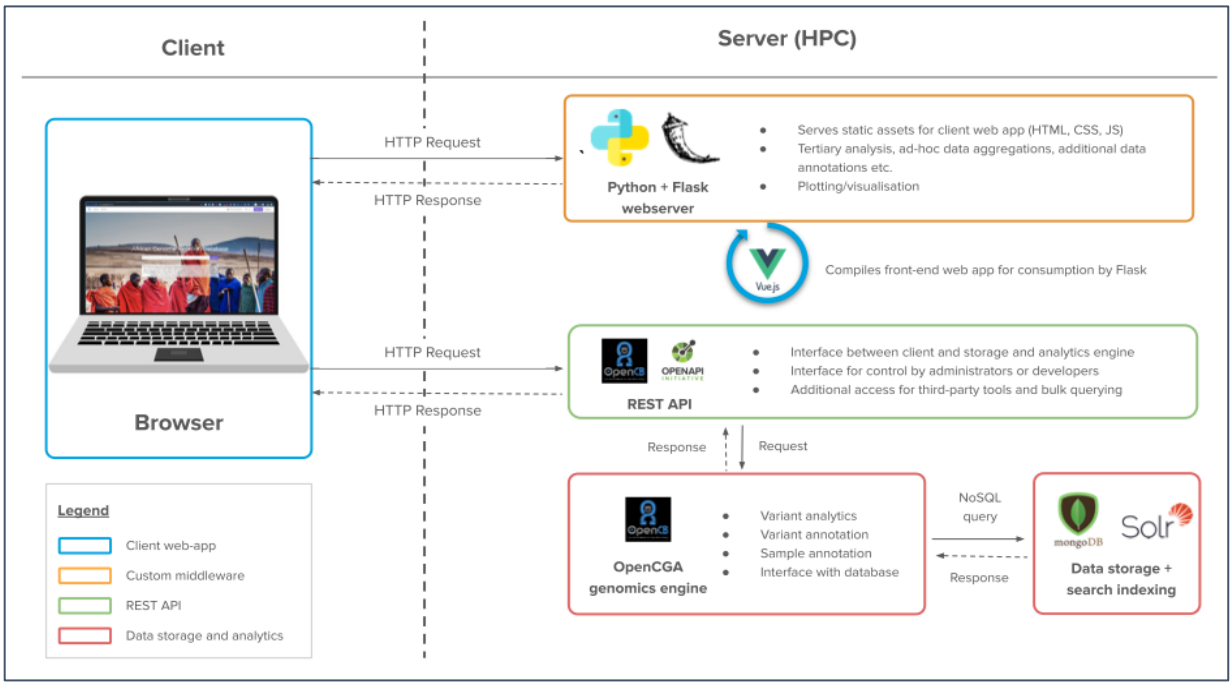

Figure 4: Overview of the AGVD application architecture, coloured by application component. At the root of the application lies OpenCGA – an open source genomics data storage and analysis engine – which is supported by a MongoDB NoSQL database and Apache Solr search indexing service (red boxes). A REST API (green box) facilitates communication between the client web application and the genomics engine in a language-agnostic manner via standardised HTTP requests. The client web application is served by a layer of middleware (orange box), which also supplements the genomics engine with additional data analysis and visualisation functionality.

4.1.1 Client web application

The user interface of the AGVD is available to end-users by means of a web application which is accessible via most modern web browsers, such as Chrome. The web application was developed using the progressive JavaScript frontend framework, Vue.js (available: https://vuejs.org/v2), along with supporting web languages such as HTML5, CSS3 and JavaScript/ECMAScript 2018. Together, this architecture is entirely open source, is widely adopted by the web development community and is well supported by modern browsers. In addition, the Vue.js framework is also lightweight and has been heavily optimized for speed, making it a logical choice for inclusion in the frontend development process (Song, Zhang and Xie, 2019; Vue.js org, 2020).

The frontend web application was developed incrementally by the addition of individual "Vue components". Vue components are standalone, composable, single file modules that closely align with Web Component standards (WebComponents.org, 2020). Each component encapsulates the complete template, styling and data model methods required for rendering of the Document Object Model (DOM). Throughout development, components were also designed to allow nesting of child components under parent components. Taken together, this design pattern ensures modularity within the application and reusability between software components.

Vue components were either custom built for purpose or installed from external libraries via the yarn v1.16.0 package installer and dependency manager (available: http://yarnpkg.com/getting-started/install). The external library Buefy v0.9.3 (available: https://buefy.org/) was used extensively for its lightweight and easy to integrate components, along with the Bulma v0.9.0 CSS framework (available: https://bulma.io/), which provides convenient access to styled HTML DOM elements. Custom components were written to provide additional functionality where necessary as well as for the purposes of maintaining a consistent UI across the entire application. Both custom components as well as external components have been implemented in a way that is mobile friendly, thus ensuring cross-platform compatibility. All of the above tools are open source and therefore did not require a license for use.

Before deployment of the web application, the codebase was minified and compiled into static HTML, CSS and JavaScript files. This step was necessary in order to generate a set of standalone assets that could be easily served as a single page web application by the

middleware. Splitting the code up in this way – whereby the client application is developed entirely separately from the backend and middleware application layers – was done in order to maintain separation of concerns and simplify the development process. Bundling assets in this way is also important for optimizing the size and runtime speed of the web application, since the web application is stored entirely on the client's browser after the initial page request. In addition, the bundling process was also used to resolve dependencies between view components and third-party libraries, allowing for "lazy loading" of components where possible as well as removing redundant modules. Webpack v4 was used as the bundling tool (available: https://webpack.js.org/guides/installation/) alongside the Babel v7 JavaScript compiler (available: https://babeljs.io/), and was configured as part of the Vue CLI tool set (available: https://cli.vuejs.org/).

4.1.2 Custom middleware

The client web application is served as a static asset bundle by a layer of middleware that was also custom built for purpose. The middleware serves a dual-purpose as both the web server of the client web application and as an auxiliary set of functions which complement the functionality afforded by the genomics engine. This component of the application was written in the Python v3.6 programming language (available: https://www.python.org/), using the Flask v1.1 micro framework (available: https://flask.palletsprojects.com/en/1.1.x/). Django was initially used as a web development framework during early development phases of the AGVD but was later abandoned in favour of the flexibility and minimal overhead afforded by Flask – two important design considerations of the project.

The middleware component exposes available request endpoints via a simple REST API, which can be called upon asynchronously by the web application to retrieve data or perform data processing and visualization capabilities. The external Flask plugin, Flask-RESTPlus v0.13 (available: https://flask-restplus.readthedocs.io/en/stable/) was used for supporting REST API functionality, and was selected for its intuitive API and active development (LibHunt, 2020). All available endpoints were automatically documented and served by the Flask application as an interactive "Swagger" web page, using the docstrings available for each function. This was performed using the swagger-flask plugin according to the specifications set out in the OpenAPI V3 spec (OAI, 2016). By exposing available endpoints in this manner, future iterations of the project may benefit from a standardised and well documented

framework that can be leveraged for adding additional request endpoints – for example adding additional graphing or analytical functionality. In addition, if the future intent is to expose this functionality more directly to end-users, then having a standardised API will greatly improve both the discoverability as well as ease of use of the API by external users.

Similarly to the way the client application was structured, the middleware was divided up into individual modules, each composed with related functions or common API endpoints; this was achieved by following the Flask Blueprint design schema (Flask, 2019). Structuring the code base in this manner was important for maintaining extensibility of the project as new features are added to the application and the code base develops in size. The process of refactoring the code base in this way was carried out in accordance with the recommended best practices of the Flask framework (Flask, 2020). All of the frameworks, libraries and plugins described in this section are open source and therefore freely available for use.

4.1.3 Data storage and analytics

One of the objectives of the research phase of the project was to discover and assess suitable candidates for genomic data storage and analytics which could feasibly be included as part of the AGVD. As part of these outcomes, OpenCGA v1.3.11 (available: https://github.com/opencb/opencga/releases/tag/v1.3.11) was selected for inclusion as part of the backend of the AGVD after comparing the available feature set, documentation, level of active development as well as external usage of OpenCGA, GLOW and Hail as alternative genomic storage and analytics engines. OpenCGA ("the genomics engine") satisfied the majority of the requirements for large data storage and analytics of the application, as defined in the Software requirements outlined in *3. Developing Software Requirements*. For instance, OpenCGA is an open source application which is highly scalable to larger data and provides a wide array of data storage and analytics functionality for use in large scale genomics projects, for instance allele frequency calculations (see *2.2.3 OpenCGA for an overview*) (Medina and OpenCB, 2019; OpenCB, 2020). For the pilot release of the AGVD, the genomics engine was used primarily for storing and managing variant data as well as adding a variety of external annotations to these variants (see *4.5 Features of the AGVD*).

OpenCGA is accessible primarily via a command line interface (CLI), as well as a REST API. The CLI and REST API are both suitable candidates for the majority of data storage and analytics operations required by the AGVD and these interfaces are closely aligned with one

another. However, several documentation and implementation issues were experienced with the CLI; thus, the REST API was ultimately selected as the primary interface between the front-end web application and the genomics engine. This interface also confers many advantages for the ease of use of the application as well as interoperability between application layers, as will be discussed in *4.1.4 REST API*.

4.1.4 REST API

The AGVD client application was built to interface with the genomics data engine by means of a REST API which is exposed under the "/webservices/rest" endpoint of the root domain. All available endpoints are documented via an interactive web page at this location and are aligned with the OpenAPI v3 spec (OAI, 2016); this pattern also aligns closely with the design schema used for the custom middleware REST API, as outlined previously. The REST API provides several advantages over the conventional command line interface for accessing OpenCGA. For example, a more interactive UI, the ability to perform bulk querying and the exposure to (authorised and authenticated) users and developers without the need to log in to the server.

The REST API is available as a convenient accompaniment to the OpenCGA engine. As a result, development of this interface can be largely relinquished to the OpenCGA development team. This component is therefore expected to benefit from synchronicity between development efforts focused on the underlying engine itself and the interface used to access it; this is particularly important since future updates to AGVD may include newer versions of OpenCGA. Where additional data querying, visualisation, or analytics capabilities are required that are not satisfied by OpenCGA, the custom layer of middleware serves to supplement the available feature set.

The REST API is served using the Tomcat v8 web server (available: https://tomcat.apache.org/download-80.cgi), but can be substituted with an equivalent web server, should infrastructure limitations apply to future iterations of the project. Briefly, the following endpoints that are exposed by the REST API under the root domain were used extensively by the AGVD client application: "/users", "/projects", "/studies", "/analysis and "/meta".

4.2 Data ingestion and annotation

The 1kGP phase 3 genotype dataset and accompanying metadata was selected as a proof of concept dataset for inclusion in the pilot phase of the AGVD and downloaded according to the protocol outlined in *2.1.3 Data extraction*. For the pilot phase of the AGVD, only open access data was added to the database. However, tiered access to study data in future iterations of the AGVD requires that users be added to the database and their access permissions defined upfront. Therefore, we implemented access restrictions on the data as a proof of concept for future versions. The process of installing the AGVD as well as creating users are therefore prerequisites to many of the steps outlined in this section and are described in in *4.3.2 Application configuration.* The following section outlines the processes followed in order to prepare, load, and annotate variants from the proof of concept dataset into the AGVD following successful installation and configuration.

4.2.1 Metadata curation

A custom population metadata curation effort was followed to enhance certain of the graphing functionalities performed by the AGVD. The data is incorporated into all tables and graphs which represent different populations and their allele frequencies, in the current release. For example, the data is used to better resolve membership of populations to their super populations as listed independently by each study annotation (e.g. 1kGP). Additionally, the world map view of allele frequencies by population uses this data to colour populations by ancestry (as listed by the external study) and plot them by geographical location.

The process of curating this data involved compiling a CSV file with rows corresponding to cohorts (as named by each respective study) and columns corresponding to various cohort descriptors that may be useful during data visualisation. For example, brief cohort descriptions (e.g. “Yoruba in Ibadan, Nigeria”), super population membership (e.g. “AFR” - African) and the study the cohort derived from (e.g. “1000G”) were all captured in standalone columns and were obtained during manual synthesis. The study metadata was particularly important for ensuring uniqueness between cohorts that share cohort ids between studies, which may otherwise be ambiguous. For example, “AFR” is used to denote African ancestry in both the 1000 Genomes Project and gnomAD Genomes studies, which this metadata helps to delineate.

Additionally, information describing the geographical coordinates to be used during plotting were also captured in their respective columns. The rationale for why each coordinate was

selected to represent a population was captured in an additional column. Briefly however, coordinates were generally extracted directly from the study or cell line literature when available, or otherwise selected based on a convenient geographical mid-point or based on population density. The resulting CSV annotation files can be located in the "data" directory accompanying the codebase, and a summarised version will be made available to end-users within the site documentation in future.

4.2.2 Creating Catalogue entries

In order to modify the data catalogue in OpenCGA, the "test" user was first logged in via the CLI by executing the following command, replacing the placeholder text in parentheses:

```
./bin/opencga.sh users login -u {{user}} -p <<< {{password}}
```

Following successful user authentication, a new "Project" Catalogue entry was made in OpenCGA for the GRCh37 reference assembly. The following command was executed:

```
./bin/opencga.sh projects create -a reference_grch37 \
-n "Reference studies GRCh37" --organism-scientific-name "Homo sapiens" \
--organism-assembly "GRCh37"
```

A "Study" Catalogue entry was then created for managing the collection of files from the 1kGP dataset within OpenCGA. For the pilot phase, a single study was defined to store all publicly accessible 1kGP phase 3 data loaded in the AGVD. However, several different studies may be defined in future phases of the project according to different levels of access or research groups as required. The following command was executed:

```
./bin/opencga.sh studies create -a 1kG_phase3 \
-n "1000 Genomes Project - Phase 3" --project reference_grch37
```

4.2.3 Loading variant data

Before loading files into the database, the Catalogue daemon was executed as a background service according to the following command:

```
./bin/opencga-admin.sh catalog daemon --start
```

This service is required to be running before all data loading and indexing operations in order to execute submitted jobs.

In order to populate the AGVD with variant data, genotype files were first registered in the data catalogue. The VCF files which corresponded to the following chromosomes were included in order to validate and demonstrate utility of the AGVD, as outlined in *5. Validation and Utility of the AGVD*: chr1 chr11 and chr22. All chromosome files were added separately due to the limitations of the OpenCGA CLI with respect to handling multiple files at once. Each compressed genotype file was registered using the following command to register each file on the server to the catalog study:

```
./bin/opencga.sh files link -I \
{{ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz}} \
-s 1kG_phase3
```

Once registered in the data catalogue, variants from genotype files were normalised (i.e. variants were left-aligned and represented parsimoniously) and loaded in the database according to a two-step process. First, variants from each file were normalised and serialised into Avro format for later ingestion into the NoSQL database (MongoDB). The following command was executed on each registered file name (note that complete file paths are not necessary after a file has been linked):

```
./bin/opencga.sh variant index --file \
{{ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz}} \
--transform -o outDir
```

Secondly, transformed files were loaded into the database by executing the following command on the registered file name:

```
./bin/opencga.sh variant index -file \
{{ALL.chr21.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz}} \
--load -o outDir -Dload.split-data=true
```

Importantly, the additional flag "-Dload.split-data=true" was supplied for all but the first file during the creation and loading of the file index due to subsequent chromosomes containing identical sample names. This flag is not required in cases where samples are not duplicated across files, such as in cases where VCF files are split across individuals, rather than chromosomes.

A validation step was conducted on all chromosome files once loaded in order to confirm that variants were loaded successfully. A query was performed using the REST API and the results were compared against the original genotype files. The following URL query was performed via the REST API and the results examined for each chromosome:

```
{{domain root}}/opencga-
1.3.11/webservices/rest/v1/analysis/variant/metadata?project=reference_grch37&study
=1kG_phase3
```

The “numVariants” and “numSamples” fields from within the “studies.files.stats” JSON field reported on the number of variants and samples in the database for each loaded chromosome file. In each case, the correct number of samples (i.e. 2504) was reported, as expected from the 1kGP study design and, additionally, listed in the genotype columns of the chromosome files. The number of reported variants was verified against the number of variants in the genotype files by executing the command (example for chromosome 22):

```
zcat < ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz \
| grep "^22\t" | awk '$5 ~ /,/ {print $5}'  | sed "s|[ACTG]||g" | sort | uniq -c
```

The following command was subsequently run:

```
zcat < ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz \
| grep "^22\t" | wc -l
```

Respectively, these commands returned the number of records containing multiple variant alleles and the total number of records containing any number of alleles. These values were used to calculate the total number of unique variants and were compared to “numVariants” for equality. Validation was successful for all chromosome files loaded into the database.

4.2.4 Annotating variants

Following the successful loading of genotypes in the database for all chromosomes, annotation was carried out for all loaded variants. Annotation of variants was performed automatically in OpenCGA via the default annotator, the CellBase REST API, and is discussed in more detail in *4.5.5 Diverse and flexible variant annotations*. To annotate variants, the following command was executed:

```
./bin/opencga-analysis.sh variant annotate -s 1kG_phase3 \
-o /tmp/temporal_annotation/ --log-file annotate.log
```

Log files generated during annotation were retained for the purposes of maintaining a data audit trail and for possible troubleshooting if required in future. Annotation was carried out separately after loading and indexing each chromosome’s VCF file. This iterative approach was not required, but was useful during testing of the database and the development of new features, since annotation of a single chromosome sometimes required up to 24 hours to complete.

4.2.5 Calculating statistics

Cohort level statistics were calculated on indexed variants using available OpenCGA command-line tools. This step was performed in order to generate allele frequency and genotype frequency data for the default cohort "ALL". Specifically, the reference allele, alternate allele, and minor allele frequencies of each variant were calculated, alongside the genotype frequencies. Alternate allele frequencies (AAF) for each variant are calculated in OpenCGA as the number of individuals with the alternate allele divided by the total count of alleles for that variant observed across all individuals. For biallelic variants, this is reduced to the count of the alternate allele versus the count of the reference allele and the annotated minor allele frequency (MAF) will be equal to AAF. In the case of multiallelic variants however, the denominator consists of the reference allele count plus the sum of counts of each alternate allele. Additionally, the minor allele annotated during this process may not necessarily be identical to the alternate allele, in which case the calculated MAF and AAF may also differ. Genotype frequencies were calculated in a similar manner – observed zygosity information is used to generate counts, and then frequencies for each of the following genotypes: "homozygous reference", "homozygous alternate" and "heterozygous" (which also includes phase information for phased samples within the 1kGP dataset).

The above statistics were performed on variants in the database by executing the OpenCGA command line tool "opencga-analysis.sh" in the following manner:

```
./bin/opencga-analysis.sh variant stats -s 1kG_phase3 -o /tmp/temporal_statistics --cohort-ids ALL --log-file stats.log
```

An additional parameter, "--file", was also specified during calculation of statistics for the first chromosome during development of the application in order to write out calculated statistics to file. This was useful for assessing the output from the command and the subsequent development of features using this data. Following the first chromosome, statistics were collectively calculated on the remainder of the variants in the database.

4.2.6 Defining custom cohorts

Example custom cohorts were created and samples assigned to them in order to calculate AAFs across different African populations and demonstrate proof of concept for the pilot phase. A

custom Python script was written to simplify the process of manual cohort creation and circumvented many of the issues initially experienced when attempting to manually create cohorts for multiple samples via OpenCGA directly. For example, the current REST API implementation used in the pilot phase of the AGVD limits sample ID requests to batches of 100 at a time, while the CLI implementation did not work as expected. Instead, the custom script reduces a series of queries and manual data entry operations to a single, standardised command. The script "assign_cohorts.py" was added to the "scripts" directory within the codebase for re-use whenever new variants, samples or populations are added to the AGVD.

Additional validation is also performed on user input and available parameters are well annotated for users in order to standardise and improve reproducibility. Additionally, assign_cohorts.py logs the progress and results of the workflow which can help with error detection during execution. The tool uses a panel file to create cohorts in OpenCGA and map them to their respective samples. The original 1kGP Phase 3 panel file, "integrated_call_samples_v3.20130502.ALL.panel" is included alongside the codebase and is used to infer the relationship between samples and cohorts by default, but a custom panel can also be provided by the user. Panel files should be in tab-separated format and contain at least a "sample" column containing sample IDs as they occur in the loaded VCF files, as well as at least one additional metadata column. With the above in mind, assign_cohorts.py first registers empty cohorts in the Metadata Catalogue according to the supplied information. Following cohort creation, batch requests are executed via the OpenCGA REST API to fetch sample IDs for relevant samples stored in the database. Finally, sample IDs are assigned to cohorts in the Metadata Catalogue according to the specified panel file. In addition to reliably executing the correct OpenCGA commands, the custom script also logs progress and validates user input, therefore helping to standardise the workflow and ensure reproducible research.

Creating and assigning cohorts to African samples from the 1kGP dataset for chromosome 1 was therefore achieved by executing assign_cohorts.py as follows from within the root directory of the AGVD code repository:

```
python assign_cohorts.py "LWK,ACB,GWD,ESN,MSL,ASW,YRI" 1kG_phase3 \
--cohort-description "User defined cohort from custom script" \
--cohort-suffix "manual-cohort"
```

Once samples were assigned to cohorts, alternate allele frequencies were calculated for all variants in the database. First, it was necessary to retrieve information about existing cohorts in the Metadata Catalogue. The following command was executed to generate this information:

```
./bin/opencga.sh cohorts search --output-format JSON > cohorts-search-output.json
```

Cohort IDs were then extracted using a custom Python script, "extract_cohort_ids.py", available under the "scripts" directory of the code repository. Finally, AAFs were calculated for all custom cohorts by executing the following command,

```
./bin/opencga-analysis.sh variant stats -s 1kG_phase3 --file-id \
ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz \
-o /tmp/temporal_statistics --log-file statscreate-cmd.log --create \
--cohort-ids {{cohort1,cohort2,cohort3}},
```

while supplying the generated cohort IDs to the "--cohort-ids" parameter in comma-separated format. Importantly, the cohort ID corresponding to the default "ALL" cohort was removed from the list, since default statistics were calculated previously. Additionally, this command was executed in batches of 12 cohort IDs at a time, due to limitations with the current OpenCGA implementation.

4.3 Deployment and availability

4.3.1 Server Configuration

The AGVD pilot is stored and served via a HyperV Virtual Machine (VM) from within H3ABioNet infrastructure. The VM has the following resources allocated: 4 cores, 16 GB RAM and 2TB data disk. The OS running on the VM is Ubuntu 18.04. These resources were sufficient for running the pilot release of the AGVD. Future versions of the AGVD would likely look to serve the application from a dedicated machine with additional resources, as deemed sufficient by the project team. Additionally, the VM is behind a reverse proxy, which provides an additional layer of security and prompts users for credentials in order to access the platform. This is a temporary measure used during testing of the pilot phase and will not be required for the eventual public-facing release. The user credentials that can be used to access the platform via the reverse proxy are listed in *4.3.3 Availability*.

4.3.2 Application configuration

Configuration of OpenCGA was carried out following the installation of OpenCGA and all accompanying software dependencies, and prior to performing any data management

operations in OpenCGA. In order to create a data catalogue in OpenCGA the following command was executed, replacing the placeholder text in parentheses:

```
./bin/opencga-admin.sh catalog install --algorithm HS256 --secret-key {{secret}} \
<<< {{admin_password}}
```

A test user was then created using the following command, replacing the text in parentheses:

```
./bin/opencga-admin.sh users create -p -u test --email test.email@gmail.com \
--name "John Doe" --user-password {{user_password}} <<< {{admin_password}}
```

The field "sessionDuration" also was set to "1200" (seconds) in the config file "client-configuration.yml" to increase the length of time that a user session persists before being de-authenticated from OpenCGA. Additionally, the "cellbase" field within "storage-configuration.yml" was set to the host http://bioinfo.hpc.cam.ac.uk/cellbase/ and the "database" field was set to match the IP address and port of the local MongoDB database service.

Finally, before executing the Python Flask application "FLASK_ENV" environment variable was set to "production" and a build and compile operation of the web application was performed by executing the command:

```
yarn build
```

The remaining steps of execution were carried out as per the usage instructions accompanying the code repository.

4.3.3 Availability

Once all prerequisite installation, configuration and data management steps were complete, the AGVD was deployed from the server and is available to end-users via the following URL:

```
https://agvd-dev.h3abionet.org
```

The OpenCGA REST API is available from the following URL:

```
https://agvd-dev.h3abionet.org/opencga-1.3.11/webservices/
```

And the custom REST API is available from the following URL:

```
https://agvd-dev.h3abionet.org/api/
```

Upon accessing the AGVD, the reverse proxy requires that the following test credentials be entered for the first time in order to be routed to the AGVD successfully:

```
User: "agvd"
```

```
Password: "Mu6peug2uWai6eigh2ni"
```

Additional user credentials are required in order to log in to the AGVD, which can be found from a banner message on the login view, for convenience.

The source code is stored and version controlled using the online code repository and collaboration platform GitHub. The source code can be accessed by project members who have access to the repository, or external collaborators who have applied for and have been granted permissions by the repository administrators. This can be accessed via the following URL:

```
https://github.com/h3abionet/agvd-app/
```

4.4 Compliance with the H3ABioNet Web Resource Development Policy

An important consideration to make when designing any new software resource is to align as closely as possible with existing standards and best practices. It is our opinion that this is particularly important within the Bioinformatics domain in the post-genomic era, which has historically experienced a lack of standardization between available bioinformatics tools and processes to date (‘Standardizing data’, 2008; Endrullat, 2017). The AGVD is an H3ABioNet project and, as such, sought to align with the H3ABioNet Web Resource Development Policy 1.0 protocol. The following sub-sections denote key focus areas for standardization within this policy and describe how the AGVD attains compliance or deviates from each.

4.4.1 Documentation

The AGVD is documented in a variety of ways. Principally, a set of AGVD end-user documentation has been written to describe general usage instructions of the platform. This is available to the public directly via the AGVD portal, accessible under the “User Docs” link from within the navigation menu. This standalone documentation site describes the basic functionality of the web application and serves to guide end-users to better understand how to navigate the site. It also serves to highlight the intended use and relevance of the AGVD, as per the recommendations set out in the policy. In addition, it also documents expected input and output data for certain examples, as well as their biological interpretation. This information can be used alongside the various tooltip messages that are available upon mouse hover

throughout the site. The documentation site is automatically compiled using markdown files into a bundle of static assets that is served via the code repository instance.

Help is also available to users via the "Contact" view. The Contact view may be useful for users who have additional queries about the platform or the data used therein, or to leave additional feedback, such as requests for new features. As referenced in the policy, the H3ABioNet helpdesk (available: https://helpdesk.h3abionet.org) is listed as a resource for users to request assistance from. The view also provides contact information of the principle collaborators on the project, including the software developers. This may help to facilitate a faster issue resolution process, should they experience any critical issues with the platform.

Furthermore, the AGVD includes a set of hyperlinks to external resources where additional information can be gathered, or further exploratory research conducted. These references are included as items in the navigation menu at the top of the viewport for easy access. At the time of writing, this includes a menu item for the H3ABioNet website (available: https://www.h3abionet.org/), where users can find out additional information about H3ABioNet, the Database and Resources Work Package (under which the project is being coordinated) or the H3Africa consortium.

Finally, the AGVD speaks to the suggestion of including internal documentation, as set out by the policy. In this regard, the AGVD has been well documented for internal users and developers via the code repository which houses the code, as alluded to previously and detailed in subsequent sections. The code repository is not publicly accessible at time of writing. Documentation is available via various markdown files, such as *README.md* which describe project development and deployment processes. This convention of including one or more markdown files at the top level of the repository is commonly used in the software development industry. Markdown files are both human readable as well as easily amenable to automated documentation processes. As such, we suggest that future iterations of the project add a consolidated build process to automatically generate a static documentation site, in a similar manner to the end-user facing documentation site described above. This could be achieved, for example, by using the Sphinx Python tool (available: https://www.sphinx-doc.org/en/master/) which is both simple to set up and highly configurable.

4.4.2 Coding and code repositories

All custom code comprising the AGVD was written in Python, in accordance with the aforementioned policy. This approach was preferred over extending the functionality of third-party tools in their native languages (for example, writing plugins for OpenCGA directly in Java) both for the purposes of maintaining compliance, as well as for simplifying the development process and to support for future updates. The source code is also actively being maintained as part of a GitHub source code repository under the H3ABioNet GitHub organization, as per the recommendations set out in the policy. These aspects have already been outlined in previous sections.

In terms of the Python web framework used for development, the AGVD deviates slightly from that proposed in the policy. The AGVD makes use of the Flask microframework instead of Django. Although the usage of Django is recommended, this suggestion is primarily for the purposes of maintaining familiarity between different development teams within the organisation and does not speak to the simplicity or efficiency of the developed application. The policy also does not preclude the usage of other frameworks. Django was strongly considered as a viable framework during early iterations of the project but was later replaced with Flask. This was performed order to reduce the amount of boilerplate code and unnecessary modules included in the source code, since Flask offers a more minimal, incrementally extensible approach to web application development. This is especially true in the case of REST API development, for which Flask is well suited. Additionally, the Flask API was able to provide more flexibility in terms of integrating with third-party tools, such as NoSQL databases, for which the Django ecosystem was often more prescriptive. Jointly, these considerations resulted in the decision to continue development with Flask, rather than Django.

In terms of coding style, standardisation and linting, which is not detailed in the policy, all Python code was written to align closely with external, well regarded standards by the Python community. Importantly, this includes aligning with PEP8 – the official coding standard for Python (Rossum, 2019). PEP8 includes standards for naming conventions, code structure and improving readability of Python code, which can improve co-development of the project – particularly for larger development teams as the project grows in size. Additionally, the Python code was regularly formatted using the Black v19 code formatting tool prior to committing code changes (available: https://pypi.org/project/black/). The latter results in code that is formatted almost identically across completely different Python projects, thus lowering the burden on developers to understand and contribute to the source code. This process also helps

to reduce the size of each "code diff" – the highlighted differences between changes in the code base – which can simplify the downstream code review process. Finally, the JavaScript code used in the front-end client application is also subject to code linting by the ESLint v4.19.1 code linter (available: https://eslint.org/). This process has been configured to execute automatically prior to the committing of code changes.

Based on the advantages discussed above, we therefore propose that future editions of the H3ABioNet Web Resource Development Policy include a suggestion that these standards and processes be considered by future projects.

4.4.3 Version control

Source code for the AGVD is version controlled in accordance with the version control framework set out in policy. This process has been documented as part of the internal source code documentation which accompanies the repo. Briefly however, the process flow for developers to respect is as follows: create an issue via GitHub issue tracking, update the code base with changes on a new branch, submit a "Pull Request" to one or more reviewers (co-developers on the project), merge the code base into the master branch (after code has been modified, if requested, and then approved) and finally, create a new tagged release corresponding to these changes. Where pertinent, additional detail surrounding this process will be discussed below.

Version control follows the semantic versioning schema. The standard dictates when to increment software versions and what the version format should consist of (i.e. the "major", followed by "minor", followed by patch version number). This standard has been internationally adopted and was originally authored by the co-founder of GitHub (Preston-Werner, 2018). In keeping with this standard, the major version of the AGVD application will be incremented whenever larger, backwards-incompatible changes have been made, the minor version is incremented when backwards-compatible functionality has been added, and the patch version is incremented when smaller, backwards-compatible fixes have been added to the code base. An example of a major version increment may be when upgrading the version of OpenCGA used in the backend, which may incur breaking changes to the REST API. Versioning the AGVD according to this predictable structure helps to ensure that changes which may have a critical impact on the code base or external usage thereof are easier to track.

It also aids in interoperability between different versions of the AGVD, which may conceivably coincide (during, for example, development or testing).

It is also worth highlighting that assigning of new versions to the AGVD is facilitated by the bumpversion v0.6.0 CLI tool (available: https://pypi.org/project/bumpversion/). This process is to be carried out in the developer's local development environment after obtaining the recently merged code base and by following the rules set out by the semantic versioning standard. The AGVD has been configured to integrate well with bumpversion, such that various files within the code base are automatically updated with a reference to the new version. This includes a text file, *VERSION*, at the top level of the project repo, which is a common development practice. Additionally, a new git tag is created with each new version.

The AGVD is also to be versioned according to changes applied to the data contained within the platform. However, since the data models used by the AGVD are primarily managed via OpenCGA using the MongoDB backend, the AGVD does not define policies for database migration or detailed schema information. Instead, this information can be consulted via the external OpenCGA documentation, as referenced in the AGVD code repository. In the case that the datasets themselves change (for example, as new data becomes available from external resources) or aggregation or statistical analysis steps performed on the data change, then the platform version is to be incremented by following the guidelines set out previously for the code base. For example, if the data is to be extended with additional datasets, aggregated and analysed in future iterations, the major version of the project will be incremented. The need for versioning based on changes to the data will therefore be performed mostly on an ad-hoc basis, but will be reviewed at least quarterly by the greater AGVD project team of the DR work package.

4.4.4 Containerisation and hosting

The AGVD is being hosted on internal infrastructure, as described in *4.3 Deployment and availability*. The policy also encourages web resources to be containerised where possible. The AGVD is not currently available as a single, encapsulated container due to the relative ease of installation of the web platform in its current state and also due to the project being in an early development phase as of the pilot release. However, it is recommended that the application be containerised in future as the application grows in complexity, for example as a Docker container (available: https://www.docker.com/). A possible approach to achieve this would be

to use an existing OpenCGA docker image as a base container and then extend this with the commands required to install and configure the AGVD appropriately. As per the policy, this image could then be hosted via the private H3ABioNet container registry available at: https://quay.io/organization/h3abionet_org.

4.4.5 Branding

In compliance with the *H3ABioNet Web Resource Development Policy 1.0,* the AGVD makes use of a branding scheme which is consistent with H3ABioNet. As suggested by the policy, the site includes various external hyperlinks to the H3ABioNet landing page (available: https://www.h3abionet.org/). These include a banner which is clearly visible via the landing page of the AGVD, as well as in the footer on certain pages of the site and additionally via the menu within the top navigation bar. Additionally, the H3ABioNet logo is also visible in the site footer and from the user facing documentation site, which has been previously described in *4.4.1 Documentation*.

4.4.6 Copyright and licensing

The AGVD does not infringe upon any copyrighted material or use and distribute any data or assets without accreditation of the author, where required. All images used as part of the design of the web application are freely available provided that the authors receive accreditation. The authors of such materials have been accredited appropriately in the footer of the AGVD.

4.4.7 Security

The pilot release aimed to serve primarily as a proof of concept application and database using openly accessible data and did not attempt to address major security or scalability concerns. However, certain basic security measures have already been implemented. For example, data from the AGVD is only accessible to end users via the web application and users do not have access to the backend database or server. Additionally, where datasets are not publicly accessible, access to this data within the AGVD is restricted to registered users who have been authenticated and authorised against the User catalogue in the database. Only users with the appropriate access levels as well as server administrators have the appropriate clearance level

to perform CRUD (i.e. create, read, update, delete) operations on the database. Additionally, access to the web application by end users is made via a proxy server. See *4.5.9 Authentication and authorisation* for more about the data security features of the AGVD and *4.3 Deployment and availability* for more information about the infrastructure and server configuration.

4.4.8 Data protection compliance

In compliance with the *H3ABioNet Web Resource Development Policy,* a banner disclaimer was added to the Login view describing what user data will be stored, how it will be used and why it will be used in that manner. The banner is always displayed when visiting this view and only disappears once explicitly acknowledged by the user. The type of user data that is to be stored for new users is currently limited to fields such as username, email and organisation(s)/affiliation(s). All such information is currently supplied only during manual creation of new users of the AGVD, which is performed by database administrator roles on the backend server. Additional information about users may also be stored in future, which may necessitate an update of the text in this banner.

Apart from user information, no additional types of sensitive data are stored on the AGVD, such as patient data or other kinds of identifiable sample data. However, it is likely that future versions of the AGVD may need to address additional local as well cross-border data protection concerns when storing and sharing data as both the number of users and the data stored within the AGVD increase. For example, storage of European data may require compliance with the General Data Protection Regulation (GDPR) (see: https://gdpr-info.eu/).

4.4.9 Ethical considerations

The pilot release of the AGVD contains only publicly accessible data from the 1000 Genomes Project (1kGP). It does not store or share any sensitive patient information, phenotype data, or other user identifiable information, except for cohort population and gender designations. The identities of individual samples contained within the 1kGP were anonymised as part of the original study. However, to ensure that the project meets the ethical requirements as defined in the *H3ABioNet Web Resource Development Policy*, ethics approval was obtained from the UCT Faculty of Health Sciences Human Research Ethics Committee (HREC) and granted on 2020-12-28 (HREC REF: 813/2020).

4.4.10 Testing

A simple testing framework has been added to the AGVD in order to facilitate unit testing on the functional elements of the Python Flask application. Testing is performed using the Python pytest package with tox for automated creation and destruction of isolated testing environments, including the execution of unit tests. Unit tests do not provide good coverage over all aspects of the code. It is suggested that future work includes leveraging the current testing implementation to better test all aspects of the code base. This may also include, for example, automated end-to-end integration testing of the user interface using Selenium WebDriver for Python (available via the pip package manager). Additional testing of the web application by end-users has also been performed informally during fortnightly project team meetings. However, following the deployment of the pilot version, the project team will look to formalise user testing. This process will look to include both internal as well as external users who do not form part of the core project team.

4.4.11 Release plan and updating

The AGVD is governed by a basic update schedule to ensure longevity of the project. Currently, the project is expected to receive updates on a routine, monthly basis. The scope of these updates will be dictated mostly by the request for new features or issue fixes as identified by end-user feedback and developer input, but will also largely be dependent on the number of developers actively contributing to the project at any one time. Although development activity on the project is currently limited to a small cohort of developers, the project aims to engender additional interest from stakeholders during and after the release of the pilot release (v1.0.0). Additional software developers and bioinformaticians will help to ensure faster turnaround times on product releases, such that the project may benefit from repeated iterations of development, testing and feedback. As per the policy, all new versions will be tagged and accompanying release notes describing the changes made will be published.

4.4.12 Sustainability

In accordance with the *H3ABioNet Web Resource Development Policy*, the lifespan of the AGVD is planned to extend beyond the lifespan of the current project described in this text. The pilot release was purposely designed to showcase only a proof-of-concept application with

a subset of features in order to show case utility, and never intended to offer an exhaustive feature set or to address scalability concerns. Rather, additional use cases and features are planned to be added in future iterations of the project. This is possible because the current release forms part of a larger project, "African Genome Variation Database" within the Databases and Resources (DR) Work Package (WP) of H3ABioNet. The project consists of several key collaborators within H3ABioNet who are actively involved in tasks such as data procurement, analytics and annotation, code development and deployment, as well as writing formalised SOPs to ensure that the data ingested by the AGVD is collected and shared appropriately. Individuals are also involved in securing funding at the project level and engaging with key stakeholders such as Principal Investigators (PIs) across the continent and abroad. There also exists a reasonable degree of overlap between the skillsets of these individuals which can help to ensure continuation of the project. The lifespan of the AGVD within the broader context of the DR-WP is therefore expected to continue past the lifespan of the current project.

4.4.13 Tracking usage and impact

The pilot release of the AGVD includes a logging implementation as part of the application middleware, using the Flask built-in 'logging' module. The application primarily logs HTTP requests as well as information and error messages generated from the custom middleware and the REST API that accompanies it. These messages are logged both to console, as well as a rotating file handler on an automated schedule of one roll-over log file per day, occurring at midnight (UTC time). Logs are retained indefinitely on the server and reside within the application "logs" directory.

Importantly, user searches performed via the web platform are also logged to a file for tracking user behaviour on the AGVD platform. Each time a search is performed, the original search term as well as the resulting category and formatting of the raw search term (as rendered by the application) are automatically logged to file. This is a separate file from the application log and is also available under the "logs" directory of the root application. Searches are saved as a single line per search for simplified parsing by batch scripts or automated notifications to the platform developers in future.

Although not yet extensive in the pilot version, the logging implementation can easily be expanded upon in future releases to capture more detailed user actions and include automated

error reporting. Additional types of usage tracking are also recommended to be implemented in later versions of the AGVD. It is suggested that, at a minimum, data also be collected on organic search results for the AGVD from search indexing sites, cross-linking patterns between available social media platforms and external H3Africa portals, site traffic statistics (both the amount as well as the source of visitors), time spent on a particular view or scope of the site, the proportion of new versus return visitors as well as more detailed usage information regarding user interactions with specific features of the AGVD.

4.5 Features of the AGVD

4.5.1 Modularity and integration with third-party software

The AGVD was designed to leverage existing bioinformatics solutions in a manner that does not encumber the flexibility or limit the inclusion of additional features. The design included the need for a custom client-facing web application which would integrate well with a host of third-party tools and services, such as the underlying genomics engine, OpenCGA, as well as custom routines. These individual components of the application have already been detailed in the preceding sections. However, the manner in which each of these components was integrated into a single, cohesive application is worthy of further discussion.

Notably, the AGVD was designed to decouple the primary concerns of the application (namely, the UI, data analytics and storage, and sundry features) into individual domains, with communication between being facilitated by language agnostic REST APIs. This design supports the integration of both third-party software as well as bespoke solutions which have been custom built for the AGVD. The latter was a particularly important design consideration throughout the build process, since no single comprehensive solution existed at the time of writing; this necessitated customization of the source code and resulting feature set. For instance, one may choose to leverage an existing database backend for the storage of genetic variant data but require custom data aggregations by sample type, phenotype, gender, or arbitrary grouping. This is true in the case of the AGVD, for which OpenCGA serves as the genomic data storage and analytics engine, with custom population-level grouping serviced by custom-built features. To address these needs, a layer of middleware was developed to encompass all custom features in a way that is only loosely coupled with the remaining components.

An additional advantage of this decoupled design is that it allows for the majority of the data analytics development efforts to be relinquished to third-party tools. The benefit of this is twofold; firstly, the AGVD can optionally opt-in to software updates made to each of the third-party tools used, and secondly, it allows for a great deal of flexibility in which third-party components (or versions thereof) can be ingested into the application. To expand on the latter: it is relatively straightforward for individual software constituents to be substituted with different components in future without a great deal of impact to the remaining code base, as long as they incorporate a well standardised API. It would be particularly straightforward in the case of external version updates to existing software used by the AGVD, in which case the API changes would likely be minimal in comparison to adding entirely new tools. As a probable example, if an alternative version of OpenCGA with improved population-level aggregation were to be released in future then it would be possible to upgrade the AGVD to incorporate it, so long as the client and middleware application have been updated to handle any changes to the input/output format of the data consumed.

The layout of the application also heavily emphasizes the importance of extensibility and future collaboration. This is imperative for ensuring that the project life cycle does not end with the culmination of the present iteration. From a software development standpoint, the source code of the web application has been logically separated into atomic components, as alluded to in previous sections. It is worth emphasising that each of these components can be re-used in other areas of the web application, or replaced with other components published on the web. Together, this affords future developers a great deal of simplicity and variety in how they may choose to extend the source code of the client web application. Also, since the Vue.js framework is so performant, developers should be able to add new View components in this manner with little concern about optimization.

4.5.2 Reproducible research and integration with external resources

The AGVD was designed with reproducible research in mind. To that end, the web application component was developed to make use of any HTTP query parameters provided as part of the URL address when resolving routes to their corresponding views (Berners-Lee, Fielding and Masinter, 2005). Query parameters are used by the application to ensure that a consistent set of results are returned for a given set of query parameters in the URL. Such query parameters are automatically appended to the URL by the application during searching and filtering

operations and default parameters are added where values have not been provided. This has the effect of producing an entirely descriptive set of query parameters for each query performed. See *Box 4*, below for an example of URL query generated from a search and filter operation.

Box 4: Example of reusable HTTP Query parameters used by the AGVD to route to searches for variants on chromosome 11

```
http://ROOTDOMAIN/variant?variantID=&gene=&type=&coordinates=chr11&population
FrequencyAlt=&annotated=false
```

By leveraging the standardised search methodology described, the AGVD ensures that research outcomes can be easily shared and reproduced by different researchers. Researchers are only required to share the URL address they arrived at in order for their search results to be reproduced. One additional benefit of maintaining a consistent URL query scheme is that it also enhances integration and interoperability between different online resources. This is particularly true in the case of additional H3ABioNet resources, against which the current project sought close alignment. As an example, variants in the Genomic Medicine portal (currently under development by H3ABioNet collaborators) may in future include a hyperlink to the variant as it occurs in the AGVD, and vice versa. This is only possible because of close alignment between these two projects and predictable URL query schemas. Currently however, hyperlinks to only the home pages of the Genomic Medicine portal and H3ABioNet have been added to the AGVD in order to encourage cross-site usage by prospective users.

Finally, tables, plots and sundry outputs available from the platform can be exported by interacting with their controls. Currently, this includes both the bar plot and world map plot of population alternate allele frequencies. However, future work on the project will aim to extend this functionality for all data outputs and to compile all results to a convenient report for the purposes of data provenance and results dissemination. All figures are exported in PNG format and are of a high resolution, which makes them amenable for inclusion in publications.

4.5.3 Search variants

The landing page of the AGVD serves to describe the purpose and utility of the application, as well as to enable users to search for variants (*fig. 5 A*). The latter is facilitated by a fully functional search bar, as per the outcomes of the requirements generation process. The search

bar allows users to search for variants in one of several formats: Reference Sequence (RefSeq) identifier (e.g. *rs73885319*), variant locus and allele change (e.g. *22:36661906 A>G*), Ensembl gene or transcript identifier (e.g. *ENSG00000100342*, *ENST00000397278*), gene symbol (e.g. *APOL1*), or genomic location or range (e.g. *22:36661906-36661960*).

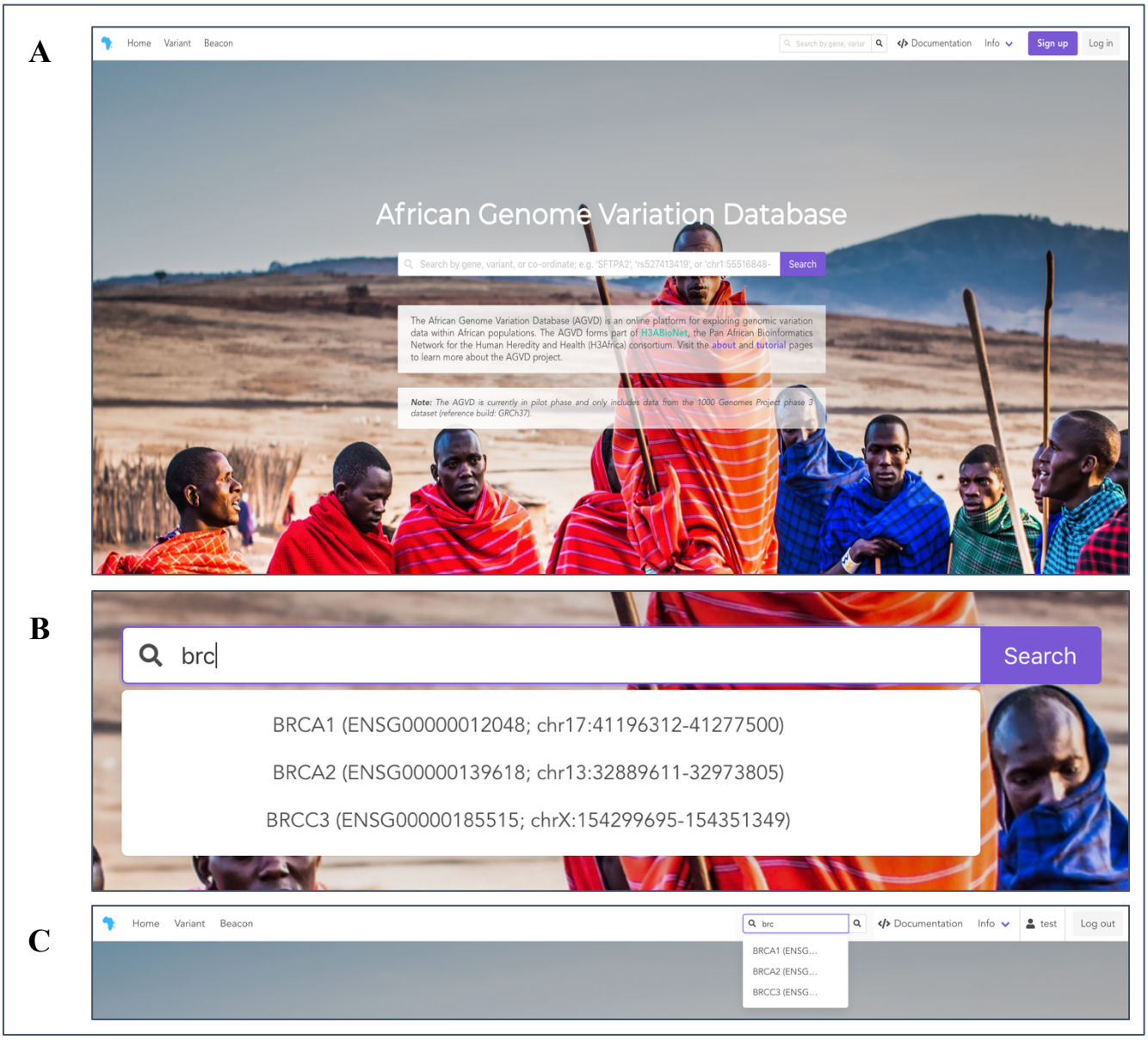

*Figure 5: Overview of the home page of the AGVD (**A**), which includes a search bar to search by gene, variant or genomic co-ordinates. The search bar also includes a "type-hint" feature with the ability to auto-complete searches (**B**), which can greatly improve the user experience. A "site-wide" search can also be performed using the search bar from the top of the navigation menu (**C**).*

The search bar also includes an additional "type-hint" feature. During typing of the search term, the AGVD will suggest a list of possible matches to the user's search, along with a set of accompanying annotations when available (*fig. 5 B, C*). This is facilitated by debounced asynchronous requests to the external CellBase lookup API, available at: http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/rest/v4/hsapiens/feature. Querying an external database for search suggestions delivers an improved user experience (UX), since not all searches are likely to be available in the database as part of the pilot release of the AGVD, which only contains the 1000 Genomes Project phase 3 dataset. This helps to give users confirmation of, or improvements to, their intended search term using a well maintained variant annotation database, which can be used in turn to search the AGVD (Bleda *et al.*, 2012).

The search bar also includes an auto-complete feature which incorporates information from the annotated search suggestions. This enables the user to choose to automatically complete their search term with the suggested term that matches most closely with it, even across all of the sundry annotation data available within the search suggestions list. Finally, the ability to search from any page within the AGVD has also been made available to users by means of a search bar within the navigation menu (*fig. 5 C*).

4.5.4 Browse variant results

The "Variant" view of the AGVD offers the majority of the functionality currently available to end-users of the platform (*fig. 6 A*). This view serves to display the results of the initial search made via either of the search bars described in *4.5.3 Search variants*, if such an action was taken by the user. Alternatively, the variant page also allows the user to browse all available variants contained within the AGVD without using specific search terms, which can be useful in instances where a more exploratory analysis approach is desired. In either case, the user also has the ability to perform various filtering operations on these results within the Variant view (filtering operations will be described in more detail in the subsequent section). Collectively, all results are displayed in the form of a "variant summary table", which was developed in accordance with the functional requirements generated during the design phase of the project (*fig. 6 B*).

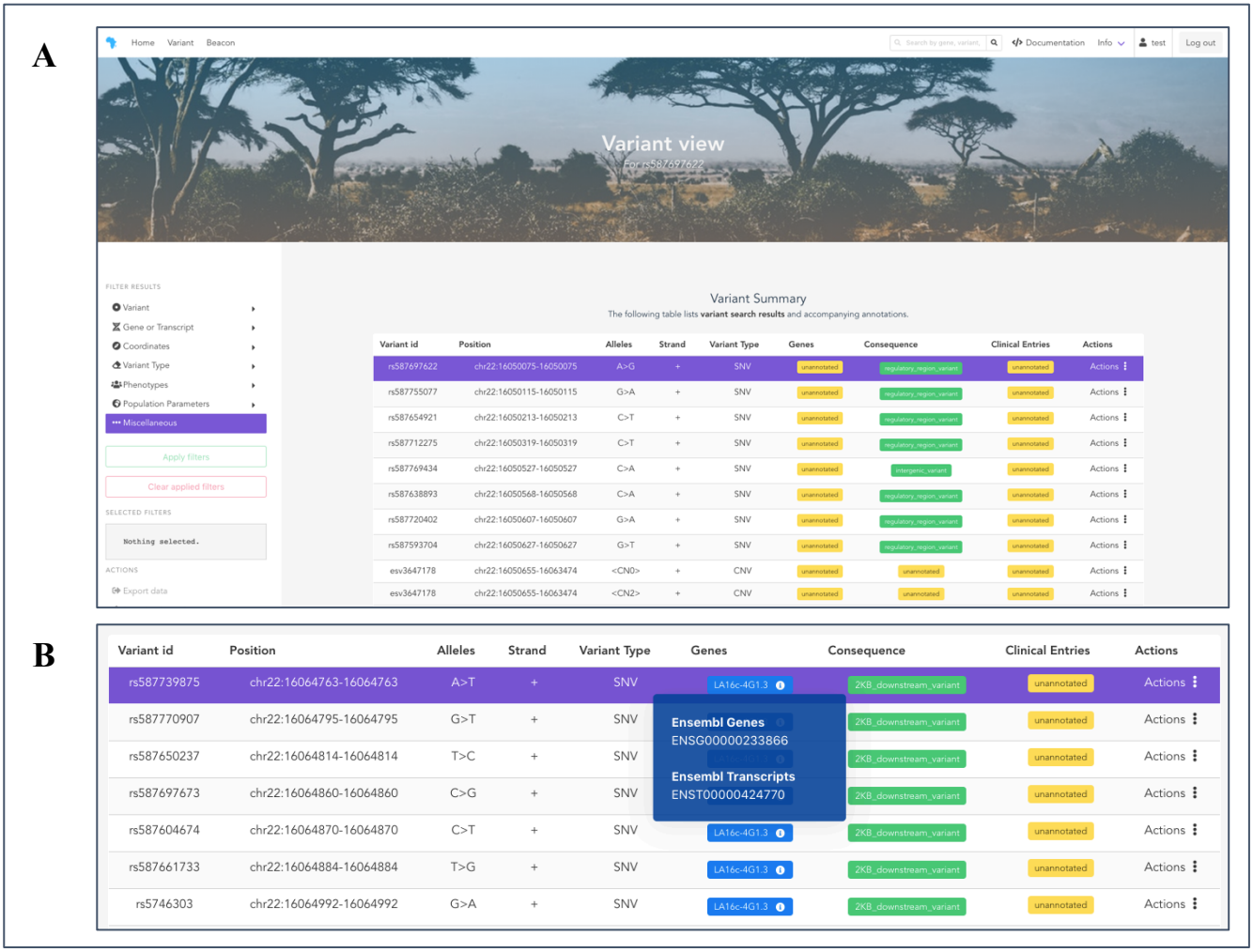

*Figure 6: Overview of the variant view of the AGVD (**A**), and variant summary table (**B**), highlighting example annotations available for the variants displayed. In this example, the variant selected is of type "SNV", with a reference SNP ID of rs587739875, and belongs to the gene LA16c-4G1.3 with an associated Ensembl gene ID ENSG00000233866 and Ensembl transcript ID ENST00000424770. Additional annotations, such as externally annotated clinical entries, as well as external actions which can be performed on the variant are also available for certain variants in the variant summary table.*

The "Variant Summary" table contains one row for each unique variant (as described by associated variant IDs, if present, or alternatively by using locus and allele information) stored in the AGVD (*fig. 6 B*). In cases where no results are available in the AGVD (for instance, when a RefSeq ID that has been searched is not yet included within the current annotation used by the AGVD) then an informative notification is displayed to the user. The following information is currently displayed within each row in the variant table, when available: the

genomic location of the nucleotide change(s) with respect to the reference genome build used as part of the study (i.e. *GRCh37* for the pilot release of the AGVD); the reference and alternate allele; strand designation; associated variant IDs as annotated from external databases, such as RefSeq IDs from dbSNP or esvIDs from DGVa (Sherry *et al.*, 2001; Lappalainen *et al.*, 2013); variant type, such as "SNV" or "CNV"; gene and transcript annotations from Ensembl (Yates *et al.*, 2020); Sequence Ontology (SO) consequence type, such as "intergenic variant" (Eilbeck *et al.*, 2005); available clinical annotations, such as ClinVar and COSMIC accessions (if present) (Landrum *et al.*, 2018); as well as a list of internal and external actions that can be performed on the variant entry, such as performing a Beacon query within the AGVD (see *4.5.8 Query Beacons*) or external query via an appropriate H3ABioNet web portal (once integrated in future releases; see *4.5.2 Reproducible research and integration with external resources*). See *Figure 6 B* for an overview of the main functionality provided by the variant table.

4.5.5 Diverse and flexible variant annotations

The AGVD supports a range of different variant annotations, some of which have already been outlined. In some cases, these have been derived from the original VCF files which store the variants and are loaded into the OpenCGA genomics backend during the data ingestion phase, as described in previous sections. Examples of such annotations include variant IDs, which occur as part of the original 1kGP phase 3 genotype files loaded in the database, or otherwise may be annotated using third party tools (such as the "annotate" command from the bcftools package; available from: http://samtools.github.io/bcftools/howtos/install.html) for VCF files which do not already have these annotations present. Similarly, genomic locus, alleles and strand information may all optionally be consumed by OpenCGA using pre-existing fields available in the VCF files loaded. This is true for any additional "INFO" fields present in the VCF files, as OpenCGA is able to utilize pre-existing annotations while annotating variants in the database.

Optionally, these annotations may instead be added retroactively after file ingestion as part of the annotation process conducted by OpenCGA. This can be performed on the CLI or via the REST API by supplying either a custom annotation file or by using the CellBase variant API, as has been performed during the data annotation phase of this project. These various

annotation features available from OpenCGA afford a great deal of flexibility during the data curation process of the AGVD.

For the pilot phase of the AGVD, all additional annotations displayed in the variant table were generated using the CellBase REST API annotator (available: http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/rest/v4/hsapiens) during the annotation phase (see: *4.2.4 Annotating variants*) and were not curated by the authors of the current text (Bleda *et al.*, 2012). This includes the following columns in the table: "Genes", "Consequence" and "Clinical Entries". Gene and transcript annotations derive from the Ensembl database, and are currently limited to the gene symbol, gene ID and associated transcript IDs (Yates *et al.*, 2020). Additional information such as the location and function of the gene are also available but have not yet been incorporated in the current release of the AGVD. Detailed gene and transcript annotations are visible upon mouse hover over the gene name within the "Genes" column (*fig. 6 B*). Sequence Ontologies (SO), which describe the effect of genetic variation in the context of the genetic sequence and function, are derived from Ensembl and InterPro, using ontologies imported from OBO Foundry (Eilbeck *et al.*, 2005; Smith *et al.*, 2007; Bleda *et al.*, 2012; Hunter *et al.*, 2012). Additional variant annotations from either clinically relevant or otherwise prominent databases are available under the "Clinical significance" column of the variant summary table. Such annotations were compiled by CellBase using dbSNP, ClinVar, COSMIC, NHGRI-EBI GWAS Catalog, HGMD, OMIM, UniProt, OpenAccess GWAS Database and Ensembl as annotation sources (Bleda *et al.*, 2012).

4.5.6 Population level variant annotations

Additional annotations are also available in the AGVD which are not displayed in the variant summary table. In the current release, this includes externally annotated population allele frequencies. For each variant in the database, the OpenCGA annotator attempts to perform a lookup of pre-computed allele frequencies generated from external studies, for both the reference and alternate allele. The version of OpenCGA used as part of the current release of the AGVD includes the following external studies during the annotation process: 1000 Genomes Project ("1kG_phase3"), HapMap ("HAPMAP"), Ensembl ("ENSEMBL"), the Exome Aggregation Consortium ("EXAC") and the gnomAD genomes and exomes datasets ("GNOMAD_GENOMES", "GNOMAD_EXOMES"), NHLBI Exome Sequencing Project Exome Variant Server ("ESP6500"), Genome of the Netherlands ("GoNL"), the Human

Metabolome Gene/Protein Database (“MGP”), DiscovEHR (“DISCOVEHR”) and UK10K (“UK10K”). Each study is available as an independent tab, and can be used to conveniently compare AFs between different studies (*fig. 7 A*) (Bleda *et al.*, 2012).

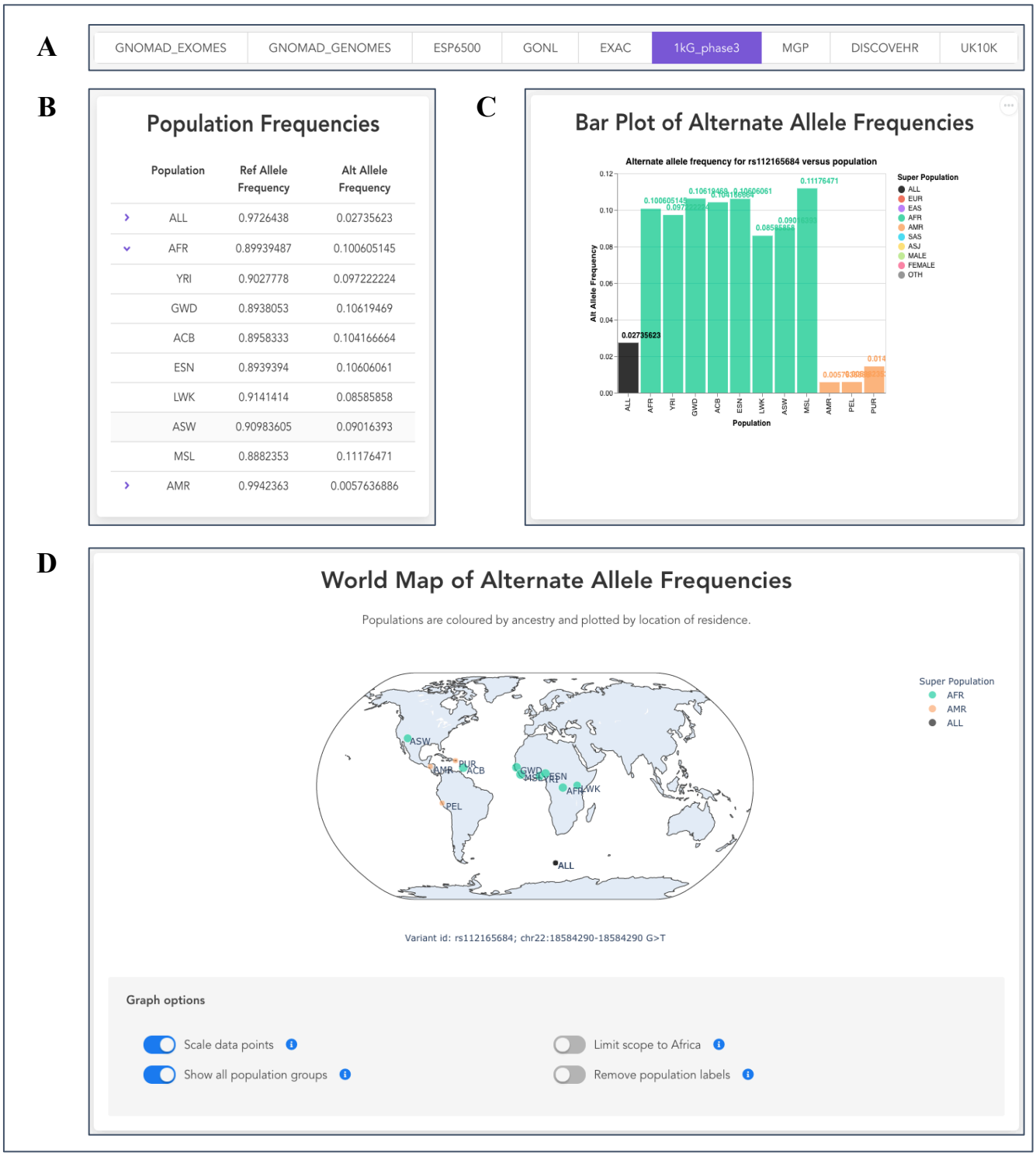

*Figure 7: Different representations of reference and alternate allele frequency (AAF) across different populations and studies annotated in the AGVD. Users can switch between the available studies that were included as part of the annotation phase by navigating between studies tabs (**A**). The reference and alternate allele for each population grouping is displayed in an interactive table where super population rows can be expanded for a more stratified view of populations (**B**). The bar plot representation is grouped and coloured by ancestry, and is useful for visually comparing between populations (**C**). The world map represents populations at different geographical locations, where the area of each data point corresponds to the AAF for a given population and is accompanied by several interactive graphing options (**D**).*

Alternate allele frequencies (AAFs) are displayed in three different formats for each study: both a table and bar plot of AAF across the available populations included in the selected study, as well as a world map representation. The "Population Frequencies" table is useful for displaying AAF grouped by super population, when present, and is more amenable to simple copy and paste operations by the user (*fig. 7 B*). Also included in the table are the reference allele frequencies for each population grouping. The bar plot representation includes the same AAF data as the table but is plotted as a bar graph for easier comparison between populations. Bars are coloured, as well as grouped, according to super population listed, which further improves the user's ability to visually inspect differences in AAF between populations (*fig. 7 C*).

Additionally, a separate tab "AGVD" is also displayed as part of the studies tab navigation for certain variants for which this annotation is available. This study reports the pre-calculated AF values for custom cohorts defined as part of the statistics calculation process (see *4.2.6 Defining custom cohorts*). The "AGVD" study allows users to view frequencies of variants included in the data build of the AGVD and to compare these frequencies against the other available studies, such as gnomAD and 1kGP (note that currently only 1kGP data is included in the pilot release, so annotations will be identical between the 1kGP and AGVD study annotations). This may be useful to researchers in future once additional data have been added to the AGVD. This feature also grants the AGVD the ability to define and annotate variants for deeply stratified populations, which may be useful in the African context where there exists a great deal of population substructure. This feature was not initially planned as part of the pilot release and as such, is currently only available for certain variants of chromosome 1 as a proof of concept for broader implementation across integrated data sets in future.

The world map representation provides a more interactive experience for the user to compare AAF between different populations (*fig. 7 D*). This component displays an orthographic projection of the Earth with data points corresponding to population AAF. Data points are coloured by ancestry and plotted by proxied location of residence. Location of residence serves as an estimate of the approximate location where the individuals sampled currently reside, and was manually curated as part of the data curation process (see *4.2.1 Metadata curation*).

The default mode renders the area of each data point proportionally to the absolute frequency (i.e. within the limits of 0-1) of the AAF in the given population. This is useful for more accurately comparing the AAF between different graphs, which have either been rendered from different studies for a given variant, or from entirely different variants. However, data points

can optionally represent frequencies that have been scaled relative only to other data points visible on the plot. This latter view can be more helpful in instances where rare alleles (AAF < 0.01) are directly compared in order to magnify small differences in frequency. At the time of writing, additional plotting options include: toggling the display of data labels; toggling the display of only the African continent; and whether or not to include super population and miscellaneous groupings in the plot which cannot be reasonably mapped to any single geographical origin, such as the "OTH" ("other") gnomAD population designation.

The plots described above are highly interactive, which can improve user engagement with the data during exploratory data analysis. The bar plot interactivity toolkit includes the ability to change the scale of the axes as well as to gather additional information about each data point upon mouse hover. In addition to the interactive graphing options, the world map can also be panned and zoomed, and populations can be de-selected from the legend if not desired by the user. Copies of all graphs generated by the AGVD can be downloaded by the user by using each component's respective interactive hover menus.

4.5.7 Filter variants

In addition to providing search and browse functionality, the AGVD also facilitates filtering variant results by several different parameters. Filter parameters represent a superset of the available search types that can be queried via the search bar, with some additional filter types also available. These are made accessible by interacting with the accordion menu of the filter panel. Available filter menus in the pilot release include "Variant", "Gene or Transcript", "Coordinates", "Variant Type", "Population Parameters" and "Miscellaneous". Each filter menu, once expanded, includes explanatory text to guide end-users about acceptable types of parameters that can be entered, the correct format to enter them, as well as examples of valid search terms for each type of filter parameter. Figure 8 A demonstrates an example of a complex filtering operation in which multiple filters are applied at once.

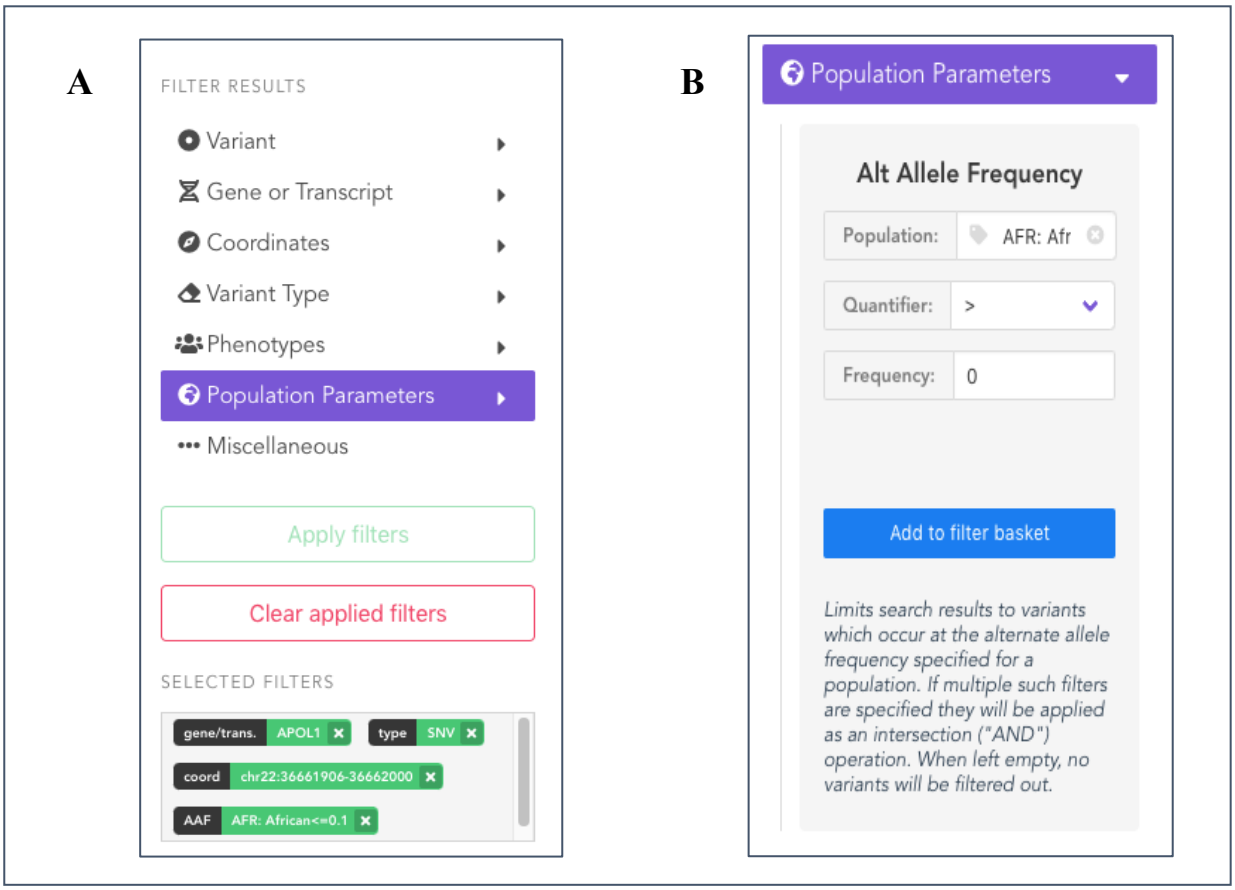

*Figure 8: Filter box available from the Variant view of the AGVD. Various filters can be applied to limit the variants displayed in the variant results table, and are added as keyword tags to a filter basket (**A**). Users can also perform advanced filtering of variants based on their allele frequency within a specified population (**B**).*

The "Variant" filter menu allows users to filter by external variant identifier (e.g. RefSeq ID) as well as a specific locus and allele change (e.g. '22:36661906 A>G'). The "Gene or Transcript" menu allows users to enter one or more gene symbols (e.g. APOL1) and Ensembl gene identifiers (e.g. ENSG00000100342) as well as Ensembl transcript identifiers (e.g. ENST00000397278). The "Coordinates" menu provides the ability to filter by either a specific genomic location or coordinate range (e.g. 22:36649056-36663576). All of the above filters are applied together as a union operation, rather than an intersection. This may be useful in generating a single list of results across various different regions, variant identifiers and locations, for example. The "Variant Type" menu consists of a dropdown multiple selection box consisting of variant types that users can filter by (e.g. SNV). Currently, users can also perform limited filtering of only annotated variants, however, this does not exclude variants for which certain annotations are available in the database but not rendered in the variant results

table. This operation, as well as future sundry filtering operations, can be accessed via the "Miscellaneous" menu.

The "Population Parameters" menu may be particularly useful in the study of rare diseases. This menu can be used to filter variants by alternate allele frequency for a given cohort or population (*fig. 8 B*). This filter consists of three input fields: "Population" (e.g. Yoruba), "Quantifier" (i.e. ">" or "<=") and "Frequency" (e.g. 0.01). The combination of these fields is applied as a single filter against the results in order to return only variants which occur at the specified frequency threshold. The available list of populations which can be selected are generated from a compiled list of cohorts from each study annotation included in the AGVD (see *4.2.1 Metadata curation*). This list is made available as a search input for the convenience of the end user, such that only the available listed populations can be selected from, but can be searched for upon user input. Input validation is also performed on entered combinations to prevent mutually exclusive filters from being applied at once.

4.5.8 Query Beacons

The Beacon protocol (see: https://beacon-project.io/) is a project of the Global Alliance for Genomics and Health (GA4GH) and was developed in order to facilitate and encourage the sharing of genomic data. The protocol serves as a framework for individuals and organisations to search other organisations for the presence of a particular variant within their databases. Information from individuals is de-identified and only aggregate or quantitative information about the variant is returned, thereby protecting the privacy of individuals. The Beacon network comprises multiple Beacons from various different institutions, which are all available for querying via a well-defined API – the Beacon API. Each Beacon is a simple web service that is able to respond to an HTTP request which includes the genome build, chromosome, location and allele of interest. The web service will typically respond to the request with either a "yes" or "no" to indicate whether the variant is present within their own database. Additional information about the variant may also be included as part of the response, for example allele frequency information or variant pathogenicity scores (Fiume *et al.*, 2019).

The AGVD leverages the Beacon protocol to enhance the variant search and discovery feature set available to users of the platform. The "Beacon" view of the AGVD facilitates Beacon querying on a particular variant via the available search bar. A search bar accepts as input the genome reference build, chromosome, location and allele of interest to the user (*fig. 9 A*). This

information is then compiled to batches of asynchronous requests by the AGVD which are queried across the Beacon Network (available: www.beacon-network.org). The results of the Beacon query are displayed as a list of Beacon services, each reporting a status of either "Present" or "Not Present". Furthermore, a button "Additional Info" will be displayed below the "Present" status in cases where the Beacon returned supplementary information about the variant. This button can be clicked to collapse or expand a detailed table describing this information as key-value pairs (*fig. 9 C*). Results can also optionally be filtered by organisation or status using the filter panel on the left of the results panel (*fig. 9 B*).

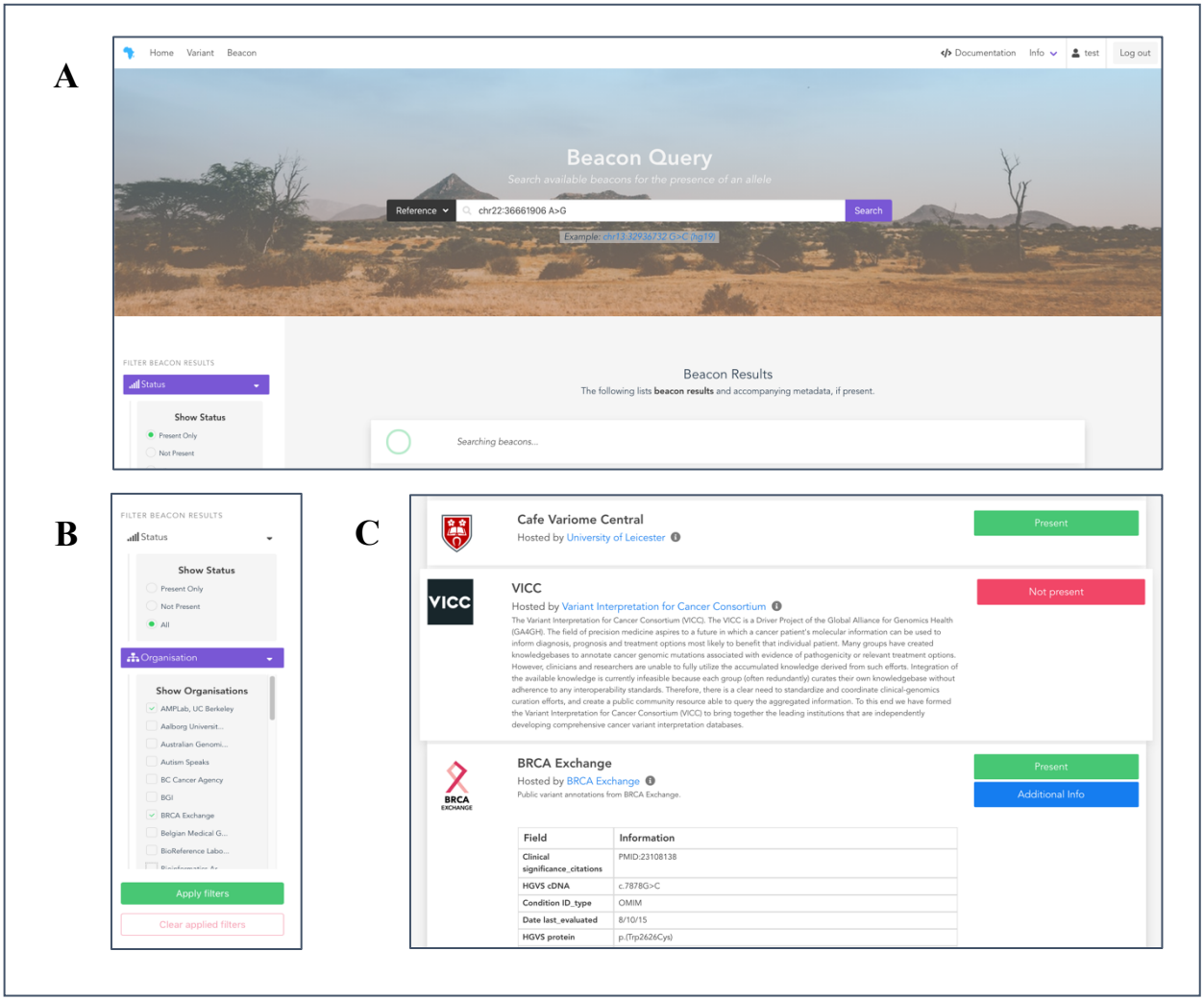

*Figure 9: Beacon view of the AGVD. Users can perform a query across the Beacon Network for a particular variant by supplying a genome reference build, chromosome, location and allele information into the search bar within the Beacon view (**A**). Users can optionally filter Beacons by status or organisation (**B**). The results of the Beacon query are displayed prominently as one row per Beacon, alongside the status information for each Beacon (i.e. "Present" or "Not Present" within the database of that Beacon) and additional information about the variant when available (**C**).*

4.5.9 Authentication and authorisation

The pilot release of the AGVD includes proof of concept user authentication and authorisation capabilities as a means to secure the data stored within the database. Even though the 1kGP data used in the pilot release is an open access data set without restriction, this level of security will be necessary for future releases of the AGVD which are likely to store data from various contributors and across various levels of sensitivity. Authentication and authorisation are currently facilitated by the OpenCGA "User" catalogue and the configuration thereof is described in *4.3.2 Application configuration.* Authorisation can be restricted at the level of the "Study" as well as "Dataset" within OpenCGA using the OpenCGA tool suite. Future versions of the AGVD will likely integrate with third-party institutional access for finer grained control of access permissions, such as ELIXIR AAI (Linden *et al.*, 2018).

Currently, the pilot release requires users to be authenticated and authorised in order to access the Variant view and perform queries on the variant data stored in the database. Users can be authenticated by entering a user name and password combination from within the Login view. An example user account with an unrestricted access level to the pilot data set can be used by test users during the pilot phase. The account details for this user account are listed in a banner within the Login view for ease of testing. In the pilot release no option exists for end-users to register themselves via the AGVD and access is strictly controlled by system administrators on the backend.

4.5.10 User profile and data monitoring

A complementary feature to the user authentication capabilities of the AGVD is the user profile dashboard (*fig. 10*). The dashboard is available from the Profile view where the user is routed automatically upon successful login. The dashboard currently offers a limited feature set in the pilot release but has the potential to include a number of additional metrics that may be useful for monitoring the status of the data that a given user has access to as well as managing database catalogues. For example, a useful metric to include would be the total numbers of variants available within the database by cohort, variant type, or frequency threshold. Similarly, summary information about the quantity of data currently available across different file types or sequencing methods (for example, micro-array vs WGS data) would be useful for data monitoring purposes and to help pre-empt scalability concerns.

Currently, the dashboard reports on the total number of variants loaded in the AGVD for which the authenticated user has access. Additionally, the dashboard reports the number of samples loaded in the AGVD for which the user has access. Importantly, information about the user is also displayed, such as a registered email address and institution if supplied during the user registration process. User information also includes a report of the projects and studies that the user has been assigned to and can therefore access data from. Notably, status information about the platform is also reported, which can be useful for notifying users when components of the platform are unavailable (for example, when performing maintenance on components of the application). This currently includes a live status of the genomic storage and analysis engine used by the AGVD – OpenCGA – as well as the REST API. Both of these applications are also annotated by the name, description and version of each program, which may be useful for reproducibility.

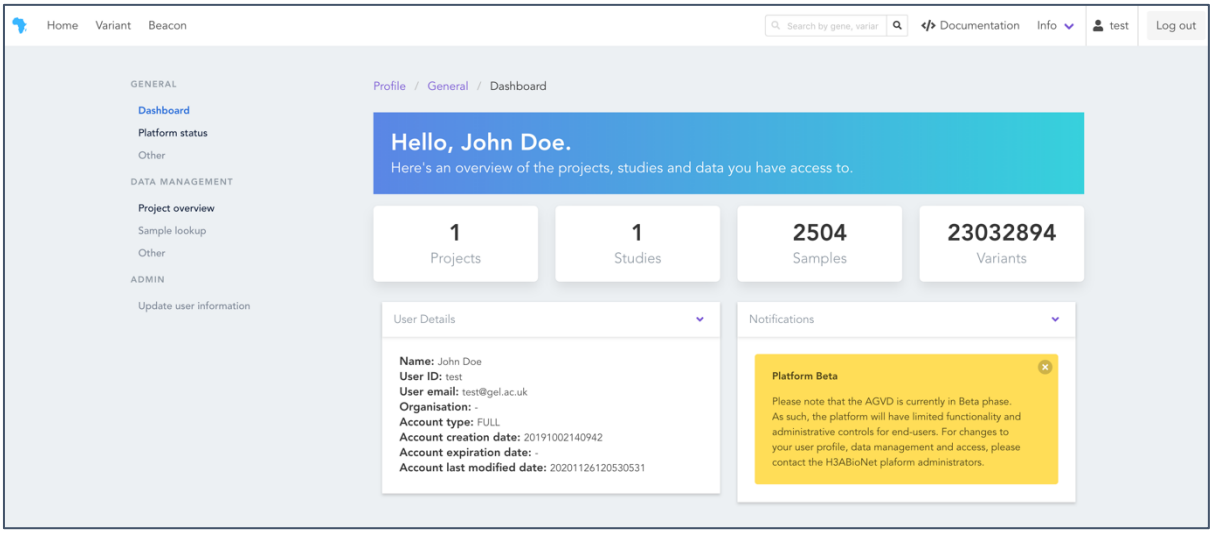

Figure 10: Overview of the Profile view of the AGVD, which includes a user dashboard with various metrics describing the current state of data stored within the AGVD as well as user and platform status information.

5. Data Validation and Demonstration of Potential Research Utility of the AGVD

Following the research, design and development phases of the AGVD, we then sought to demonstrate proof of concept for the pilot release. This process included three primary objectives: to validate that the platform behaviour was in alignment with design expectations, to validate the accuracy of reported variant annotations, as well as to demonstrate potential utility of the AGVD within a health research context. Three User Scenarios which were generated as part of the application design phase were used as a means to assess these objectives (see: *3. Developing User Requirements*, *Box 3)*. In alignment with these USs, two diseases with particular relevance to African ancestries were selected for validation and utility purposes – malaria and HAT. Variants which have been reported as protective against these diseases within African ancestries were used to perform queries against the AGVD and to cross reference these results against both current literature as well as external clinical and genome databases, such as ClinVar and gnomAD.

5.1 Malaria

5.1.1 The DARC locus

The Duffy blood group describes a group of antigens on the surface of red blood cells (RBCs) which are receptors for certain chemokines produced during an inflammatory response. The Duffy Antigen/Chemokine Receptor (DARC) is also a receptor for *P. vivax* and individuals who do not express the glycoprotein on their RBCs are resistant to RBC invasion by *P. vivax* (Dean, 2005; Hedrick, 2011). The Duffy null phenotype describes the absence of the FY-a and FY-b Duffy antigens and may be caused by several variants of DARC previously reported in the literature (McManus *et al.*, 2017). With this in mind, and in alignment with *US #9*, we set out to search the AGVD for variants of the DARC locus in an effort to explore annotated variants with known clinical significance to malaria.

Shortlisting clinically relevant variants based on their annotations in the AGVD

From the "variant" page of the AGVD, the filter term "DARC" was added to the "Gene or Transcript" filter box with no additional parameters and the filter was applied to generate a list of results. The variant summary table contained a total of 52 variants as part of the search

results, as determined by manual counting and using the table pagination options to navigate between results. The total number of search results is not currently reported as part of the AGVD pilot release due to limitations in the OpenCGA backend implementation. However, this feature will form part of the future work on this project. Of the variants reported in the search results, 10 were reported to have clinical annotations, as per the "Clinical Entries" column. Three variants had accessions in ClinVar: rs2814778 (with ClinVar accessions: *18395*, *RCV000000006*, *RCV000000007* and *RCV000000008*), rs12075 (with ClinVar accessions: *17728* and *RCV000000005*) and rs34599082 (with ClinVar accessions: *18396* and *RCV000000009*).

The AGVD reports accurate clinical annotations for variants in the database

From the variant summary table, rs2814778 is a Thymine to Cytosine SNV located at position chr1:159174683 (against the GRCh37 reference set used in the AGVD pilot release). It has a reported consequence type of "5 prime UTR variant", although three different genes have been listed with consequences: CADM3, CTA-134P22.2 and DARC. This ambiguity highlights a weakness in the current way that SO terms are reported by the AGVD and future work will look to improve on reporting specific gene and transcript consequences. However, using the available REST API, a further search query was performed to quickly resolve the labelled SO consequence to the DARC gene (Ensembl ID: ENSG00000213088), within two transcripts: ENST00000537147 and ENST00000368122. Using the AGVD reported ClinVar accession numbers to perform a search in ClinVar, the above annotations could be verified against the annotations reported by ClinVar for rs2814778 (*Table 4*) (ClinVar, 2019a, 2019d, 2019b, 2019c).

Table 4: Validation of annotations available for rs2814778 from the AGVD across external sources. ACKR1: Atypical Chemokine Receptor 1; AGVD: African Genome Variation Database; CADM3: Cell Adhesion Molecule 3; DARC: Duffy Antigen/Chemokine Receptor; NDL: not directly listed in annotation source; SO: Sequence Ontology.

Accession Number	Annotation Source	Genomic Location (hg19)	Variant Type	Allele Change	Gene	Transcript (Ensembl ID/NCBI RefSeq)	Consequence	SO Accession
-	AGVD	chr1:159174683	SNV	T>C	CADM3	ENST00000368124	Downstream variant	SO:0001632
						ENST00000368125	2KB downstream variant	SO:0002083
						ENST00000497636	Downstream variant	SO:0001632
					CTA-134P22.2	ENST00000415675	Upstream variant	SO:0001631
						ENST00000609696	Non-coding transcript variant	SO:0001619
							Intron variant	SO:0001627
					DARC	ENST00000537147	5 prime UTR variant	SO:0001623
						ENST00000368122	5 prime UTR variant	SO:0001623
						ENST00000435307	2KB upstream variant	SO:0001636

						ENST00000368121	2KB upstream variant	SO:0001636
18395	ClinVar	1: 159174683	SNV	T>C	ACKR1	NM_002036.4	5 prime UTR variant	NDL
						NM_001122951.3	5 prime UTR variant	NDL
RCV000000006	ClinVar	1: 159174683	SNV	T>C	ACKR1	NM_002036.4	5 prime UTR variant	SO:0001623
						NM_001122951.3	5 prime UTR variant	SO:0001623
RCV000000007	ClinVar	1: 159174683	SNV	T>C	ACKR1	NM_002036.4	5 prime UTR variant	SO:0001623
						NM_001122951.3	5 prime UTR variant	SO:0001623
RCV000000008	ClinVar	1: 159174683	SNV	T>C	ACKR1	NM_002036.4	5 prime UTR variant	SO:0001623
						NM_001122951.3	5 prime UTR variant	SO:0001623

It is worth noting that the gene symbol "ACKR1", rather than "DARC", was annotated in all ClinVar accessions used during validation. The discrepancy between the gene symbols listed in the AGVD and ClinVar was not altogether unexpected since the associated gene, locus and alleles have historically held several overlapping designations in common nomenclature. Most commonly, gene symbols include ACKR1, DARC, FY and CD234 (Höher, Fiegenbaum and Almeida, 2018). ACKR1 is the currently approved gene symbol from HGNC, replacing DARC and FY, and is used by RefSeq (NCBI) and GenCode (Ensembl) (HGNC, 2019; Ensembl, 2020; NCBI, 2020). Nonetheless, all symbols have been used in contemporary studies (McManus *et al.*, 2017; Höher, Fiegenbaum and Almeida, 2018; Golassa *et al.*, 2020). With this in mind, we performed a separate search for "ACKR1", "CD234" and "FY" from the AGVD in an effort to test the robustness of the search and annotation features of the platform. None of the additional searches returned any search results, which highlights a need for improved annotations which are inclusive of various gene, locus and allele symbols and common aliases. Such limitations will be addressed in future work.

5.1.2 The HbS Allele

Although several malaria protective HBB variants have previously been identified, relatively few have consistently been replicated by GWAS (Damena *et al.*, 2019). The HbS allele, rs334, is one example of a well described HBB variant which has demonstrated strong association with resistance to severe *P. falciparum* malaria and was therefore selected as the primary focus of this investigation (Aidoo *et al.*, 2002; Ghansah *et al.*, 2012). We sought to validate the data querying functionality of the AGVD by performing a search and filter operation for rs334 and related HBB variants, in alignment with *US #4*. Additionally, the AGVD was used to generate a table of alternate allele frequencies for these variants and examine differences between AAF values for populations living in malaria endemic versus non-malaria endemic regions. Furthermore, these findings were then compared against known AAF values reported in the literature.

The AGVD facilitates accurate variant querying and annotation

With the above aims in mind, we first navigated to the home page of the AGVD and entered the term "rs334" into the search bar in order to perform a search for a characterised variant. A single result was returned and prominently displayed in the Variant Summary table for further exploration (*fig. 11 A*). Identical results were also displayed when using the alternate search bar from within the navigation menu. The variant identifier for the single result matched the expected search term, "rs334". Furthermore, the reported location of this variant was position 5248232 on chromosome 11 (reference build: GRCh37) and a Thymine to Adenine substitution was reported, indicating a variant of type SNV at this location. The above annotations were cross-referenced against gnomAD and their accuracy confirmed, which was expected since these annotations were derived directly from the 1kGP input files added to the AGVD (*fig. 11 B*) (Karczewski *et al.*, 2020).

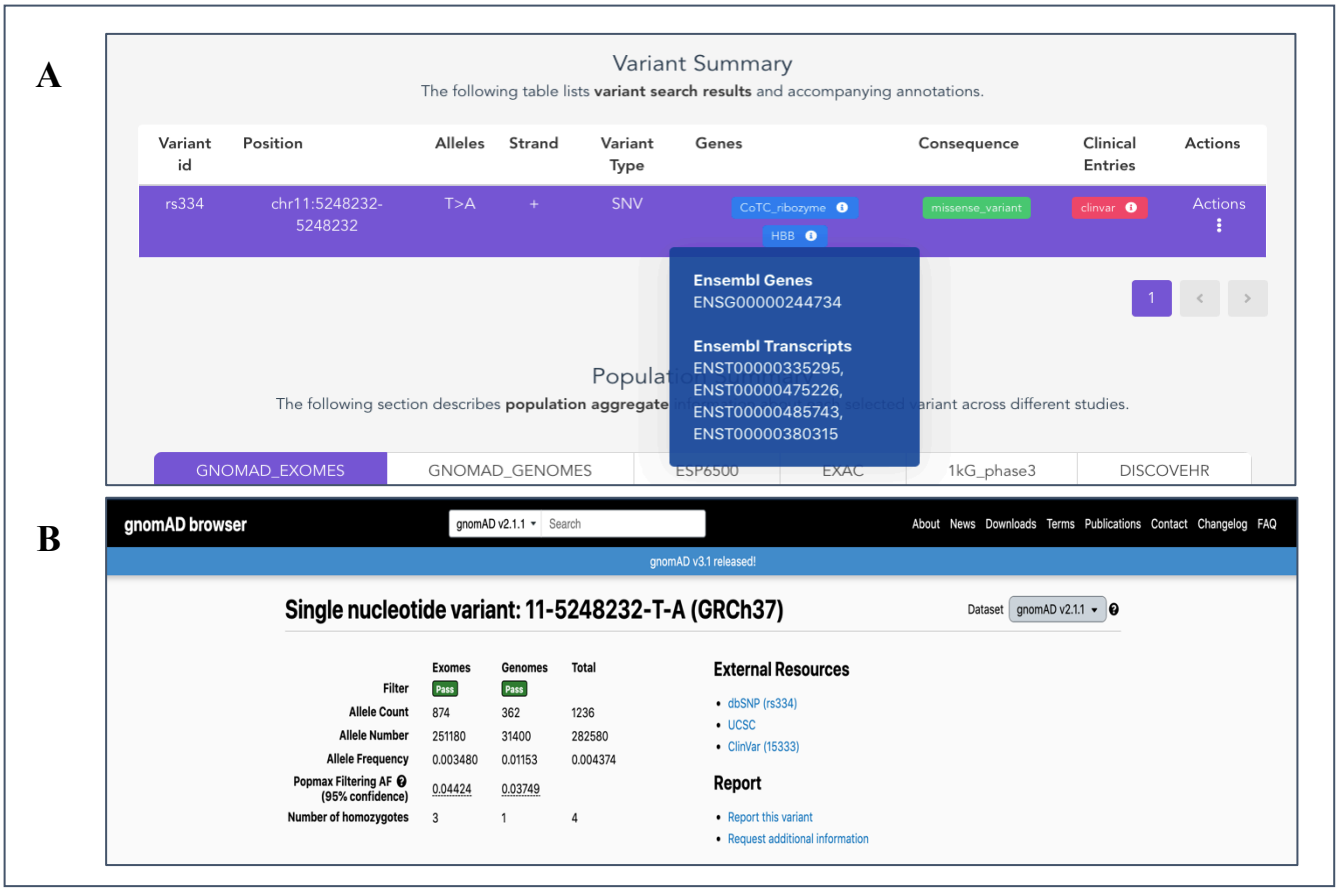

*Figure 11: Available variants and associated variant level annotations following a search for the term "rs334" from the home page of (**A**) the AGVD, as displayed via the Variant Summary table, and (**B**) the gnomAD database. The variant RefSeq identifier (rs334), chromosomal location (5248232 on chromosome 11), allele change (T-A) and variant type (SNV) reported were identical between databases. AGVD: African Genome Variation Database; gnomAD: Genome Aggregation Database.*

Collectively, these results indicate that the data search functionality afforded by AGVD offers a suitable degree of accuracy within this context under initial test conditions. A lack of any unintended, false positive search results also suggests that the data querying functionality available from the AGVD offers a high degree of precision under test conditions. The same approach was also applied to two additional HBB malaria protective alleles: HbC (rs33930165) and HbE (rs33950507), with similar results (Edison *et al.*, 2005; Ghansah *et al.*, 2012; Ha *et al.*, 2019). Although the test dataset for the AGVD pilot phase was intended to be representative of large numbers of variants across a diverse range of genomes, it remains possible that the accuracy and precision of the search functionality may not be as robust for larger datasets, particularly when partial string matches may occur. Future work should therefore include additional validation across a diverse range of datasets, studies and degrees of annotation.

The AGVD reports common frequencies for malaria protective alleles in African ancestries

Next, population allele frequencies of rs334 were compared against previously reported allele frequencies in the literature, as outlined in *1.3.1 Communicable diseases*. From the initial search results for "rs334", the "Population Frequencies Table" and "Bar Plot of Alternate Allele Frequencies" displayed the alternate allele frequencies for the HbS allele across different populations for various annotated studies (*Table 5* and *fig. 12,* respectively). The "1kG_phase3" study annotation was used as part of the downstream comparison.

From the results, the aggregated AAF across the African super population (AFR) was approximately 10%, the AAF was approximately 0.7% for the Admixed American (AMR) super population, while the alternate allele was not reported for the remaining populations included as part of the 1kGP. The highest AAF reported was 13.89% for the Yoruba of Nigeria (YRI) and greater than 10% AAF was reported for an additional four populations (in descending order): Mende in Sierra Leone (MSL; AAF 12.35%); Esan in Nigeria (ESN; AAF 12.12%); Gambian in Western Divisions in the Gambia (GWD; AAF 11.50%) and Luhya in Webuye, Kenya (LWK; AAF 10.10%). The two AMR populations which observed the allele were the Colombians from Medellin, Colombia (CLM; AAF 1.06%) and the Puerto Ricans from Puerto Rico (PUR; AAF 1.44%).

Table 5: Population allele frequencies for rs334 for the 1000 Genomes Project Phase 3 study annotation of the AGVD. ACB: African Caribbeans in Barbados; AFR: African; AGVD: African Genome Variation Database; AMR: Admixed American; ASW: Americans of African Ancestry in SW USA; CLM: Colombians from Medellin, Colombia; ESN: Esan in Nigeria; GWD: Gambian in Western Divisions in the Gambia; LWK: Luhya in Webuye, Kenya; MSL: Mende in Sierra Leone; PUR: Puerto Ricans from Puerto Rico; YRI: Yoruba in Ibadan, Nigeria.

Super population	Population	Reference allele frequency	Alternate allele frequency
AFR	-	0.9001513	0.09984872
	ACB	0.953125	0.046875
	ASW	0.9836066	0.016393442
	ESN	0.8787879	0.121212125
	GWD	0.88495576	0.11504425
	LWK	0.8989899	0.1010101
	MSL	0.87647057	0.12352941
	YRI	0.8611111	0.1388889
AMR	-	0.9927954	0.007204611
	CLM	0.9893617	0.010638298
	PUR	0.9855769	0.014423077

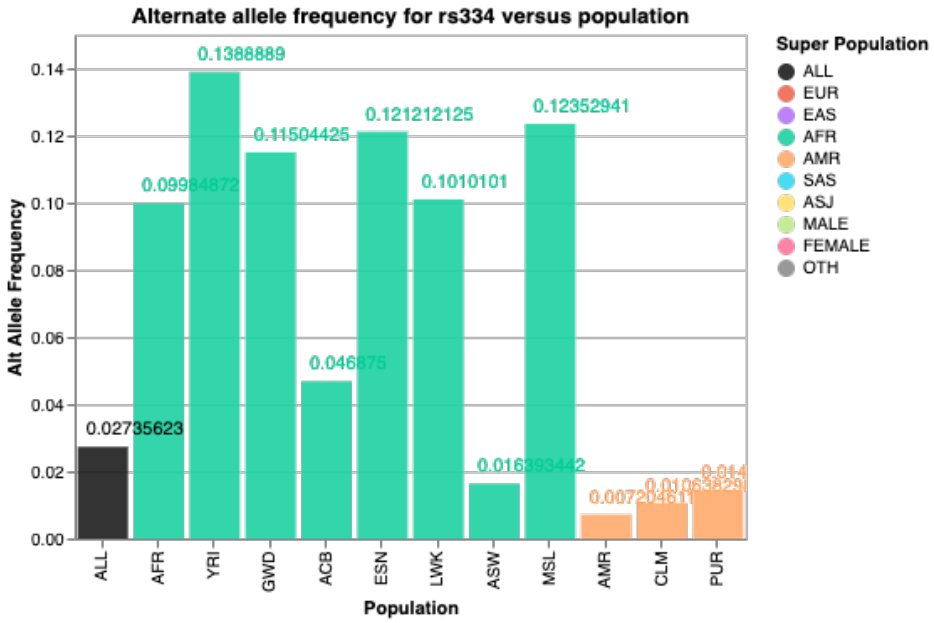

Figure 12: Bar plot representation of population allele frequencies for rs334 for the 1000 Genomes Project phase 3 study annotation of the AGVD. ACB: African Caribbeans in Barbados; AFR: African; AGVD: African Genome Variation Database; AMR: Admixed American; ASW: Americans of African Ancestry in SW USA; CLM: Colombians from Medellin, Colombia; ESN: Esan in Nigeria; GWD: Gambian in Western Divisions in the Gambia; LWK: Luhya in Webuye, Kenya; MSL: Mende in Sierra Leone; PUR: Puerto Ricans from Puerto Rico; YRI: Yoruba in Ibadan, Nigeria.

These findings broadly agree with the literature, in which studies have demonstrated an HbS AF of greater than 0.5% throughout most of SSA and up to 20% in many parts of SSA, with lower frequencies reported outside of malaria-endemic areas (Piel *et al.*, 2010; Hedrick, 2011). Additionally, Piel *et al.* predict greater than 10% AAF within most regions across Senegal to northern Angola; this is in close alignment with the values reported for GWD, MSL, YRI and ESN by the AGVD. Notably however, the allele was not observed in central and Latin America in the study and was observed in parts of the Mediterranean, Middle East and India at low AFs (approximately < 10%), which contrasts with the results from the AGVD. These regions were historically hypoendemic, mesoendemic or hyperendemic for pre-intervention malaria but the association between malaria endemicity and HbS was highest in Africa and limited in Asia. Additionally, the association was most significant for holoendemic regions, which were limited mostly to Africa (Piel *et al.*, 2010).

Populations of recent African ancestry living outside of Africa include ACB and ASW, for which the AGVD reported AAFs (4.69% and 1.64%, respectively) that were significantly lower than continental African populations and altogether absent from Piel *et al.* (although the latter may possibly be ascribed to differences between sampling methods, including the level of stratification between populations). This value is also significantly lower than the approximately 7.3% HbS incidence reported for African American populations in an independent study (Ojodu *et al.*, 2014). However, these results were highly variable depending on the population studied and ranged between approximately 4-10%, as expected from the high variability of ancestry from West Africa observed for these populations (Lema *et al.*, 2009). Lower frequencies of HbS within non-continental African populations can likely be ascribed to a combination of population admixture, genetic drift and negative selection, whereby the heterozygous state no longer confers any protective advantage in the absence of the selective pressure (i.e. exposure to the malarial pathogen) and balancing selection is disrupted (Roberts and Cavalli-Sforza, 1996; Piel *et al.*, 2010).

5.2 Human African Trypanosomiasis

5.2.1 Variants of APOL1

Two haplotypes of APOL1, G1 and G2, contribute protection against infection by *T.b. rhodesiense* and improved outcomes from *T.b. gambiense* but have demonstrated increased susceptibility to chronic kidney disease in individuals of African ancestry, as outlined previously (Sumaili *et al.*, 2018; Kamoto *et al.*, 2019). Additionally, G1 and G2 differ with respect to the level of protection they confer against parasitaemia and are also highly variable across different SSA populations (Cooper *et al.*, 2017). With this in mind, and in alignment with *US #8*, the AGVD was used to generate an interactive map of SSA population allele frequencies for variants of APOL1, including G1 and G2 haplotypes, and compare these to known values in the literature.

The AGVD facilitates accurate variant querying and annotation

Using the AGVD, both variants of the G1 haplotype – rs73885319 and rs60910145 – were queried by means of the "filter box" from the Variant view. Both RefSeq identifiers were entered into the "Variant" filter box, separated by a tab, with no additional filters applied. Both values were entered together under the assumption that a union operation would be performed on the set of terms, as would be the desired behaviour for most end-users. As expected, a corresponding row was displayed in the variant results table for each variant, with no additional rows returned. From the annotation, rs73885319 corresponds to the single nucleotide change "chr22:36661906 A>G" (GRCh37), while rs60910145 represents the change "chr22:36662034 T>G" (GRCh37); both annotations were in alignment with previous research (Cooper *et al.*, 2017).

The AGVD generates interactive G1 allele frequency maps across multiple study annotations

Next, rs73885319 was selected from the variant results table and an interactive map of allele frequencies across different populations was immediately displayed below. The diameter of data points was scaled by a consistent factor by toggling on the scaling option from the graph options below the map in order to allow for like-for-like comparison between graphs generated from different variants. Labels describing the data points were removed due to the high overlap between data points of nearby populations. The following study tabs were selected and their graphs compared: "1kG_phase3", "GNOMAD_GENOMES" and "EXAC" (*fig. 13*).

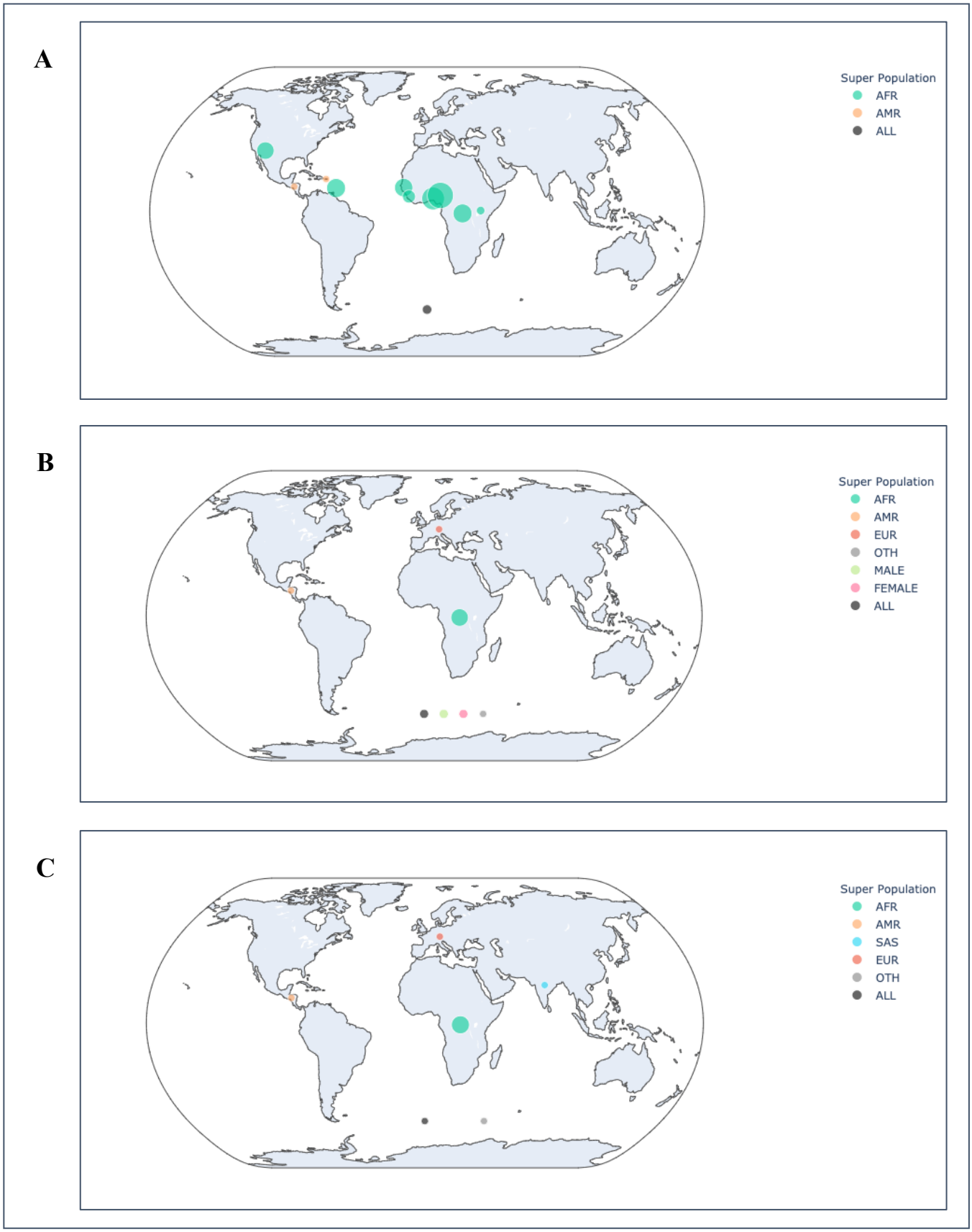

*Figure 13: World map of the allele frequencies of rs73885319 across different populations included within each of the following study annotations in the AGVD: (**A**)1000 Genomes Project, (**B**) gnomAD genomes and (**C**) ExAC. Populations are coloured by ancestry as per the super population grouping defined within each study.*

From the results of the 1kGP study annotation (*fig. 13 A*), all 7 African (AFR) populations included in the study demonstrated a non-zero allele frequency for rs73885319, including two non-continental African populations of recent descent from West Africa (ASW and ACB). Furthermore, the variant was common in all 7 of these populations, with an AF of >= 5%. Additionally, one population of admixed American ancestry (AMR), PUR, also observed the allele. However, the allele was not common within this population – having an AAF of 2.88% – and an overall AAF of 0.86% was observed for the AMR superpopulation. The AAF within the AFR superpopulation was 25.95% and the highest AAF was observed for the ESN, with a frequency of 49.49%. In contrast, the LWK in East Africa demonstrated a frequency of only 5.56%. Of the six populations of West African ancestry (ACB, ASW, ESN, GWD, MSL and YRI), the lowest frequency was observed for the MSL (AAF=12.35%).

According to the gnomAD genomes study annotation (*fig. 13 B*), the AAF for the aggregate African cohort was 21.97%, which was lower than the 1kGP study (25.95%). This difference, although small, might be partly attributed to an increase in the number of African-American samples included as part of the gnomAD v2 genomes dataset in comparison to 1kGP (4,359 versus 661) (Auton *et al.*, 2015; Karczewski *et al.*, 2020). The gnomAD dataset also includes diseased individuals from various case-control studies and therefore remains a largely distinct dataset, which makes comparisons between these studies challenging (Karczewski *et al.*, 2020). Similar values were observed between gnomAD genomes and 1kGP for the AMR superpopulation (0.95% versus 0.86%, respectively). In contrast, the allele was observed in one European population of the gnomAD genomes study (Non-Finnish European; NFE) but absent from Europeans within the 1kGP study. However, the allele was ultra-rare within this population (AAF=0.013%) and could possibly have arisen from admixture with African ancestries, although this was not tested by the current study.

The ExAC study annotation (*fig. 13 C*) includes allele frequencies calculated across greater than 60,000 exomes (Karczewski *et al.*, 2017). The AAF value for AFR within the ExAC study occurred between the values observed for 1kGP and gnomAD genomes, at a frequency of 23.29%. As with the gnomAD genomes study, the rs73885319 variant was also rare within AMR and NFE populations according to the ExAC annotation (AAF of 0.58% and 0.0011%, respectively). However, the frequency for NFR represents an approximately ten-fold decrease in relation to the frequency observed within the gnomAD genomes study. The variant was also observed at an ultra-rare frequency within SAS (AAF=0.0061%) but was not observed at appreciable levels according to the remaining two studies.

5.3 Alignment with the design goals of the AGVD

The above investigations serve to demonstrate both feasibility and validation of the AGVD against the intended design goals of the application as set out in *3. Software User Requirements*. Here, we have applied examples of potential use cases of the AGVD within an African context in order to assess the usefulness of platform features for end-users, the accuracy of search and filter operations and variant annotation sets, as well the visualisation and comparison of alternate allele frequencies for variants of interest.

Although we initially only set out to provide functionality to address three selected User Scenarios (*US #4*, *US #8* and *US #9*), features available within the AGVD pilot also provide partial functionality to address additional use cases of the application, and will form part of the future work of the project. For instance, the ability to filter variants by AAF within a specified population (a User Requirement of *US #5*) is already possible within the current implementation even though it was initially out of the scope of the project. Additionally, users can perform Beacon queries on variants from the AGVD, which partially fulfils the requirements of US #7 and basic user authentication is facilitated by the AGVD, in partial fulfilment of the requirements of US #7.

6. Conclusion

Clinical and population genetic databases greatly underrepresent diverse African populations, despite the high level of genetic and phenotypic variability observed in African ancestries. This has had important clinical consequences and the disparity is only fairly recently being acknowledged by contemporary genomic studies. In this work we set out to design, develop and deploy a proof of concept application for the effective storage, management and visualisation of African variant data that highlights the African data that is available at a more focused level than is provided in existing population genetic databases.

Our research included identifying and comparing bioinformatic software tools as well as open access African variant data sets in order to ultimately design and implement a bespoke software platform to meet our project objectives. Design of the software application was conducted by following software development industry standards as well as aligning with H3ABioNet software development best practices. Part of this process involved the generation of several User Scenarios, the requirements of which were fulfilled by the application and serve to demonstrate its utility. Here we present a pilot version of the African Genome Variation Database (AGVD). The AGVD is an open source database and web application written in Python, HTML and JavaScript, and leverages the scalable open access genomic analysis engine OpenCGA for variant storage, annotation and calculation of alternate allele frequencies.

Open access genotype data from the 1000 Genomes Project was used for the purposes of testing and validation of the platform. To this end, we demonstrated that the AGVD offers robust and accurate search and filter functionality for variants stored in the database. We further demonstrated utility of the AGVD within the context of African health research by exploring annotations and allele frequencies of clinically relevant variants for Malaria and HAT and compared these to known values in the literature. To meet these ends, several previously defined User Scenarios, which were generated during the design phase of the project, were modelled as case studies. These case studies served as the basis for acceptance criteria against which the success of project objectives was informally assessed and also to demonstrate utility in a real-world setting. Utility of the application was additionally recognised during a software demonstration to the H3Africa Scientific Advisory Board (SAB) in September 2020. However, since no formalised review process was followed during this demonstration of utility, the degree to which it was successful cannot be further quantified and largely remains the opinion of the author.

Following on from its pilot phase, the AGVD looks to be the first large-scale African owned database for managing and visualising African variant data. The AGVD also improves on existing genomic databases by providing a framework for the inclusion of more deeply stratified African cohorts and performing concomitant annotation and statistical calculations on these data. It also provides integrated access to Beacon queries for querying a large repository of additional datasets globally.

Future work will look to integrate additional open access African variant data identified as part of this study as well as tiered access to restricted data; the functionality of which is currently limited in the pilot release. Additionally, custom cohorts will be created to provide increased granularity across different African populations in order to better reflect population level genetic differences. Future versions of the AGVD will also look to provide on the fly allele frequency calculations across user-defined cohorts, while also being mindful of ethical considerations, such as the appropriate level of resolution that should be permitted when making such comparisons. Further changes to the platform will also include expanding the available annotation set applied to variants and including these in search and filter operations (for example improved alignment with HGNC and HGVS gene and variant nomenclature), which are currently limited in the pilot release (den Dunnen *et al.*, 2016). Additional statistics will also be added in future versions, such as the calculation of LD between loci of interest as well as Fst values for the comparison of population genetic differentiation.

In future, it is also important to define formalised acceptance criteria against which the success of the system in meeting its stated objectives can be objectively measured. This process should include, at minimum, assessment of each of the individual features of the application by independent test users, according to a standardised rubric. The scoring system might for example include a list of the three selected USs and associated FRs – identified in *Chapter 3* and later applied as use cases in *Chapter 5* – alongside a simple checklist defining whether or not the original design specification was met. Each reviewer could assess each US in turn and decide upon a general “success” or “failure” remark at the level of the US based upon how completely the FRs were fulfilled. Following this approach, successful USs would be those that were agreed upon by the majority of the test users.

Performance and scalability are also important considerations for a genomics database and software suite to be broadly useful to researchers, especially given the dearth of African variant data anticipated in the coming years. Future work will therefore also look at assessing the performance of the system under several test conditions. At minimum, the AGVD should be

benchmarked during the data ingestion and annotation process (since this is the most resource intensive phase of the application build) using appropriate tools which asses the time and hardware resources utilised during this process. Resource consumption (each of CPU cores, memory, disk space and time) could for example be plotted against size for the input data ingested across a range of incrementally larger data sets, and an assessment of the application's scalability could then be deduced based on the slope of this graph. A similar assessment will also be conducted on the performance of the web interface and database querying features in a simulated production environment with multiple concurrent test users.

Finally, improvements will be made to the end-user as well as developer documentation. This will include the addition of a tutorial page as well as more detailed information about the current data build, analytical methods used, a change log and planned update schedule. By addressing the current shortcomings of the pilot phase, the AGVD has the potential to serve as a powerful tool in the research of human diversity and health.

References

Affairs, A. (2020a) ‘Scenarios | Usability.gov’, *Usability.gov*. Available at: https://www.usability.gov/how-to-and-tools/methods/scenarios.html.

Affairs, A. (2020b) ‘Website Requirements | Usability.gov’, *Usability.gov*. Available at: https://www.usability.gov/how-to-and-tools/methods/requirements.html#:~:text=User Requirements describe how user,to complete on your site.

Agile Business Consortium (2014) ‘MoSCoW Prioritisation’, *The DSDM Agile Project Framework (2014 Onwards)*. Available at: https://www.agilebusiness.org/page/ProjectFramework_10_MoSCoWPrioritisation (Accessed: 3 October 2020).

Aidoo, M. *et al.* (2002) ‘Protective effects of the sickle cell gene against malaria morbidity and mortality’, *Lancet*. Elsevier Limited, 359(9314), pp. 1311–1312. doi: 10.1016/S0140-6736(02)08273-9.

Alemán, A. *et al.* (2014) ‘A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies’, *Nucleic Acids Research*, 42(W1), pp. 88–93. doi: 10.1093/nar/gku407.

Allison, A. C. (1954) ‘Protection afforded by sickle-cell trait against subtertian malarial infection’, *British Medical Journal*. BMJ Publishing Group, 1(4857), pp. 290–294. doi: 10.1136/bmj.1.4857.290.

Almal, S. H. and Padh, H. (2012) ‘Implications of gene copy-number variation in health and diseases’, *Journal of Human Genetics*. Nature Publishing Group, pp. 6–13. doi: 10.1038/jhg.2011.108.

Altshuler, D. L. *et al.* (2010) ‘A map of human genome variation from population-scale sequencing’, *Nature*, 467(7319), pp. 1061–1073. doi: 10.1038/nature09534.

Altshuler, D. M. *et al.* (2010) ‘Integrating common and rare genetic variation in diverse human populations’, *Nature*. Nature Publishing Group, 467(7311), pp. 52–58. doi: 10.1038/nature09298.

Altshuler, D. M. *et al.* (2012) ‘An integrated map of genetic variation from 1,092 human genomes’, *Nature*. Nature Publishing Group, 491(7422), pp. 56–65. doi: 10.1038/nature11632.

Arms, W. Y. (2014) ‘Cornell University - CS 5150 Software Engineering - Scenarios and Use

Cases'. Available at: https://www.cs.cornell.edu/courses/cs5150/2014fa/slides/D2-use-cases.pdf.

Aron, S. *et al.* (2017) 'H3ABioNet: Developing Sustainable Bioinformatics Capacity in Africa', *EMBnet.journal*. EMBnet Stichting, 23(0), p. 886. doi: 10.14806/ej.23.0.886.

Ashley-Koch, A., Yang, Q. and Olney, R. S. (2000) 'Sickle Hemoglobin (Hb S) Allele and Sickle Cell Disease: A HuGE Review', *American Journal of Epidemiology*. Oxford University Press, 151(9), pp. 839–845. doi: 10.1093/oxfordjournals.aje.a010288.

Ashorobi, D. and Bhatt, R. (2019) 'Sickle Cell Trait'. StatPearls Publishing.

Assembly, W. H. (2006) '*Sickle-cell anaemia: report by the Secretariat*'. World Health Organization. Available at: https://apps.who.int/iris/handle/10665/20890.

Atkinson, Q. D. (2011) 'Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa', *Science*, 332(6027), pp. 346–349. doi: 10.1126/science.1199295.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*. Nature Publishing Group, pp. 68–74. doi: 10.1038/nature15393.

Bae, C. J. (2013) *Archaic Homo sapiens*, *Nature Education Knowledge 4(8):4*. Available at: https://www.nature.com/scitable/knowledge/library/archaic-homo-sapiens-103852137/ (Accessed: 13 July 2020).

Band, G. *et al.* (2015) 'A novel locus of resistance to severe malaria in a region of ancient balancing selection', *Nature*. Nature Publishing Group, 526(7572), pp. 253–257. doi: 10.1038/nature15390.

Belmont, J. W. *et al.* (2005) 'A haplotype map of the human genome', *Nature*. Nature Publishing Group, 437(7063), pp. 1299–1320. doi: 10.1038/nature04226.

Bergström, A. *et al.* (2020) 'Insights into human genetic variation and population history from 929 diverse genomes', *Science*. American Association for the Advancement of Science, 367(6484). doi: 10.1126/science.aay5012.

Berners-Lee, T., Fielding, R. and Masinter, L. (2005) 'RFC 3986, Uniform Resource Identifier (URI): Generic Syntax'. Available at: https://tools.ietf.org/html/rfc3986#section-3.4 (Accessed: 19 November 2020).

Bleda, M. *et al.* (2012) 'CellBase, a comprehensive collection of RESTful web services for

retrieving relevant biological information from heterogeneous sources', *Nucleic Acids Research*. Oxford Academic, 40(W1), pp. W609–W614. doi: 10.1093/nar/gks575.

Bowcock, A. M. *et al.* (1994) 'High resolution of human evolutionary trees with polymorphic microsatellites', *Nature*, 368(6470), pp. 455–457. doi: 10.1038/368455a0.

Brown, F. H., McDougall, I. and Fleagle, J. G. (2012) 'Correlation of the KHS Tuff of the Kibish Formation to volcanic ash layers at other sites, and the age of early Homo sapiens (Omo I and Omo II)', *Journal of Human Evolution*. Academic Press, 63(4), pp. 577–585. doi: 10.1016/j.jhevol.2012.05.014.

Buck, L. T. and Stringer, C. B. (2014) 'Homo heidelbergensis', *Current Biology*. Cell Press, pp. R214–R215. doi: 10.1016/j.cub.2013.12.048.

Campbell, M. C. and Tishkoff, S. A. (2008) 'African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping', *Annual Review of Genomics and Human Genetics*, 9(1), pp. 403–433. doi: 10.1146/annurev.genom.9.081307.164258.

Cann, R. L., Stoneking, M. and Wilson, A. C. (1987) 'Mitochondrial DNA and human evolution', *Nature*. Nature Publishing Group, 325(6099), pp. 31–36. doi: 10.1038/325031a0.

Carter, R. and Mendis, K. N. (2002) 'Evolutionary and historical aspects of the burden of malaria', *Clinical Microbiology Reviews*. American Society for Microbiology Journals, pp. 564–594. doi: 10.1128/CMR.15.4.564-594.2002.

Caswell-Jin, J. L. *et al.* (2018) 'Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk', *Genetics in Medicine*. Nature Publishing Group, 20(2), pp. 234–239. doi: 10.1038/gim.2017.96.

Cavalli-Sforza, L. L. and Feldman, M. W. (2003) 'The application of molecular genetic approaches to the study of human evolution', *Nature Genetics*. Nature Publishing Group, pp. 266–275. doi: 10.1038/ng1113.

CDC (2020) *African Trypanosomiasis - General Information*. Available at: https://www.cdc.gov/parasites/sleepingsickness/gen_info/faqs.html (Accessed: 26 May 2020).

Chan, E. K. F. *et al.* (2019) 'Human origins in a southern African palaeo-wetland and first migrations', *Nature*. Springer US, 575(7781), pp. 185–189. doi: 10.1038/s41586-019-1714-1.

Choudhury, A. *et al.* (2020) 'High-depth African genomes inform human migration and

health', *Nature*. NLM (Medline), 586(7831), pp. 741–748. doi: 10.1038/s41586-020-2859-7.

Clarke, L. *et al.* (2017) 'The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data', *Nucleic Acids Research*, 45(D1), pp. D854–D859. doi: 10.1093/nar/gkw829.

Clegg, D. and Barker, R. (1994) *CASE Method Fast-Track: A RAD Approach*. 1st edn. Addison-Wesley.

ClinVar (2019a) *Accession: RCV000000006.5*. Available at: https://www.ncbi.nlm.nih.gov/clinvar/RCV000000006.5/ (Accessed: 29 November 2020).

ClinVar (2019b) *Accession: RCV000000007.3*. Available at: https://www.ncbi.nlm.nih.gov/clinvar/RCV000000007.3/ (Accessed: 29 November 2020).

ClinVar (2019c) *Accession: RCV000000008.4*. Available at: https://www.ncbi.nlm.nih.gov/clinvar/RCV000000008.4/ (Accessed: 29 November 2020).

ClinVar (2019d) *Accession: VCV000018395.1*. Available at: https://www.ncbi.nlm.nih.gov/clinvar/variation/18395/#id_first (Accessed: 29 November 2020).

Cohen, K. . *et al.* (2020) *The ICS International Chronostratigraphic Chart 2020/01. International Commission on Stratigraphy, IUGS*, *International Commission on Stratigraphy, IUGS*. Available at: http://www.stratigraphy.org/ICSchart/ChronostratChart2020-01.pdf (Accessed: 3 April 2020).

Coluzzi, M. (1999) 'The clay feet of the malaria giant and its African roots: Hypotheses and inferences about origin, spread and control of Plasmodium falciparum', *Parassitologia*, 41(1–3), pp. 277–283.

Conniff, R. (2019) *Lush Okavango Delta Pinpointed as Ancestral Homeland of All Living Humans - Scientific American*, *Evolution*. Available at: https://www.scientificamerican.com/article/lush-okavango-delta-pinpointed-as-ancestral-homeland-of-all-living-humans/ (Accessed: 6 March 2021).

Cook, C. E. *et al.* (2016) 'The European Bioinformatics Institute in 2016: Data growth and integration', *Nucleic Acids Research*, 44(D1), pp. D20–D26. doi: 10.1093/nar/gkv1352.

Cooper, A. *et al.* (2017) 'APOL1 renal risk variants have contrasting resistance and susceptibility associations with African trypanosomiasis', *eLife*. eLife Sciences Publications

Ltd, 6. doi: 10.7554/eLife.25461.

Crawford, N. G. *et al.* (2017) 'Loci associated with skin pigmentation identified in African populations', *Science*, 358(6365). doi: 10.1126/science.aan8433.

Cruciani, F. *et al.* (2011) 'A revised root for the human y chromosomal phylogenetic tree: The origin of patrilineal diversity in Africa', *American Journal of Human Genetics*. The American Society of Human Genetics, 88(6), pp. 814–818. doi: 10.1016/j.ajhg.2011.05.002.

Damena, D. *et al.* (2019) 'Genome-wide association studies of severe P. falciparum malaria susceptibility: Progress, pitfalls and prospects', *BMC Medical Genomics*. BioMed Central Ltd., pp. 1–14. doi: 10.1186/s12920-019-0564-x.

Dean, L. (2005) 'The ABO blood group - Blood Groups and Red Cell Antigens - NCBI Bookshelf', *Blood groups and Red cell antigen*, pp. 1–8. Available at: https://www.ncbi.nlm.nih.gov/books/NBK2271/ (Accessed: 28 November 2020).

Devlin, B. and Risch, N. (1995) 'A comparison of linkage disequilibrium measures for fine-scale mapping', *Genomics*. Academic Press, 29(2), pp. 311–322. doi: 10.1006/geno.1995.9003.

den Dunnen, J. T. *et al.* (2016) 'HGVS Recommendations for the Description of Sequence Variants: 2016 Update', *Human Mutation*. John Wiley and Sons Inc., 37(6), pp. 564–569. doi: 10.1002/humu.22981.

Eberhard, D. M., Simons, G. F. and Fennig, C. D. (2020) 'Ethnologue: Languages of the World. Twenty-third edition.', *Ethnologue.* Available at: https://www.ethnologue.com/.

Edison, E. S. *et al.* (2005) 'Hyperbilirubinemia in homozygous HbE disease is associated with the UGT1A1 gene polymorphism', *Hemoglobin*. Hemoglobin, 29(3), pp. 189–195. doi: 10.1081/HEM-200066314.

Edwards, S. L. *et al.* (2013) 'Beyond GWASs: Illuminating the Dark Road from Association to Function', *The American Journal of Human Genetics*. Cell Press, 93(5), pp. 779–797. doi: 10.1016/J.AJHG.2013.10.012.

Eilbeck, K. *et al.* (2005) 'The Sequence Ontology: a tool for the unification of genome annotations', *Genome Biology*, 6(5), p. R44. doi: 10.1186/gb-2005-6-5-r44.

Eldredge, N. and Gould, S. J. (1972) 'Punctuated equilibria: an alternative to phyletic gradualism', *Thomas J. M. Schopf (ed.), Models in Paleobiology. Freeman Cooper. pp. 82–*

115.

Ellegren, H. (2004) 'Microsatellites: Simple sequences with complex evolution', *Nature Reviews Genetics*. Nature Publishing Group, pp. 435–445. doi: 10.1038/nrg1348.

Endicott, P., Ho, S. Y. W. and Stringer, C. (2010) 'Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins', *Journal of Human Evolution*, 59(1), pp. 87–95. doi: 10.1016/j.jhevol.2010.04.005.

Endrullat, C. (2017) *Standardization in next-generation sequencing - Issues and approaches of establishing standards in a highly dynamic environment*. doi: 10.7287/peerj.preprints.2771.

Ensembl (2020) *Gene: ACKR1 (ENSG00000213088) - Summary - Homo_sapiens - Ensembl genome browser 101*. Available at: https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000213088;r=1:159203307-159206500 (Accessed: 30 November 2020).

EVS (2014) *Exome Variant Server*, *NHLBI GO Exome Sequencing Project (ESP)*. Available at: https://evs.gs.washington.edu/EVS/ (Accessed: 18 January 2021).

Fiume, M. *et al.* (2019) 'Federated discovery and sharing of genomic data using Beacons', *Nature Biotechnology*. Nature Publishing Group, pp. 220–224. doi: 10.1038/s41587-019-0046-x.

Flask (2019) *Modular Applications with Blueprints*. Available at: https://flask.palletsprojects.com/en/1.1.x/blueprints/ (Accessed: 14 October 2020).

Flask (2020) *Larger Applications*. Available at: https://flask.palletsprojects.com/en/1.1.x/patterns/packages/#working-with-blueprints (Accessed: 14 October 2020).

Flint, J. *et al.* (1998) 'The population genetics of the haemoglobinopathies', *Bailliere's Clinical Haematology*. Bailliere Tindall Ltd, 11(1), pp. 1–51. doi: 10.1016/S0950-3536(98)80069-3.

Franco, J. R. *et al.* (2020) 'Monitoring the elimination of human African trypanosomiasis at continental and country level: Update to 2018', *PLOS Neglected Tropical Diseases*. Edited by E. Matovu, 14(5), p. e0008261. doi: 10.1371/journal.pntd.0008261.

Frazer, K. A. *et al.* (2007) 'A second generation human haplotype map of over 3.1 million SNPs', *Nature*. Nature Publishing Group, 449(7164), pp. 851–861. doi: 10.1038/nature06258.

Gallego Llorente, M., Eriksson, A. and Siska, V. (2015) ‘Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa’, *Science*, 350(6262), pp. 821–822.

Ganna, A. *et al.* (2016) ‘Ultra-rare disruptive and damaging mutations influence educational attainment in the general population’, *Nature Neuroscience*. Nature Publishing Group, pp. 1563–1565. doi: 10.1038/nn.4404.

Genovese, G. *et al.* (2010) ‘Association of trypanolytic ApoL1 variants with kidney disease in African Americans’, *Science*. American Association for the Advancement of Science, 329(5993), pp. 841–845. doi: 10.1126/science.1193032.

Ghansah, A. *et al.* (2012) ‘Haplotype analyses of haemoglobin C and haemoglobin S and the dynamics of the evolutionary response to malaria in Kassena-Nankana district of Ghana’, *PLoS ONE*. Public Library of Science, 7(4). doi: 10.1371/journal.pone.0034565.

Golassa, L. *et al.* (2020) ‘The biology of unconventional invasion of Duffy-negative reticulocytes by Plasmodium vivax and its implication in malaria epidemiology and public health’, *Malaria Journal*. BioMed Central Ltd, p. 299. doi: 10.1186/s12936-020-03372-9.

Gomez, F., Hirbo, J. and Tishkoff, S. A. (2014) ‘Genetic variation and adaptation in Africa: Implications for human evolution and disease’, *Cold Spring Harbor Perspectives in Biology*, 6(7), pp. 1–22. doi: 10.1101/cshperspect.a008524.

Gonder, M. K. *et al.* (2007) ‘Whole-mtDNA genome sequence analysis of ancient african lineages’, *Molecular Biology and Evolution*, 24(3), pp. 757–768. doi: 10.1093/molbev/msl209.

Gurdasani, D. *et al.* (2015) ‘The African Genome Variation Project shapes medical genetics in Africa’, *Nature*. Nature Publishing Group, 517(7534), pp. 327–332. doi: 10.1038/nature13997.

Guthery, S. L. *et al.* (2007) ‘The structure of common genetic variation in United States populations’, *American Journal of Human Genetics*. University of Chicago Press, 81(6), pp. 1221–1231. doi: 10.1086/522239.

H3ABioNet (2020) *Organization - H3ABioNet - Pan African Bioinformatics Network for the Human Heredity and Health in Africa*. Available at: https://www.h3abionet.org/about/organization (Accessed: 20 January 2021).

Ha, J. *et al.* (2019) ‘Hemoglobin E, malaria and natural selection’, *Evolution, Medicine, and Public Health*. Oxford University Press (OUP), 2019(1), pp. 232–241. doi: 10.1093/emph/eoz034.

Hammer, M. F. (1995) ‘A recent common ancestry for human Y chromosomes’, *Nature*. Nature Publishing Group, 378(6555), pp. 376–378. doi: 10.1038/378376a0.

Hammond, A. S., Royer, D. F. and Fleagle, J. G. (2017) ‘The Omo-Kibish I pelvis’, *Journal of Human Evolution*, 108, pp. 199–219. doi: 10.1016/j.jhevol.2017.04.004.

Harvati, K. (2007) ‘100 years of Homo Heidelbergensis - Life and times of a controversial taxon’, *Mitteilungen der Gesellschaft für Urgeschichte*, 16, pp. 85–94.

Hedrick, P. W. (2011) ‘Population genetics of malaria resistance in humans’, *Heredity*. Nature Publishing Group, pp. 283–304. doi: 10.1038/hdy.2011.16.

Hedrick, P. W. (2012) ‘Resistance to malaria in humans: The impact of strong, recent selection’, *Malaria Journal*. BioMed Central, p. 349. doi: 10.1186/1475-2875-11-349.

Hemminki, K., Försti, A. and Bermejo, J. L. (2008) ‘The “common disease-common variant” hypothesis and familial risks’, *PLoS ONE*. Public Library of Science, 3(6). doi: 10.1371/journal.pone.0002504.

Henn, B. M. *et al.* (2011) ‘Hunter-gatherer genomic diversity suggests a southern African origin for modern humans’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(13), pp. 5154–5162. doi: 10.1073/pnas.1017511108.

HGNC (2019) *ACKR1 gene symbol report | HUGO Gene Nomenclature Committee*, *HGNC*. Available at: https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:4035 (Accessed: 30 November 2020).

Höher, G., Fiegenbaum, M. and Almeida, S. (2018) ‘Molecular basis of the Duffy blood group system’, *Blood Transfusion*. Edizioni SIMTI, pp. 93–100. doi: 10.2450/2017.0119-16.

Holsinger, K. E. and Weir, B. S. (2009) ‘Genetics in geographically structured populations: Defining, estimating and interpreting FST’, *Nature Reviews Genetics*. Nature Publishing Group, pp. 639–650. doi: 10.1038/nrg2611.

Holt, B. M. (2015) ‘Anatomically Modern Homo sapiens’, in *Basics in Human Evolution*. Elsevier Inc., pp. 177–192. doi: 10.1016/B978-0-12-802652-6.00013-X.

Hublin, J. J. (2009) ‘The origin of Neandertals’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, pp. 16022–16027. doi: 10.1073/pnas.0904119106.

Hublin, J. J. *et al.* (2017) ‘New fossils from Jebel Irhoud, Morocco and the pan-African origin

of Homo sapiens', *Nature*. Nature Publishing Group, 546(7657), pp. 289–292. doi: 10.1038/nature22336.

Hunter, S. *et al.* (2012) 'InterPro in 2011: New developments in the family and domain prediction database', *Nucleic Acids Research*. Oxford University Press, 40(D1), pp. D306–D312. doi: 10.1093/nar/gkr948.

Hürsch, W. L. and Lopes, C. V. (1995) *Separation of Concerns*. doi: 10.1.1.29.5223.

Jakobsson, M. *et al.* (2008) 'Genotype, haplotype and copy-number variation in worldwide human populations', *Nature*, 451(7181), pp. 998–1003. doi: 10.1038/nature06742.

Jin, L. and Su, B. (2000) 'Natives or immigrants: Modern human origin in East Asia', *Nature Reviews Genetics*. Nature Publishing Group, pp. 126–133. doi: 10.1038/35038565.

Jorde, L. B. *et al.* (1997) 'Microsatellite diversity and the demographic history of modern humans', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 94(7), pp. 3100–3103. doi: 10.1073/pnas.94.7.3100.

Jorde, L. B. (2000) 'Linkage disequilibrium and the search for complex disease genes', *Genome Research*. Cold Spring Harbor Laboratory Press, pp. 1435–1444. doi: 10.1101/gr.144500.

Kamoto, K. *et al.* (2019) 'Association of APOL1 renal disease risk alleles with Trypanosoma brucei rhodesiense infection outcomes in the northern part of Malawi', *PLoS Neglected Tropical Diseases*. Edited by A. Benkahla. Public Library of Science, 13(8), p. e0007603. doi: 10.1371/journal.pntd.0007603.

Karczewski, K. J. *et al.* (2017) 'The ExAC browser: Displaying reference data information from over 60 000 exomes', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D840–D845. doi: 10.1093/nar/gkw971.

Karczewski, K. J. *et al.* (2020) 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*. Nature Research, 581(7809), pp. 434–443. doi: 10.1038/s41586-020-2308-7.

Kimura, M. and Ohta, T. (1973) 'The age of a neutral mutant persisting in a finite population.', *Genetics*, 75(1), pp. 199–212. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4762875 (Accessed: 28 April 2020).

Kinney, N. *et al.* (2019) 'Abundance of ethnically biased microsatellites in human gene

regions', *PLOS ONE*. Edited by A. Palsson. Public Library of Science, 14(12), p. e0225216. doi: 10.1371/journal.pone.0225216.

Ko, W. Y. *et al.* (2013) 'Identifying darwinian selection acting on different human apol1 variants among diverse african populations', *American Journal of Human Genetics*. Elsevier, 93(1), pp. 54–66. doi: 10.1016/j.ajhg.2013.05.014.

Krissaane, I. *et al.* (2020) 'Scalability and cost-effectiveness analysis of whole genome-wide association studies on Google Cloud Platform and Amazon Web Services', *Journal of the American Medical Informatics Association*. Oxford University Press (OUP), 27(9), pp. 1425–1430. doi: 10.1093/jamia/ocaa068.

Kulski, J. K. (2016) 'Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications', in *Next Generation Sequencing - Advances, Applications and Challenges*. doi: 10.5772/61964.

Kuper, R. and Kröpalin, S. (2006) 'Climate-controlled holocene occupation in the Sahara: Motor of Africa's evolution', *Science*. Science, 313(5788), pp. 803–807. doi: 10.1126/science.1130989.

Landrum, M. J. *et al.* (2018) 'ClinVar: improving access to variant interpretations and supporting evidence.', *Nucleic acids research*, 46(D1), pp. D1062–D1067. doi: 10.1093/nar/gkx1153.

Landry, L. G. *et al.* (2018) 'Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice', *Health Affairs*, 37(5), pp. 780–785. doi: 10.1377/hlthaff.2017.1595.

Lappalainen, I. *et al.* (2013) 'DbVar and DGVa: Public archives for genomic structural variation', *Nucleic Acids Research*. Oxford University Press, 41(D1), p. D936. doi: 10.1093/nar/gks1213.

Lee, C. (2020) *Opinion: Greater Diversity Is Needed in Human Genomic Data*, *The Scientist Magazine*□. Available at: https://www.the-scientist.com/critic-at-large/diversify-our-human-genomic-data-66308 (Accessed: 24 November 2019).

Leffler, E. M. *et al.* (2017) 'Resistance to malaria through structural variation of red blood cell invasion receptors', *Science*, 356(6343), pp. 1140–1152. doi: 10.1126/science.aam6393.

Lema, G. *et al.* (2009) 'The Genetic Structure and History of Africans and African Americans', *Science*, 324(5930), pp. 1035–1044. doi: 10.1126/science.1172257.The.

Levy-Sakin, M. *et al.* (2019) 'Genome maps across 26 human populations reveal population-specific patterns of structural variation', *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–14. doi: 10.1038/s41467-019-08992-7.

Lewontin, R. C. and Kojima, K. (1960) 'THE EVOLUTIONARY DYNAMICS OF COMPLEX POLYMORPHISMS', *Evolution*. John Wiley & Sons, Ltd, 14(4), pp. 458–472. doi: 10.1111/j.1558-5646.1960.tb03113.x.

Li, J. Z. *et al.* (2008) 'Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation', *Science*, 319(5866), pp. 1100–1104. doi: 10.1126/science.1153717.

LibHunt (2020) *flask-restful vs Flask RestPlus*. Available at: https://python.libhunt.com/compare-flask-restful-vs-flask-restplus (Accessed: 14 October 2020).

Lieberman, D. E., McBratney, B. M. and Krovitz, G. (2002) 'The evolution and development of cranial form in Homo sapiens', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 99(3), pp. 1134–1139. doi: 10.1073/pnas.022440799.

Liew, A. *et al.* (2017) 'Hypertrophic Cardiomyopathy—Past, Present and Future', *Journal of Clinical Medicine*. MDPI AG, 6(12), p. 118. doi: 10.3390/jcm6120118.

Limou, S. *et al.* (2014) 'APOL1 Kidney Risk Alleles: Population Genetics and Disease Associations', *Advances in Chronic Kidney Disease*. W.B. Saunders, pp. 426–433. doi: 10.1053/j.ackd.2014.06.005.

Limou, S. *et al.* (2015) 'Sequencing rare and common APOL1 coding variants to determine kidney disease risk', *Kidney International*. Elsevier Masson SAS, 88(4), pp. 754–763. doi: 10.1038/ki.2015.151.

Linden, M. *et al.* (2018) 'Common elixir service for researcher authentication and authorisation [version 1; referees: 3 approved, 1 approved with reservations]', *F1000Research*. F1000 Research Ltd, 7, p. 1199. doi: 10.12688/f1000research.15161.1.

Lonjou, C. *et al.* (2003a) 'Linkage disequilibrium in human populations.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 100(10), pp. 6069–74. doi: 10.1073/pnas.1031521100.

Lonjou, C. *et al.* (2003b) 'Linkage disequilibrium in human populations.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences,

100(10), pp. 6069–74. doi: 10.1073/pnas.1031521100.

Loosdrecht, M. Van De *et al.* (2018) 'Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations', *Science*, 552, pp. 548–552.

Lopez, J. *et al.* (2017) 'HGVA: The Human Genome Variation Archive', *Nucleic Acids Research*. Oxford University Press, 45(W1), pp. W189–W194. doi: 10.1093/nar/gkx445.

Maakaron, J. E. and Taher, A. T. (2020) *Sickle Cell Anemia: Practice Essentials, Background, Genetics*, *Medscape*. Available at: https://emedicine.medscape.com/article/205926-overview#a4 (Accessed: 15 July 2020).

MacDonald, J. R. *et al.* (2014) 'The Database of Genomic Variants: A curated collection of structural variation in the human genome', *Nucleic Acids Research*. Nucleic Acids Res, 42(D1). doi: 10.1093/nar/gkt958.

Mallick, S. *et al.* (2016) 'The Simons Genome Diversity Project: 300 genomes from 142 diverse populations', *Nature*, 538(7624), pp. 201–206. doi: 10.1038/nature18964.

Manrai, A. K. *et al.* (2016) 'Genetic Misdiagnoses and the Potential for Health Disparities', *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 375(7), pp. 655–665. doi: 10.1056/nejmsa1507092.

Manzi, G. (2011) 'Before the Emergence of Homo sapiens: Overview on the Early-to-Middle Pleistocene Fossil Record (with a Proposal about Homo heidelbergensis at the subspecific level)', *International journal of evolutionary biology*, 2011, p. 582678. doi: 10.4061/2011/582678.

McEvoy, B. P. *et al.* (2011) 'Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs', *Genome Research*. Cold Spring Harbor Laboratory Press, 21(6), pp. 821–829. doi: 10.1101/gr.119636.110.

McManus, K. F. *et al.* (2017) 'Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans', *PLoS Genetics*. Public Library of Science, 13(3), p. e1006560. doi: 10.1371/journal.pgen.1006560.

Medina, N. and OpenCB (2019) *Welcome to OpenCGA*. Available at: http://docs.opencb.org/display/opencga (Accessed: 14 October 2020).

Mendez, F. L. *et al.* (2013) 'An African American paternal lineage adds an extremely ancient root to the human y chromosome phylogenetic tree', *American Journal of Human Genetics*.

The American Society of Human Genetics, 92(3), pp. 454–459. doi: 10.1016/j.ajhg.2013.02.002.

Mills, M. C. and Rahal, C. (2019) 'A scientometric review of genome-wide association studies', *Communications Biology*, 2(1). doi: 10.1038/s42003-018-0261-x.

Morales, J. *et al.* (2018) 'A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog', *Genome Biology*. BioMed Central Ltd. doi: 10.1186/s13059-018-1396-2.

Mounier, A. and Mirazón Lahr, M. (2019) 'Deciphering African late middle Pleistocene hominin diversity and the origin of our species', *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–13. doi: 10.1038/s41467-019-11213-w.

Mulder, N. *et al.* (2017) 'Genomic Research Data Generation, Analysis and Sharing – Challenges in the African Setting', *Data Science Journal*, 16, pp. 1–15. doi: 10.5334/dsj-2017-049.

Mulder, N. *et al.* (2018) 'H3Africa: Current perspectives', *Pharmacogenomics and Personalized Medicine*. Dove Medical Press Ltd, pp. 59–66. doi: 10.2147/PGPM.S141546.

Mulder, N. J. *et al.* (2016) 'H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa', *Genome Research*. doi: 10.1101/gr.196295.115.

NCBI (2020) *ACKR1 atypical chemokine receptor 1 (Duffy blood group) [Homo sapiens (human)] - Gene - NCBI*. Available at: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=2532 (Accessed: 30 November 2020).

Neale, B. (2018) *UK Biobank — Neale lab*. Available at: http://www.nealelab.is/uk-biobank (Accessed: 13 January 2021).

Nei, M. (1986) 'Definition and estimation of fixation indices', *Evolution*. Wiley-Blackwell, 40(3), pp. 643–645. doi: 10.1111/j.1558-5646.1986.tb00516.x.

Ngo Bitoungui, V. J. *et al.* (2015) 'Beta-Globin Gene Haplotypes among Cameroonians and Review of the Global Distribution: Is There a Case for a Single Sickle Mutation Origin in Africa?', *OMICS A Journal of Integrative Biology*. Mary Ann Liebert Inc., pp. 171–179. doi: 10.1089/omi.2014.0134.

Nielsen, R. *et al.* (2017) 'Tracing the peopling of the world through genomics', *Nature*,

541(7637), pp. 302–310. doi: 10.1038/nature21347.

Nothaft, F. A. *et al.* (2019) *Introducing Glow: An Open-Source Toolkit for Large-Scale Genomic Analysis - The Databricks Blog*. Available at: https://databricks.com/blog/2019/10/18/introducing-glow-an-open-source-toolkit-for-large-scale-genomic-analysis.html (Accessed: 13 January 2021).

OAI (2016) *OpenAPI specification*, *swagger.io*. Available at: http://spec.openapis.org/oas/v3.0.3 (Accessed: 14 October 2020).

Ojodu, J. *et al.* (2014) 'Incidence of sickle cell trait--United States, 2010.', *MMWR. Morbidity and mortality weekly report*. Centers for Disease Control and Prevention, 63(49), pp. 1155–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25503918 (Accessed: 19 February 2021).

OpenCB (2020) *Welcome to OpenCB | OpenCB*. Available at: http://www.opencb.org/ (Accessed: 14 October 2020).

Otto, T. D. *et al.* (2018) 'Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria', *Nature Microbiology*. Nature Publishing Group, 3(6), pp. 687–697. doi: 10.1038/s41564-018-0162-2.

van Oven, M. and Kayser, M. (2009) 'Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation.', *Human mutation*, 30(2), pp. 386–394. doi: 10.1002/humu.20921.

Owens, A. T. and Reza, N. (2020) *Diagnosis of Hypertrophic Cardiomyopathy: What Every Cardiologist Needs to Know*, *American College of Cardiology*. Available at: https://www.acc.org/latest-in-cardiology/articles/2020/02/25/06/34/diagnosis-of-hypertrophic-cardiomyopathy#sort=%40commonsortdate descending (Accessed: 29 October 2020).

Parker, J. (2020) 'Business, User, and System Requirements - Enfocus Solutions Inc', *Enfocus Solutions Inc*. Available at: https://enfocussolutions.com/business-user-and-system-requirements/#:~:text=User requirements are generally signed,input for creating system requirements.&text=A functional requirement specifies something,this is a functional requirement.

Parsa, A. *et al.* (2013) 'APOL1 risk variants, race, and progression of chronic kidney disease', *New England Journal of Medicine*. Massachussetts Medical Society, 369(23), pp. 2183–2196. doi: 10.1056/NEJMoa1310345.

Pays, E. *et al.* (2014) 'The molecular arms race between African trypanosomes and humans', *Nature Reviews Microbiology*. Nature Publishing Group, pp. 575–584. doi: 10.1038/nrmicro3298.

Pemberton, T. J., DeGiorgio, M. and Rosenberg, N. A. (2013) 'Population structure in a comprehensive genomic data set on human microsatellite variation', *G3: Genes, Genomes, Genetics*, 3(5), pp. 891–907. doi: 10.1534/g3.113.005728.

Piel, F. B. *et al.* (2010) 'Global distribution of the sickle cell gene and geographical conxfirmation of the malaria hypothesis', *Nature Communications*. Nature Publishing Group, 1(8), p. 104. doi: 10.1038/ncomms1104.

Pinto, J. C. *et al.* (2016) 'Food and pathogen adaptations in the Angolan Namib desert: Tracing the spread of lactase persistence and human African trypanosomiasis resistance into southwestern Africa', *American Journal of Physical Anthropology*. Wiley-Liss Inc., 161(3), pp. 436–447. doi: 10.1002/ajpa.23042.

Popejoy, A. B. and Fullerton, S. M. (2016) 'Genomics is failing on diversity', *Nature*. Nature Publishing Group, pp. 161–164. doi: 10.1038/538161a.

Porras-Hurtado, L. *et al.* (2013) 'An overview of STRUCTURE: Applications, parameter settings, and supporting software', *Frontiers in Genetics*. Frontiers Media SA, 4(MAY). doi: 10.3389/fgene.2013.00098.

Preston-Werner, T. (2018) *Semantic Versioning 2.0.0 | Semantic Versioning*. Available at: https://semver.org/ (Accessed: 19 October 2020).

Pritchard, J. K. and Przeworski, M. (2001) 'Linkage disequilibrium in humans: Models and data', *American Journal of Human Genetics*. University of Chicago Press, pp. 1–14. doi: 10.1086/321275.

Pule, G. D. *et al.* (2017) 'Beta-globin gene haplotypes and selected Malaria-associated variants among black Southern African populations', *Global health, epidemiology and genomics*. NLM (Medline), 2, p. e17. doi: 10.1017/gheg.2017.14.

Pulit, S. L., Voight, B. and de Bakker, P. I. W. (2010) 'Multiethnic genetic association studies improve power for locus discovery', *PLoS ONE*. Edited by M. N. Weedon. Public Library of Science, 5(9), pp. 1–9. doi: 10.1371/journal.pone.0012600.

Ramachandran, S. *et al.* (2005) 'Support from the relationship of genetic and geographic in human populations for a serial founder effect originating in Africa', *Proceedings of the*

National Academy of Sciences of the United States of America, 102(44), pp. 15942–15947. doi: 10.1073/pnas.0507611102.

Ramakrishnan, A. P. (2013) 'Linkage Disequilibrium', in *Brenner's Encyclopedia of Genetics: Second Edition*. Elsevier Inc., pp. 252–253. doi: 10.1016/B978-0-12-374984-0.00870-6.

Reed, F. A. and Tishkoff, S. A. (2006) 'African human diversity, origins and migrations', *Current Opinion in Genetics and Development*. Elsevier Current Trends, pp. 597–605. doi: 10.1016/j.gde.2006.10.008.

Regional Committee for Africa (2011) *Sickle-Cell Disease: a strategy for the WHO African Region*. Malabo, Equatorial Guinea.

Relethford, J. H. (2008) 'Genetic evidence and the modern human origins debate', *Heredity*. Nature Publishing Group, pp. 555–563. doi: 10.1038/hdy.2008.14.

Richards, S. *et al.* (2015) 'Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genetics in Medicine*. Nature Publishing Group, 17(5), pp. 405–424. doi: 10.1038/gim.2015.30.

Richter, D. *et al.* (2017) 'The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age', *Nature*. Nature Publishing Group, 546(7657), pp. 293–296. doi: 10.1038/nature22335.

Rightmire, G. P. (1998) 'Human evolution in the middle pleistocene: The role of homo heidelbergensis', *Evolutionary Anthropology*. John Wiley & Sons, Ltd, 6(6), pp. 218–227. doi: 10.1002/(SICI)1520-6505(1998)6:6<218::AID-EVAN4>3.0.CO;2-6.

Rightmire, G. P. (2009) 'Middle and later Pleistocene hominins in Africa and Southwest Asia', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, pp. 16046–16050. doi: 10.1073/pnas.0903930106.

Roberts, D. F. and Cavalli-Sforza, L. L. (1996) 'The History and Geography of Human Genes.', *The Journal of the Royal Anthropological Institute*, 2(1). doi: 10.2307/3034645.

Rosa, A. *et al.* (2004) 'MtDNA Profile of West Africa Guineans: Towards a Better Understanding of the Senegambia Region', *Annals of Human Genetics*, 68(4), pp. 340–352. doi: 10.1046/j.1529-8817.2004.00100.x.

Rosenberg, N. A. *et al.* (2002) 'Genetic structure of human populations', *Science*. American

Association for the Advancement of Science, 298(5602), pp. 2381–2385. doi: 10.1126/science.1078311.

Rossum, G. Van (2019) *PEP 8 -- Style Guide for Python Code | Python.org*, *python.org*. Available at: https://www.python.org/dev/peps/pep-0008/ (Accessed: 19 October 2020).

Rotimi, C. N. *et al.* (2017) 'The genomic landscape of African populations in health and disease', *Human Molecular Genetics*, 26(R2), pp. R225–R236. doi: 10.1093/hmg/ddx253.

Scerri, E. M. L. *et al.* (2018) 'Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter?', *Trends in ecology & evolution*. Elsevier Ltd, 33(8), pp. 582–594. doi: 10.1016/j.tree.2018.05.005.

Schlebusch, C. M. *et al.* (2017) 'Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago', *Science*, 358(6363), pp. 652–655. doi: 10.1126/science.aao6266.

Schoetensack, O. (1908) 'Der Unterkiefer des Homo Heidelbergensis aus den Sanden von Mauer bei Heidelberg. Ein Beitrag zur Paläontologie des Menschen', *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, 1(1), pp. 408–410. doi: 10.1007/bf01990624.

Scozzari, R. *et al.* (2012) 'Molecular Dissection of the Basal Clades in the Human Y Chromosome Phylogenetic Tree', *PLoS ONE*, 7(11). doi: 10.1371/journal.pone.0049170.

Sherman, R. M. *et al.* (2019) 'Assembly of a pan-genome from deep sequencing of 910 humans of African descent', *Nature Genetics*. Nature Publishing Group, pp. 30–35. doi: 10.1038/s41588-018-0273-y.

Sherry, S. T. *et al.* (2001) 'DbSNP: The NCBI database of genetic variation', *Nucleic Acids Research*. doi: 10.1093/nar/29.1.308.

Sherwani, S. I. *et al.* (2016) 'Significance of HbA1c test in diagnosis and prognosis of diabetic patients', *Biomarker Insights*. Libertas Academica Ltd., pp. 95–104. doi: 10.4137/Bmi.s38440.

Simarro, P. *et al.* (2014) 'Epidemiology of human African trypanosomiasis', *Clinical Epidemiology*. Dove Medical Press Ltd, p. 257. doi: 10.2147/CLEP.S39728.

Sirugo, G., Williams, S. M. and Tishkoff, S. A. (2019) 'The Missing Diversity in Human Genetic Studies', *Cell*, pp. 26–31. doi: 10.1016/j.cell.2019.02.048.

Slatkin, M. (2008) 'Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future', *Nature Reviews Genetics*. NIH Public Access, pp. 477–485. doi:

10.1038/nrg2361.

Smith, B. *et al.* (2007) 'The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration', *Nature Biotechnology*. Nature Publishing Group, pp. 1251–1255. doi: 10.1038/nbt1346.

Song, J., Zhang, M. and Xie, H. (2019) 'Design and implementation of a Vue.js-based college teaching system', *International Journal of Emerging Technologies in Learning*, 14(13), pp. 59–69. doi: 10.3991/ijet.v14i13.10709.

Southwood, D. and Ranganathan, S. (2019) 'Genome Databases and Browsers', in *Encyclopedia of Bioinformatics and Computational Biology*, pp. 251–256. doi: 10.1016/b978-0-12-809633-8.20754-1.

'Standardizing data' (2008) *Nature Cell Biology*, pp. 1123–1124. doi: 10.1038/ncb1008-1123.

Stringer, C. (2002) 'Modern human origins: Progress and prospects', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 357(1420), pp. 563–579. doi: 10.1098/rstb.2001.1057.

Stringer, C. (2012a) 'Evolution: What makes a modern human', *Nature*. Nature Publishing Group, pp. 33–35. doi: 10.1038/485033a.

Stringer, C. (2012b) 'The status of Homo heidelbergensis (Schoetensack 1908)', *Evolutionary Anthropology*. John Wiley & Sons, Ltd, 21(3), pp. 101–107. doi: 10.1002/evan.21311.

Stringer, C. (2016) 'The origin and evolution of homo sapiens', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1698). doi: 10.1098/rstb.2015.0237.

Stringer, C. B. and Andrews, P. (1988) 'Genetic and fossil evidence for the origin of modern humans', *Science*. American Association for the Advancement of Science, 239(4846), pp. 1263–1268. doi: 10.1126/science.3125610.

Sudmant, P. H. *et al.* (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*. Nature Publishing Group, 526(7571), pp. 75–81. doi: 10.1038/nature15394.

Sumaili, E. K. *et al.* (2018) 'G1 is the major APOL1 risk allele for hypertension-attributed nephropathy in Central Africa', *Clinical Kidney Journal*, 12(2), pp. 188–195. doi: 10.1093/ckj/sfy073.

Tattersall, I. (1986) 'Species recognition in human paleontology', *Journal of Human Evolution*. Academic Press, 15(3), pp. 165–175. doi: 10.1016/S0047-2484(86)80043-4.

Thomson, R. *et al.* (2014) 'Evolution of the primate trypanolytic factor APOL1', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(20), p. E2130. doi: 10.1073/pnas.1400699111.

Torres, F., Barbosa, M. and Maciel, P. (2015) 'Recurrent copy number variations as risk factors for neurodevelopmental disorders: Critical overview and analysis of clinical implications', *Journal of Medical Genetics*. BMJ Publishing Group, 53(2), pp. 73–90. doi: 10.1136/jmedgenet-2015-103366.

Trampuz, A. *et al.* (2003) 'Clinical review: Severe malaria', *Critical Care*. BioMed Central, pp. 315–323. doi: 10.1186/cc2183.

Tryka, K. A. *et al.* (2014) 'NCBI's database of genotypes and phenotypes: DbGaP', *Nucleic Acids Research*. Oxford Academic, 42(D1), pp. D975–D979. doi: 10.1093/nar/gkt1211.

Tsaras, G. *et al.* (2009) 'Complications Associated with Sickle Cell Trait: A Brief Narrative Review', *American Journal of Medicine*. Elsevier, pp. 507–512. doi: 10.1016/j.amjmed.2008.12.020.

Tuohy, C. V. *et al.* (2020) 'Hypertrophic cardiomyopathy: the future of treatment', *European Journal of Heart Failure*. John Wiley and Sons Ltd, pp. 228–240. doi: 10.1002/ejhf.1715.

Tuttle, R. H. (2020) *Human Evolution: The emergence of Homo sapiens*, *Encyclopædia Britannica*. Available at: https://www.britannica.com/science/human-evolution (Accessed: 13 July 2020).

US Department of Defense Systems Management College (2001) 'Systems Engineering Fundamentals', *22060-5565*, (January), p. 222. Available at: http://ocw.mit.edu/courses/aeronautics-and-astronautics/16-885j-aircraft-systems-engineering-fall-2005/readings/sefguide_01_01.pdf.

Vue.js org (2020) *Comparison with Other Frameworks - Runtime Performance*. Available at: https://vuejs.org/v2/guide/comparison.html#Runtime-Performance (Accessed: 12 October 2020).

WebComponents.org (2020) *Specifications - WebComponents.org*. Available at: https://www.webcomponents.org/introduction#specifications (Accessed: 12 October 2020).

Weir, B. S. and Cockerham, C. C. (1984) 'Estimating F-Statistics for the Analysis of Population Structure', *Evolution*. JSTOR, 38(6), p. 1358. doi: 10.2307/2408641.

Weischenfeldt, J. *et al.* (2013) 'Phenotypic impact of genomic structural variation: insights from and for human disease', *Nature Reviews Genetics*. Nature Publishing Group, 14(2), pp. 125–138. doi: 10.1038/nrg3373.

Wetterstrand, K. (2019) *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*, *NHGRI*. Available at: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (Accessed: 24 July 2020).

Wheeler, E. *et al.* (2017) 'Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis', *PLoS Medicine*. Edited by E. Gregg. Public Library of Science, 14(9), p. e1002383. doi: 10.1371/journal.pmed.1002383.

WHO (2020a) *Malaria Fact Sheet*. Available at: https://www.who.int/news-room/fact-sheets/detail/malaria (Accessed: 30 April 2020).

WHO (2020b) *Trypanosomiasis, human African (sleeping sickness)*. Available at: https://www.who.int/en/news-room/fact-sheets/detail/trypanosomiasis-human-african-(sleeping-sickness) (Accessed: 26 May 2020).

Williams, T. N. *et al.* (2005) 'Sickle Cell Trait and the Risk of Plasmodium falciparum Malaria and Other Childhood Diseases', *The Journal of Infectious Diseases*. Oxford University Press (OUP), 192(1), pp. 178–186. doi: 10.1086/430744.

Willoughby, P. R. (2007) *The evolution of modern humans in Africa*. AltaMira Press.

Wright, S. (1949) 'The genetical structure of populations', *Annals of Eugenics*. John Wiley & Sons, Ltd, 15(1), pp. 323–354. doi: 10.1111/j.1469-1809.1949.tb02451.x.

Wu, C. and Buyya, R. (2015) 'Business Needs', in *Cloud Data Centers and Cost Modeling*. Elsevier, pp. 43–95. doi: 10.1016/b978-0-12-801413-4.00002-7.

Wu, K. J. (2019) *Lack of diversity in genetic research could be costing us our health*, *NOVA*. Available at: https://www.pbs.org/wgbh/nova/article/lack-diversity-genetic-research-could-be-costing-us-our-health/ (Accessed: 30 October 2020).

Yates, A. D. *et al.* (2020) 'Ensembl 2020', *Nucleic Acids Research*. Oxford University Press, 48(D1), pp. D682–D688. doi: 10.1093/nar/gkz966.

Zhang, F. *et al.* (2009) 'Copy Number Variation in Human Health, Disease, and Evolution', *Annual Review of Genomics and Human Genetics*. Annual Reviews, 10(1), pp. 451–481. doi:

10.1146/annurev.genom.9.081307.164217.

Zhivotovsky, L. A., Rosenberg, N. A. and Feldman, M. W. (2003) ‘Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers’, *American Journal of Human Genetics*. Elsevier, 72(5), pp. 1171–1186. doi: 10.1086/375120.

Appendices

Supplementary material can be accessed from the following external URL:

https://drive.google.com/drive/folders/17VmM1HYEbejKbcr0P5L0cbDEe7KPn_Cx?usp=sharing

Appendix A – PubMed literature search

Description: *Literature review of African genomic studies*

Supplementary file name: *Appendix A - PubMed Literature Search.xlsx*

Appendix B – AGVD User Requirements Documentation

Description: *Application design document of the AGVD; lists the User Scenarios, User Requirements and Functional Requirements generated during the design phase of the project*

Supplementary file name: *Appendix B - AGVD User Requirements Documentation.docx*
