# Creating and analysing an African pan-genome

Jessica Jean Bourn

Supervisor: Professor Nicola Mulder


**Submitted for the degree of Master of Science in Medicine specialising in Bioinformatics**

Division of Computational Biology

Department of Integrative Biomedical Sciences

Faculty of Health Sciences

University of Cape Town

*14 May 2022*

# Declaration

I, Jessica Jean Bourn, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

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

Signed by candidate

Signature: ……

Date: 14 May 2022

# Acknowledgements

Goodness gracious me, I can't believe I made it to the end. What a strange two years it has been.

First and foremost, I would of course like to thank my supervisor, Professor Nicky Mulder, for piquing my interest in the project in the first place, for letting me run with it in my own directions, for the advice when I needed it, and for the knowledge she instilled in me over two years. A huge thank you also for funding the second year of my degree through H3ABioNet, without which I might not have completed my degree at all.

I also need to thank Shaun Aron, Chris Fields and Gerrit Botha, who I had the pleasure of working with through H3ABioNet. Although they were not directly involved in my research, their knowledge of computational biology and the field of human genomics that they shared with me was immense, and massively influenced how I undertook my research. They all taught me so much, however indirectly, and I will forever be grateful to them for trusting me to be part of their team.

I must also thank Gerrit for his help in coming to grips with a high-performance computing cluster; his guidance in the first few months of my degree was a blessing. Thanks must also go to Dane Kennedy from the Ilifu high performance computing facility for his quick replies, simple explanations, and his interest in my research when I was having problems. Similarly, thank you to the entire Ilifu group and the resources they provide – the cluster has become my sort of safe space for logic and efficiency, and the computational biologist in me is going to miss it immensely.

For financing my degree in addition to Nicky, thank you to UCT and the National Research Foundation.

And lastly, I of course have to thank all the people in my life who helped me get through these last two Covid-filled years working from home. Mom-a-loms and pop-a-lops, thank you for believing in me and letting me complain about my degree whenever I came round to steal your food and alcohol. Pops, I appreciate your input to my dissertation so much – you are my hero. I promise to pay you back when I get a real job. To Leo, thank you for putting up with my grumpiness and making me laugh even when I didn't want to. Thank you also to Jony and Jeremy, for all the years of fun, but mostly for distracting me from my write-up when I needed it. And finally, thank you to all the animals in my life – all 13 of my foster dogs (Slim, Melon, George, Beethoven, Deka, Zena, Bibi, Flint, Casey, Luna, Teddy, Gino and Yoda) and my cats Marmite and Fig. They were very often the only reason I got out of bed in the morning, and I can't even begin to express my love for all of them in words.

Onwards and upwards!

# Abstract

The human reference genome is currently a core resource for understanding the role of genetics in human health, disease, and variation, and has been invaluable in the development of clinical and computational tools for these purposes. However, the limited number of individual genomes used to create the reference has resulted in an underrepresentation of the extensive genetic diversity present in different human populations. Since an important use of the reference genome is to identify genetic variants that may be implicated in disease, this lack of diversity could limit the scientific utility of the reference for ethnic groups that are poorly represented in it. As a result, adaptations to the reference genome structure have been proposed. One such proposal has been the use of multiple reference genomes, each of which represent different human populations. A logical and highly practical method of achieving this is through the use of a pan-genome, which is a curated collection of all the DNA sequences that are found within a population under study. Despite the fact that African populations exhibit the greatest genetic diversity and variation in the world, the many and sometimes ancient ethnolinguistic groups from Africa are among those least represented within the reference genome. Consequently, this study aimed to explore the feasibility of creating and analysing an African pan-genome, and to begin developing tools to achieve this.

Several distinct African regional ancestral groups – namely east African Nilo-Saharan, east African Afro-Asiatic, far west Niger-Congo, central west Niger-Congo, Bantu-speaking Niger-Congo, central African rainforest hunter-gatherer, and the Khoe and San – have previously been identified, and this study included and analysed samples from each group in order to assemble a more inclusive and representative pan-genome. A software pipeline developed by Duan *et al.* (2019), termed the HUman Pan-genome ANalysis (HUPAN) pipeline, was used here to assemble the African pan-genome. As the HUPAN pipeline was originally designed to analyse only single populations, the inclusion of multiple populations required modifications and improvements, which were implemented following the testing and analysis of the pipeline using a smaller dataset of whole genome sequences. Subsequently, a final dataset of 168 African high- and medium-coverage whole genome sequences representing the seven separate regional ancestral groups was submitted to the adapted HUPAN pipeline. For each group, nucleotide sequences that were absent from the human reference genome were assembled and extracted, which resulted in the identification of 43.37 Mbp of non-redundant non-reference genomic sequence and 31 novel predicted protein-coding genes from African individuals. Alignment to other pan-genome sequences, whole genomes from different human populations, and the complete telomere-to-telomere human genome validated a large portion of the sequences as non-reference and confirmed that the dataset contained sequences specific to African populations.

However, the gene presence-absence variation analysis of the pan-genome within all 168 samples revealed patterns of gene presence and absence that were strongly correlated to the sample dataset of origin, rather than to the ancestral group of origin. This hindered the identification of genuine genetic variation specific to the groups analysed. Further, it appears that previous pan-genomic research has not investigated the degree to which the genetic variation identified is dataset-specific or truly population-specific. Consequently, the failure to acknowledge and account for the effects of spurious inter-dataset variation in previous pan-genomic research indicates that those analyses may be incomplete or ambiguous. This, therefore, calls into question the methods currently used for pan-genomic research, and highlights that robust, standardised methods for human pan-genome research must be agreed on to ensure that comprehensive population-specific pan-genomes are produced in the future.

Despite this inherent weakness of pan-genomic research, this study successfully enabled the creation and analysis of a comprehensive and inclusive African pan-genome. Unique sets of non-reference sequences specific to African regional ancestral groups were identified and obtained, enabling the assembly of a non-redundant set of pan-African non-reference sequences. Furthermore, certain complex but previously unconsidered aspects of pan-genome research were identified and explored, and these observations may play a role in the advancement of pan-genome research in future.

# Preface

The human reference genome has been essential for the analysis and understanding of the genetic causes of human health, disease and variation in recent decades. However, certain limitations of the reference genome have prompted investigations into population-specific pan-genomes and their utility as tools within human genomics research. This dissertation explores the creation of an African pan-genome using an adapted version of the HUman Pan-genome ANalysis (HUPAN) pipeline developed by Duan *et al.* (2019). This research has resulted in the first pan-genome representing the African continent and has provided insights into the process and requirements of assembling a pan-genome that had potentially not previously been discussed.

The first chapter contains a literature review which explores two overarching topics. The first topic examines both the reasons for and the effects of the limitations of the current human reference genome. It then summarises various methods and adaptations that have been proposed to address these limitations and concludes with a focus on pan-genomes. The second topic explores the lack of genomic research on the African continent. Both the reasons for and the effects of the exclusion of African individuals from genomic research are analysed.

The second chapter details the setting up and testing of the HUPAN pipeline using a sample dataset of 95 African whole genome sequences from the 1000 Genomes Project dataset. The test dataset enabled the identification of certain weaknesses within the pipeline that were consequently improved, and adaptations to ensure that samples from multiple human populations could be easily submitted to the pipeline were developed for future research.

The third chapter details the assembly of an African pan-genome using the adapted HUPAN pipeline. Using 168 samples from seven diverse regional ancestral groups represented in Africa, copious amounts of DNA not present in the reference genome was extracted, analysed for population-specific differences, and compared to other genomic datasets. This work not only provided insight into the analyses and conclusions drawn from other pan-genomic research, but also revealed important aspects essential for the creation of pan-genomes that have potentially been previously unconsidered.

The fourth chapter summarises the findings from this research and explores what future work might be needed to further advance and improve the African pan-genome.

This study was approved by the Human Research Ethics Committee (HREC) of the University of Cape Town, South Africa (HREC reference number: 023/2021).

# Contents

# List of common abbreviations

**1kGP:** 1000 Genomes Project

**BantuNC:** Bantu-speaking Niger-Congo

**bp:** base pairs

**CARF:** central African rainforest hunter-gatherer

**CDS:** coding DNA sequence

**contig:** contiguous sequence

**CWNC:** central west Niger-Congo

**EAAA:** east African Afro-Asiatic

**EANS:** east African Nilo-Saharan

**EGA:** European Genome-phenome Archive

**EHC:** Ethiopian high coverage

**ELC:** Ethiopian low coverage

**FWNC:** far west Niger-Congo

**Gbp:** gigabase pairs

**GRCh38:** build 38 of the human reference genome

**H3Africa:** Human Heredity and Health in Africa Consortium

**HGDP:** Human Genome Diversity Project

**HUPAN:** HUman Pan-genome ANalysis

**IGSR:** International Genome Sample Resource

**indels:** insertions and deletions

**kbp:** kilobase pairs

**KhoeSan:** Khoe and San

**KSP:** Khoe-San Project

**Mbp:** megabase pairs

**NCBI:** The National Centre for Biotechnology Information

**PAV:** presence-absence variation

**PCA:** principal component analysis

**RNA-seq:** RNA sequencing

**SGA:** String Graph Assembler

**SGDP:** Simons Genome Diversity Project

**SNP:** single nucleotide polymorphism

**SNV:** single nucleotide variant

**T2T-CHM13:** telomere-to-telomere complete hydatidiform mole

**TrypanoGEN:** Trypanosomiasis Genomics Network of the H3Africa Consortium

# 1. Literature review: human pan-genomes

## 1.1. The human reference genome

The completion of the Human Genome Project was a momentous achievement in the development of genomic technology that resulted in the first draft sequence of the human genome (Venter et al., 2001). The sequence was termed the human reference genome, and the construction and release of this first draft created novel opportunities for human genomic research, as it provided a standardised reference that could be used to compare and analyse human genetic information. The initial sequence covered around 94% of the human genome (Lander et al., 2001) and subsequent releases have resulted in notable improvements. The most recent release by the Genome Reference Consortium in 2013, termed GRCh38 as it was the 38$^{th}$ build, corrected nucleotides and misassembled regions, added missing sequences to leave fewer than 1000 gaps in the sequence and, importantly, increased the diversity of the reference. It is purportedly the "most accurately sequenced human genome in the world (Ballouz et al., 2019; Guo et al., 2017).

Despite this, further improvements to the human reference genome are essential. The reference is the foundation for understanding the role of genetics in human health, disease and variation, and has been invaluable in the development of clinical and computational tools for these purposes (Rosenfeld et al., 2012; U.S. Department of Energy Office of Science, 2008). It forms the basis of our ability to identify genetic diseases through the comparison of healthy and diseased individuals' genomes. Further, it is the future of personalised human genomics as it could result in diagnoses and clinical interventions that are specific to individuals based on their genomic sequences (Gonzaga-Jauregui et al., 2012). The reference genome has also enhanced the field of paleogenomics, resulting in important insights into ancient human migrations, ancestral relatedness between different groups and the evolution of population-specific phenotypes (Günther & Nettelblad, 2019). The reference genome's role as a universal coordinate system that provides standardised annotations and a frame of reference is also vital to most human genomics research (Novak et al., 2017). However, in order to maintain and further improve its crucial role in the field of human genomics, the limitations of the human reference genome and the techniques it is used for must be constantly considered, and some aspects may need to be re-evaluated.

### 1.1.1. Limitations of the human reference genome

According to The Genome Reference Consortium (n.d.) documentation on GRCh38, 70% of the reference sequence was obtained from a single male, likely of mixed African-American and European

decent. A further 23% of the sequence was obtained from only 10 individuals and the remaining 7% from over 50 samples. Yet with such a limited number of individuals contributing to the reference sequence, there can be no doubt that the entirety of the human population's genetic variation is inadequately represented, and attempts to do so may be impractical. In addition, the inability of certain sequences in individual genomes to map to the reference suggests that these sequences may belong to known gap regions of the reference (Kidd et al., 2010; R. Li et al., 2010). These aspects confirm the incompleteness of the reference genome and highlight its shortcomings as a standardised reference. With invaluable amounts of genetic research being based on the notion that the reference is the standard human genome, understanding the effects of its limited diversity is of paramount importance.

Ballouz et al. (2019) have suggested that, because of the small number of contributing samples, the reference genome is more akin to a haploid personal genome than to a human population-wide consensus genome. This is because genetic loci and alleles within the reference were not purposefully chosen as the most common to the human population by any metric, but were rather joined together randomly from multiple individuals' personal genomes. Resultingly, the human reference genome can be considered a mosaic of personal genomes that does not necessarily represent any population or even one specific individual (Ballouz et al., 2019). Further, although the individuals used to assemble the reference genome were apparently healthy, there is the possibility that they carried rare or even disease-associated alleles which were subsequently incorporated into the reference (Magi et al., 2015). This would clearly affect its function as a standard for all human genomic research. For example, a crucial use of the reference is in the alignment and mapping of genomic reads from individual genomes to provide genetic context to the sequences. However, due to the reference genome's known missing sequences, its lack of genetic diversity, and because of the presence of rare alleles within it, mapping sequences to the reference genome introduces a weakness termed reference allele bias (Novak et al., 2017; Paten et al., 2017).

Reference allele bias is the highly prevalent and well-studied tendency to disfavour sequences and alleles that are not present in the reference, thereby over-representing data that more closely match the reference genome (Ballouz et al., 2019; Paten et al., 2017). In terms of read alignment, this means that sequences with alleles that differ from the reference have a lower chance of being aligned correctly. This can have confounding effects on downstream analyses, such as in the measurement of allele-specific expression of genes using RNA-sequencing data, which is known to favour reads matching to the reference allele and results in a significant overestimation of the expression of these alleles (Stevenson et al., 2013). Additionally, reference allele bias has a notable effect on variant calling

and discovery (Ros-Freixedes et al., 2018). Here, sequences are aligned and compared to the reference in order to identify genetic variation, usually in the context of a disease (Hoffman-Andrews, 2017). This process allows the identification of single nucleotide variants and polymorphisms (SNVs and SNPs), small insertions and deletions (indels) and larger genetic structural variants of more than 50 base pairs (bp) such as inversions, translocations and copy number variants (Collins et al., 2020; Magi et al., 2015). However, owing to the makeup of the reference genome, it is not always possible to determine whether these genetic differences are in fact relevant to the phenotype in question. There is strong evidence that the reference contains a considerable number of rare alleles, which will result in the incorrect calling of variants that may be commonly present in a population. Conversely, rare variants might not be discovered if they are shared by the reference genome, thereby preventing the identification of variants that are implicated in genetic disease (Ballouz et al., 2019; Magi et al., 2015). This potentially hinders our ability to convert genetic knowledge into medically relevant genetic therapeutics or interventions (Sirugo et al., 2019). This observation is of major relevance with regard to different human subpopulations, as population-specific genetic variation is well established; thus, the lack of genetic diversity could potentially be limiting the scientific utility of the reference for ethnic groups that are poorly represented in it (Kidd et al., 2008).

### 1.1.2. Efforts to supplement the human reference genome

Following the recognition of the human reference genome's limitations, initiatives such as the 1000 Genomes Project, the 100,000 Genomes Project, and the Human Pangenome Project were born, in order to systematically identify and document human genetic variation across multiple populations (The 100,000 Genomes Project Pilot Investigators, 2021; The 1000 Genomes Project Consortium, 2015; T. Wang et al., 2022). Accordingly, in recent years, the 1000 Genomes Project Consortium (2015) stated that they had characterised over 88 million genetic variants. These included 84.7 million SNPs, 3.6 million small indels, and 60 000 large structural variants, which more than doubled the genetic variation the same group had identified in 2012. The above genetic variant databases, among others, have been partially used in the most recent genome assembly releases, GRCh37 and GRCh38, which have included some representation of larger structural variants in the form of alternate haplotype sequences (Novak et al., 2017; Paten et al., 2014). These are described relative to the primary sequence by mapping both ends to loci on the reference but presenting an alternate sequence in the middle (Paten et al., 2017). Yet despite the comprehensive catalogues of human variation and the tentative inclusion of some of it in the reference, there remain difficulties in representing variation within the reference genome structure.

Currently, the alternate sequences included in the reference genome represent a very small number of known variants. Including all the genetic variants with a confirmed prevalence of more than 1% would mean the addition of alternate haplotype sequences over the entire reference genome, with hundreds of these sequences overlapping single genetic locations (Paten et al., 2017). In addition to its limited practicality, such a multi-branched – or "scaffold"– structure also means that the reference is no longer a single haploid sequence. This in turn affects the way that current sequence alignment and analysis tools function, as most were designed to map to a haploid reference and are therefore misled when sequences align to multiple locations over alternate loci (Church et al., 2015).

The problem of variant representation is significantly less challenging for smaller variants such as SNPs and indels of less than 50 bp. Large databases of these genetic variants such as dbSNP (Sherry et al., 2001) can be used in conjunction with the reference genome and included in genome analyses using SNP-aware aligners. However, there is no such similar solution for larger structural variants (Sherman et al., 2019; Sherman & Salzberg, 2020). Early research focused on small variants such as SNPs, as it was assumed that their abundance resulted in their being the most significant source of genetic diversity. However, later studies predicted that larger structural variants played a more important role than previously believed (Feuk et al., 2006). It is now evident that structural variants are as important as SNPs in their contribution to genetic variation, and that they encompass more polymorphic base pairs than single and small nucleotide variants (Church et al., 2011; Kidd et al., 2008). Yet there are numerous difficulties in identifying and genotyping structural variants. They are hard to characterise using short-read whole genome sequencing methods because they are often found within repetitive sequences of DNA that cannot be spanned by short reads, and because they have a large range of sizes and types (Alkan et al., 2011; Antaki et al., 2018). Long-read whole genome sequencing methods somewhat overcome these issues as repetitive and long regions of genomes are able to be resolved by the larger read length. However, the lower efficiency and higher cost of this approach currently limit its practicality (Kosugi et al., 2019). Consequently, and until long-read sequenced databases become more widely accessible, we must aim to address the lack of genetic diversity within the reference by considering and further developing methods that utilise short-read sequencing technologies.

Another significant concern regarding the reference genome is the use of a linear coordinate system to describe the position of each base in the sequence. Prior to the addition of the alternate sequences in GRCh37, each chromosome had only one linear sequence or "genetic path" from beginning to end which provided a simple and efficient way of labelling bases and genes. However, as the reference is currently represented, coordinates that fall within regions containing alternate sequences will be

ambiguous, as there will be multiple possible paths to get to each coordinate (Rand et al., 2017). This means that there is currently no standard process for representing and describing genes that fall within an alternate sequence.

### 1.1.3. Adaptations to the reference genome

To begin addressing the known limitations of the reference genome, various research groups have proposed adaptations to it. A number of these improvements would require only minor changes to the reference genome structure, while others would necessitate massive modifications to both the reference and the tools and techniques that use it. Examples of small scale changes that have been proposed include masked, enhanced or diploid references, all of which could reduce reference allele bias and improve mapping and alignment to the reference (Ballouz et al., 2019). A masked reference is one in which repetitive and low-complexity DNA or rare alleles in the reference are effectively hidden from alignment algorithms, thereby increasing the likelihood that short reads will align to the correct positions without being misled by the repetitive regions (Frith, 2011). Enhanced and diploid genomes insert additional sequences into the reference; these sequences can be the alternate alleles at known SNP and indel loci or larger structural variants (Satya et al., 2012). Multiple research groups have shown that this can also be achieved with personalised reference genomes, which incorporate genetic variants that were previously discovered in the samples under study (Groza et al., 2020; Yuan & Qin, 2012). However, adaptations such as these are not comprehensive solutions. Masking repetitive regions of DNA effectively leads to blindness in a significant portion of the human genome and greatly devalues the evaluation of gene expression (Slotkin, 2018). Similarly, enhanced and diploid genomes are limited to incorporating only known small genetic variants and act as complements to the current reference rather than solutions to its limitations (Magi et al., 2015; Satya et al., 2012). In recent years, short-term and incomplete fixes such as these have been overshadowed by proposals for multiple reference genomes that each represent a different human population.

#### *1.1.3.1. Population-specific reference genomes*

The 1000 Genomes Project Consortium (2015), in their bid to identify and catalogue all known human genetic variation, sequenced and assembled the genomes of over 2000 individuals from 26 different populations. Their results demonstrated and ultimately confirmed that the genetic variation between human populations is extensive. Notably, when comparing the collected variation to the reference genome, they found that 86% of non-reference variants were found in only single continental groups. Clearly, assembling reference sequences for different human subpopulations would drastically improve the diversity that can be represented. To assess this, draft reference genomes for Chinese,

Korean and Danish populations, among many others, have already been assembled (Cho et al., 2016; Maretty et al., 2017; Shi et al., 2016), but none are yet as comprehensive or high-quality as the current reference genome (Schneider et al., 2017). A further limitation of many of these subpopulation references is that they are assembled from only a single individual, examples being the Chinese and Ashkenazi genomes (Shi et al., 2016; Shumate et al., 2020). Thus, although these references may be more specific for their respective populations, representation of genetic variation within that population is not drastically improved and the reference can still be considered a haploid personal genome. Other population reference genomes have made use of multiple individuals in an attempt to incorporate more population-specific variation. Examples include a Korean reference genome, which has a consensus variome to complement the reference (Cho et al., 2016), a Danish reference genome which is made up of a collection of 150 *de novo* assembled genomes (Maretty et al., 2017), and a set of major allele reference sequences for multiple populations created using 1000 Genomes Project data (Dewey et al., 2011). However, if total variation in a population is to be simply represented in a reference genome, then consensus sequences or collections of multiple unconnected genomes are not comprehensive solutions.

#### *1.1.3.2. Pan-genomes*

A pan-genome is another reference genome format that is able to represent population-specific variation. Unlike other reference structures, which use consensus sequences or provide partial representation of genetic variation for a population, a pan-genome is a collection of all the DNA sequences that are found within the population under study, as depicted in Figure 1.1 (R. Li et al., 2010; Sherman et al., 2019; Sherman & Salzberg, 2020). They are non-redundant and composed of the core genome, which represents the DNA sequences that every individual in the population shares, and the accessory or distributed genome, which represents the sequences that are only found within subgroups or even in single individuals within the population (Duan et al., 2019; Marroni et al., 2014). Importantly, with the inclusion of the distributed genome, a pan-genome can enable the representation of not only population-specific variation, but also individual-specific variation within that population.

The potential utility of a human pan-genome was first analysed by Li *et al*. (2010) using one Asian and one African individual genome; when compared to the then-current human reference genome (build 36), they found that around 5 megabase pairs (Mbp) of DNA from the *de novo* assembled genomes was not present within the reference. They also estimated that a pan-genome for the full human population would contain around 19 to 40 Mbp of potentially novel DNA not found within the reference. More recent studies appear to show that this was a notable underestimation, as a

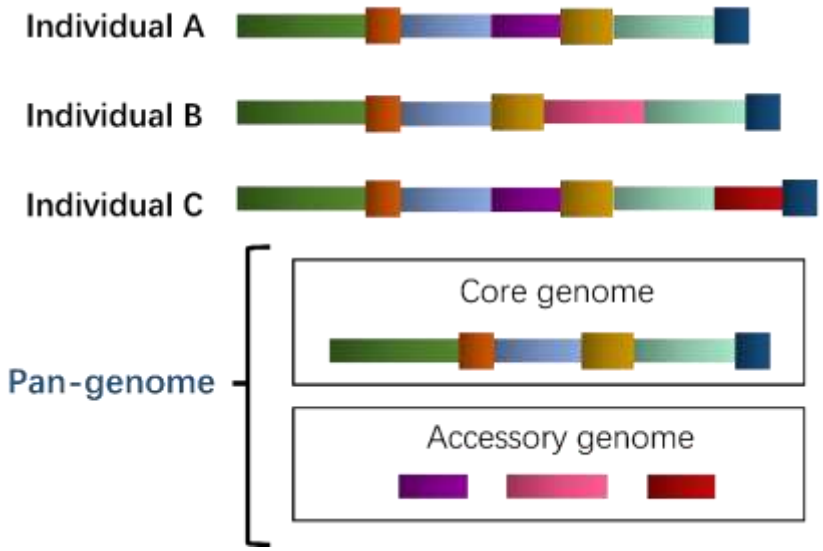


**Figure 1.1. Representation of a pan-genome, which is made up of the core and accessory genomes.** A pan-genome is an ordered combination of all the DNA sequences present in every individual that was included in the creation of the pan-genome. Adapted from Sherman & Salzberg (2020).

subsequent study estimated that a single theoretical human diploid genome could have up to 16 Mbp of additional DNA when compared to the reference (Huddleston et al., 2017), meaning a full population pan-genome would likely contain significant amounts of sequence not found in the reference. Recently, due to the growing abundance of human whole genome sequences, multiple human pan-genomes have been assembled for specific populations. Li *et al.* (2021) used 486 Han Chinese individuals to obtain over 270 Mbp of novel non-reference sequence, while Wong *et al*. (2020) used 338 human whole genome assemblies from diverse populations to identify over 120 000 novel insertions that were then integrated into the human reference genome. This improved reference, termed the Human Diversity reference, subsequently allowed hundreds of thousands of previously unmapped reads from single genomes to map back to the reference, thereby presenting the possibility of improved read mapping, genome assembly and variant calling. Sherman *et al.* (2019) created an African-descent pan-genome using 910 individuals and discovered almost 300 Mbp of novel DNA not found within the reference; notably, only 81 Mbp of the novel sequence was present in more than one individual. This indicates that most of the non-reference DNA identified was individual-specific, demonstrating one advantage of the pan-genome structure.

A common method for constructing a pan-genome is that of anchored pseudo *de novo* assembly (Faber-Hammond & Brown, 2016; Sherman et al., 2019). Here, any reads from previously sequenced genomes that can successfully map to the reference genome are discarded, and only the remaining unmapped reads are *de novo* assembled. These assembled reads are then incorporated into the already completed sequences of the human reference genome, allowing the discovery and inclusion

of large structural variation. This method is advantageous in that it can produce a human pan-genome using the foundation of the human reference genome and so has computational demands that are readily achievable (Sherman et al., 2019). An alternative method of pan-genome production is the *de novo* assembly of many whole genomes, which are then constructed into pan-genome sequences based on available reference genomes and previously assembled sequences. Such a method has been used by Hu *et al.* (2017) to create EUPAN, a eukaryotic pan-genome analysis pipeline that was used to effectively analyse over 450 rice genomes. However, the EUPAN method cannot be applied to the human pan-genome, as the immense size of the human genome (3.1 gigabase pairs) means the *de novo* assembly of many human genomes has massive computational demands. Hundreds of gigabytes of memory are required to assemble even a single human genome (C. Ye et al., 2012), which is prohibitively expensive in terms of computational power and time. As a result, Duan *et al.* (2019) have developed a modified pipeline called HUPAN (HUman Pan-genome ANalysis) that uses a low-memory requirement program to *de novo* assemble human genomes and identifies non-reference sequences using both totally and partially unaligned genomic reads. This approach should enable the discovery of more non-reference sequences than would be possible using anchored pseudo *de novo* assembly, which is essential for the creation of a comprehensive pan-genome. Ultimately, however, the future of *de novo* assembly of human genomes on a population-scale lies within the development of long-read sequencing technologies (Miga & Wang, 2021), and a current goal of the field is to advance these technologies in accuracy and cost-effectiveness. Until this can be achieved, however, methods such as those discussed above are appropriate intermediates that will further enable important discoveries in human genomics.

Overall, the use of pan-genomes to represent genetic variation at a population scale is an entirely worthwhile pursuit. The lack of diversity in the reference should not be left unaddressed, and the pervasiveness of reference allele bias needs to be mitigated to ensure that reference-based genomic research is equally relevant for all human populations. Population-specific pan-genome references enable these advancements and, while the creation and implementation of an entirely novel reference model is a massive task, it is one that the genomics research community should feel compelled to undertake.

## 1.2. Genomic research in Africa

Since the first release of the human reference genome, a significant amount of research has gone into identifying and cataloguing human genetic variation. Through this research, it has been established that the African continent exhibits more genetic diversity and variation, both within and between populations, than any other in the world (Gurdasani et al., 2015; Reed & Tishkoff, 2006; The 1000

Genomes Project Consortium, 2015). This is consistent with the "Out of Africa" hypothesis, which proposes that modern humans have existed in Africa longer than in any other region and with few genetic bottlenecks, resulting in higher levels of genetic diversity (Tishkoff et al., 2009). However, despite the extensive and often ancient genetic variation within Africa, multiple analyses have highlighted that the majority of human genomics research has focussed on individuals of European descent. An early analysis of all catalogued genome-wide association studies (GWAS) by Need & Goldstein (2009) showed that 10 times as many GWAS were being performed on European ancestry populations than on all other populations combined, and that individuals of European descent made up 96% of all participants in GWAS word-wide. A follow up study in 2016 revealed that the percentage of non-European ancestry individuals in GWAS had increased from 4% to 19%, yet more than 70% of this growth was attributed to a massive increase in samples from Asian countries. This analysis also showed that, when combined, those of African, Latin American or Hispanic ancestry, and native or indigenous peoples contributed only 4% to all individuals analysed in GWAS (Popejoy & Fullerton, 2016). Landry *et al.* (2018) further reported that, of the 2817 GWAS from the Genome-Wide Association Study Catalog, 67% were based on those of European descent, 29% on Asian populations and only the remaining 4% focused on individuals from the other under-represented ethnicities and groups. Similarly, The 1000 Genomes Project Consortium (2015) spent years producing one of the most extensive and comprehensive catalogues of human genetic variation, yet only 7 out of an estimated 2000 African ethnolinguistic groups were included in the study (Reed & Tishkoff, 2006).

### 1.2.1. Reasons for the disparities in African genomics

There are numerous reasons for this population-level disparity in genomic data. The first and most apparent is the significant difference in the availability of research and clinical facilities in African countries compared to more developed nations in North America, Europe and Asia (The H3Africa Consortium et al., 2014). Well-established research facilities, which are predominantly based in developed western countries, tend to utilise groups and cohorts in those same geographical areas (Popejoy & Fullerton, 2016) due to cost reduction and ease of access. Further, datasets that have included African individuals have often focused more on ease of accessibility than equal representation, resulting in entire African ethnolinguistic groups being completely excluded from genomic databases (Rotimi et al., 2017). Other aspects, such as ongoing wars, a public mistrust for genomic research on non-European populations, limited funding opportunities for research in Africa, and a shortage of African scientists participating in genomic research, have also contributed to the considerable lack of African genomic data (Miga & Wang, 2021; Rotimi et al., 2017; The H3Africa Consortium et al., 2014).

However, socio-economic factors are not solely responsible in this regard. Within genomics research, the ability to discover variants of interest is significantly greater in exclusively European datasets because of the homogeneity of the genomes. When genetically homogenous populations are utilised in these types of analyses, population stratification and admixture play much smaller roles in the outcomes of the studies, which allows researchers to more confidently determine whether a variant is associated with the trait in question (Need & Goldstein, 2009). African populations are significantly less genetically homogenous and often have a greater degree of admixture, meaning they are more often excluded from genomic studies in favour of populations that require less intensive statistical calculations and corrections. In a similar vein, the reduced linkage disequilibrium in African populations – which occurs because anatomically modern humans have existed on the African continent for longer than in any other geographical area – further increases the complexity of genetic variant discovery in Africa (Campbell et al., 2014; Rotimi et al., 2017). Non-African populations are more likely to inherit entire sequences of linked variants due to genetic bottlenecks, and this means that knowing specific variants in an individual can reliably inform researchers of what many other nearby variants are likely to be (Need & Goldstein, 2009). However, the reduced linkage of variants in those of African descent makes it more challenging to fully genotype individuals, as the variants under study may only reliably indicate the state of a few nearby variants rather than entire blocks (The International HapMap Consortium, 2007). This means that the current method for variant discovery may be less applicable to African populations and will result in fewer variants being reliably discovered. Finally, studies that have examined genetic diversity within Africa have often focused on specific and very large populations or ethnolinguistic groups, and studies including smaller but no less important African populations have been limited by sample size (Gurdasani et al., 2015). Yet despite all of these factors, it is undoubtedly the responsibility of the genomics research community to overcome these obstacles so that human genomic research remains representative and accurate for all human populations irrespective of which continent their ancestors originated.

### 1.2.2. Consequences of the under-representation of African genomic data

As previously discussed, understanding and investigating human genetic variation is essential if we are to gain insight into human health and disease. When certain populations are excluded from these analyses, they are also excluded from the potential beneficial outcomes that may result from the research. One such relevant example of this was shown by Manrai *et al.* (2016), where the researchers clearly showed that multiple "causal variants" were actually common in African American individuals, but since the studies to identify these variants had not intentionally included individuals of African American descent, the variants had been misclassified. According to the study, this resulted in multiple

African America individuals receiving false positive reports for pathogenic variants, and following the study these variants were reclassified as benign. This is just one small example of the effects of excluding diverse populations from genomic research, the results of which are then applied to the human population as a whole. Various other studies have shown that polygenic risk score predictions are many fold more accurate for European populations than for African ones (Martin et al., 2019) and that potentially causative variants for previously known disorders are found more rarely in populations of African descent, suggesting that there are a great number of disease-causing genes or variants that are yet to be discovered in African populations (Sloan-Heggen et al., 2016; Yan et al., 2016). These results indicate both that there are consensus disease-causing variants that in fact do not cause disease in African populations, and that there are also African-specific causal variants that are not included in the consensus lists. For certain drugs, the efficacy and potential to cause adverse effects can also be dependent on population-specific genetic variation, which further highlights the necessity of including diverse populations in genomic studies (Sirugo et al., 2019). Taken together, these considerations mean that researchers' abilities to translate genetic findings into diagnostic or therapeutic interventions for African ancestry groups may be greatly hindered, and commensurate measures must be taken in the genomics community to prevent and reverse this phenomenon. Overall, the under-representation of many genetically diverse populations in genomics research hinders our ability to fully understand human genetic diseases, and has the potential to further increase health inequality between developed and developing nations.

# 2. Evaluating and adapting the HUPAN pipeline

## 2.1. Introduction

The HUman Pan-genome ANalysis (HUPAN) pipeline is a system that was created by Duan *et al.* (2019) for the purpose of creating a Han Chinese pan-genome. As explored in section 1.1.3, the human reference genome excludes notable amounts of genetic variation from diverse populations. As such, the HUPAN authors aimed to create a system with the ability to identify genetic sequences present in the Han Chinese population, but which had been excluded from the reference genome.

### 2.1.1. The HUPAN pipeline rationale

Given the rise of second or next generation sequencing (NGS) technologies, in combination with the rapidly decreasing cost of sequencing whole genomes, there has been an explosion in the production of human genomic data (Kodama et al., 2012; Mardis, 2017). Furthermore, the development of many highly efficient assembly algorithms has enabled the assembly of many more human whole genomes (Paszkiewicz & Studholme, 2010), which has greatly facilitated the understanding and identification of human genetic variation. Genomes have commonly been assembled from short-read sequences (150 to 400 bp) produced by second generation sequencing technologies such as those of Illumina and Ion Torrent (Mardis, 2017). In recent years, the introduction of third generation or long-read sequencing technologies, such as those developed by Pacific Biosciences and Oxford Nanopore, has enabled the assembly of complex and repetitive regions that were previously unable to be assembled using short-read sequences. The contiguous sequences (contigs) assembled from long-read sequences have therefore allowed the analysis of complex regions of the human genome that were previously unresolved (T. Hu et al., 2021; Nurk et al., 2021). This growing database of human whole genome sequences is a prerequisite for human pan-genomic research, and the HUPAN authors, therefore, created a system that utilises the vast amount of human genomic data already available to create a pan-genome (Yu & Wei, 2020). Further, the pipeline is able to accept unassembled sequences in both short- and long-read form.

The HUPAN authors took the approach of *de novo* assembly of individual genomes in the construction of their pan-genome. The process of *de novo* sequence assembly uses overlapping identical base pairs to align and merge short sequence reads together to form long contigs that represent the original DNA templates (Paszkiewicz & Studholme, 2010). This is performed without the use of a reference genome, which is in contrast to reference assembly methods, whereby reads are mapped back to a reference genome in order to identify the correct sequence order before being merged to form the whole

genome. *De novo* assembly of genomes or novel sequences is the preferred method for the creation of a pan-genome, since a pan-genome specifically contains sequences that may be absent from the reference genome and reference assembly methods may hinder the identification of these sequences.

However, as discussed in section 1.1.3.2, another method termed pseudo *de novo* assembly can be utilised, whereby only reads that cannot be mapped to the reference genome are *de novo* assembled (Faber-Hammond & Brown, 2016). This technique is often utilised due to its lower computational requirement compared to full *de novo* assembly, as significantly fewer sequences will need to be assembled. The HUPAN authors have investigated the difference between assembling all or only unmapped reads with simulated sequencing data, and were able to show that pseudo *de novo* assembly seemed to both underestimate the total size of novel sequence, and also resulted in more misassembles within the unmapped reads (Duan et al., 2019). The HUPAN authors therefore decided to pursue a pipeline that utilises only full *de novo* assembly. This is a different approach to other pan-genome assembly pipelines, such as those developed by Sherman *et al.* (2019) and Li *et al.* (2021). These pipelines prioritise computational accessibility by utilising pseudo *de novo* assembly, but – as reported by Duan *et al.* (2019) – potentially at the cost of more reliably assembled sequences. The HUPAN authors note, however, that the additional challenges in computational complexity presented by their pipeline can be fully addressed with the use of a high-performance computational facility and by using high-quality software with low memory requirements. Further, given their immense size, the fully assembled human genomes are not used through the entirety of the pipeline; rather, following the computationally expensive *de novo* assembly step, the pipeline limits the amount of compute resources needed by extracting only the assembled contigs that cannot align to the human reference genome above a 90% sequence identity threshold. These contigs, termed 'non-reference sequences', are further reduced in size through redundancy removal and then taken through the rest of the HUPAN pipeline before being combined with the human reference genome to create the population-specific pan-genome. Using this method, the authors have ensured that only the first assembly step is computationally expensive while the rest of the pipeline is comparable to pan-genome pipelines that utilise pseudo *de novo* assembly of unmapped reads. As a result, the HUPAN pipeline produces high quality non-reference sequences with high confidence in the quality of the assembly, while only requiring large computational power in the early stages of the pipeline.

### 2.1.2. The structure of the HUPAN pipeline

The HUPAN pipeline consists of sequential steps that result in different data types and analyses, with the output from each step being utilised in the following step. A system diagram of the pipeline can be seen in Figure 2.1, which briefly details the processes used to obtain the nucleotide sequences absent from the reference genome, termed non-reference sequences.

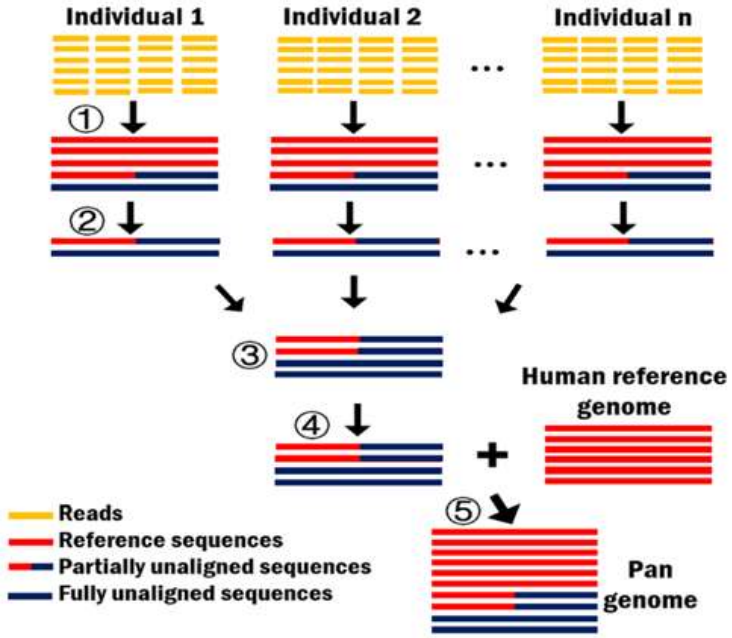


**Figure 2.1. Simple system diagram of the HUPAN pipeline with an input of n samples.** For step 1, the sample reads are *de novo* assembled, and contigs already represented in the human reference genome are removed in step 2. The non-reference contigs are merged in step 3 and combined with the human reference genome sequences in step 4. In step 5, the final pan-genome is analysed and utilised. Adapted from Duan *et al.* (2019).

Figure 2.2 presents a more detailed depiction of the pipeline; as shown in this figure, there are two distinct stages. There are additional preliminary steps for quality control examination and trimming of the raw sequencing reads using FastQC (Andrews, 2010) and Trimmomatic (Bolger et al., 2014), which are not shown in the diagram. The first stage shown, stage A, is the construction of the pan-genome, whereby each sample is *de novo* assembled by one of two assembler programs, String Graph Assembler (SGA) (Simpson & Durbin, 2012) or SOAPdenovo2 (Luo et al., 2012). These two programs were chosen due to their notably lower memory requirements for assembling a human genome compared to other well-known assemblers. The assembled contigs are then aligned to the human reference genome using the NUCmer tool from the MUMmer package (Kurtz et al., 2004) and any contigs that align at or above a set threshold of 95% identity and 95% coverage are discarded, as

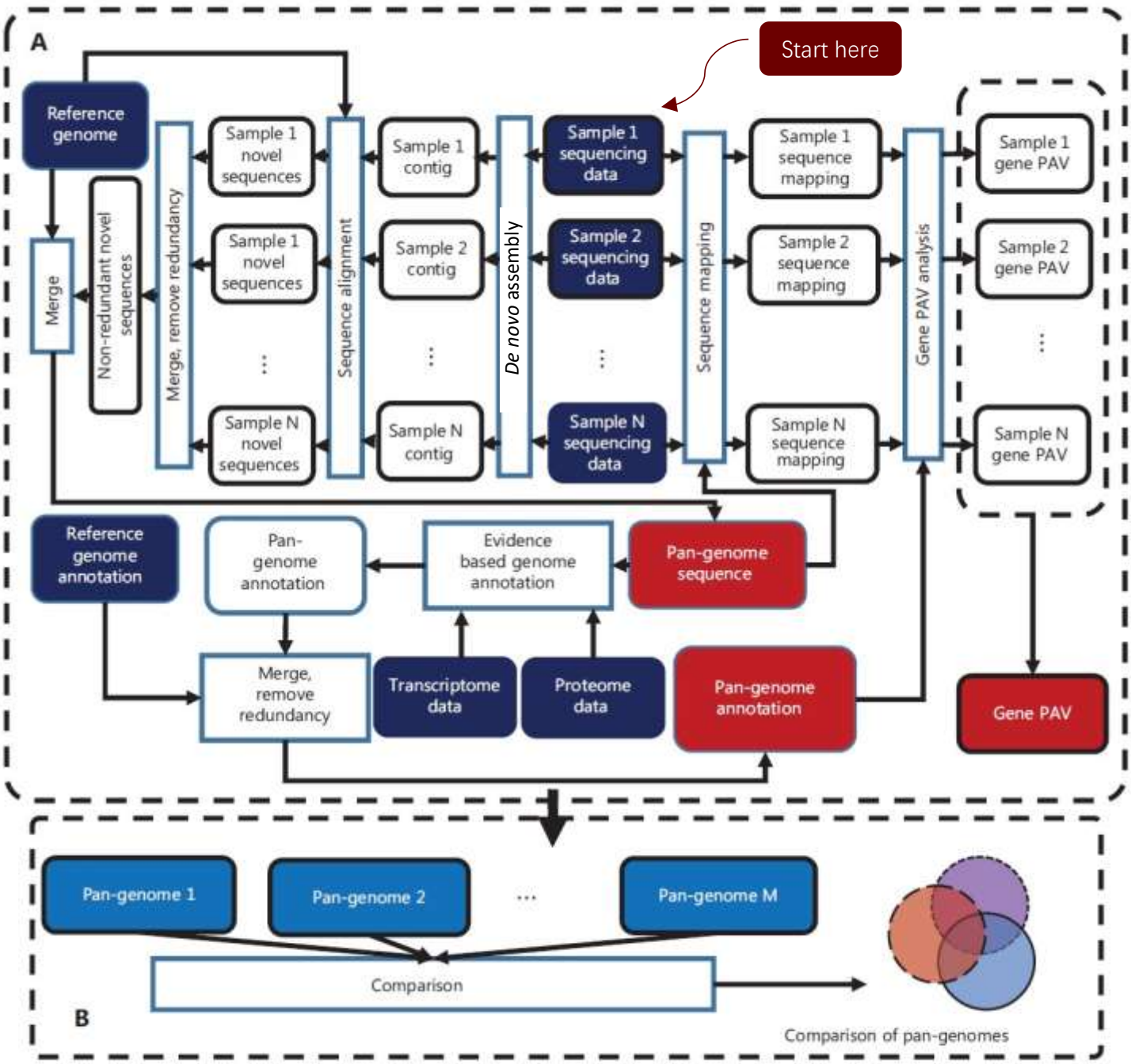


**Figure 2.2. Full schematic diagram of the HUPAN pipeline**. The pipeline was designed to be performed in two stages. In the first stage (A), raw sequenced reads are trimmed and filtered (not shown) and the genomes are *de novo* assembled. The assembled contigs are aligned to the reference to identify novel or unaligned contigs for each sample, and these are then merged and redundancy is removed. The non-redundant novel sequences are combined with the reference genome to create the pan-genome, which is then annotated. The raw sequence reads for each sample can then be mapped to the pan-genome to create a gene presence-absence variation profile for all the samples. In the second stage (B), which is performed independently from the pipeline, various pan-genomes (or simply whole genomes) are aligned and compared to identify population-specific features of the resulting pan-genome. Adapted from Yu & Wei, 2020. PAV: presence-absence variation.

these can be considered already represented within the reference genome. QUAST (Gurevich et al., 2013) then utilises either NUCmer or Minimap2 (H. Li, 2018) to evaluate the remaining contigs and identifies both fully and partially unaligned sequences for each sample. The sequences from each sample are then merged and redundancy is removed using CD-HIT (Fu et al., 2012). Contaminating sequences are identified and discarded using The National Centre for Biotechnology Information (NCBI) BLAST software (J. Ye et al., 2006), not shown in Figure 2.2. This is a strict filtering step, as any sequence aligning with more than 60% identity to an NCBI sequence classified as either bacterial, viral, fungal, archaeal or as a non-primate eukaryotic sequence is removed from further analysis. The de-contaminated non-reference sequences are merged and redundancy is removed a second time, and the final set of novel non-redundant non-reference sequences are then combined with the reference genome to create the population-specific pan-genome. Genome annotation is also performed on the novel sequences using *ab initio* methods in MAKER (Holt & Yandell, 2011), which is followed by multiple strict HUPAN filtering criteria. These criteria ensure the novel genes predicted by MAKER are longer than 100 bp, have less than 80% identity to any other novel gene predicted, have less than 50% identity to any genes or gene transcripts in the human reference genome, have both a start and a stop codon, and are made up of less than 50% repeat sequences. The raw sequencing reads from each sample used in the first step of the pipeline are then aligned to the pan-genome using Burrows-Wheeler Aligner (BWA) (H. Li & Durbin, 2009) or Bowtie 2 (Langmead & Salzberg, 2012) and SAMtools (H. Li et al., 2009) and the gene presence-absence variation (PAV) profile for each sample is generated and compared using both the human reference genome primary sequence annotations and the novel gene annotations predicted by MAKER. In stage B, the population-specific pan-genome is compared to other pan- or whole genomes of other populations to identify novel sequences that are potentially population-specific. However, this functionality is not included in the HUPAN pipeline and is performed by the user independently.

### 2.1.3. Analysing multiple populations within the HUPAN pipeline

Previous pan-genome studies have focused on building pan-genomes for single population groups with the express aim of creating a population-specific reference sequence. In the case of the HUPAN study, for example, the authors included only individuals of Han Chinese descent. Consequently, most publicly available pan-genome analysis pipelines have thus far only been tested on single population datasets. However, since human pan-genome research is a relatively novel field, accepted and standardised methods for creating and analysing pan-genomes have yet to be established. This can be clearly seen in the entirely different methods used by Duan *et al.* (2019) in the HUPAN pipeline compared to those used recently by Sherman *et al.* (2019) and Li *et al.* (2021); the variation in terms

of alignment thresholds, redundancy cut-offs, contamination removal thresholds and even the definition of what constitutes the core genome is highly inconsistent between studies. Ultimately, this means that novel sequence identified by independent pan-genome studies likely cannot be reliably compared, as the definitions and thresholds used for the creation of each pan-genome are often incompatible. Therefore, to examine this aspect of pan-genome assembly and to observe the effects of including multiple populations in one study, a dataset that included multiple different populations was assembled. The 1000 Genomes Project (1kGP) most recent release of high-coverage whole genome sequences (Byrska-Bishop et al., 2021) included five African populations, those being Esan in Nigeria (ESN), Gambia in the Western Division (Mandinka) (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSN), and Yoruba in Ibadan, Nigeria (YRI). These samples are also publicly accessible and are therefore ideal as a test dataset to examine the functionality of the HUPAN pipeline.

## 2.2. Materials and methods

### 2.2.1. Installing the HUPAN software

The HUPAN software (Duan et al., 2019) was downloaded from the HUPAN GitHub repository (https://github.com/SJTU-CGM/HUPAN). R software (version 3.1 or later) (R Core Team, 2020) was required for the correct installation of HUPAN, which was available (version 3.6.1) on the Ilifu high performance computing cluster (http://www.ilifu.ac.za/) within a Singularity (Kurtzer et al., 2017) container. The instructions to compile and install HUPAN were performed as specified in the provided GitHub repository.

Singularity images for SGA (version 0.10.15) (Simpson & Durbin, 2012), the MUMmer package for NUCmer (both version 3.1) (Kurtz et al., 2004), QUAST (versions 4.5 and 5.0.2) (Gurevich et al., 2013), CD-HIT (version 4.8.1) (Fu et al., 2012, p.), NCBI's BLAST (version 2.9.0) (J. Ye et al., 2006), MAKER (version 3.01.03) (Holt & Yandell, 2011) – which made use of SNAP (version 2013-11-29) (Korf, 2004), AUGUSTUS (version 3.4.0) (Stanke et al., 2004), Exonerate (version 2.2.0) (Slater & Birney, 2005), EVidenceModeler (version 1.1.1) (Haas et al., 2008) and RepeatMasker (version 4.1.0) (Nishimura, 2000) – as well as Bowtie 2 (version 2.4.1) (Langmead & Salzberg, 2012) and SAMtools (version 1.9) (H. Li et al., 2009) were either obtained or created by myself depending on the software's availability on the Ilifu cluster. Some of the Docker recipes for these images can be obtained from a publicly accessibly GitHub repository (https://github.com/grbot/containers).

The datasets required to run the HUPAN pipeline were downloaded as specified in the HUPAN GitHub repository and stored on the Ilifu high performance computing cluster. NCBI's non-redundant nucleotide database was downloaded on 17 February 2021 and NCBI's taxonomy database and

associated accession numbers file were downloaded on 23 February 2021. The human reference genome required for the HUPAN pipeline was obtained from the GATK (McKenna et al., 2010) Resource Bundle using FTP server access on 6 January 2016. The reference genome used is build 38, henceforth referred to as GRCh38, and contains all primary sequences, alternate contigs and decoy sequences, but excludes patch sequences. Detailed information on the sequence can be viewed at https://gatk.broadinstitute.org/hc/en-us/articles/360035890951.

### 2.2.2. Obtaining the test dataset and testing the HUPAN pipeline

A test dataset of 100 genomes (Appendix A) was obtained from The International Genome Sample Resource (https://www.internationalgenome.org/data-portal/data-collection/30x-grch38). Samples were obtained from the high-coverage 1000 Genomes Project (1kGP) dataset and thus had a known coverage of 30X. Twenty genomes from each of the five African populations included in the 1kGP dataset (ESN, GWD, LWK, MSN and YRI) were downloaded in FASTQ format and were assembled using SGA in the HUPAN pipeline. Following multiple assembly failures, five samples (HG02465, HG02588, HG02891, NA19031 and NA19461) could not be assembled and were excluded from the dataset.

The samples were submitted to the pipeline exactly as specified in the HUPAN GitHub repository. Where it was required for the software to run successfully on the Ilifu high performance cluster, wall time and memory requirements were increased and the HUPAN scripts updated on a forked version on the HUPAN pipeline, available at https://github.com/BournSupremacy/HUPAN. The samples were run through the HUPAN pipeline up until and including step 6i (as designated in the original HUPAN GitHub repository), which is the step where the non-reference sequences are merged and redundancy is removed.

### 2.2.3. Analysis of the non-reference sequences

All analyses of the final sets of non-reference sequences were performed using Python (version 3.8.3) scripts within JupyterNotebooks in JupyterHub (single-user server version 2.2.0) (Granger & Perez, 2021), accessible through the Ilifu high performance computing cluster. Alignment of the 1kGP non-reference sequences to GRCh38 was performed using NUCmer in the MUMmer package with default parameters (`-b 200 -c 65 -g 90 -l 20`). Analysis of the alignment outputs was performed using a Python script which can be accessed at https://github.com/BournSupremacy/BioinformaticsTools/tree/main/alignment. The script for the 1kGP and GRCh38 alignment plot is accessible at https://github.com/BournSupremacy/BioinformaticsTools/blob/main/PythonPlots/1kGPvb38.ipynb. All other plot scripts are available at https://github.com/BournSupremacy/BioinformaticsTools/tree/

main/PythonPlots. The JupyterNotebook script for principal component analysis was adapted from a script written by Ayton Meintjies (University of Cape Town).

## 2.3. Results

### 2.3.1. Assembly of the test dataset and analysis of the pipeline infrastructure

Following the failure of SGA to assemble five samples, the final 1kGP test dataset consisted of 95 samples in total (Table 2.1). Three samples that could not be assembled belonged to the GWD population, and the other two samples that could not be assembled belonged to the LWK population.

**Table 2.1. Summary of the populations and samples present in the dataset used to test the HUPAN pipeline.** All 95 genomes were obtained from the 1000 Genomes Project high-coverage (30X) dataset.

| Population | Ethnic group and country of origin | Number of samples in the final test dataset |
|---|---|---|
| **ESN** | Esan from Nigeria | 20 |
| **GWD** | Mandinka from the Western Division of The Gambia | 17 |
| **LWK** | Luhya from Kenya | 18 |
| **MSL** | Mende from Sierra Leone | 20 |
| **YRI** | Yoruba from Nigeria | 20 |
| **Total** | | **95** |

Having run all of the samples through the HUPAN pipeline up to and including the stage of obtaining the final set of non-redundant non-reference sequences, we were able to understand and analyse the structure of the pipeline and thereby identify potential challenges in the early stages. As depicted in Table 2.2, the longest and most computationally expensive step in terms of memory within the pipeline was the *de novo* assembly using SGA, as expected based on reporting from the HUPAN authors. All other steps were reasonably fast and entirely computationally feasible for the Ilifu high performance computing structure. However, one inefficiency of the HUPAN pipeline that became clear through testing was that the pipeline lacks the functionality of identifying and alerting the user to failed jobs or steps. This therefore requires the user to manually monitor and confirm each job has completed successfully using the cluster scheduling and managing systems before moving onto the next step. Additionally, for single steps that perform multiple commands and produce multiple different outputs, the pipeline does not cache intermediate results. As a consequence, if a job fails at some point during its execution – perhaps due to a computational node failure or insufficient memory or time, among other possible causes – any previous computational progress of the job is lost, and the

**Table 2.2. The maximum memory and time requirements of each stage in the unmodified HUPAN pipeline when run on the Ilifu high performance computing cluster.** The 1kGP test dataset was used to test the pipeline. The maximum time used is given per sample where samples were required to run individually, and given as time for the entire step once the data from individual samples were merged. Individual samples were run in parallel, hence, the time to run all of the samples was equivalent to the time taken to run the most computationally expensive sample. The maximum memory used for each step was previously determined and was therefore suitably requested within the pipeline.

| Pipeline step | Maximum time used | Maximum memory requested |
|---|---|---|
| Assembly | 15 days per sample | 50 GB |
| Aligning to the reference genome | 36 hours per sample | 7 GB |
| Removing highly similar contigs | < 1 day for all samples | 2 GB |
| Assessing candidate contigs | 1 hour per sample | 35 GB |
| Collecting all unaligned contigs | 1 hour per sample | 3 GB |
| Merging unaligned contigs and removing redundancy | 12 hours | 3 GB |
| Identifying contamination | 24 hours | 3 GB |
| Obtaining taxonomic classifications of all sequences | 1 hour | 3 GB |
| Removing contaminating sequences | 1 hour | 3 GB |
| Combining full and partially unaligned sequences and removing redundancy | 12 hours | 3 GB |

entire step must be rerun. This requires the user to request and use additional computational time and memory that would be unnecessary had the intermediate results of the job been cached. This was of particular importance for the assembly step, as a 14-day job is exceptionally computationally expensive and the lack of cached intermediate data resulted in many re-runs of failed samples. This ultimately greatly increased the amount of time and memory required to complete the assembly of all 95 samples.

### 2.3.2. Analysis of the non-reference sequences obtained from the test dataset

#### *2.3.2.1. First test run of the pipeline*

Following the identification of the non-redundant non-reference sequences using the HUPAN pipeline, the sequences were briefly analysed to confirm that the results were comparable to those presented in the HUPAN publication by Duan *et al.* (2019). The analyses performed by the HUPAN authors on the Han Chinese non-reference sequences were taken directly from the HUPAN publication and are shown in Figure 2.3. The same analyses were performed on the 1kGP test dataset non-reference sequences and are displayed in Figure 2.4, for ease of comparison with the Han Chinese results.

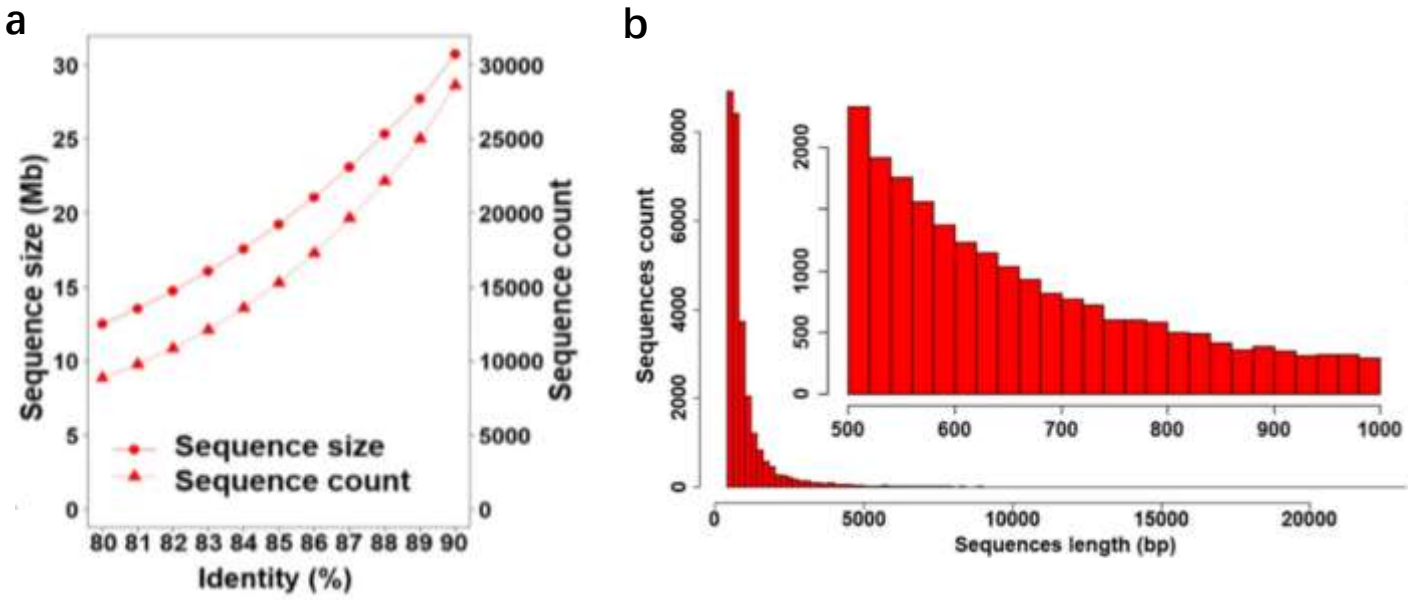


**Figure 2.3. Analysis of the HUPAN Han Chinese non-reference sequences, adapted from the original HUPAN publication (Duan *et al.*, 2019).** The HUPAN research used a dataset consisting of 185 Han Chinese samples. Sequence count denotes number of contigs. **(a)** The total size (in Mbp) and count of non-redundant non-reference sequences as the CD-HIT threshold for identity is increased. **(b)** The contig length distribution profile of the final non-redundant non-reference sequences obtained at a 90% identity threshold. Inset plot showing expanded length distributions between 500 and 1000 bp.

When comparing Figure 2.3 and Figure 2.4, it is clear that, while the overall form and trends of the data appeared similar, there was a major difference with regard to the total amount of non-reference sequence identified. In total, from 185 samples at a 90% identity cut-off for clustering, the HUPAN authors obtained 30.72 megabase pairs (Mbp) of novel sequence which was comprised of 28 622 contigs (Fig. 2.3a). In comparison, the first run of the 95-sample test dataset, at the same redundancy identity threshold of 90%, resulted in 67.09 Mbp of novel sequence made up of 44 472 contigs (Fig. 2.4a). This is a substantially higher amount despite the smaller sample size.

The effect of changing the CD-HIT (Fu et al., 2012) redundancy removal identity threshold was also examined and compared to the Han Chinese dataset. CD-HIT functions by applying the user-defined identity threshold to sort nucleotide sequences that meet this threshold into groups termed “clusters”. These clusters are simply collections of nucleotide sequences that have base pair compositions at or above the given identity threshold to the longest sequence in that cluster. For example, at a threshold of 95% identity, CD-HIT will identify and group clusters of sequences which have a minimum of 95% of their bases identical to the longest sequence in the same cluster. The CD-HIT algorithm then takes the longest sequence from each cluster and designates this sequence as the “representative” contig. This contig and its nucleotide sequence are then used in the subsequent steps

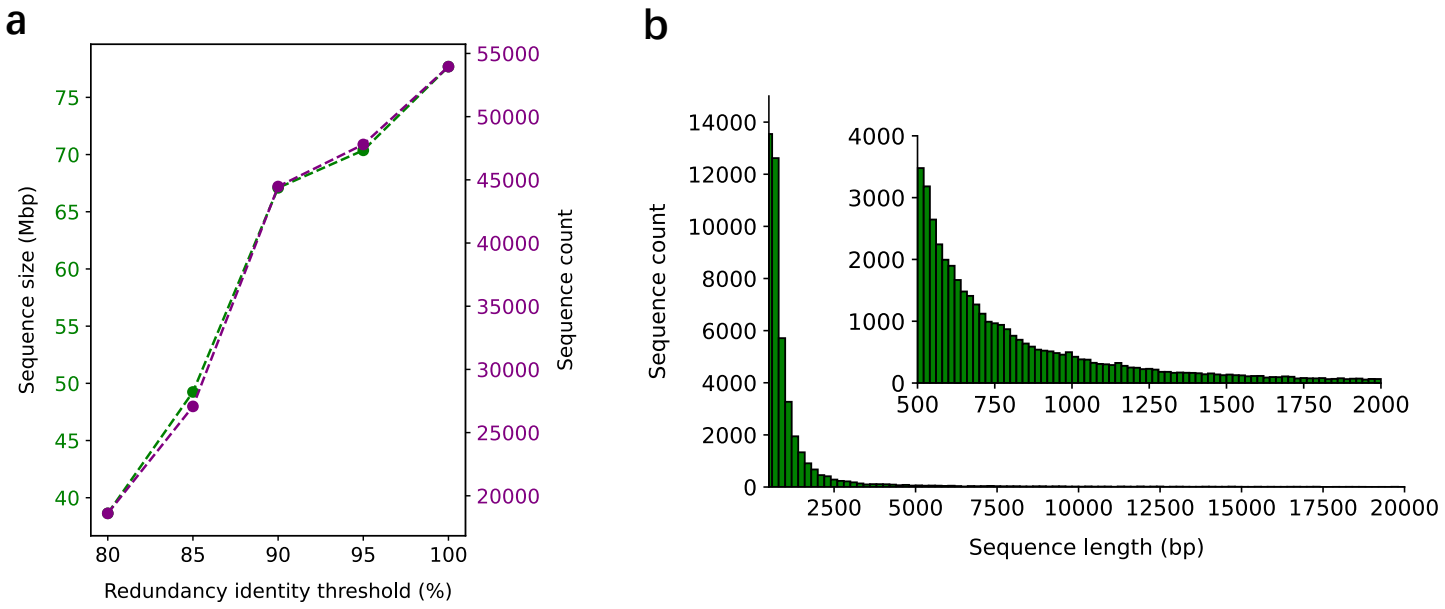


**Figure 2.4. Analysis of the 1kGP 95-sample dataset non-reference sequences produced by the first test run of the HUPAN pipeline.** Sequence count denotes number of contigs. **(a)** The total size (in Mbp) and count of non-redundant non-reference sequences as the CD-HIT threshold for identity is increased. **(b)** The contig length distribution profile of the final non-redundant non-reference sequences obtained at a 90% identity threshold. Inset plot showing expanded length distributions between 500 and 2000 bp.

of the pipeline to represent all the sequences that fell within that cluster. The HUPAN authors examined the effect of changing the CD-HIT identity threshold from 80% to 90% but did not include analysis for thresholds higher than this; we wanted to analyse identity thresholds higher than 90%, so data for thresholds of 95% and 100% identity were included for the 1kGP dataset.

When comparing the effect of changing the identity threshold on the total amount of sequence, the Han Chinese dataset obtained significantly less sequence than the 1kGP test dataset at identity thresholds of 80% and 85%, consistent with the results observed at the 90% identity threshold. However, overall, both datasets showed similar trends when the CD-HIT clustering identity threshold was increased from 80% to 90%, in that the total number and size of the novel sequence appeared to increase exponentially. Within this range, the amount of non-reference sequence increased from 12.51 Mbp to 30.72 Mbp for the Han Chinese dataset, and increased from 38.63 Mbp to 67.09 Mbp for the 1kGP dataset. However, when the identity threshold was increased from 90% to 95% for the 1kGP test dataset, the number and size of non-reference sequences only grew by 3328 contigs and 3.28 Mbp. This was a notably smaller growth than the results from the previous 5% increases in identity thresholds, but because the HUPAN authors did not include this analysis in their research, it was unclear whether this was an expected result. However, given the only slight increase in the amount of sequence with the change in identity threshold from 90% to 95%, we decided to use the

same 90% identity threshold as the HUPAN authors for redundancy removal. Importantly, this ensures that the results obtained here can be fairly compared to those presented in the HUPAN publication, which is essential for analysing the abilities of the pipeline using the test dataset.

At the 90% identity redundancy threshold, the distribution of contig lengths for both the HUPAN data and the first run of the 1kGP dataset were also broadly similar (Fig. 2.3b and Fig. 2.4b). For both, there is a clear abundance of shorter contigs – defined as being shorter than 1 kilobase pair (kbp) – which made up 73.61% of all contigs in the Han Chinese dataset, and 71.67% of all contigs in the first run of the 1kGP dataset. However, the Han Chinese dataset had only 94 contigs longer than or equal to 10 kbp, while the 1kGP dataset had 956 contigs over this threshold, indicating that the 1kGP dataset had produced a notably higher number of long contigs.

Having confirmed that the 1kGP test dataset results were comparable to those produced by the HUPAN authors, the 1kGP non-reference sequences were aligned to GRCh38 using NUCmer (Kurtz et al., 2004) as a confirmation that the sequences obtained could be classified as novel or non-reference (Figure 2.5). However, as shown in the figure, the total amount of alignment between the 1kGP

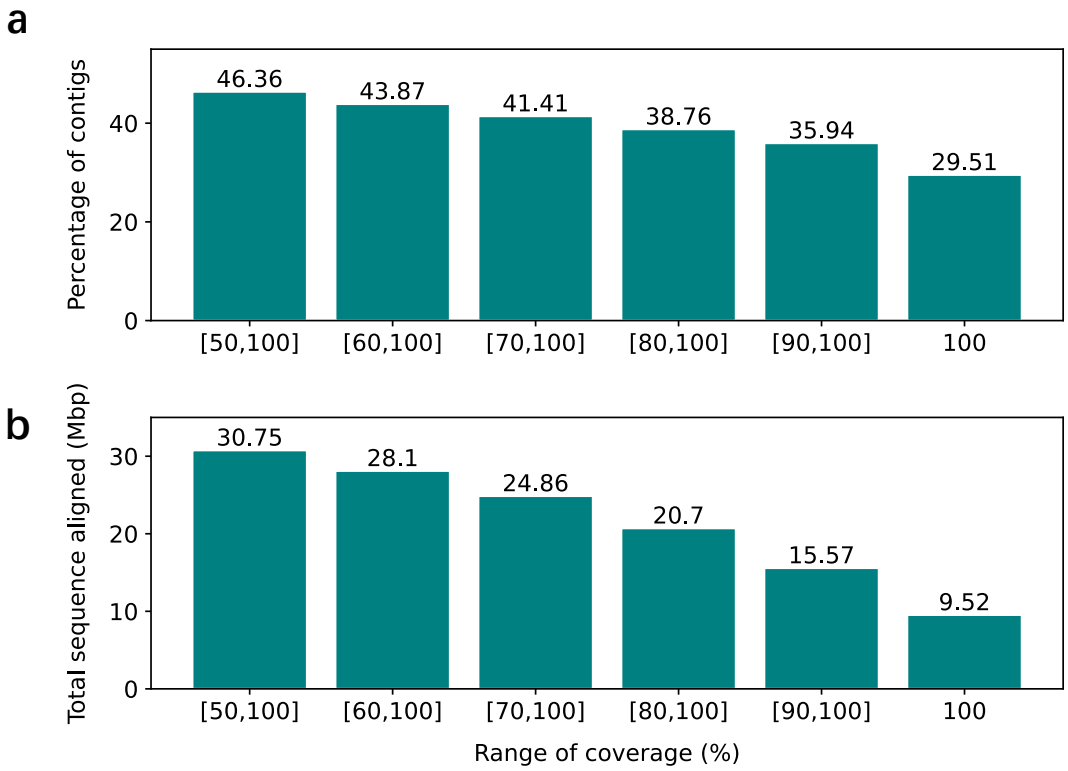


**Figure 2.5. Alignment of the first run 1kGP test dataset non-reference sequences to GRCh38 over increasing coverage thresholds.** All alignments shown have a sequence identity of ⩾ 90%. **(a)** The percentage of non-reference contigs that aligned to GRCh38 at different coverage thresholds. **(b)** The total length of non-reference sequence in Mbp that aligned to GRCh38 at different coverage thresholds.

non-reference sequences and GRCh38 was significantly higher than expected. In fact, almost 30% of the non-reference contigs aligned to the reference genome at minimum thresholds of 90% identity and 100% coverage (Fig. 2.5a). Alignment identity, similarly defined in CD-HIT, is the percentage of base pairs that match exactly within an aligned region. In contrast, alignment coverage is defined as the percentage of the total sequence length that falls within the aligned region. Hence, higher coverage indicates that a larger portion of the whole sequence is aligned, while higher identity indicates the sequences match more exactly within the alignment. When both coverage and identity are above 90%, this suggests the sequences align very well and therefore have strong homology to each other. However, according to the pipeline's documentation and specified function, sequences with such strong alignment to the reference genome should have been removed in the early stages of the pipeline, and these alignment results should therefore not be possible.

It was initially hypothesised that the contigs displaying the highly specific alignment were the partially unaligned contigs, which are defined as having at least one alignment of any length to the reference genome and at least one unaligned region longer than 500 bp. However, upon further investigation, it became apparent that the sequences showing high similarity with the reference had been classified as fully unaligned within the pipeline by QUAST (Gurevich et al., 2013), and were therefore defined as having no alignment to the reference genome. This indicated that an error had potentially occurred within the first few steps of the pipeline. Accordingly, to determine where and how this had occurred, the spurious contigs were traced back through the beginning of the HUPAN pipeline until the cause was identified. Through this process, it was determined that the QUAST software (version 4.5) used in the fourth step of the pipeline to assess and identify all fully and partially unaligned sequences was utilising a naming convention that was extensively ambiguous. This version of QUAST classifies all sequences that align below a default minimum identity threshold of 95% as "fully unaligned" contigs, despite this classification being entirely misleading and plainly incorrect. This default was unchanged in the HUPAN pipeline software, meaning any sequence that aligned to the reference genome with less than 95% identity was confusingly labelled as fully unaligned. This explained why such a significant portion of the non-reference sequences aligned to the reference genome at such high coverage and identity thresholds; despite being termed "fully unaligned", some sequences were in fact very well aligned to the reference genome but still below the default threshold of 95% identity set by this version of QUAST, and were consequently included in the final set of non-reference sequences.

Following this revelation, the latest available version of QUAST (5.0.2) was tested on two samples to examine whether this feature was retained in later versions of the QUAST software. These tests confirmed that the newer version of QUAST was no longer classifying the spurious sequences as fully

unaligned and was, therefore, correctly removing them from the final set of non-reference sequences. Accordingly, the HUPAN pipeline was adapted to utilise QUAST version 5.0.2 instead of version 4.5.

#### *2.3.2.2. Second test run of the pipeline*

Following the discovery of the QUAST misclassifications, it was necessary to re-run the 1kGP dataset through the modified HUPAN pipeline. The first stages of the pipeline were completed again, and the same 90% identity redundancy removal threshold was used to obtain the final non-reference sequences. The sequences were then compared to both those produced by the earlier QUAST software and those published by the HUPAN authors for the Han Chinese dataset. This comparison is summarised in Table 2.3 and the analysis of the 1kGP non-reference sequences produced by the corrected second run of the HUPAN pipeline are shown in Figure 2.6.

As reported in Table 2.3, the total number of contigs making up the 1kGP test dataset non-reference sequences decreased by over 28 500 following the correction of the QUAST analysis. The total amount of sequence decreased by almost 20 Mbp, which is a similarly considerable difference. These values indicate that the QUAST misclassification had a crucial impact on the final dataset of non-reference sequences. There were also clear differences in the overall trends of the non-reference sequences compared to those produced by the original QUAST analysis, although the effect of changing the CD-HIT clustering threshold appeared to be similar (Fig. 2.4a and Fig. 2.6a). Leaving aside the total amount of novel sequence, both 1kGP datasets produced comparable curves for this CD-HIT analysis.

**Table 2.3. Comparison of the counts and lengths of the non-reference sequences obtained from the Han Chinese dataset, the 1kGP test dataset utilising QUAST version 4.5 in the HUPAN pipeline, and the 1kGP test dataset utilising QUAST version 5.0.2 in the HUPAN pipeline.** The same redundancy removal threshold of 90% identity was used for all three non-reference sequence datasets.

| Metric | Han Chinese non-reference sequences | 1kGP non-reference sequences using QUAST v. 4.5 | 1kGP non-reference sequences using QUAST v. 5.0.2 |
|---|---|---|---|
| Total number of contigs | 28 197 | 44 472 | 15 923 |
| Number of contigs ≥ 1000 bp | 7 096 | 12 597 | 8 236 |
| Number of contigs ≥ 5000 bp | 470 | 1 806 | 1 813 |
| Number of contigs ≥ 10000 bp | 98 | 956 | 1 035 |
| Total length (Mbp) | 29.50 | 67.09 | 47.48 |
| Total length ≥ 1000 bp (Mbp) | 15.56 | 46.00 | 42.12 |
| Total length ≥ 5000 bp (Mbp) | 3.93 | 29.66 | 27.23 |
| Total length ≥ 10000 bp (Mbp) | 1.39 | 21.17 | 24.21 |
| Longest contig (bp) | 27 453 | 122 943 | 122 943 |

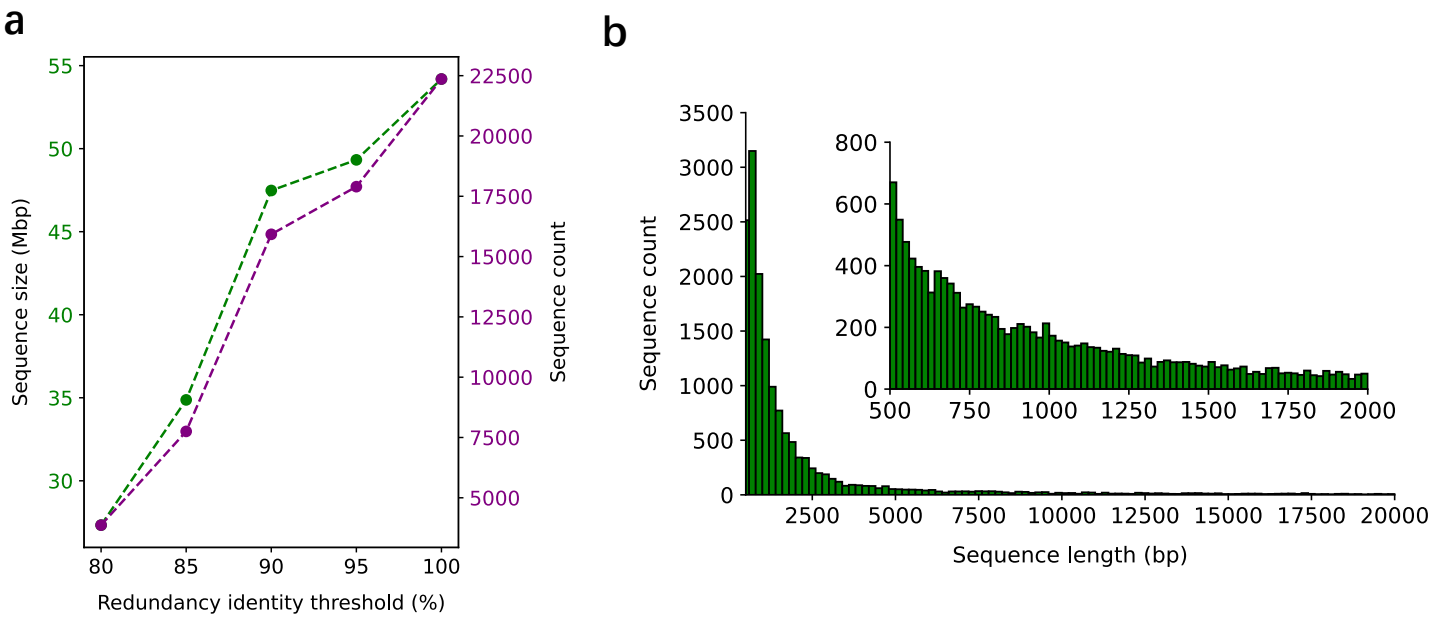


**Figure 2.6. Analysis of the 1kGP 95-sample dataset non-reference sequences produced by the HUPAN pipeline following correction of the QUAST error.** Sequence count denotes number of contigs. **(a)** The total size (in Mbp) and count of non-redundant non-reference sequences as the CD-HIT threshold for identity is increased. **(b)** The contig length distribution profile of the final non-redundant non-reference sequences obtained at a 90% identity threshold. Inset plot showing expanded length distributions between 500 and 2000 bp

This therefore indicated that CD-HIT was functioning equivalently during the merging of both sets of non-reference sequences, which provided confidence in the consistency of the software. However, when comparing the contig length distribution between the 1kGP dataset before and after updating the QUAST software, the results were significantly different (Fig. 2.4b and Fig. 2.6b). There were substantially fewer numbers of shorter sequences (< 1 kbp) in the corrected 1kGP dataset than in the first 1kGP dataset, but seemingly a similar number of longer sequences. This can be seen by examining both the decreased steepness of the contig length histogram in Figure 2.6b when compared to Figure 2.4b, but also by comparing total values. For example, as stated previously, 71.67% of all contigs in the first 1kGP non-reference sequences were shorter than 1 kbp and only 2.15% were longer than or equal to 10 kbp. However, as the corrected dataset had far fewer shorter contigs, only 51.72% of all contigs were less than 1 kbp in length, and 6.5% were longer than or equal to 10 kbp. This discrepancy can also be seen in Table 2.3, where there is a clear difference in the total number of contigs and total sequence length, but similar numbers and lengths for contigs longer than both 5 kbp and 10 kbp. This distinct difference indicates that a large majority of the spurious sequences included by the older QUAST software were between 500 and 1000 bp. Further, when compared to the two 1kGP datasets, the Han Chinese non-reference sequences obtained by the HUPAN authors appear to have distinct

similarities to only the first run of the 1kGP dataset, with a high proportion of contigs shorter than 1 kbp as stated previously. Only 0.35% of these sequences, however, are ≥ 10 kbp, which is a much lower proportion than either of the 1kGP datasets, although closer to that of the initial 1kGP sequences. Thus, it appears that the Han Chinese sequences, similarly to the first 1kGP dataset, display highly distinct contig length distribution differences compared to the corrected 1kGP dataset.

#### *2.3.2.3. Investigation of population-specific clustering in the test dataset*

Having analysed the final and corrected set of non-reference sequences from the 1kGP dataset, the effect of including multiple populations was assessed. Due to the nature of population-specific genetic differences explored in section 1.1.3.1, and the findings from previous population pan-genome studies discussed in section 1.1.3.2, it is known that genetic sequences that are present in only single populations exist and are identifiable. Consequently, we predicted that samples from the same population would be likely to produce the same or similar non-reference sequences through the HUPAN pipeline. As described above (section 2.3.2.1), CD-HIT was used to identify and group contigs with a sequence identity of 90% or higher to remove redundancy; therefore, by examining the clusters of sequences produced by CD-HIT and identifying the samples that contributed sequences to each cluster, we are able to determine whether samples from a single population occur within the same clusters more frequently than with other populations. This frequency, termed cluster sample membership, was therefore analysed as displayed in Figure 2.7. Importantly, however, the final set of 1kGP non-reference sequences had gone through two independent rounds of redundancy removal and had therefore lost population-specific information in two separate steps of the pipeline. Given this multi-stage merging of sequence, analysis could only be easily performed on the data following the first round of redundancy removal. Hence, the results shown in Figure 2.7 are analyses of the 23 732 clusters produced following the first redundancy removal step of the HUPAN pipeline.

First, a principal component analysis (PCA) was performed on the sample membership of the 23 732 clusters produced by CD-HIT. PCA is a statistical method that reduces complex, multi-dimensional data into fewer dimensions, thereby allowing the identification of patterns or factors that result in the largest amount of variation. These factors are termed the principal components. By plotting the two strongest factors of variation – principal components 1 and 2 – any patterns in sample variation related to aspects such as phenotypes or differential treatments, if there are any, can be easily identified. Here, the differentiating aspect of the samples was their population of origin, so it was predicted that the factors of variation identified within the PCA would be correlated with population-specific differences, resulting in the separation of samples based on population. Surprisingly, however, the sample membership PCA of the first two principal components showed no obvious population-specific

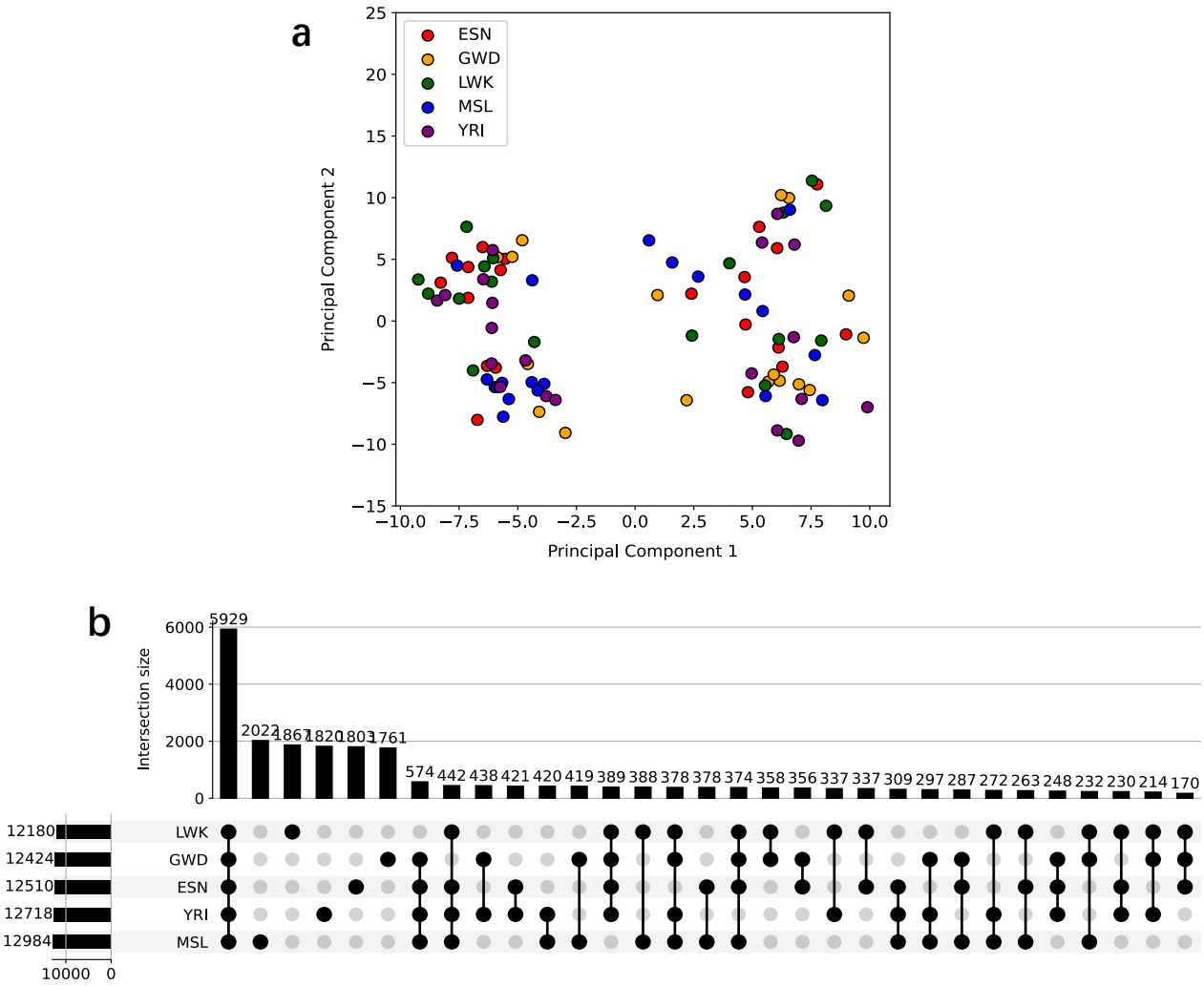


**Figure 2.7. Analysis of the sample membership of the 23 732 nucleotide clusters produced by CD-HIT following the first round of redundancy removal in the 1kGP test dataset. (a)** Principal component analysis of the sample membership of the clusters produced showing the first two principal components. **(b)** Upset plot of the sample membership of the clusters produced. Intersection size indicates how many times that unique combination of populations was found within the same cluster. All intersection size values add up to 23 732. Values next to the population names indicate how many clusters that population contributed towards.

clustering (Fig. 2.7a). All samples from all five populations appeared to spread out evenly and similarly over the first two principal components, with no clear patterns of population-specific separation. This suggested that the patterns of variation identified within the CD-HIT clusters were not correlated to the population of each sample.

An upset plot examining the same 23 732 clusters resulting from the first round of redundancy removal was also produced and is shown in Figure 2.7b. Upset plots are able to depict the frequency

with which different variables co-occur by creating a unique set for each combination of variables, and then determining the frequency count of that set. Here, the variables were the five different African populations, and Figure 2.7b was able to demonstrate the frequency with which samples from the different populations occurred together within the CD-HIT clusters. This analysis showed that samples from all five populations occurring together in a single cluster was by far the most common combination, with 5 929 clusters out of the 23 732 clusters produced having sequences from all five populations contributing towards them. Like the PCA, this indicated a lack of population-specific genetic variation. However, there were also many clusters that contained sequences from samples in only single populations; for example, there were 2022 clusters that consisted of sequences from only samples within the MSL population. Similar counts of between 1761 and 1867 were identified for the other four single populations. Upon further analysis, however, it became clear that a very large majority of these population-specific clusters were clusters that had been created from only a single contig. This means that the contig in question had no similarity to any other non-reference sequences above the given identity threshold of 90% and thus CD-HIT did not cluster it with any other sequences. This resulted in many "clusters" that consisted of only a single sequence, which were then identified in Figure 2.7b as being unique to a single population. These singleton clusters, however, clearly suggest individual-specific genetic variation rather than population-specific variation. This therefore explains the lack of population-specific sample clustering in Figure 2.7a, despite the seemingly large amount of population-specific CD-HIT clusters shown in Figure 2.7b.

## 2.4. Discussion

Following the two test runs of the HUPAN pipeline (Duan et al., 2019), it was evident that the software was easy to run and the outputs simple to understand. However, one aspect of the pipeline that was not sufficiently reliable or straightforward was that of assembly using SGA (Simpson & Durbin, 2012). Many samples failed to assemble multiple times due to either external cluster-wide failures or even sample specific differences in computational requirements. The lack of cached intermediate data and the extensive time taken for an assembly to complete successfully therefore resulted in a computationally expensive and unnecessary bottleneck in the pipeline. As a result, for any further research performed using the HUPAN pipeline, this inefficiency should be addressed, and adaptations investigated to ensure the pipeline is suitable for use on all high-performance computing environments.

The second significant aspect of the HUPAN pipeline that needed to be reconsidered was that of the QUAST (Gurevich et al., 2013) software. Given that version 4.5 of QUAST was producing objectively misleading outputs based on the definition of a non-reference sequence, it was necessary to update

the pipeline to utilise a version of QUAST that functioned as expected by its users. The version chosen was 5.0.2, which was the latest version at the time. This version of QUAST utilises Minimap2 (H. Li, 2018) for alignment, whereas older versions use NUCmer from the MUMmer package (Kurtz et al., 2004). This was the only major QUAST update that was consequential for its use within the HUPAN pipeline. Upon investigation and communication with the QUAST software developers – which can be viewed at https://github.com/ablab/quast/discussions/208) – it was confirmed that the authors were aware of this misleading classification of "fully unaligned" sequences, but did not consider it sufficiently problematic to warrant changing. This 95% identity default, as discussed in section 2.3.2.1, was therefore still utilised in QUAST version 5.0.2, meaning the difference in outputs between the older and newer QUAST software was unrelated to the identity threshold.

Upon analysis of the QUAST outputs, it was discovered that the change of aligner from NUCmer to Minimap2 was the main cause of the discrepancy in results. Minimap2 is more likely to identify sequences that align well – but not fully – to the reference sequence as "misassembled" rather than "fully unaligned", as it assumes that variation within 10% to 20% identity of the human reference genome is likely the result of an error by the assembler. However, it correctly classifies sequences with very little or no alignment to the reference genome as "fully unaligned". NUCmer, on the other hand, simply classifies all sequences that fall below an alignment identity threshold of 95% as "fully unaligned", despite this being a misleading label. The change from NUCmer to Minimap2 resulted in many of the well-aligned sequences from the first 1kGP dataset being excluded from the second, while the sequences with little or no alignment to the reference genome that had been correctly classified as "fully unaligned" by NUCmer were retained.

This understanding of the QUAST software necessitated changing the HUPAN pipeline to utilise the newer version of QUAST. Due to the change from NUCmer to Minimap2, the QUAST output formats were slightly different and needed to be moderately reformatted to be used in the following pipeline step, but these necessary adaptations were relatively straightforward. Yet despite the adaptations being negligeable, the implications of the use of the older QUAST software are not.

For the 1kGP dataset, the total amount of sequence misleadingly included by version 4.5 of QUAST was almost 20 Mbp. This is a significant amount of sequence that could have produced incorrect and unfounded results had the alignment to GRCh38 not been performed as a quality check. Most importantly, this earlier version of QUAST is the version that is included in the HUPAN software download, and the version that the HUPAN authors specifically state should be used, as per the current (April 2022) HUPAN GitHub page (https://github.com/SJTU-CGM/HUPAN). Further, in Figure 3c of the HUPAN publication (Duan et al., 2019), the authors show that a significant portion of the Han

Chinese non-reference sequences align very well to the human reference genome. The result is not explained or discussed in detail in the publication, but this figure indicates that more than 40% of the Han Chinese non-reference sequence (representing around 10 Mbp) is removed from the dataset when the maximum identity threshold for alignment to the reference genome is decreased from 90% to only 80%. This therefore means that 40% of the Han Chinese non-reference sequences align to the human reference genome between 80% and 90% identity. This result is highly similar to the findings made for the first run of the 1kGP test dataset, where 38.76% of the non-reference sequences aligned to the reference genome above an 80% identity threshold. It is, therefore, highly likely that the HUPAN authors utilised the earlier version of the QUAST software to identify the Han Chinese non-reference sequences. This, too, would explain the similarities in the length distributions between the Han Chinese non-reference sequences and the non-reference sequences produced in the first 1kGP dataset, and why the corrected set of 1kGP non-reference sequences had lesser similarity to either dataset. Ultimately, if the HUPAN authors did indeed utilise QUAST version 4.5 within their pipeline, this means a notable amount of the Han Chinese non-reference sequences identified have high similarity to sequences within the human reference genome. Although this finding does not negate or question the value of the research done by the HUPAN authors, this aspect of the non-reference sequences is not clearly discussed in the HUPAN publication and resulted in a misleading definition for a “fully unaligned non-reference sequence”. Researchers who do not perform an in-depth analysis of the HUPAN research may therefore be similarly misled in their understanding of the non-reference sequences obtained. As a result, we will utilise only QUAST version 5.0.2 for further pan-genome research with the HUPAN pipeline, as the aim of this study is to identify African genetic sequences that are largely or completely absent from the human reference genome.

Through updating the HUPAN pipeline to utilise the latest and working version of QUAST, it was evident that the pipeline is fully able to identify and extract non-reference sequences from high-coverage human genomes. Notably, although there was a substantial reduction in the amount of novel sequence identified once the newer QUAST version was used, the identification of over 47 Mbp of novel sequence that is not present in the reference genome is a significant result. Furthermore, the total amount of novel sequence extracted from only 95 samples from the 1kGP dataset (47.48 Mbp) was notably higher than the amount produced by the HUPAN authors using 275 Han Chinese samples, which they stated was 29.5 Mbp (Duan et al., 2019). This is even more significant when we consider that the Han Chinese non-reference dataset contains a notable amount of sequence that is largely present within the reference genome, as discussed above. The discrepancy between the 1kGP and Han Chinese datasets is likely due to the inclusion of five different 1kGP African populations, as previous research has shown that including multiple genetically distinct populations in genomic

analyses increases the amount of genetic variation (Choudhury et al., 2020; Sudmant et al., 2015). Therefore, as is the case for this research, including multiple populations will result in an increase in the amount of novel sequence identified. Since the HUPAN publication's samples all belonged to a single genetically distinct population, there was predictably greater redundancy in the non-reference sequences. This was subsequently removed using CD-HIT, resulting in less sequence overall despite the inclusion of more samples.

When considering the redundancy removal step, the HUPAN authors decided to cluster sequences together if they had an alignment identity of 90% or higher. The rationale for this cut-off was not explained in the HUPAN publication, and as mentioned previously, the identity cut-off for redundancy removal in previous pan-genome research varies depending on the research group. For example, Li *et al.* (2021) also chose a 90% identity cut-off for unplaced (i.e. fully unaligned) contigs, but used identity and coverage cut-offs ranging from 90% to 99% and 80% to 95% respectively for placed or partially aligned contigs. Conversely, Sherman *et al.* (2019) used an identity cut-off of 97% or above for placed contigs. The differences in cut-off thresholds for redundancy removal will clearly have an impact on the amount of sequence present in the final dataset, and so the effect of changing this threshold in the HUPAN pipeline was examined. Within the 1kGP dataset, there was a more than 25% decrease in the total amount of sequence when the cut-off was changed from 90% to 85%, indicating a severe loss of sequence and therefore likely valuable and important genetic information. However, the increase of the identity threshold from 90% to 95% only increased the total sequence amount by around 4%, showing a distinct lessening of the redundancy effect at the higher identity threshold. This slight effect indicated that not significantly more sequence would be gained by increasing the identity threshold. Based on these analyses, we decided to utilise an identity cut-off of 90% for clustering using CD-HIT. Furthermore, as this research aims to directly compare the results and outputs with those obtained from the Han Chinese dataset, using a different identity threshold to the HUPAN publication would render this comparison inconsistent. As a result, it was decided that the 90% identity threshold would be maintained for this research to ensure uniformity in the final comparisons to the HUPAN dataset.

The final aspect that was examined in the HUPAN pipeline was the effect of using samples from multiple different populations. As already discussed, the inclusion of five different populations is likely the cause of the large amount of novel sequence identified compared to the HUPAN publication. Despite this, however, the structure of the pipeline also appears to hinder analysis into population-specific genetic differences. The HUPAN pipeline was designed to include samples from only single populations, and thus population information is not kept or analysed within the pipeline. This therefore made population-specific analysis of the non-reference sequences cumbersome and

impractical, and required extensive additional scripting to achieve. This was particularly evident in the analysis of the final non-reference sequences. As all of the samples underwent two separate redundancy removal steps together, the amount of population-specific genetic differences that were merged and ultimately excluded from the pan-genome was likely large. This may, therefore, be the reason for the lack of population-specific non-reference sequences and CD-HIT clusters, which were conspicuously absent from the non-reference sequence analyses. Further, extracting the population data of the samples after the second round of redundancy removal was complex, resulting in analyses only being performed on the data following the first round of redundancy removal. Given that this research aims to examine population-specific differences in the novel sequences identified, the current format of the pipeline is a direct hinderance to this analysis. Consequently, it is necessary to consider how best to enable population-specific analyses and the HUPAN pipeline must be adapted accordingly before further research with the pipeline can be performed.

# 3. Creating and analysing an African pan-genome

## 3.1. Introduction

### 3.1.1. Major adaptations to the HUPAN pipeline

Following the testing of the first half of the HUPAN pipeline (Duan et al., 2019), we next aimed to utilise the pipeline to create an African pan-genome using whole genome sequences from diverse African populations. However, to ensure efficient and sustainable use of the software for future research, two limitations of the pipeline first needed to be addressed: the bottleneck caused by the assembler SGA (Simpson & Durbin, 2012), and the lack of population-specific analyses in the pipeline.

The assembly of the test dataset samples by SGA was slow, computationally expensive, and inefficient due to its lack of cached intermediate data. Therefore, various adaptations that could address these concerns were assessed, and Nextflow emerged as the clearly superior option. Nextflow is an *in silico* workflow management system and language that has been specifically designed to be scalable and portable, while ensuring traceable execution of jobs on numerous different scheduling systems (Di Tommaso et al., 2017). Crucially, it allows pipelines written in other languages – such as Perl in the case of HUPAN – to be easily adapted into a Nextflow workflow. Of equal importance, each process designated in a Nextflow workflow produces and stages its own files, meaning that intermediate outputs are cached and can be used to resume workflows from a point of failure. This ensures that no data is lost and no computationally expensive tasks need to be re-run. These features of Nextflow are ideal for the improvement of the assembly step in the HUPAN pipeline; the traceable workflows allow the identification of any bottlenecking jobs, and enable the assignment of suitable amounts of time and compute resources for each separate command. Most importantly for the assembly step, however, is the ability to resume a job from its point of failure. We therefore decided that the SGA assembly step in the HUPAN pipeline should be converted into its own basic Nextflow pipeline, which can be run separately from the original HUPAN software. Crucially, the resulting assemblies can still easily be used in the following HUPAN steps without requiring any reformatting.

The next adaptation to consider was that required for population-specific analysis. In order to include this functionality, two possibilities were considered; either additional scripts and functions could be written to analyse sample IDs and assign the associated population data for the analysis of each step, or each population of interest could be run through the pipeline as separate datasets and the results combined in a single additional step. The former option would require comprehensive scripting to extract this population-specific information, and the scripts would likely be specific for the populations

and samples used in this research. Thus, should other research groups also aim to use the HUPAN pipeline for population-specific analysis, scripting to analyse their own populations and samples would be required. In contrast, the second option would require only one additional step that would utilise the merging and redundancy removal functionality already implemented in the pipeline. This would be easily adaptable for other research groups, although it would also result in the user having to run multiple instances of the pipeline simultaneously, with each instance analysing separate populations. The research performed here is supported by the Ilifu high-performance computing group, which utilises a SLURM job management and scheduling system (Yoo et al., 2003) on a multi-threaded Unix-based cluster. Thus, multiple instances of the same pipeline can be used efficiently, and the same commands submitted for different datasets using simple command line Bash scripting.

Given the relative ease with which the second method can be achieved, we decided that the African pan-genome generated from this research should be assembled by submitting genetically distinct populations to the pipeline separately. This will thereby produce discrete datasets of population-specific non-reference sequences that can then be merged to obtain pan-African sequences. The merging step will be performed once each population's final set of non-reference sequences have been obtained and will thereafter utilise the redundancy removal functionality of CD-HIT that is already incorporated in the pipeline. Additionally, population-specific information will easily be recorded and later analysed as only one round of inter-population clustering will be performed to obtain the pan-African non-reference sequences. In this way, the final non-reference sequences will be collected together as though all the samples had been submitted to the pipeline as one dataset, but the population-specific analysis of these non-reference sequences will be achievable and considerably more informative. The updated schematic diagram of the first stage of the HUPAN pipeline, including the novel Nextflow assembly pipeline and the population-specific instances of the pipeline, is shown in Figure 3.1.

### 3.1.2. The genetic context of African populations

#### *3.1.2.1. African ethnolinguistic groups and ancestral lineages*

As discussed in section 1.2, African populations show extensive intra-population genetic variation when compared to populations from other continents (Gurdasani et al., 2015; Tishkoff et al., 2009). This is equally true of Africa's diverse ethnolinguistic groups, as genomic research has identified comparatively high levels of diversity between populations and population substructures (Rotimi et al., 2017). More than 2000 distinct ethnolinguistic groups are present on the African continent, which make up almost 30% of the world's known languages (Eberhard et al., 2022). This genetic and ethnolinguistic diversity is almost certainly due to the longer time span anatomically modern

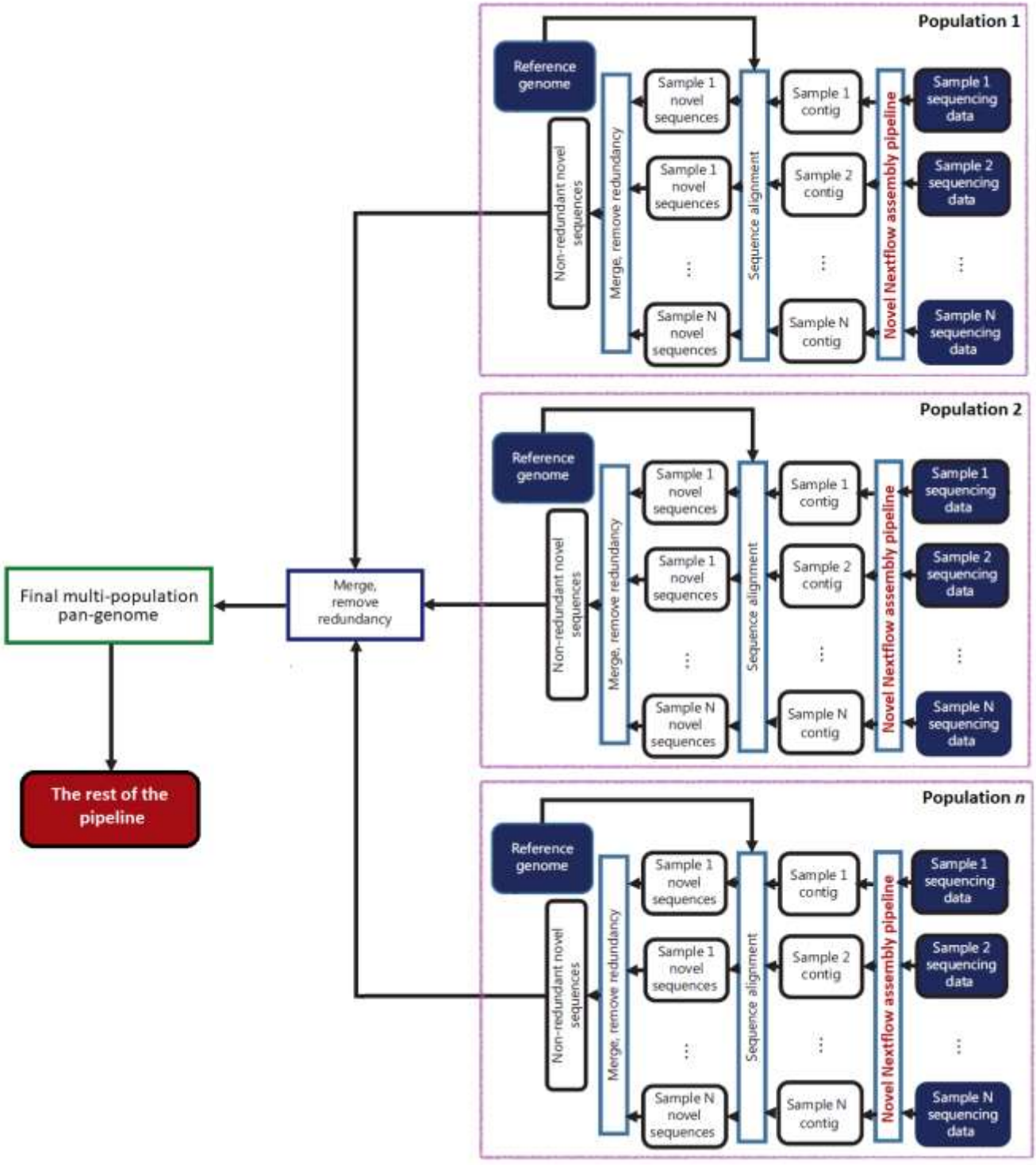


**Figure 3.1. System diagram of the first stage of the adapted HUPAN pipeline (Duan *et al.*, 2019) showing novel techniques used.** The adapted first stage of the HUPAN pipeline requires that each population or ethnolinguistic group included in the resulting pan-genome be submitted to the pipeline as discrete datasets, as depicted by the purple squares. Each population will be run through the first steps of the pipeline separately and will result in non-reference sequences specific to that population. Subsequently, all sets of non-reference sequences will be merged and redundancy will be removed to obtain the final set of non-reference sequences representing all populations present. These sequences are finally combined with the reference genome to create a multi-population pan-genome.

humans have existed in Africa. Traditionally, populations in Africa have been described in the context of four major language families: Afro-Asiatic (AA) which is spoken primarily by agriculturists and agro-pastoralists in northern and eastern Africa; Nilo-Saharan (NS) which is spoken dominantly by pastoralists in eastern and central Africa; Niger-Congo (NC) which is spoken over a very broad region of Africa and is the largest language phylum in the world; and the Khoe and San languages – commonly referred together as Khoisan, Khoesan or Khoe-San – which are languages which contain click consonants and are spoken in eastern and southern Africa by hunter-gatherer and herder populations (Campbell & Tishkoff, 2008; Fan et al., 2019; C. Schlebusch, 2010; C. M. Schlebusch et al., 2020; Tishkoff et al., 2009). The NC language family further contains the Bantu languages whose speakers are extensively dispersed over the African continent due to the Bantu expansion, which was a series of southern and eastern migrations of Bantu-speakers from western central Africa between 4000 and 5000 years ago (Fan et al., 2019; Patin et al., 2017). A great extent of admixture additionally exists between the four language families due to numerous migration events and the far-reaching genetic influence of the Bantu expansion (Choudhury et al., 2020; Tishkoff et al., 2009). As a result, most African populations have highly diverse ancestries and are genetically heterogenous (Fan et al., 2019). Yet despite this vast amount of ethnolinguistic diversity, the lack of African individuals included in human genomic research has only recently been identified as inadequate and exclusionary, as discussed in section 1.2. As a result, genomic analysis of the sub-populations within these four major language families is in its infancy.

Recently, an extensive study that utilised high-coverage whole genome sequences from over 400 individuals comprising 50 ethnolinguistic groups within Africa was performed by Choudhury *et al.* (2020). This research was able to identify over 3 million previously undescribed genomic variants that were predominantly found in individuals from ethnolinguistic groups that had never before been genetically characterised. Notably, using principal component analysis of whole genome sequencing data, this study was able to show that sub-populations within the major language families formed clear clusters relative to each other, and thus share distinct genetic signatures (Fig. 3.2). This confirmed the findings from similar analyses performed previously by other research groups (Fan et al., 2019; Gurdasani et al., 2015). However, due to the breadth of the most recent study, the principal component analysis was able to further distinguish groups of populations within the larger NC language family. It was further shown that the groups separate together based on geographical proximity, as evidenced in Figure 3.2b. In summary, this research indicated that the samples belonging to the NC language family could be further separated into far west, central west, and southern Bantu-speaking Niger-Congo groups.

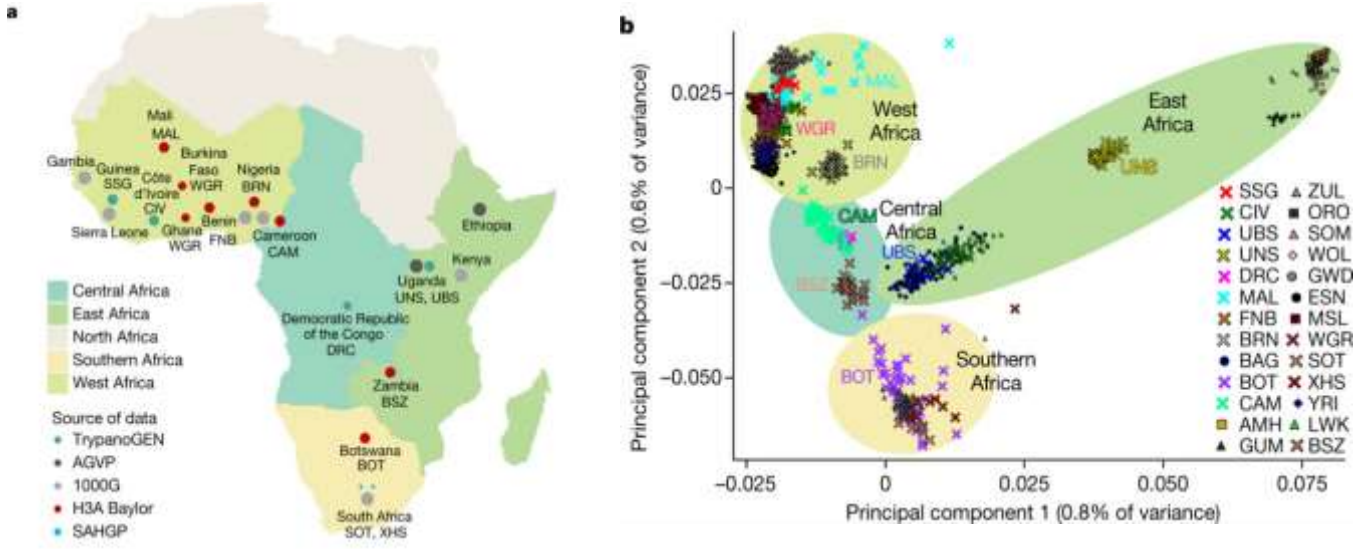


**Figure 3.2. Analysis by Choudhury *et al.* (2020) of whole genome sequence data from 426 individuals of African descent. (a)** Geographical regions, countries of origin and datasets of origin of all samples. **(b)** Principal component analysis of whole genome sequence data showing the first two principal components. Shaded ovals encompassing clusters represent regions in the same colours in (a).

This work by Choudhury *et al.* (2020) did not include individuals belonging to Khoe-San populations, but previous research has shown that sequences obtained from Khoe-San samples separate strongly and distinctly from the other African language families (Fan et al., 2019; Gurdasani et al., 2015). Similarly, the hunter-gatherer populations found in the central African rainforests – previously but derogatorily referred to as Pygmies – have also been shown to form distinct genetic clusters separate from the other ethnolinguistic families in Africa. However, some genetic relatedness to the southern African Khoe-San hunter-gatherer populations has also been suggested (Choudhury et al., 2020; Tishkoff et al., 2009).

When taken together, these early genomic studies of African populations over recent years have shown that there are distinct, measurable genetic differences between at least seven overarching ancestral lineages that can be roughly grouped by region: Afro-Asiatic, Nilo-Saharan, central African rainforest hunter-gatherer, Khoe-San, and the three Niger-Congo groups from far west Africa, central west Africa and southern Bantu-speaking. This classification is by no means exhaustive or complete and should be considered a framework that requires continuous improvement as more studies uncover the intricacies of African population genetic substructures. However, given that this framework unambiguously includes individuals from many diverse and geographically distant populations, it is appropriate and useful for the purposes of this study. Consequently, high quality whole genome sequences from populations within each of these seven regional ancestral groups must be included in any research that aims to assemble a more inclusive and extensive African pan-genome.

#### *3.1.2.2. Data accessibility of African whole genome sequences*

Another aspect that needs to be considered in the assembly of an African pan-genome is that of data availability for the selected regional ancestral groups. As noted in section 1.2, African individuals had been largely excluded from genomic research until recently (Landry et al., 2018; Need & Goldstein, 2009), and this has unquestionably had an effect on the amount of African genomic data available for research purposes. As a result, finding and then accessing genomic data of African individuals is a challenge that also needs to be addressed.

However, in addition to the limited African samples available for research, the quality of the whole genome sequences that are available also presents a challenge for the use of the HUPAN pipeline. The HUPAN authors state that the pipeline is specifically for the use of deep sequenced whole genome sequences. As defined by Sims *et al.* (2014), depth (or coverage) in sequencing is "the average number of times that a particular nucleotide is represented in a collection of random raw sequences". For example, if a whole genome sequence is sequenced at a depth of 10X, this means the average number of times that every single nucleotide base in that sample has been sequenced is 10. Of course, using average sequence depth could mean that certain regions are sequenced many times and other regions sequenced only a few, but higher targeted sequencing depths still ensure that more bases are sequenced multiple times. Logically, then, a more deeply sequenced sample means that each nucleotide can be assigned at a higher degree of confidence, since incorrect assignments become less likely as each nucleotide is sequenced multiple times. For example, when compared to the results of the low-coverage (7.4X) 1kGP dataset, the 1kGP high-coverage (30X) dataset of the exact same samples revealed an additional 18.6 million SNPs (a 1.24-fold increase), 9.8 million indels (a 4.05-fold increase) and around 100 000 structural variants (a 2.47-fold increase). These results clearly indicated the improved accuracy and confidence that high-coverage sequences are able to provide (Byrska-Bishop et al., 2021). Thus, high-coverage whole genome sequences – defined as being sequenced to a depth of 30X or higher – are considered better quality and more likely to completely represent the entire genome of the sample in question. However, the availability of high-coverage whole genome sequences from many different African populations is considerably less than those for individuals of European or Asian descent. Although efforts to rectify this disparity are underway, many high-coverage African genomes that are available tend to be from close geographical regions, such as in the west of Africa, leaving whole swathes of the African continent completely uncharacterised (The H3Africa Consortium et al., 2014).

In light of these difficulties in data access, a particularly important consideration in the creation of the African pan-genome is the number of samples required. Different sample numbers have been used in

all previous pan-genome research, yet comprehensive analysis to identify the optimal number of samples for the creation of a population-specific pan-genome has not been performed. For example, in the HUPAN research, Duan *et al.* (2019) showed that the amount of additional non-reference sequence identified increased by less than 4 megabase pairs (Mbp) from the inclusion of the 50$^{th}$ to the 100$^{th}$ sample. In a different analysis, Li *et al.* (2021) demonstrated that they could obtain 90% of the identified common non-reference sequences using 279 samples out of the original dataset of 486. Clearly, there is as yet no standardised method for determining the sample size required for a comprehensive pan-genome, and the answer will be highly dependent on the overall quality and coverage of the dataset used, as well as the genetic diversity of the population in question (discussed in section 3.1.3). Because of this, the HUPAN authors used a dataset of 30X samples, while Li's group used an over 50X dataset. These more deeply sequenced samples are known to produce better assemblies with fewer errors and misassemblies, thus resulting in more easily extractable non-reference sequences. Consequently, this implies that research using lower coverage samples will be required to include more samples to obtain the same level of significance as studies using higher coverage samples (Y. Li et al., 2011). Ideally, researchers should attempt to reach the point where the total amount of non-reference sequence increases only slightly with the addition of more samples. In the case of African pan-genomic research, however, this conclusion is once again likely irreconcilable with the total amount of available high-coverage African whole genomes sequences.

Further, in order to ensure each ancestral group is represented equivalently in the final non-reference sequences and resulting pan-genome, it is necessary to have a similar number of samples representing each group. This poses an additional challenge given the unequal distribution of sampled genomes in Africa. As a consequence, the sample number and coverage issue can only be resolved once the availability of high-coverage African whole genome sequences is fully investigated. However, since the HUPAN research identified the 50$^{th}$ sample as a potential lower limit for high-coverage samples, we have tentatively aimed for a similar number of samples in each regional ancestral group. Should the number of high-coverage African samples for each group not be able to meet this number, the use of medium-coverage (10X) or even low-coverage datasets must be considered, but strict quality assessment steps will be taken to ensure that the reliability of the data is not compromised with the inclusion of these samples.

### 3.1.3. Findings from previous pan-genome research

As previously discussed (section 1.1.3.2), human pan-genome efforts have been ongoing for more than a decade (R. Li et al., 2010) but there has only recently been a convergence on the methods used to assemble them. Of the recent human pan-genome publications, the HUPAN pipeline is the only one

to utilise *de novo* assembly of entire genome sequences, rather than anchored pseudo *de novo* assembly of unaligned reads. Among recent pan-genome publications, an immediate and noticeable difference between the results of these two methods is the total amount of novel or non-reference sequence identified. Notably, the HUPAN authors discovered only 29.5 Mbp of novel sequence from a Han Chinese dataset, while in contrast, Sherman *et al.* (2019) identified 296.5 Mbp from an African-descent cohort, and Li *et al.* (2021) identified 276 Mbp from a different Han Chinese cohort, both using the method of anchored pseudo *de novo* assembly. While both of these publications utilised far more samples (totals of 910 and 486 respectively) than the 275 samples used in the HUPAN research, the almost 10-fold difference in amount of novel sequence identified cannot solely be attributed to the sample size.

The methodology for obtaining non-reference sequences is not the only difference between these bodies of research, and these additional disparities likely also play a large role in the anomalies in results. Li *et al.* (2021) used samples sequenced at an average coverage of 53.6X, which is almost double that of the HUPAN samples. This will certainly improve the quality of the assembled genomes, and thus sequences that do not align to the reference are more likely to be identified. Their methodology also included any unaligned sequences longer than 200 bp, while the HUPAN pipeline only extracts unaligned sequences longer than 500 bp. The distribution of novel sequences identified by pan-genome research utilising short-read sequenced samples is known to be skewed towards contigs shorter than 1 kilobase pair (kbp), so it can be assumed that considerably more novel sequence would have been identified had the sequence length threshold been lower in the HUPAN pipeline. Additionally, Li's group removed repetitive portions of sequences before clustering, but kept any non-repetitive portion on the same contig if it was longer than 200 bp; since repetitive regions are more likely to form clusters due to their similar identity, removing repetitive portions before clustering will certainly result in less sequence being merged during redundancy removal. All of these factors will undoubtedly have major effects on the final amount of non-reference sequence identified, and could explain the difference in sequence amount identified by the Li and HUPAN groups despite their use of samples from the same Han Chinese population. Sherman *et al.* (2019), on the other hand, used samples sequenced at a similar coverage to the HUPAN cohort (30-40X), but only included sequences longer than 1 kbp in their final dataset, which would reduce the total amount of novel sequence identified compared to the HUPAN research. However, this group also used an identity threshold of 98% for placed contig clustering and 95% for unplaced (i.e., fully unaligned) contig clustering, thereby potentially increasing the total sequence obtained compared to the HUPAN dataset. Another crucially important factor that is independent of methodology and that could explain the large amount of sequence identified in this research is the highly diverse and admixed samples from various African

and African-descent populations used, as described by the authors. Individuals included in the dataset originated from 19 different populations, of which nine self-identified as African American, two were from the African continent, and the rest were from Central America and the Caribbean. This degree of population diversity included by Sherman *et al.* (2019) in the study might explain the large amount of novel sequence identified. In the research performed here, our new African pan-genome will be compared to that of Sherman *et al.* (2019), and any similarities or differences will be highlighted and discussed.

Conversely, one characteristic of the non-reference sequences that was similar between the research performed by the HUPAN, Li and Sherman groups was the abundance of repeat sequences. Over 75% of the HUPAN fully unaligned sequences were classified as repetitive, while the Sherman group categorised 88% of the unplaced contigs as repetitive, totalling almost 257 Mbp (Sherman et al., 2019). The Li group identified only 5.8% of the unplaced contigs as repeats due to the removal of repetitive sequences early in the workflow, but this still translated into 16 Mbp of repeat sequence in total (Q. Li et al., 2021). Repetitive DNA is comprised of sequences that have either similar or identical sequences in other regions of the genome, and these make up around 50% of the human reference genome sequence (Treangen & Salzberg, 2011). Previously considered to be "junk" or "filler" DNA, repeat sequences are now known to be highly important in the organisation and evolution of genomes (Shapiro & Sternberg, 2005). However, their repetitive nature poses a computational problem for alignment and assembly algorithms as it causes ambiguity in the analyses. This is true for assembly in particular; it is often the case that the length of a sequenced read is shorter than the repeat region itself, which therefore increases the likelihood of the sequence being misassembled (Zerbino & Birney, 2008). As a result, repeat sequences have often been left out of previous genomic analyses (Slotkin, 2018; Treangen & Salzberg, 2011) and, thus, the overrepresentation of repeats among sequences that cannot align to the reference is expected. However, the increasing use of paired-end reads – which are sequenced reads that cover two ends of the same DNA molecule and thereby improve confidence in the validity of the sequence produced by the sequencer – in combination with long-read sequencing has improved the ability of assembly algorithms to resolve repeat sequences (Korbel et al., 2007; Miller et al., 2008). In the work presented here, the data used are exclusively paired-end sequences. It is therefore expected that many repeat regions will be resolved and extracted, and the repeat content of the African non-reference sequences produced in this research will likely be similar to that produced in other pan-genome work.

#### *3.1.3.1. Partial validation of non-reference sequences in pan-genome research*

Each of the pan-genome publications focussed on here had their own method for validation of the non-reference sequences. The HUPAN authors used *ab initio* methods to predict 167 novel genes from the non-reference sequences, of which 70% were shown to be expressed using two Han Chinese RNA-seq datasets (Duan et al., 2019). This confirmed the presence of those gene transcripts in at least one individual in the dataset used. Li *et al.* (2021) conducted sequence alignment using the non-reference sequences that were present in more than one individual, and showed that over 90% of all unaligned reads from all samples could be mapped back to those "common" sequences, although the thresholds used for alignment coverage and identity (defined in section 2.3.2.1) were not specified. They also showed that supplementing the human reference genome with the identified non-reference sequences improved subsequent variant calling for an individual of Han Chinese descent. Sherman *et al.* (2019) aligned the African-descent non-reference sequences to non-reference sequences obtained from European and African samples from the Simons Genome Diversity Project dataset, and showed a slightly higher presence of the African-descent non-reference sequences in the African individuals than in the Europeans. Similarly, both Duan *et al.* (2019) and Li *et al.* (2021) used alignment of the non-reference sequences against various other datasets to validate and provide genetic context to the novel sequences.

We lack access to African population-specific RNA-seq datasets, and variant calling experiments are currently beyond the scope of this research. However, alignment to various datasets and sequences will be performed and the results analysed in an attempt to partially validate the non-reference sequences identified. Additionally, since the publication of the previous pan-genome studies, a complete telomere-to-telomere assembly of the human genome has been released (Nurk et al., 2021) and this can potentially be used for an additional validation. This sequence, assembled by the Telomere-to-Telomere Consortium and termed T2T-CHM13 (telomere-to-telomere complete hydatidiform mole), addresses the problem of the missing sequence in the human reference genome and is reportedly a gapless assembly of the 22 autosomes and the X chromosome. The methods used to assemble the initial human reference genome restricted assembly to only easily resolved and mainly euchromatic regions of the human genome, thereby excluding or incorrectly assembling many repetitive and polymorphic regions. However, recent advances in long-read sequencing technologies and assembly algorithms have enabled the Telomere-to-Telomere Consortium to assemble and release the first completely gapless homozygous human genome, containing an additional 130 Mbp of sequence and over 3000 additional genes. With regard to the following work, we argue that, should there be any sequences in the African non-reference dataset that cannot align to the human reference

genome but do align to T2T-CHM13, this will validate those sequences as being absent from the reference genome, and thus non-reference with regard to GRCh38.

#### *3.1.3.2. Gene presence-absence variation profiles*

As discussed in section 1.1.3.2, pan-genomes are made up of the core genome, which every individual in the population shares, and the distributed genome, which is only found within certain members or even single individuals in the population. However, the exact method of defining a core genome varies greatly between the different bodies of pan-genome research recently published. Li *et al.* (2021) did not specifically refer to the core or distributed genomes, but they defined "common" sequences as any sequence that was present in two or more individuals. Sherman *et al.* (2019) screened the novel sequences to determine which were present or absent in each of the 910 samples, creating what is called a presence-absence variation (PAV) profile. For this, a minimum alignment threshold of 80% coverage and 90% identity was required to classify a sequence as present within a sample.

For the HUPAN pipeline, Duan *et al.* (2019) have allowed the user to define their own threshold for what constitutes a core genome. Firstly, rather than identifying sequences that are present in every individual, the HUPAN pipeline utilises the reference genome gene annotations and the novel genes predicted from the non-reference sequences to create a gene PAV profile, and not a sequence PAV profile. Although the reasons for this are not stated in the publication, this method was likely used because it is simpler and less computationally demanding to determine the presence or absence of a few thousand protein coding genes than of every sequence in the human genome, thus ensuring the pipeline is more accessible for other researchers wishing to use it. The pipeline creates the gene PAV profile by aligning the raw sequenced unassembled reads from each sample to the pan-genome sequence, and then mapping the sequence alignment results to the gene annotations to identify whether the sequence producing the annotation is covered by reads. Secondly, the user can decide whether to use coverage of coding DNA sequence (CDS) or coverage of the entire gene sequence to determine presence or absence. Although these options sound similar, there is an important distinction; gene coverage means that the entire gene sequence – including introns, exons and regulatory sequences – must be covered by a read to be defined as present, while CDS coverage means that only the regions that are translated into protein must be covered by reads. Finally, the HUPAN pipeline also allows the user to define the read coverage threshold – which is the percentage of the total sequence length that is covered by the sequenced reads of a sample – for an annotation to be considered present. In their own research, for example, the HUPAN authors used CDS coverage to determine gene presence or absence and used a minimum coverage threshold of 95%. This meant that 95% of the CDS regions within a gene had to be covered or aligned to reads from an individual

sample for that gene to be defined as present in the individual. Further, they defined a distributed gene as any gene that was absent in at least one individual. When this is repeated for every gene annotation in every sample used to create the pan-genome, a gene PAV profile is created, and the core and distributed genes can be identified. The HUPAN authors stated that they chose this threshold because there was no significant change in the number of core and distributed genes when the coverage was decreased to lower than 95%. Overall, at this threshold, they identified 606 distributed genes (Duan et al., 2019). Clearly, however, the number of core and distributed genes will differ depending on the thresholds set, and the effects of changing these thresholds are explored in the following work.

## 3.2. Materials and methods

### 3.2.1. Adapting the HUPAN pipeline

A Nextflow (Di Tommaso et al., 2017) pipeline for the assembly step using SGA (version 0.10.15) (Simpson & Durbin, 2012) was written by Gerrit Botha and fully tested and adapted by myself. The final SGA Nextflow pipeline instructions and configuration files can be accessed at https://github.com/BournSupremacy/dec-2020-hackathon-stream1/tree/main/assembly. Where it was required for the software to run successfully on the Ilifu high performance cluster (http://www.ilifu.ac.za/), wall time and memory requirements were increased and the HUPAN scripts updated on a forked version of the HUPAN pipeline, available from https://github.com/BournSupremacy/HUPAN.

### 3.2.2. Obtaining datasets of African whole genome sequences

#### *3.2.2.1. Identifying potential datasets for inclusion*

The International Genome Sample Resource (IGSR) (https://www.internationalgenome.org/home) and the European Genome-phenome Archive (EGA) (https://ega-archive.org/) were searched for datasets containing whole genome sequences of African individuals. From these searches, eight datasets were identified for potential inclusion. From the IGSR, the high-coverage 1000 Genomes Project (1kGP) dataset (Byrska-Bishop et al., 2021), the Human Genome Diversity Project (HGDP) dataset (Bergstrom et al., 2020) and the Simons Genome Diversity Project (SGDP) dataset (Mallick et al., 2016) were appropriate for this research. From the EGA, the Human Heredity and Health in Africa Consortium (H3Africa) dataset (Choudhury et al., 2020), the Trypanosomiasis Genomics Network of the H3Africa Consortium (TrypanoGEN) dataset (Ilboudo et al., 2017), the low-coverage Ethiopian Genome Project (ELC) dataset (Pagani et al., 2015), the high-coverage Ethiopian Genome Project (EHC) dataset (Pagani et al., 2015), and the high-coverage Khoe-San Project (KSP) (C. M. Schlebusch et al., 2020) dataset were appropriate for this study. For all datasets stored on the EGA, data access

agreements were signed, and access to the datasets were granted by the respective DACs or DBACs. Table 3.1 summarises the specifications of each dataset. Population information of the samples of African descent in each dataset was obtained from the metadata files. Each sample was then classified as belonging to one of the seven regional ancestral groups based on the population of origin. Where population information was omitted, the samples were removed from further analysis and excluded from the final dataset. Ultimately, a dataset of 285 African samples was created (Appendix B).

**Table 3.1. Specifications of the datasets available on the IGSR and the EGA that met the criteria for potential inclusion in the pan-African dataset.** Dataset names are specified in the text. Datasets available on the IGSR have no accession numbers associated with them. Accession numbers are given for datasets available on the European Genome-phenome Archive.

| Dataset | Average sequencing depth and paired-end read length | Accession number |
|---|---|---|
| 1kGP | 30X, 150 bp read length | Available on the IGSR |
| HGDP | 35X, 151 bp read length | Available on the IGSR |
| SGDP | 43X, 100 bp read length | Available on the IGSR |
| H3Africa | 30X, 150 bp read length | EGAS00001002976 |
| TrypanoGEN | 10X, 150 bp read length | EGAD00001005076 |
| ELC | 4X, 100 bp read length | EGAD00001000598 |
| EHC | 30X, 100 bp read length | EGAD00001000696 |
| KSP | 30X, 100 bp read length | EGAD00001006183 |

#### *3.2.2.2. Assembly and quality assessment of whole genome sequences*

Samples from the 1kGP, HGDP and SGDP datasets are publicly accessible and had been previously downloaded to the Ilifu high performance computing cluster by Gerrit Botha using Globus Online (Foster, 2011) or the European Nucleotide Archive (ENA) (Leinonen et al., 2011) FTP site on 10 April 2020, 19 October 2020, and 29 June 2020 respectively. The EGA samples were downloaded by myself using the EGA download client, pyEGA3 (version 3.4.1). The H3Africa dataset, the TrypanoGEN dataset, the EHC and ELC datasets, and the KSP dataset were downloaded on 10 September 2021, 29 June 2021, 1 April 2021 and 4 October 2021 respectively.

All samples were obtained in BAM or CRAM file format (all based on alignment to GRCh38), and Gerrit Botha's pre-built Nextflow pipeline (available from https://github.com/grbot/varcall/tree/master/cram-to-fastq) was used to convert all BAM/CRAM files to FASTQ format. All 285 samples were submitted to the SGA Nextflow assembly pipeline and assembled within 3 to 10 days. Those samples that could not be assembled within the time limit of 10 days were excluded from further analysis. I then created a Nextflow pipeline (accessible at https://github.com/BournSupremacy/Bioinformatics

Tools/tree/main/QuastMultiQC) to assess the assembled samples using QUAST (version 5.0.2) (Gurevich et al., 2013) and MultiQC (version 1.9) (Ewels et al., 2016). Particular attention was paid to the total assembled length, and therefore fraction of the reference genome length, and N50 values of the assembled samples. The human reference genome is around 3.1 gigabase pairs (Gbp) long, and a simple calculation can therefore determine what fraction of the reference genome length each sample is once assembled. The N50 is a statistic describing a type of weighted median of contig lengths making up an assembly, whereby higher N50 values indicate that the assembly contains more contigs of greater lengths. Following examination of the total assembled length of each sample, a minimum threshold of 60% of the reference genome length was chosen for the samples to be considered acceptable quality for inclusion in the final sample dataset (see Implementation and Results section 3.3.1). Thus, only samples whose total assembled length was at or above 1.86 Gbp were included in further analysis.

### 3.2.3. Preparing the HUPAN pipeline and submitting the datasets

The HUPAN pipeline had been prepared as previously described in section 2.2.1. The human reference genome (build 38, GRCh38), NCBI's non-redundant nucleotide database, taxonomy database and associated accession numbers file had also already been prepared as described in section 2.2.2. The human reference genome transcript (release 35) and annotation (release 38) files were downloaded from GENCODE (Frankish et al., 2019) on 15 July 2021 and 12 March 2021 respectively. Datasets of human cDNA expressed sequence tags (ESTs), expressed mRNA sequences (mRNA-seq) and protein sequences were downloaded using NCBI's data retrieval system Entrez (https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html) for *ab initio* gene prediction using MAKER (version 3.01.03) (Holt & Yandell, 2011). The search query used to obtain the EST dataset was `txid9606[orgn] AND gbdiv_est[prop] AND is_est[filter]`, to obtain the mRNA-seq dataset was `txid9606[orgn] AND mRNA[All Fields]` (both in the Nucleotide database), and to obtain the protein database was `txid9606[orgn]` in the Protein database. The datasets were downloaded on 21 April 2021, 22 May 2021, and 13 April 2021 respectively.

The samples were separated into ancestral group-specific directories on the Ilifu high performance computing cluster. Each group's dataset was submitted to the HUPAN pipeline separately as described in the README.md file in the forked version of the HUPAN GitHub repository https://github.com/BournSupremacy/HUPAN/blob/master/README.md). Once the non-reference sequences for each group had been obtained, the sequences were merged and redundancy removed at a 90% identity threshold using CD-HIT (version 4.8.1) (Fu et al., 2012, p.). All group-specific and pan-African sequence files were then submitted to an online version of QUAST available at http://cab.cc.spbu.ru/quast/ to

summarise and compare the sequence statistics. The seven groups' non-reference sequences were also submitted to MAKER using the HUPAN pipeline to obtain the RepeatMasker (version 4.1.0) (Nishimura, 2000) annotation of repeat sequences. The pan-African non-reference sequences were then submitted to the rest of the HUPAN pipeline to identify potential novel genes as predicted by MAKER, and the African pan-genome was created by merging the GRCh38 sequence file with the pan-African non-reference sequences. The novel predicted gene annotations were also merged with the annotations in the human reference genome gene annotation file. Following this, gene annotations in the reference genome annotation file that were not from primary reference sequences (i.e., those from patch, alternate, decoy or unplaced contigs) were removed. Genes from the Y chromosome were also removed as these are not present in female samples and so will be defined as distributed in the dataset even if they are present in every male sample.

### 3.2.4. Comparative alignments of the African non-reference sequences

Sequences used for alignment were GRCh38 (previously downloaded), the telomere-to-telomere human genome (referred to as T2T-CHM13) (Nurk et al., 2021) under GenBank accession number GCA_009914755.4, the HUPAN Han Chinese non-reference sequences assembled by Duan *et al.* (2019) from the data repository at https://cgm.sjtu.edu.cn/hupan/download.php, and the African-descent non-reference sequences assembled by Sherman *et al.* (2019) under GenBank accession number PDBU00000000.1. A Fon sample from Benin (A0018) from the H3Africa high-coverage dataset and a Han Chinese sample from China (HGDP00777) from the high-coverage HGDP dataset were also downloaded in CRAM format and converted to FASTQ format as previously described, before being assembled for alignment purposes. A Singularity (Kurtzer et al., 2017) container for pblat (version 2.5) (M. Wang & Kong, 2019) was created and used to align the pan-African non-reference sequences to the Han Chinese non-reference sequences and the African-descent non-reference sequences using the parameters `-minMatch=6 -minIdentity=50`. The Singularity container for NUCmer (version 3.1) (Kurtz et al., 2004) which had been previously created, was used to align the African non-reference sequences to GRCh38, T2T-CHM13, A0018 and HGDP00777 using default parameters (`-b 200 -c 65 -g 90 -l 20`). Analysis of the alignment outputs was performed using a Python script which can be accessed at https://github.com/BournSupremacy/BioinformaticsTools/tree/main/alignment.

### 3.2.5. Analysis of the African non-reference sequences and pan-genome

All analysis of the final sets of non-reference sequences was performed using using Python (version 3.8.3) scripts within JupyterNotebooks in JupyterHub (single-user server version 2.2.0), accessible

through the Ilifu high performance computing cluster. The scripts for all plots and analyses are accessible from https://github.com/BournSupremacy/BioinformaticsTools/tree/main/PythonPlots.

## 3.3. Implementation and results

### 3.3.1. Assembly of the final 168-sample dataset

In total, 285 whole genome sequences from individuals of African ethnicity were obtained and assembled. Each regional ancestral dataset consisted of 29 samples or more. Only samples from east Africa were included in the Afro-Asiatic and Nilo-Saharan groups due to a lack of other available datasets, so these ancestral groups were thereafter labelled east African Afro-Asiatic and east African Nilo-Saharan. For ease of reference, abbreviations were utilised for each group in the following work: BantuNC for Bantu-speaking Niger-Congo, CARF for central African rainforest hunter-gatherer, CWNC for central west Niger-Congo, EAAA for east African Afro-Asiatic, EANS for and east African Nilo-Saharan, FWNC for far west Niger-Congo, and KhoeSan for the Khoe and San. A summary of the samples is shown in Table 3.2, and a detailed description of all samples can be found in Appendix B.

All 285 samples were submitted to the SGA (Simpson & Durbin, 2012) Nextflow pipeline. Most samples were assembled within 7 days using this pipeline, with the minimum time being 3 days and the maximum time being 10 days. This was a substantial improvement from the original SGA workflow, which took a minimum of 11 days and a maximum of 15 days to assemble a single sample. However, of the 285 samples, six could not be assembled by the SGA Nextflow pipeline within the maximum allocated wall time of 10 days. Notably, all six of these samples originated from the HGDP datasets sampled from the Central African Republic or the Democratic Republic of Congo (shown in Table 3.3, discussed shortly). This potentially indicated an issue with the protocols used to produce the sequencing files for these samples, as these samples were all sequenced in the same data centre.

Following further quality assessment of the remaining 279 whole genome sequences using MultiQC (Ewels et al., 2016) (reports provided in Appendix C), many of the samples from specific datasets were identified as being insufficiently complete for the requirements of the HUPAN pipeline. This was particularly apparent when assessing the total assembled length – and therefore fraction of the reference genome – and N50 values (see Materials and Methods section 0). Unsurprisingly, the length and N50 values were tightly correlated within the assembled genomes. Of note, all of the

**Table 3.2. Summary of all samples and datasets obtained for potential inclusion in the African pan-genome.** The average sequencing depths and the paired-end read lengths of the samples in each dataset are shown in brackets in the third column. 1kGP: 1000 Genomes Project dataset, HGDP: Human Genome Diversity Project dataset, SGDP: Simons Genome Diversity Project dataset, H3Africa: H3Africa Consortium dataset, TrypanoGEN: Trypanosomiasis Genomics Network dataset, EHC: Ethiopian high-coverage dataset, ELC: Ethiopian low-coverage dataset, KSP: Khoe-San Project dataset.

| Regional ancestral group | Number of samples | Datasets used |
|---|---|---|
| Bantu-speaking Niger-Congo (BantuNC) | 37 | 18 samples from 1kGP (30X, 150 bp reads)<br>13 samples from HGDP (35X, 151 bp reads)<br>6 samples from SGDP (43X, 100 bp reads) |
| Central African rainforests (CARF) | 32 | 27 samples from HGDP (35X, 151 bp reads)<br>5 samples from SGDP (43X, 100 bp reads) |
| Central-west Niger-Congo (CWNC) | 54 | 25 samples from H3Africa (30X, 150 bp reads)<br>29 samples from TrypanoGEN (10X, 150 bp reads) |
| East-African Afro-Asiatic (EAAA) | 29 | 1 sample from SGDP (43X, 100 bp reads)<br>4 samples from EHC (30X, 100 bp reads)<br>24 samples from ELC (4X, 100 bp reads) |
| East-African Nilo-Saharan (EANS) | 66 | 6 sample from SGDP (43X, 100 bp reads)<br>1 sample from EHC (30X, 100 bp reads)<br>10 samples from ELC (4X, 100 bp reads)<br>49 samples from TrypanoGEN (10X, 150 bp reads) |
| Far-west Niger-Congo (FWNC) | 35 | 15 samples from 1kGP (30X, 150 bp reads)<br>10 samples from HGDP (35X, 151 bp reads)<br>10 samples from H3Africa (30X, 150 bp reads) |
| Khoe and San (KhoeSan) | 32 | 25 samples from KSP (30X, 100 bp reads)<br>2 sample from HGDP (35X, 151 bp reads)<br>5 sample from SGDP (43X, 100 bp reads) |
| **Total** | **285** | |

samples obtained from the Ethiopian low-coverage dataset (4X) had a total assembled length that was less than 25% of the human reference genome length, and therefore could clearly be considered incomplete and unsuitable for pan-genome analysis. Consequently, it was decided that a minimum threshold for the fraction of the reference genome length needed to be established, as any assemblies that were significantly smaller than the reference length would be lacking substantial amounts of sequence that is both absent and present within the reference genome. Thus, the likelihood of extracting valid, correctly assembled non-reference sequences is reduced in samples with notably smaller assembled lengths. Upon examination of the MultiQC data, it became apparent that many of the TrypanoGEN dataset (10X) samples had reference genome fractions of between 50% and 70%. Thus, in order to include as many samples as possible while removing those of poorer quality, it was decided that any sample whose total assembled length was less than 60% (1.86 Gbp) of the reference genome sequence length would be excluded from further analysis (see Table 3.3).

Following the implementation of this quality threshold, 97 samples were removed from the dataset, leaving 182 samples. Of the 97 samples removed, 35 were from the Ethiopian low-coverage dataset (4X), 9 were from the SGDP dataset (43X) and 53 were from the TrypanoGEN dataset (10X) (see Table 3.3). Unfortunately, and although a total dataset size of 182 is acceptable for the HUPAN pipeline in general, the failure to assemble certain genomes and the removal of low-quality samples considerably affected the numbers of samples present in certain specific regional ancestral groups. Most notably, this reduced the number of EANS samples to 20 and the number of EAAA samples to only 5. To examine the effects of this, analyses (see Appendix C) were performed on the distribution profiles of the total lengths and N50 values of the samples in these two groups before and after removing the low-quality genomes. These analyses are depicted in Figure 3.3 and showed that the sample removals considerably improved the average of the assembled length and N50 values for both ancestral groups. For the EAAA group, the average length and N50 increased by 1356.01 Mbp and 0.57 kbp respectively, and by 591.47 Mbp and 0.2 kbp respectively for the EANS group. This improvement in overall quality thereby improves confidence in the results and appears preferable to including more samples of lower quality. The remaining five regional ancestral groups, which all had fewer samples removed than the EAAA and EANS groups, had similar but smaller effects (Appendix D).

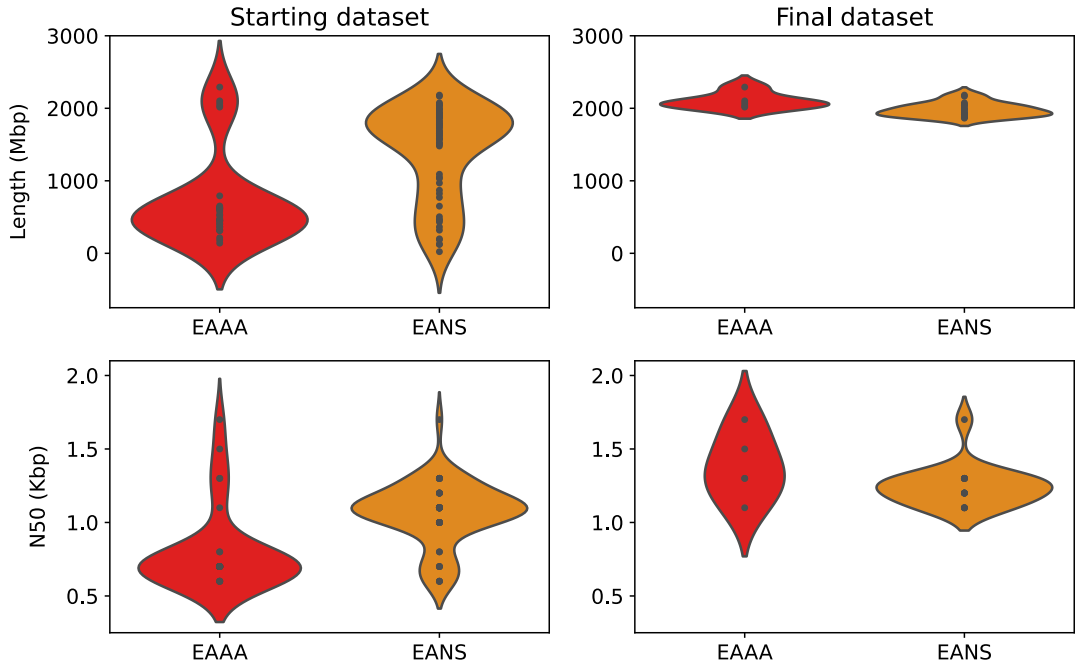


**Figure 3.3. Comparison of the distribution profiles of the total length (Mbp) and N50 values (kbp) before and after removing low-quality samples from the EAAA and EANS datasets.** The EAAA dataset had 29 samples before and 5 samples after performing quality control, and the EANS dataset had 66 samples before and 20 samples after performing quality control. The width of each violin plot indicates the density of samples at those values. Each point represents a sample. In the N50 plots, there are multiple samples with the same N50 values, resulting in overlapping points. Each violin plot shows density distributions extending past both the lowest and highest datapoints due to the nature of the plots.

The quality-controlled datasets were subsequently submitted to the pipeline and the non-reference sequences obtained, but another concern with several samples was detected at the contamination removal step. As shown in Figure 3.4, specific samples from the CARF, EANS and KhoeSan datasets were identified through the HUPAN pipeline as containing a large number of non-reference sequences that were classified as "microbial" (i.e., bacterial, fungal, archaeal, or viral) using BLAST's nucleotide database. The extraordinarily high numbers of microbial sequences in these samples compared to all other samples suggested that there may have been protocol errors in the collection or library and sequencing preparation of the genomic material, resulting in extremely high levels of contamination. Errors such as these may have critical effects on downstream analysis of the sequences and therefore

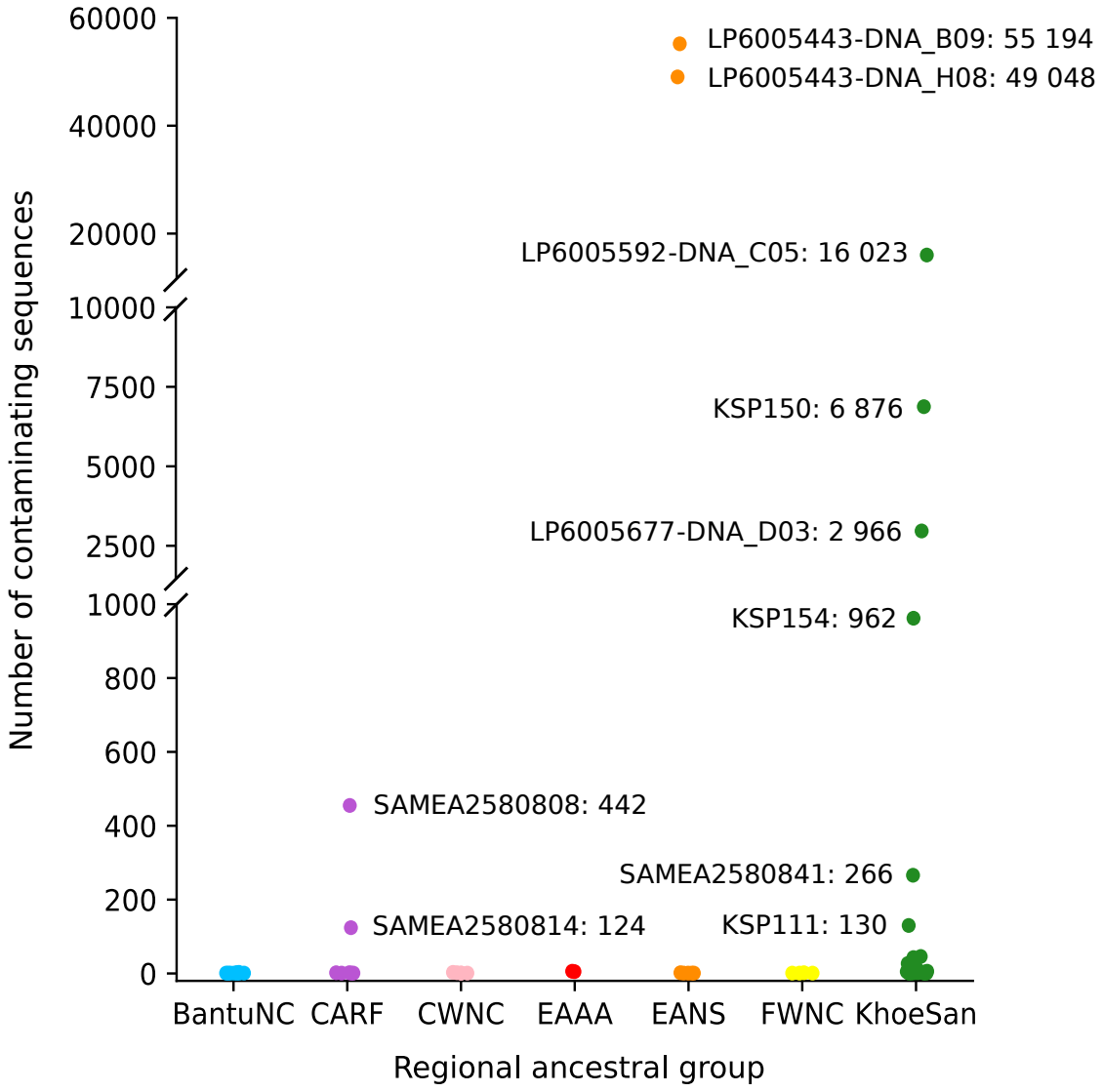


**Figure 3.4. The number of contaminating bacterial, fungal, archaeal, or viral sequences identified by BLAST within the non-reference sequences of each sample from each ancestral group.** Samples identified as containing more than 100 contaminating sequences have been labelled and the number of contaminating sequences indicated alongside.

impact the conclusions of research involving these sequences, even if the contaminating sequences that were identified are removed. As a consideration, there is also the possibility that the sequences identified are not true contaminants and may instead be foreign DNA that has been inserted into the human genome, which is a known phenomenon in mammalian genomes; however, since only certain samples from specific datasets showed these extraordinarily high levels of contamination, it is considerably more likely that the identified sequences are true contaminants that will affect the quality of the final African pan-genome. Of note, the two SGDP samples with the most contaminating sequences were both Dinka samples from Sudan, while the remaining SGDP samples were both KhoeSan samples from Namibia. Similar area- or population-specific patterns of contamination were seen in the HGDP and KSP datasets (see Appendix B), clearly suggesting protocol errors in the laboratories where these samples were prepared. Thus, although the HUPAN pipeline was designed to identify any potential contaminating sequences and remove them, all samples that were recorded

**Table 3.2. Summary of the samples used to assemble the African pan-genome.** The average sequencing depth of the samples in each dataset is shown in the final column. 1kGP: 1000 Genomes Project dataset, HGDP: Human Genome Diversity Project dataset, SGDP: Simons Genome Diversity Project dataset, H3Africa: H3Africa Consortium dataset, TrypanoGEN: Trypanosomiasis Genomics Network dataset, EHC: Ethiopian high-coverage dataset, ELC: Ethiopian low-coverage dataset, KSP: Khoe-San Project dataset.

| Population | Samples removed | Number of samples | Datasets used |
|---|---|---|---|
| BantuNC | 2 from SGDP: failed QC | 35 | 18 samples from 1kGP (30X)<br>13 samples from HGDP (35X)<br>4 samples from SGDP (43X) |
| CARF | 6 from HGDP: failed to assemble<br>2 from SGDP: failed QC<br>2 from HGDP: contamination | 22 | 19 samples from HGDP (35X)<br>3 samples from SGDP (43X) |
| CWNC | 20 from TrypanoGEN: failed QC | 34 | 25 samples from H3Africa (30X)<br>9 samples from TrypanoGEN (10X) |
| EAAA | 24 from ELC: failed QC | 5 | 1 sample from SGDP (43X)<br>4 samples from EHC (30X) |
| EANS | 4 from SGDP: failed QC<br>33 from TrypanoGEN: failed QC<br>10 from ELC: failed QC<br>2 from SGDP: contamination | 18 | 1 sample from SGDP (43X)<br>1 sample from EHC (30X)<br>16 samples from TrypanoGEN (10X) |
| FWNC | None | 35 | 15 samples from 1kGP (30X)<br>10 samples from HGDP (35X)<br>10 samples from H3Africa (30X) |
| KhoeSan | 2 from SGDP: failed QC<br>1 from HGDP: contamination<br>2 from SGDP: contamination<br>8 from KSP: contamination | 19 | 17 samples from KSP (30X)<br>1 sample from HGDP (35X)<br>1 sample from SGDP (43X) |
| **Total** | | **168** | |

as containing more than 25 contaminating microbial sequences were removed from the pan-African dataset, leaving a total of 168 samples. This final dataset utilised in the pipeline is summarised fully in Table 3.3.

### 3.3.2. Obtaining the pan-African non-reference sequences using the HUPAN pipeline

#### *3.3.2.1. Preliminary comparison of the regional ancestral group datasets*

As a result of the quality control steps, there were notably different numbers of samples in each of the seven regional ancestral groups. In addition, because the different datasets from which these samples were obtained were all sequenced at varying depths (10X to 43X), the overall quality of the samples in each group was also highly varied.

Figure 3.5 shows that the distributions profiles of total sample length and N50 values were evidently different between the groups. For example, the EAAA and EANS groups had considerably lower total assembled lengths and N50 values than the other groups, indicating that they consisted of samples

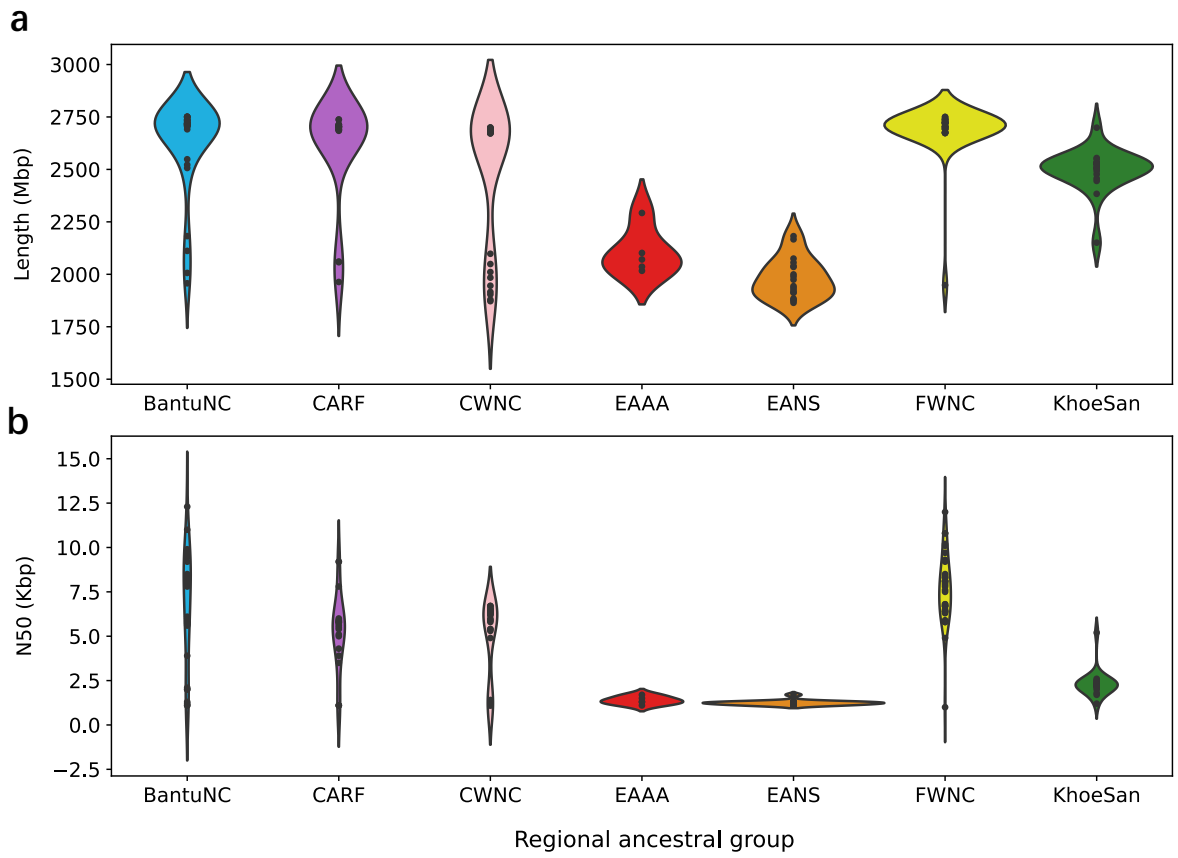


**Figure 3.5. Distribution profiles of assembly metrics of the samples comprising each regional ancestral group.** The width of each violin plot indicates the density of samples at that value. Each violin plot shows density distributions extending past both the lowest and highest datapoints due to the nature of the plots. **(a)** Length in Mbp and **(b)** N50 values in kbp of each assembled sample.

whose sequences were less comprehensive than those in other groups. Conversely, the BantuNC and FWNC groups appeared to be comprised of samples with high N50 values and total lengths close to the size of the human reference genome. This is important to note, as all further analysis of the non-reference sequences and resulting pan-genome will need to take these differences into account.

### *3.3.2.2. Obtaining regional ancestral group non-reference sequences*

The final set of 168 samples within the seven regional ancestral groups was submitted to the HUPAN pipeline and the non-reference sequences for each population were obtained. As shown in Table 3.4, removing redundancy within groups at the final stage did not greatly affect the total amount of sequence present. However, in general, the amount of sequence greatly decreased following the first redundancy removal step, indicating a large amount of similar genetic sequences within each group. Notably, the groups with the most samples showed the greatest reduction in sequence following the first round of redundancy removal, which suggested that having many more samples in single groups may not necessarily result in the identification of more novel non-reference sequence, since the amount of redundancy present will likewise increase.

**Table 3.3. The number and length of non-reference sequence identified for each regional ancestral group before and after removing redundancy using a 90% identity threshold.** Only values for the fully unaligned sequences are shown before the first redundancy removal step.

| **Population** | **Before first redundancy removal (fully unaligned sequences)** | | **Before second redundancy removal** | | **After second redundancy removal** | |
|---|---|---|---|---|---|---|
| | Number of contigs | Length (Mbp) | Number of contigs | Length (Mbp) | Number of contigs | Length (Mbp) |
| BantuNC | 78 962 | 68.75 | 17 055 | 34.19 | 12 020 | 29.39 |
| CARF | 47 868 | 41.72 | 14 074 | 24.34 | 10 334 | 20.81 |
| CWNC | 23 357 | 16.67 | 10 936 | 14.00 | 9 335 | 12.80 |
| EAAA | 8 276 | 7.43 | 5 555 | 6.85 | 4 762 | 6.17 |
| EANS | 33 188 | 25.47 | 11 601 | 12.33 | 9 238 | 10.39 |
| FWNC | 57 901 | 50.18 | 15 240 | 31.03 | 11 099 | 27.10 |
| KhoeSan | 28 903 | 21.72 | 10 904 | 12.69 | 9 003 | 11.15 |

Overall, and as expected, the groups that contained the most samples, such as the BantuNC and FWNC groups, resulted in the most non-reference sequence in both number of contigs and total length (Fig. 3.6a and b). Surprisingly, however, the CWNC group, which consisted of 34 samples, only generated the fourth highest amount of total non-reference sequence at 12.8 Mbp, less than the CARF group

which yielded a notably higher amount of sequence despite having 12 fewer samples. Likewise, the EANS and KhoeSan groups produced similar amounts of sequence to the CWNC group despite consisting of only 18 and 19 samples respectively. The smaller amount of sequence is mirrored in the N50 values of the contigs making up the non-reference sequences identified (Fig. 3.6c), as the groups with lower N50 values are those with the least amount of non-reference sequence. Overall, this indicated that the groups with the most sequence (BantuNC, CARF and FWNC) yielded higher numbers of long sequences, which explains the clear discrepancies in total sequence amount between groups. This is likely due to the different datasets used in each regional ancestral group, as the N50 trends in Figure 3.6c closely resemble those in Figure 3.5b. As expected, the EAAA group resulted in the least amount of sequence as it consisted of only five samples in total.

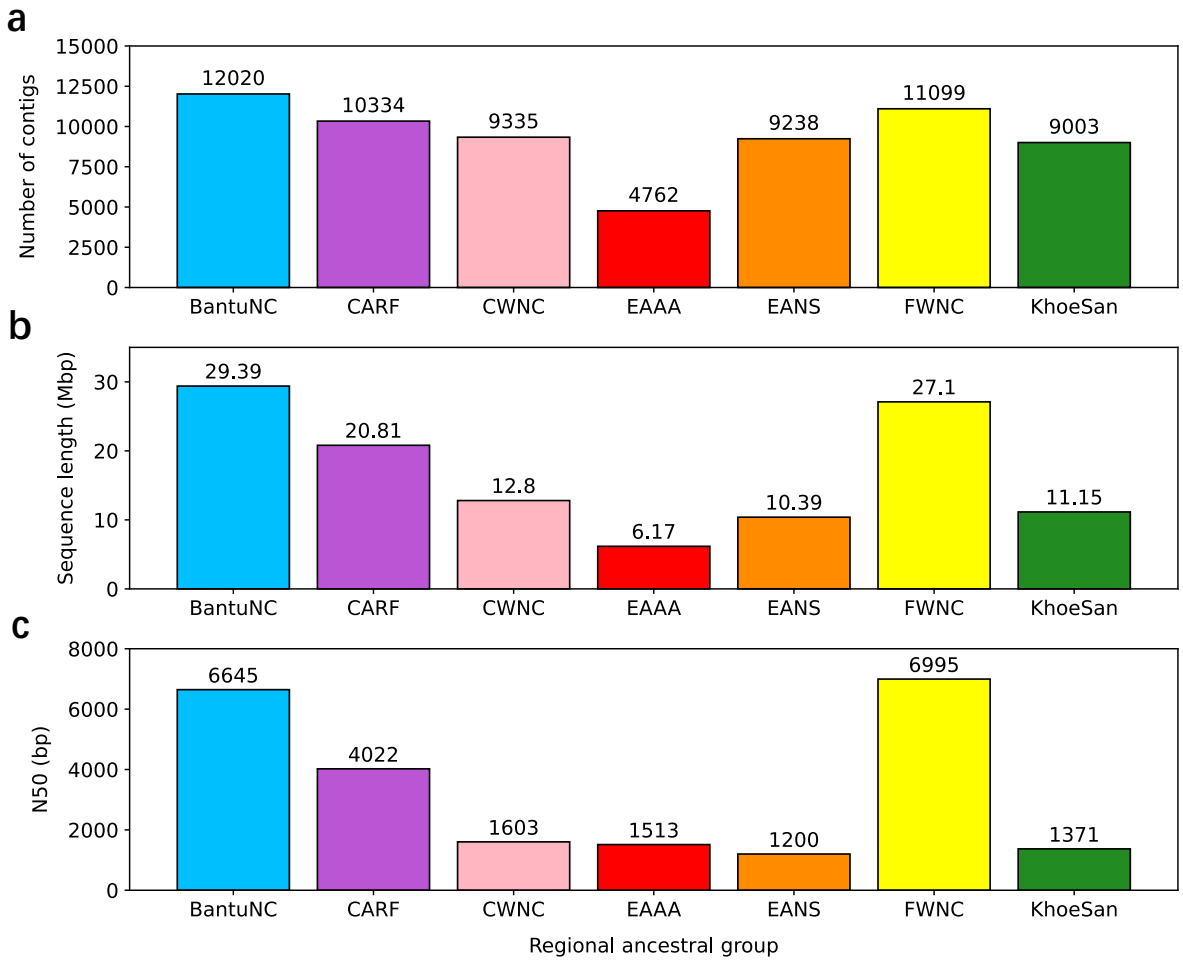


**Figure 3.6. Metrics of the non-reference sequences identified from each of the seven regional ancestral groups. (a)** The number of contigs making up the non-reference sequences. **(b)** The total length (Mbp) of non-reference sequence. **(c)** The N50 (bp) of the contigs making up the non-reference sequences.

### *3.3.2.3. Merging the pan-African non-reference sequences*

Using the novel additional step, the non-reference sequences from each population were merged and redundancy was removed at the same 90% identity threshold as previously. The metrics of the final pan-African non-reference sequences are shown in Table 3.5 and the length distribution profiles of the contigs is described in Figure 3.7. Overall, the redundancy removal step reduced the number of contigs by 73%, while the total amount of sequence was reduced by 63%. This discrepancy is not unexpected and indicates that many shorter contigs were merged with longer ones. This too is indicated by the reasonably large N50 value (7051 bp) of the merged African non-reference sequences, as none of the regional ancestral group non-reference sequence N50 values depicted in Figure 3.6c are higher than 7000 bp. This must, therefore, mean that the pan-African non-reference

**Table 3.4. Summary of the pan-African non-reference sequences obtained from merging the non-reference sequences of the seven regional ancestral groups.** Redundancy removal was performed at a 90% identity threshold using CD-HIT.

| | |
|---|---|
| **Before removing redundancy** | |
| Number of contigs | 65 791 |
| Length (bp) | 117 809 061 |
| **After removing redundancy** | |
| Number of contigs | 17 550 |
| Length (bp) | 43 372 280 |
| N50 | 7 051 |
| GC content (%) | 42.78 |

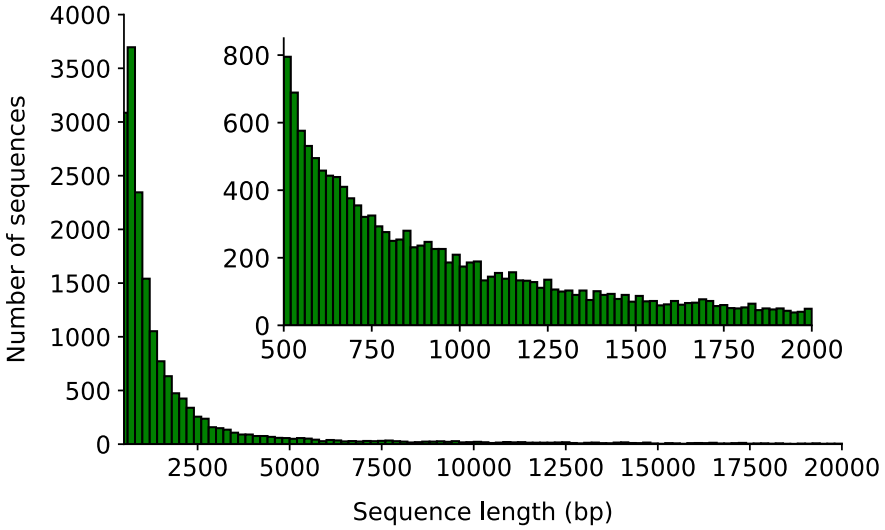


**Figure 3.7. The contig length distribution profile of the pan-African non-reference sequences obtained following redundancy removal at a 90% identity threshold.** Inset plot showing expanded length distributions between 500 and 2000 bp.

sequences have fewer shorter contigs overall than each of the ancestral groups. However, Figure 3.7 shows there that are still many contigs shorter than 1 kbp making up the African non-reference sequences and there are more of these short contigs than in the 1kGP non-reference sequences produced previously (Fig. 2.6b). Again, this is likely due to the differing datasets used, as 1kGP samples (which were used in Chapter 2) are known to have the highest N50 values of all the datasets used here. Finally, the GC content of the pan-African non-reference sequences (42.78%) is slightly higher than in GRCh38 (40.9%) but not so high that it would indicate biases in sequencing.

### 3.3.3. Analysing the pan-African non-reference sequences

#### *3.3.3.1. Exploring the non-reference sequence repeat elements*

The repeat elements within the regional ancestral groups and pan-African non-reference sequences were then investigated and compared to GRCh38 using RepeatMasker. As shown in Figure 3.8, the repeat element distribution profiles of the non-reference sequences differ extensively from that of GRCh38. Short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) are repeats of around 100 to 300 bp and 500 to 8000 bp respectively (Treangen & Salzberg, 2011) and are clearly under-represented in the non-reference sequences compared to the human reference genome. Conversely, satellites and simple repeats, which are between only 2 bp and 100 bp long, are

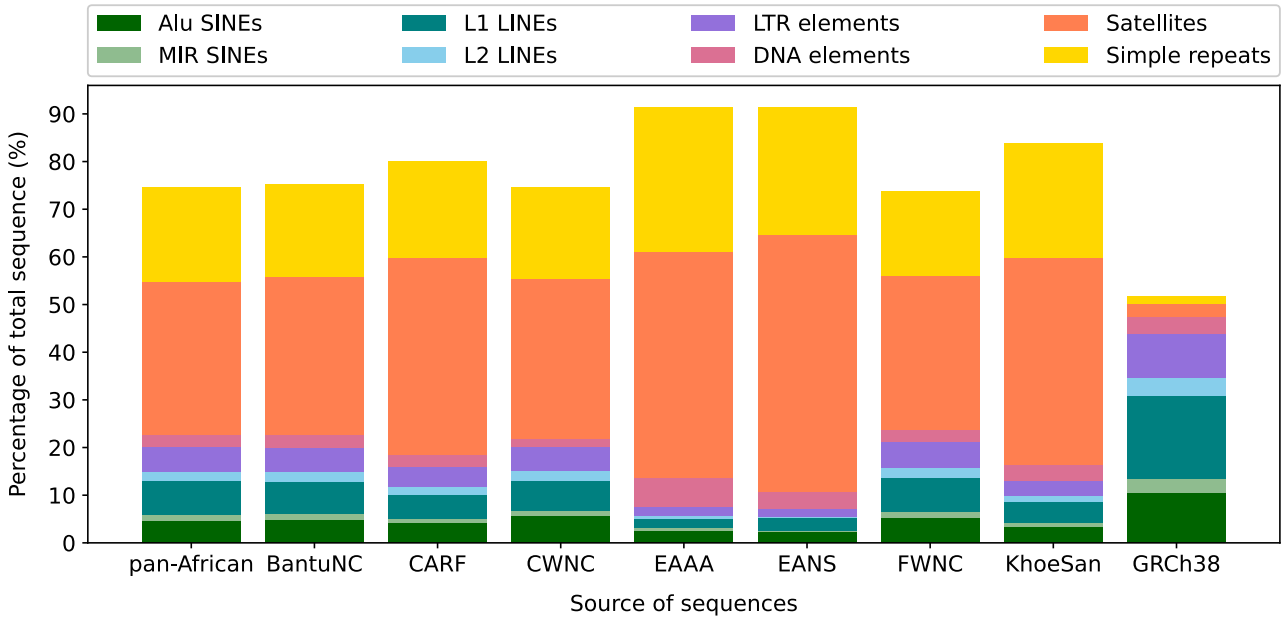


**Figure 3.8. Percentages of repeat elements identified by RepeatMasker in the pan-African non-reference sequences, the seven regional ancestral group non-reference sequences, and GRCh38.** The remaining percentage of sequence for each group were not classified as repeat elements by RepeatMasker. SINE: short interspersed nuclear element, LINE: long interspersed nuclear element, MIR: mammalian-wide interspersed repeat, LTR: long terminal repeat.

greatly overrepresented in the non-reference sequences, and together make up between approximately 50% and 75% of all non-reference sequences. As SINEs and LINEs are longer than satellites and simple repeats, they are more easily resolved by assembly algorithms and, therefore, can be aligned more easily to the reference genome. Shorter repeats, on the other hand, remain difficult to place within the reference sequence, which results in an increased likelihood of their being identified as non-reference. As expected, these results are comparable to those obtained by Duan *et al.* (2019) in the HUPAN publication and Sherman *et al.* (2019).

However, one observation that had not been previously described is the effect of including more samples on repeat element content of the non-reference sequences. As shown in Figure 3.8, the groups with the least amount of non-reference sequence – EAAA, EANS and KhoeSan – tended to have the highest percentage of repeat elements in total. This may indicate that repeat sequences are the most easily identifiable as non-reference, but as more samples are added to a dataset, the more other, non-repetitive non-reference sequences are identified. As the non-repetitive sequences are important for the discovery of novel protein-coding and regulatory regions, this observation further implies that the inclusion of more samples will ensure a more comprehensive and informative pan-genome.

#### *3.3.3.2. Identifying group-specific clustering*

We next wanted to examine whether submitting the seven regional ancestral groups to the pipeline separately resulted in any group-specific clustering of sequences during the final CD-HIT (Fu et al., 2012) redundancy removal step of the African non-reference sequences. As described in section 2.3.2.3, a principal component analysis (PCA) was performed on the sample membership of each cluster present in the African non-reference sequences following the final round of redundancy removal using CD-HIT. This resulted in distinct group-specific clustering of the samples (Fig. 3.9a), which is particularly notable in comparison to the PCA of the 1kGP non-reference sequences clusters shown in Figure 2.7a, where no population-specific clustering was evident. The only population not to exhibit localised clustering was the EAAA group, and this is likely due to the smaller number of samples in the group, making the identification of cluster patterns more difficult.

Additionally, despite the group-specific clustering of the BantuNC, CARF and FWNC groups, these groups also showed considerable cluster overlap, indicating that the samples in these three groups were more likely to yield the same or similar non-reference sequences that were then clustered together by CD-HIT during redundancy removal. This conclusion was supported by the results of the upset plot that also analysed the sample membership of the clusters produced by CD-HIT (Fig. 3.9b, full plot shown in Appendix E). As expected, the groups with the most non-reference sequence

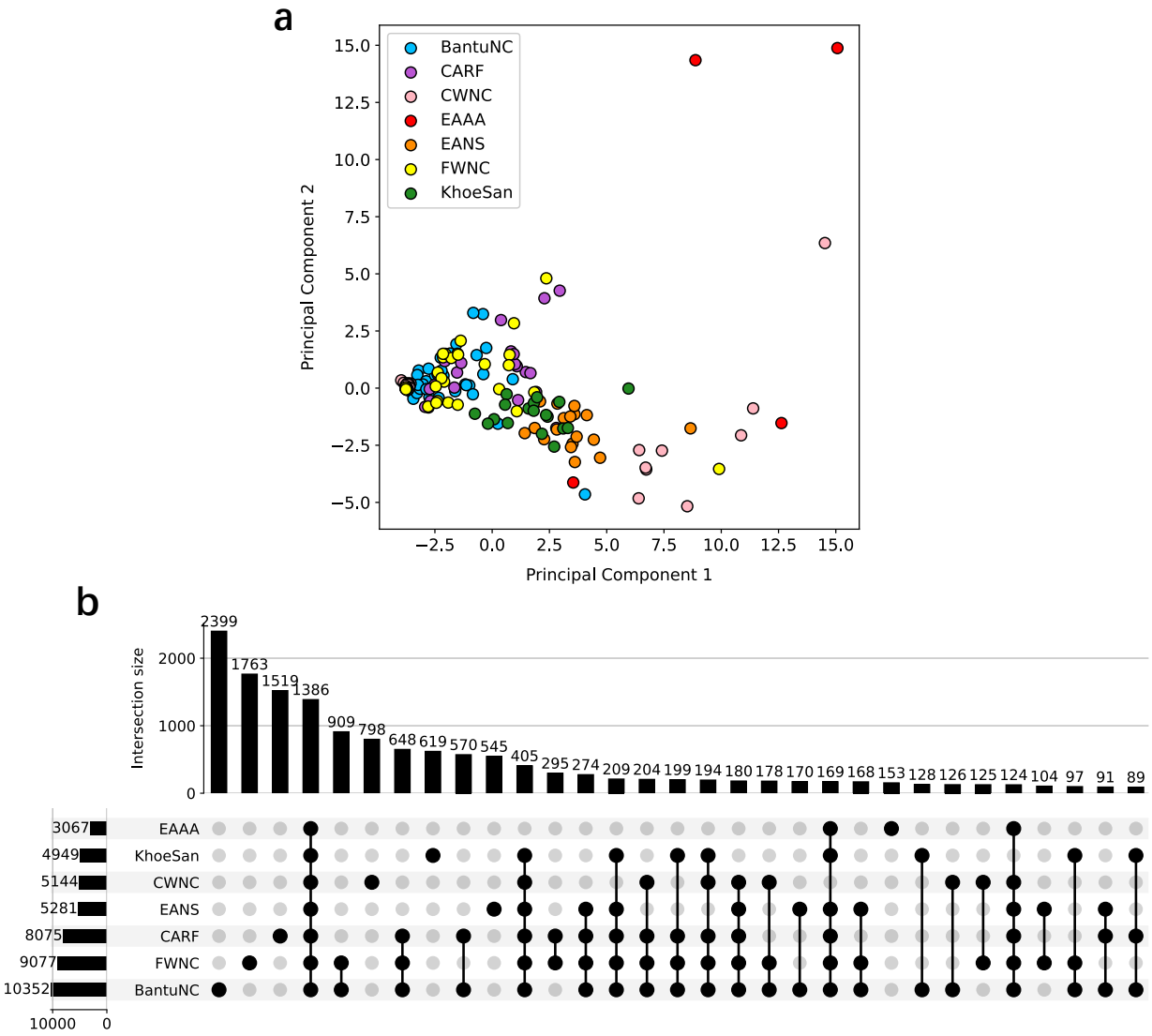


**Figure 3.9. Analysis of the sample membership of the 17 550 nucleotide clusters produced by CD-HIT following the merging and redundancy removal of the pan-African non-reference sequences.** **(a)** Principal component analysis of the sample membership of the clusters produced showing the first two principal components. **(b)** Upset plot of the sample membership of the clusters produced. Intersection size indicates how many times that unique combination of groups was found within the same cluster. Values next to the group names indicate how many clusters that group contributed towards. Combinations of clusters with an intersection size of less than 85 are not shown. The full upset plot is shown in Appendix E.

(BantuNC, CARF and FWNC) contributed to the most clusters, and formed many clusters that were comprised on non-reference sequences from samples in only those single populations. There were also 1386 clusters that consisted of sequences from samples from all seven regional ancestral groups, even though the EAAA group consisted of so little sequence. This indicated that these non-reference sequences are likely relevant to and present in many different African populations. Figure 3.9b also showed notable cluster overlap between the BantuNC, CARF and FWNC groups, reflecting what was shown in Figure 3.9a.

### *3.3.3.3. Analysing the novel predicted protein-coding genes*

The African non-reference sequences were submitted to MAKER (Holt & Yandell, 2011) for *ab initio* gene prediction and the results were filtered by the HUPAN pipeline to identify full-length novel protein-coding genes from the non-reference sequence. The MAKER output and subsequent filtering steps (described in section 2.1.2) for both the final African non-reference sequences and the HUPAN Han Chinese non-reference sequences obtained by Duan *et al.* (2019) are shown in Figure 3.10. In total, MAKER predicted 31 novel protein-coding genes from 43.37 Mbp of African non-reference sequence, while 167 novel genes were predicted from only 30.72 Mbp of novel Han Chinese sequence. Clearly, MAKER was able to predict many more novel genes from the HUPAN Han Chinese sequence,

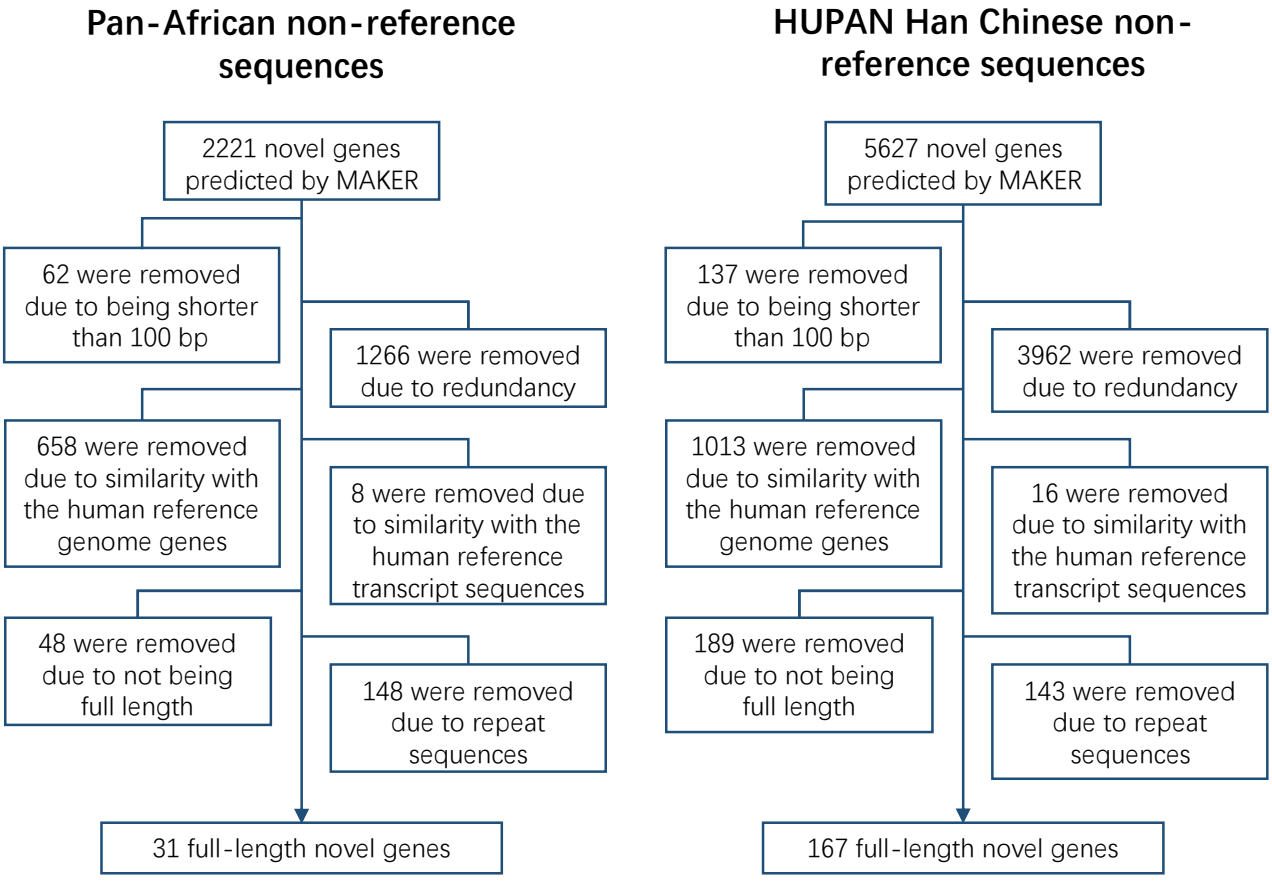


**Figure 3.10. Analysis of the filtering steps performed by the HUPAN pipeline on the novel genes predicted by MAKER (Holt & Yandell, 2011) for both the pan-African and HUPAN Han Chinese non-reference sequences.** First, genes shorter than 100 bp were removed. Redundancy in the predicted genes was then removed at an 80% sequence identity threshold using CD-HIT. Genes with sequences or transcript sequences similar to the reference genome were also removed at a threshold of 50% sequence identity. Any genes lacking either a start codon, a stop codon or both were removed. Lastly, any genes whose sequences were made up of more than 50% repeat sequence were removed, leaving the final filtered set of full-length novel genes predicted by MAKER.

despite there being almost 13 Mbp less sequence in the HUPAN dataset and the exact same settings and parameters being used for MAKER. A credible explanation for this difference is the HUPAN researchers' use of mRNA-seq gene expression data obtained from the same samples and tissues that were used to obtain the genome sequences used in the pipeline. The African dataset used in this research lacked such a sample-specific mRNA-seq dataset and instead used non-specific human mRNA-seq data downloaded from NCBI (section 3.2.3). RNA sequencing expression data is used by MAKER to assist in the identification of potential novel genes, and the use of sample-specific RNA data on nucleotide sequences from the same individuals may explain the hugely increased number of genes predicted in the HUPAN dataset compared to the pan-African one. This hypothesis is supported by the Han Chinese research performed by Li *et al.* (2021), as they too lacked sample-specific RNA-sequencing evidence and were only able to predict 53 full-length novel genes from 276 Mbp of non-reference sequence in MAKER. This shows the importance of using population- or sample-specific nucleotide and protein evidence for the prediction of novel genes using *ab initio* methods, as these methods may be less able to identify African novel genes given the smaller number of African-specific datasets available through NCBI.

Analysis of the coding DNA sequences (CDS) of the novel predicted genes was briefly performed, as MAKER predicts potential exon and intron splice sites of any novel genes identified. Of the 31 novel predicted genes, 20 had two CDS and eight had three CDS. The remaining three novel genes each had one CDS. As the CDS regions are the only parts of the gene translated into protein, these regions are highly important for inferring gene function. Therefore, identifying and characterising the CDS regions of each gene in future is essential should the function of the 31 novel genes be explored in further research.

The sample membership of sequences from which the novel genes were predicted was also analysed in order to identify which ancestral groups each of the genes were present in (Fig. 3.11). This was done by analysing the clustering information produced by CD-HIT in the redundancy removal step. Of the 31 genes, six (19.35%) were predicted from sequences that were present in all of the regional ancestral groups. This is especially notable as only 7.9% of the African non-reference sequences were found in all seven ancestral groups, and indicated that the sequences present in all groups were more likely to carry novel genes. Another two novel genes were predicted from sequences present in all groups except the EAAA group, and it is possible that these genes may have been detected in the EAAA group had more samples been included in that group.

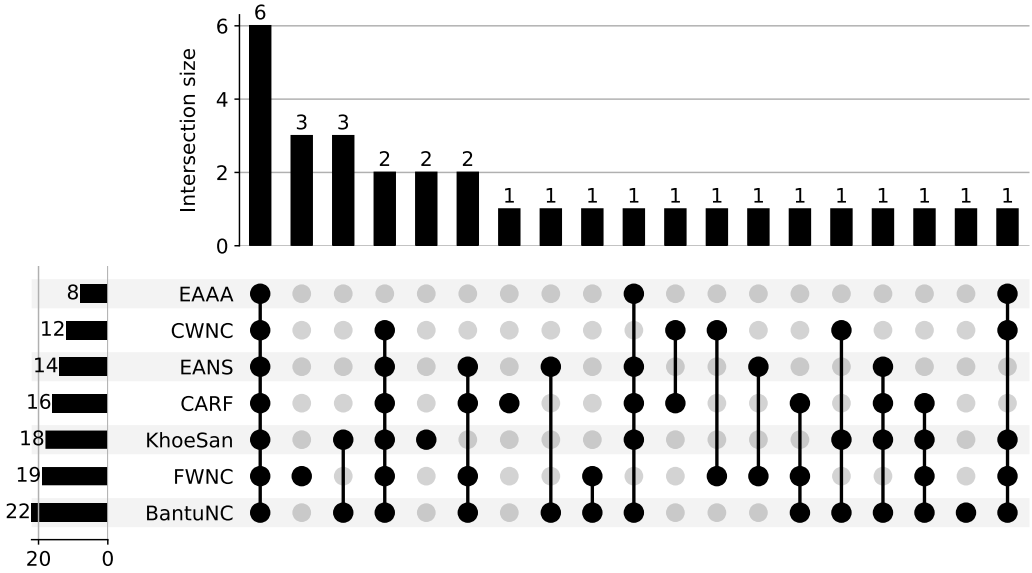


**Figure 3.11. Upset plot showing the sample membership of the non-reference sequences from which the 31 full-length novel genes were predicted by MAKER.** Intersection size indicates how many times that unique combination of groups was found within a cluster that produced a full-length novel gene. All intersection size values add up to 31. Values next to the group names indicate how many clusters that group contributed towards.

### 3.3.4. Validation of the pan-African non-reference sequences using alignments

Before the creation of the final African pan-genome, the pan-African non-reference sequences were aligned to other datasets and genomes to identify whether they could be partially validated. The non-reference sequences were first aligned reciprocally to both the HUPAN Han Chinese non-reference sequences assembled by Duan *et al.* (2019) and the African-descent non-reference sequences assembled by Sherman *et al.* (2019), and the results were recorded for all alignments with a sequence identity of 90% or higher (Fig. 3.12).

For both reciprocal alignments, the pan-African and African-descent non-reference sequences had more base pairs aligned than between the pan-African and Han Chinese non-reference sequences. Of note, at a threshold of 90% identity and 100% coverage, 2.75 Mbp of the pan-African non-reference sequences aligned to the African-descent non-reference sequences, while only 0.74 Mbp aligned to the Han Chinese non-reference sequences, implying the presence of African-specific sequence (Fig. 3.12a). For the reciprocal alignment at the same threshold, however, both the African-descent and Han Chinese non-reference sequences aligned to the pan-African non-reference sequences with similar amounts of sequence at 4.23 Mbp and 4.86 Mbp respectively (Fig. 3.12b). Although this may seem contradictory, it can potentially be explained by the composition of the sets of non-reference sequences. The African-descent non-reference sequences have a minimum length of 1 kbp, while the

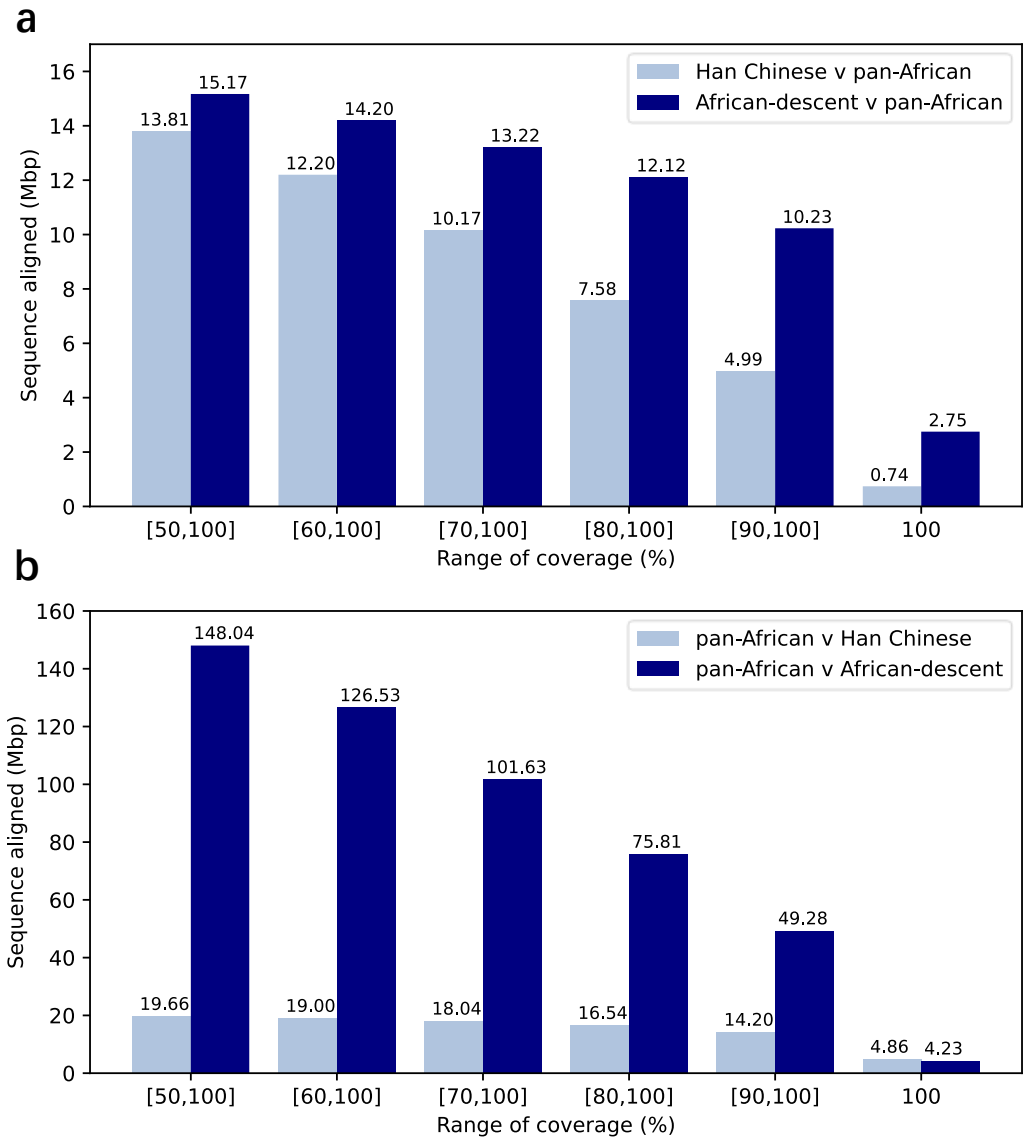


**Figure 3.12. Reciprocal alignments of the pan-African non-reference sequences to the HUPAN Han Chinese non-reference sequences assembled by Duan *et al.* (2019) and the African-descent non-reference sequences assembled by Sherman *et al.* (2019) over an increasing range of coverage.** All alignments shown have a sequence identity of ⩾ 90%. The figure legends denote target v query sequence. The total amount of sequence in Mbp aligned for each range of coverage is shown above each bar. **(a)** The total length of the pan-African non-reference sequences that aligned to the Han Chinese and African-descent non-reference sequences. **(b)** The total length of the Han Chinese and African-descent non-reference sequences that aligned to the pan-African non-reference sequences.

Han Chinese sequences have a minimum length of only 500 bp. Since the contig length distribution of the pan-African non-reference sequences is also skewed towards sequences between 500 bp and 1 kbp long, it is easier for the Han Chinese sequences to reach the coverage threshold of 100% than the African-descent sequences. This is reflected in the sudden increase of African-descent non-reference sequence aligned as the coverage threshold is decreased, while the amount of Han Chinese

sequence aligned increases only slightly in comparison. In addition, the amount of sequence aligned at thresholds of 100% identity and 100% coverage for each alignment was calculated (not shown); the pan-African sequences had 0.17 Mbp and 0.93 Mbp aligned to the Han Chinese and African-descent sequences respectively, confirming that there is almost 1 Mbp of sequence that is identical between the pan-African and African-descent non-reference sequences despite the differences in contig length distributions.

Interestingly, the total amount of African-descent sequence that aligned to the pan-African non-reference sequence at a threshold of 90% identity and 80% coverage was over 75 Mbp. As the total amount of pan-African non-reference sequence is only just over 43 Mbp, this indicated that multiple different African-descent non-reference contigs had aligned to the same contigs within the pan-African dataset. This in turn suggests there may be a larger amount of redundancy within the African-descent dataset, and is a potential explanation for the large amount of non-reference sequence obtained by that research group (Sherman et al., 2019). Additionally, although the reciprocal pan-African and Han Chinese non-reference sequence alignments resulted in less total sequence aligned, there were still notable levels of sequence alignment at thresholds of 90% identity and 80% coverage, indicating a relatively strong homology (7.58 Mbp and 16.54 Mbp respectively). This may suggest that, as this aligned sequence is common to both the Han Chinese and African populations, it may be non-reference sequence that is also relevant to human populations outside the Asian and African continents. The remaining non-reference sequence that could not be aligned may therefore be more likely African or African-descent specific.

In a further attempt to validate the pan-African non-reference sequences, both the pan-African and HUPAN Han Chinese non-reference sequences were aligned to the whole genome sequences of a Fon sample from Benin from the H3Africa dataset and a Han Chinese sample from China from the HGDP dataset. The results for alignments with a sequence identity of 90% or higher are shown in Figure 3.13. In the alignment to the Han sample, as expected, the Han Chinese non-reference sequences aligned better than the pan-African sequences at all coverage thresholds above 50% (Fig. 3.13a), suggesting more Han-specific sequence in the HUPAN Han Chinese non-reference sequences. It was also expected that the pan-African non-reference sequences would show better alignment to the Fon sample than the Han Chinese non-reference sequences, and this was the case for coverage values at 80% or lower; however, for coverage above 80%, the Han Chinese non-reference sequence aligned better than the pan-African sequences (Fig. 3.13b). Although this was unexpected, it may be because at least 10 Mbp of the HUPAN Han Chinese sequences are well represented in the reference genome, as discussed in section 2.4. Some of these sequences are, therefore, likely also present in both the Han

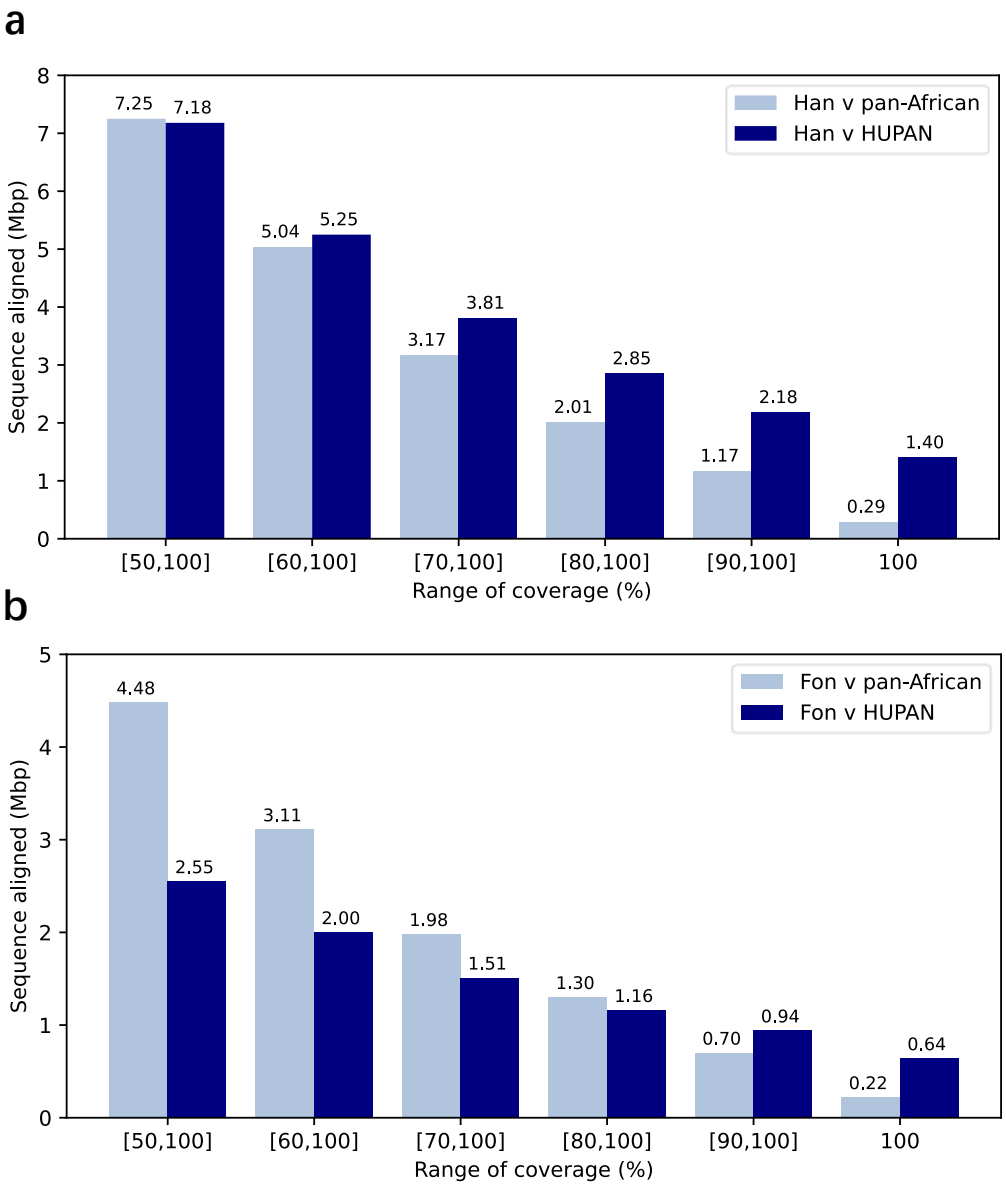


**Figure 3.13. Alignments of the pan-African non-reference sequences and the HUPAN Han Chinese non-reference sequences assembled by Duan *et al.* (2019) to the whole genome sequences of a Han sample from China and a Fon sample from Benin over an increasing range of coverage.** All alignments shown have a sequence identity of ≥ 90%. The figure legends denote target v query sequence. The total amount of sequence in Mbp aligned for each range of coverage is shown above each bar. **(a)** The total length of the pan-African and Han Chinese non-reference sequences that aligned to the Han sample whole genome sequence. **(b)** The total length of the pan-African and Han Chinese non-reference sequences that aligned to the Fon sample whole genome sequence.

and Fon samples used for alignment, resulting in the HUPAN Han Chinese non-reference sequences aligning better to both samples than the pan-African non-reference sequences. Despite the presence of reference sequences in the HUPAN dataset, however, the pan-African sequences still displayed better alignment to the Fon sample at most coverage values compared to the Han Chinese non-reference sequences. This again suggests that the pan-African dataset contains African-specific

sequence and implies this sequence could potentially be used in future for real-world research applications with respect to African samples.

In terms of the total amount of sequence aligned, both non-reference sequence datasets had notably more sequence aligned to the Han Chinese sample than to the Fon sample, despite the presence of potential Han Chinese- and pan-African-specific sequence. This is likely due to the fact that the samples were obtained from two different datasets; the Han sample is from the HGDP dataset (35X) and the Fon sample is from the H3Africa dataset (30X). Because of the this, the final assembled samples likely differed slightly in total length and composition as observed previously, resulting in differing amounts of alignment. However, the overall alignment trends discussed previously remain valid, and this observation merely suggests that samples from different populations but the same dataset should potentially be utilised in future to further confirm the observed alignment results.

For the final alignment, the pan-African non-reference sequences were aligned to both GRCh38 and the complete telomere-to-telomere human genome (T2T-CHM13) assembled by Nurk *et al.* (2021). As shown in Figure 3.14, 24.6 Mbp (totalling 56.72%) of pan-African non-reference sequence aligned to T2T-CHM13 at thresholds of 90% identity and 100% coverage, while only 0.63 Mbp aligned to GRCh38 at the same thresholds. Similarly, at thresholds of 100% identity and 100% coverage (not shown), the pan-African sequences had 0.79 Mbp aligned to T2T-CHM13, but 0 Mbp aligned to GRCh38. The lack of alignment of the pan-African non-reference sequences to GRCh38 was entirely expected and acted

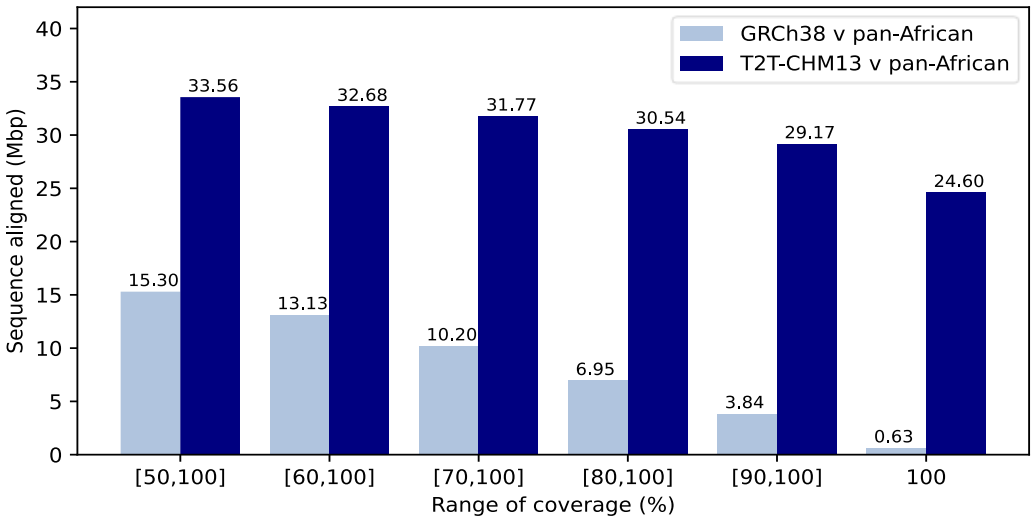


**Figure 3.14. Total aligned length of the pan-African non-reference sequences to the human reference genome (GRCh38) and the complete telomere-to-telomere human genome (T2T-CHM13) over an increasing range of coverage.** All alignments shown have a sequence identity of ⩾ 90%. The figure legend denoted target v query sequence. The total amount of sequence in Mbp aligned for each range of coverage is shown above each bar.

as validation that the pan-African sequences obtained can indeed be classified as non-reference. The alignment of notable amounts of the pan-African non-reference sequence to T2T-CHM13 but not to GRCh38 further supports this conclusion.

### 3.3.5. Analysing the African pan-genome

Following the validation of the pan-African non-reference sequences, the sequences were merged with the human reference genome to create the sequence file of the African pan-genome. The 31 novel predicted gene annotations were also merged with the protein-coding primary sequence annotations of the human reference genome. In total, excluding the 64 genes found on the Y chromosome (see section 3.2.3), there were 19 909 protein-coding genes in the pan-genome annotation file. This was made up of the 19 878 genes located in the human reference genome primary sequences and the 31 genes predicted from the pan-African non-reference sequence. The pan-genome annotation file was then used to analyse the gene presence-absence variation (PAV) profiles of the 168 African samples and therefore identify the core and distributed or accessory genes.

As discussed in section 3.1.3.2, the HUPAN authors chose a CDS coverage of 95% to identify 606 distributed genes within the Han Chinese pan-genome (Duan et al., 2019). Before choosing a threshold by which to determine the African distributed genome, however, we first examined the effects of changing the coverage threshold from 85% to 95%, and the effect of using either CDS coverage, gene coverage or both to identify whether genes were present within a sample or not. The results of this analysis are displayed in Table 3.6.

**Table 3.6. The effect of changing the coverage threshold of either coding sequence regions, whole genes, or both on the number of core and distributed genes within the African pan-genome.** CDS: coding DNA sequence.

| Threshold | Number of core genes | Number of distributed genes |
|---|---|---|
| CDS coverage ≥ 85% | 18 592 | 1 317 |
| Gene coverage ≥ 85% | 19 015 | 894 |
| Both CDS and gene coverage ≥ 85% | 18 485 | 1 424 |
| CDS coverage ≥ 90% | 17 741 | 2 168 |
| Gene coverage ≥ 90% | 18 617 | 1 292 |
| Both CDS and gene coverage ≥ 90% | 17 554 | 2 355 |
| CDS coverage ≥ 95% | 15 353 | 4 556 |
| Gene coverage ≥ 95% | 17 291 | 2 618 |
| Both CDS and gene coverage ≥ 95% | 14 899 | 5 010 |

For all three of CDS coverage, gene coverage and both, the number of core genes increased greatly as the coverage threshold was decreased from 95% to 85%, which was expected. In addition, using gene coverage to determine presence or absence resulted in the highest number of core genes in every case, while using both CDS and gene coverage thresholds simultaneously appeared to be the strictest requirement and therefore resulted in the fewest core genes.

The HUPAN authors did not explain their choice of CDS coverage to determine gene presence or absence. However, as the coverage of a whole gene sequence would indicate more complete coverage of a region than just coverage of the CDS regions, we decided to use this coverage definition for further analysis. Additionally, since using the gene coverage requirement resulted in the highest number of core genes, this coverage definition would also likely avoid potential false positive results for distributed genes. Furthermore, since an identity threshold of 90% had been used consistently in this research to confirm sequence homology and redundancy removal, a coverage threshold of 90% was also chosen here to determine presence or absence of each gene. Using this definition for presence and absence, there were a total of 18 617 core genes and 1 292 distributed genes, of which 29 were non-reference sequence novel predicted genes identified in this work. The remaining two novel predicted genes were present in all samples at both 90% and 95% gene coverage thresholds and were therefore classified as core genes. PAV profiles for the distributed genes at each of the thresholds examined in Table 3.6 are shown in Appendix F.

At a 90% gene coverage threshold, the number of genes present per sample varied from 19 472 to 19 860 out of the total 19 909 genes in the pan-genome (Fig. 3.15a), with the mean number of genes being 19 765. This distribution can be considered a reflection of the completeness of the genetic sequence for each assembled sample. Notably, there was no single gene that was absent in all 168 samples. However, a highly unexpected trend was observed when the gene PAV plot for the distributed genes was created, as shown in Figure 3.15b.

The 1292 distributed genes identified using the 90% gene coverage threshold were plotted based on presence or absence and the samples were sorted into their regional ancestral groups. The HUPAN authors were able to show that gene presence-absence of the distributed genes in their research was uniform, with most samples showing a similar gradient of genes from present to absent (Fig. 4d in Duan *et al.* (2019)). In the African pan-genome presented here, certain ancestral groups, such as BantuNC, CARF, EAAA and KhoeSan, similarly showed highly uniform PAV profiles within their groups, as shown in Figure 3.15b. However, the CWNC and FWNC groups lacked this uniformity and showed highly distinct sets of samples with gene PAV profiles that diverged greatly from the other samples in the same groups. Further, one set of samples in the CWNC group had a gene PAV profile almost

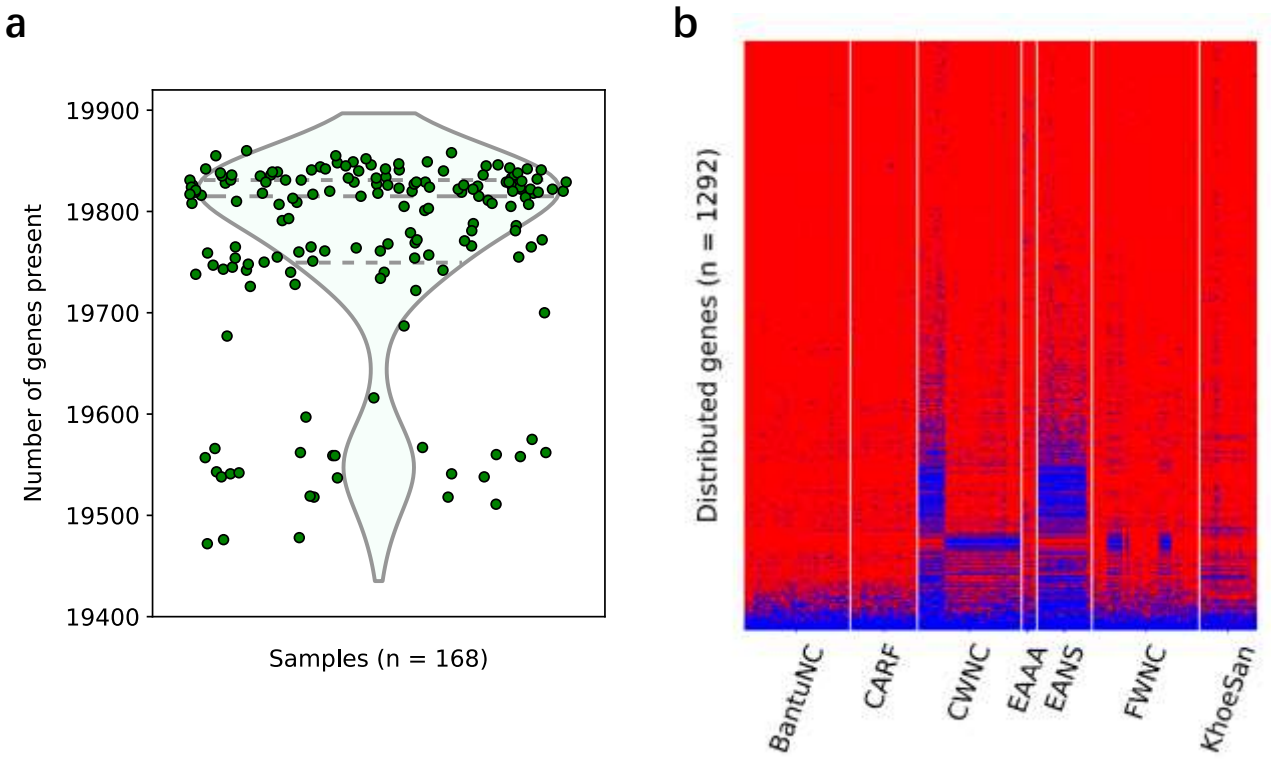


**Figure 3.15. African pan-genome gene presence-absence variation analysis of the 168 samples that were used to assemble the African pan-genome.** A gene was defined as present if 90% or more of the entire gene sequence was covered by raw sequenced reads from the sample being analysed. **(a)** Violin plot showing the distribution profile of total gene number present in the 168 samples. Each green point represents a sample. The middle dashed grey line within the violin plot shows the median value and the uppermost and lowermost dashed grey lines show the upper quartile and lower quartile, respectively. **(b)** A gene presence-absence variation profile showing the presence or absence of the 1292 distributed genes in the African pan-genome. Red signifies presence and blue signifies absence. The distributed genes have been ordered from those present in the most samples to those present in the least samples. Regional ancestral groups are separated by the white lines. The samples are ordered alphabetically by sample ID in each group.

identical to that of the EANS group, and a set in the FWNC group had a PAV profile identical to a majority of the CWNC group. This was evidently due to the alphabetical ordering of samples within each regional group in the PAV profile. Upon further analysis, it was shown that the sets of samples within different groups that showed similar gene PAV profiles were obtained from the same dataset, and thus had the same naming conventions. For example, the left-most set of samples in the CWNC group were all from the TrypanoGEN datset; the gene PAV profile of this set closely matched the PAV profile of the EANS group, in which 16 out of the 18 samples were also from the TrypanoGEN dataset. Of note, the sets of samples within this dataset very clearly displayed the most blue, and therefore had the fewest genes present in the gene PAV profiles. This indicated that the reads within these samples were the least able to reach the gene coverage threshold, and were therefore potentially not

suitable for inclusion in the African pan-genome. In a similar manner, the larger set of samples within the CWNC group were from the H3Africa dataset, as were the ten samples in the FWNC group that displayed the same gene PAV profiles. Due to the greater presence of blue within these PAV profiles, the samples from the H3Africa dataset also appeared to have fewer genes present overall than samples obtained from other datasets. The remaining samples in the FWNC group showed gene PAV patterns similar to that of the samples within the BantuNC and CARF groups, and all of these samples were obtained from the 1kG, HGDP and SGDP datasets.

These patterns can be more clearly seen in Figure 3.16, which shows the same PAV plot as Figure 3.15b with samples grouped by dataset instead of regional ancestral group. This figure shows almost complete uniformity within the datasets, which meant that the gene PAV profiles largely reflected the

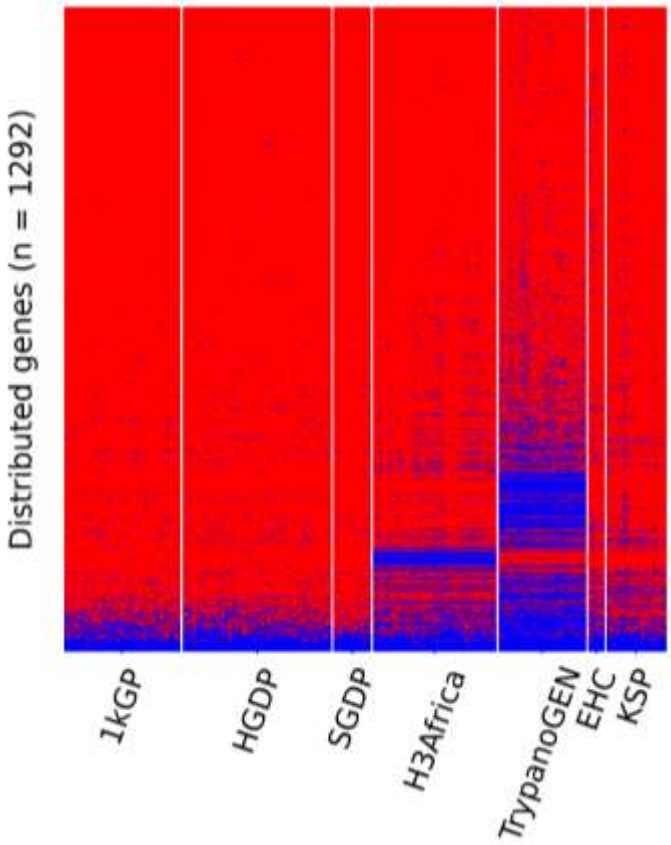


**Figure 3.16. African pan-genome gene presence-absence variation plot showing the 1292 distributed genes within the African pan-genome with samples sorted by dataset of origin.** A gene was defined as present if 90% or more of the entire gene sequence was covered by raw sequenced reads from the sample being analysed. Red signifies presence and blue signifies absence. The distributed genes are ordered from those present in the most samples to those present in the least samples. Datasets of origin are separated by white lines. The samples are ordered alphabetically by sample ID in each dataset. 1kGP: 1000 Genomes Project dataset, HGDP: Human Genome Diversity Project dataset, SGDP: Simons Genome Diversity Project, H3Africa: H3Africa Consortium dataset, TrypanoGEN: Trypanosomiasis Genomics Network dataset, EHC: Ethiopian high-coverage dataset, KSP: Khoe-San Project dataset.

sequencing coverage – or the completeness of each sample – and, therefore, their datasets of origin. This had the effect of obscuring any underlying regional ancestral group-specific patterns of variation, so identifying group-specific genetic differences was challenging, as the dataset-wide patterns could not easily be separated from true genetic differences. Further, these patterns are maintained for all thresholds of presence and absence, as seen in the figure in Appendix F. This aspect appears not to have been considered in any previously published pan-genomic research, since, to our knowledge, the work presented here is the first to include samples from multiple different datasets. Consequently, this observation is potentially a confounding factor with far-reaching implications in pan-genomic analysis that needs to be discussed and considered in future research.

## 3.4. Discussion

This research is the first to successfully assemble and analyse many whole genome sequences to create a pan-genome exclusively focused on African populations. Furthermore, once the necessary adaptations and system-specific changes had been made, the creation of an African pan-genome using the HUPAN (Duan et al., 2019) pipeline was largely straightforward and achievable. Over 43 Mb of novel non-reference sequence was identified from the African samples utilised, with each regional ancestral group yielding notable amounts of sequence despite the varying numbers and quality of samples making up each group. Further, 31 novel genes were predicted from the identified non-reference sequence, with at least six being present in all seven regional ancestral groups. These ubiquitous genes are of great interest considering the relatively small amount of non-reference sequence present in all seven groups, and certainly warrant further investigation into their potential functions and genetic locations in future research.

The modifications made to the HUPAN pipeline for this research proved to be both useful and logical. The use of the Nextflow SGA assembly pipeline greatly reduced the time taken to assemble single samples and allowed the resumption of stalled assemblies with cached data, meaning that computational resources were used efficiently. Since Nextflow pipelines are portable and easily adaptable, this modification ensures that the HUPAN pipeline can be utilised in other high-performance computing cluster environments. This means that the pipeline is more accessible to other researchers should they wish to utilise it.

The separation of regional ancestral groups within the resulting pan-genome was also a logical and efficient use of the HUPAN pipeline's infrastructure. The use of this method means this that research is among the first pan-genome studies to include and separately examine multiple genetically distinct and regionally separated populations. Furthermore, this method utilises the pre-existing HUPAN

pipeline infrastructure without requiring extensive source code alterations and is therefore easily applicable for other research groups that may also wish to perform multi-population analyses. Most importantly, this technique provides a simple way forward for creating population-specific pan-genomes for many populations, which can then be combined and merged to create region- and even continent-specific pan-genomes with very little additional computational expense. The analysis of the sample membership within the nucleotide sequence clusters produced via redundancy removal also suggests that this technique can correctly ensure population- or group-specific genetic variation is retained by submitting the groups to the pipeline separately. To support this, the BantuNC, CARF and FWNC samples showed distinct overlapping in the sequence clusters produced by CD-HIT, which is expected given the ethnolinguistic relationships between these groups. As population-specific genetic variation is expected, these findings confirm that this adaptation to the pipeline is a worthwhile pursuit in the discovery of these differences.

The extensive alignment of the African non-reference sequence to T2T-CHM13 but not to the reference genome also suggests that these sequences have been correctly identified as non-reference. Further, the 10 Mbp sequence that failed to align well to either GRCh38 or T2T-CHM13 is potentially genetic sequence that is completely novel and perhaps specific to African populations, and is therefore not represented within either human genome. Likewise, the more extensive alignment of the pan-African non-reference sequences to the African-descent non-reference sequences (Sherman et al., 2019) and the Fon sample from Benin compared to the Han Chinese non-reference sequences (Duan et al., 2019) and sample suggests that a notable amount of sequence is African-specific. Much of the identified novel sequence that aligned to both sets of non-reference sequences could potentially be some of the same sequences that aligned to T2T-CHM13, and analysis of these sequences could further validate them as non-reference sequence that may be present in many human populations on many different continents. These observations echo what previous pan-genomic research has identified and highlights the utility of population-specific reference genomes for human genetic research.

The African-descent alignments also revealed an effect of redundancy removal. Overall, 75.81 Mbp of the African-descent non-reference sequence aligned to the pan-African non-reference sequences at thresholds of 80% coverage and 90% identity. This is almost double the amount of pan-African sequence identified, which suggests that multiple African-descent contigs aligned to the same pan-African contigs. This implies the African-descent non-reference dataset contains more redundancy than the non-reference sequences identified in this work. The high redundancy cut-off (≥ 95%) of Sherman's group likely explains this, particularly since a significant amount of the African-descent

sequence was made up of repetitive elements (Sherman et al., 2019). In contrast, the pipeline used in the research presented here had three redundancy removal steps with a lower identity requirement of 90% at each step. The pan-African non-reference sequence identified here is also made up of majority repeat sequence, so it is possible that, in some cases, multiple repeat elements with similar identities were merged during the redundancy removal steps, despite being from different sequences and regions altogether. This could result in the notably smaller amount of non-reference sequence identified compared to the work by Li *et al.* (2021) and Sherman *et al.* (2019). Consequently, future pan-genome research needs to consider whether less redundancy within non-reference sequence datasets is preferable, or whether more redundancy can be tolerated in order to obtain a larger total amount of non-reference sequence. Further, because this research was the first to include multiple populations, the question of redundancy removal between populations has likely not previously been considered. Having shown that the total amount of non-reference sequence was decreased by 63% when the seven regional ancestral groups were merged and redundancy was removed at the 90% identity threshold, it is worth considering the use of more stringent redundancy removal thresholds for the merging of non-reference sequences from multiple different groups. By using lower thresholds – such as the 90% threshold used here – for single groups or populations, but higher thresholds when merging multiple groups, redundant sequences within single groups will be removed, but the important genetic variation between groups will be retained, enabling easier identification of those genetic differences and ensuring the amount of non-reference sequence identified is representative of the groups or populations included. This may allow a more nuanced and informative analysis of the genetic variation present in closely related sub-populations. Specifically for the research presented here, the use of a redundancy threshold of 95% or higher in the merging of group-specific non-reference sequences to obtain the pan-African sequences should be considered and analysed in future.

The use of multiple different datasets and African populations to create a pan-genome was also a novel aspect of this study compared to previous pan-genomic research. However, despite the successful creation of a multi-population African pan-genome, the inclusion of multiple populations highlighted certain complexities of pan-genome research, some of which previously may not have been described in the literature.

Firstly, as expected, the number of samples included per population or group had a profound effect on the amount of non-reference sequence identified. Overall, the research performed here was greatly hampered by both the lack of African datasets and by the low coverage of many of the available datasets. As a result, far fewer samples were used per regional ancestral group than was optimal for

the creation of a truly comprehensive African pan-genome. Furthermore, the number of samples present in each regional ancestral group was unavoidably different due to the uneven distribution of samples from different African populations; thus, the sequence contributions of each group to the final pan-genome were unequal. As a result, before a more comprehensive African pan-genome – or even separate African population-specific pan-genomes – can be assembled, the disparities in data availability between African individuals and those of European descent need to be addressed.

However, even within the high quality African whole genome datasets that do exist, this research was able to highlight several variables that affected the final African pan-genome. This was first evident when comparing the average assembly metrics, such as total assembled length and N50, of each of the regional ancestral groups. The BantuNC and FWNC groups were the only groups that contained samples from the 1kGP high-coverage dataset, and these groups had the largest lengths and highest N50 values. The CARF (HGDP and SGDP datasets), CWNC (H3Africa and TrypanoGEN datasets) and KhoeSan (Khoe-San Project dataset) groups consisted of samples with only slightly lower total assembled lengths, but the N50 values were noticeably smaller than the BantuNC and FWNC groups. Fundamental differences between the groups were also noticeable in later stages of the pipeline; the BantuNC and FWNC groups yielded more non-reference sequence than the CARF group, and a substantial amount more than both the CWNC and KhoeSan groups. Further, the N50 values of the non-reference sequences reflected those of the assembled samples from which they were obtained. Clearly, given the assembly metrics and the amount of non-reference sequence ultimately identified, the whole genome sequences from the 1kGP dataset can be considered more comprehensive compared to samples from the H3Africa and Khoe-San datasets, for example. This is of particular interest because these three datasets were all sequenced using Illumina short-read technologies to a depth of 30X or higher. The difference between the H3Africa dataset and the 1kGP dataset is especially remarkable, as both datasets used Illumina sequencing platforms – Illumina NovaSeq 6000 and Illumina X-Ten respectively, which are desribed by the Illumina Platform Comparison Tool (https://www.illumina.com/systems/sequencing-platforms/comparison-tool.html#/research-use-only/population-scale-wgs) as being comparable – and both purportedly have a targeted sequencing depth of 30X.

Further anomalies were identified following the assembly of the pan-genome and the creation of the gene presence-absence variation (PAV) profiles. This analysis firstly confirmed that the TrypanoGEN samples, which were sequenced to a depth of only 10X, had the highest number of absent genes in the distributed gene PAV profile. This indicates that the reads constituting those samples can be considered the least comprehensive, which is expected given the lower sequencing depth. This in turn

suggests that the 60% reference genome length threshold chosen to include the samples in the African pan-genome (section 3.3.1) was perhaps too lenient. This further implies that medium-coverage datasets may not be suitable for inclusion in pan-genome research utilising non-reference sequences, as the samples in those datasets may hinder the identification of sequences that are truly absent from the reference genome. However, the 1kGP, HGDP, SGDP, H3Africa and Khoe-San Project datasets were all sequenced to a depth of 30X or higher using Illumina technologies, yet there were still extensive differences between the gene PAV profiles of the samples in each dataset. Although less so than the TrypanoGEN dataset, the H3Africa dataset clearly also displayed greater gene absence than the 1kG, HGDP and SGDP datasets. The Khoe-San Project dataset also had noticeably more gene absence than the other three datasets, while the SGDP samples (43X) appeared to be the most complete. However, although the SGDP dataset was sequenced at the highest coverage (43X), the apparent low quality of some of the samples in this dataset is relatively surprising, since six samples from this dataset had to be removed due to extensive amounts of contamination, and eight were removed having failed the quality control threshold of 60% of the reference genome sequence length. Clearly, even within datasets, there can some degree of variation in the quality of whole genome sequences.

Overall, the differences in datasets that were initially considered of comparable quality suggest that there are factors that affect the quality of whole genome sequences other than the type and method of sequencing. These may include factors such as the protocols used for sample extraction, and laboratory-specific differences in equipment and workflows. Crucially, though, the importance of this observation to human pan-genomic research should not be understated.

This research is the first of its kind to include multiple datasets in one analysis. For example, Duan *et al.* (2019), Li *et al.* (2021) and Sherman *et al.* (2019) all used samples from only a single dataset in their pan-genome analysis. Subsequently, both the Duan and Li groups performed additional analyses on a second dataset of Han Chinese samples, but there are no published reports where these samples have been directly compared to the samples in the original datasets. Thus, it appears that previous pan-genomic research has not investigated the degree to which the genetic variation observed is dataset-specific or truly population-specific. Given the clear differences in datasets described here, the failure to acknowledge and account for the effects of spurious inter-dataset variation in previous pan-genomic research indicates those analyses may be incomplete or ambiguous. This, therefore, calls into question the methods currently used for pan-genomic research, and further highlights that robust, standardised methods for human pan-genomic research must be converged on in order to ensure that comprehensive population-specific pan-genomes are produced in the future.

Ultimately, the dataset-specific variation within this research created complications for the identification of genetic variation specific to the populations analysed. Due to the scope of this research, these challenges could not be resolved, and thus group-specific analysis was not explored in depth. Further investigation into how to mitigate the obscuring effects of using multiple datasets within one pan-genome study will first need to be performed before population-specific differences can be truly investigated in research such as this. Moreover, this limitation will need to be addressed before pan-genomes can be utilised as ubiquitously as the human reference genome. Despite this inherent weakness of pan-genomic research, however, the work presented here was able to successfully create and analyse the first comprehensive and inclusive African pan-genome. Furthermore, certain complex aspects of pan-genome research that have not previously been discussed in depth in the literature were identified and explored, and these important observations may play a role in the advancement of pan-genome research in future.

# 4. Conclusions and future work

The research performed in this study showed that the HUPAN pipeline developed by Duan *et al.* (2019) is a valuable and effective software tool. It easily allows adaptations and enables user-specified settings to be implemented in order to successfully identify genetic sequences that are absent from the human reference genome. The pipeline thus has the potential to be highly valuable in future pan-genome research. However, the inclusion of multiple genetically distinct populations and groups in this study highlighted certain aspects of pan-genome analysis that need to be addressed in future.

The first aspect was the effect of redundancy removal, which clearly plays an extensive role in determining the amount of non-reference sequence identified. The stringency of redundancy removal also appears to affect the likelihood of identifying and analysing population- or group-specific differences. This means inter- and intra-population redundancy removal may require differing methods and thresholds to ensure that population-specific differences are not obscured. Standardised methods and thresholds for redundancy removal should, therefore, be investigated and established to ensure that all pan-genome sequences identified remain valuable and informative, but not overly redundant.

The second aspect was the effect of dataset-specific variation on whole genome sequences, which has, to our knowledge, not previously been discussed in pan-genomic literature. Ultimately, the strong correlation between dataset and pan-genome variation discovered here hindered the identification of group-specific genetic differences. These observations suggest that more research that includes samples from multiple different datasets and populations will need to be performed to evaluate the true extent of this phenomenon, and techniques to mitigate the obfuscating effects of it will need to be explored in future.

Despite these limitations, this study successfully identified weaknesses in the HUPAN pipeline, which was subsequently adapted and improved, thereby enabling the successful assembly of the first multi-population African pan-genome. In addition, unique sets of non-reference sequences specific to regional ancestral groups were identified and obtained, enabling future research into the genetic variation of each group separately. The alignment results of the pan-African non-reference sequences to various other datasets and sequences, summarised in Figure 4.1, also served to validate a majority of the sequences as non-reference, and a notable portion as either African-specific or potentially completely novel. Further research into the non-reference sequences identified has the potential to improve the comprehensiveness of this African pan-genome and the genetic variation of the populations represented in it. Focus should first be on the inclusion of more samples, particularly as

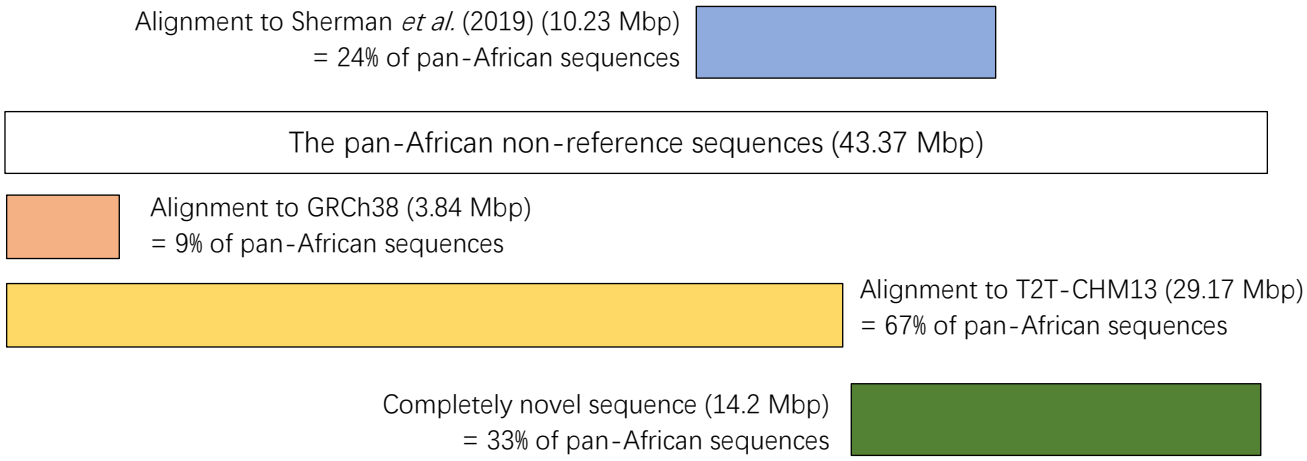


**Figure 4.1. Summary of the alignments of the pan-African non-reference sequences to GRCh38, T2T-CHM13 and the African-descent non-reference sequences produced by Sherman *et al.* (2019).** All alignments shown have both sequence identity and coverage of ⩾ 90%. The pan-African non-reference sequences are the query sequence in every case. The alignment results of the African-descent non-reference sequences is unknown, but it is likely that some of the sequence identified will align to T2T-CHM13, hence the placement.

more high quality African whole genome sequences become available. This will not only enable the inclusion of more subpopulations and the likely identification of additional novel sequence, but will also provide additional resources in the analysis and validation of the pan-genome. For example, if African-specific mRNA-sequencing datasets become available, these could in future be used to validate and potentially identify additional novel genes in the non-reference sequences obtained. The potential functions of the novel genes should also be investigated further, and this may help explain the apparent absence of these genes in the reference genome and other human populations. Variant calling of African genomes should also be performed using the African pan-genome to determine whether the inclusion of African-specific sequence positively impacts our ability to identify important African-specific genetic variation.

Overall, this study was successful in developing useful and logical adaptations to the HUPAN pipeline, which resulted in the creation of the first pan-genome representing the African continent. The research performed here will be valuable for future genomic research in Africa and provides an analysis of pan-genome assembly methods that is instructive for future human pan-genome research.

# References


1. Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*(5), 363–376.

2. Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

3. Antaki, D., Brandler, W. M., & Sebat, J. (2018). SV2: Accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics*, *34*(10), 1774–1777. https://doi.org/10.1093/bioinformatics/btx813

4. Ballouz, S., Dobin, A., & Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol*, *20*(1), 159. https://doi.org/10.1186/s13059-019-1774-4

5. Bergstrom, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanche, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484). https://doi.org/10.1126/science.aay5012

6. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

7. Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Consortium, T. H. G. S. V., Flicek, P., Germer, S., Brand, H., Hall, I. M., … Zody, M. C. (2021). *High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios* (p. 2021.02.06.430068). bioRxiv. https://doi.org/10.1101/2021.02.06.430068

8. Campbell, M. C., Hirbo, J. B., Townsend, J. P., & Tishkoff, S. A. (2014). The peopling of the African continent and the diaspora into the new world. *Current Opinion in Genetics & Development*, *29*, 120–132. https://doi.org/10.1016/j.gde.2014.09.003

9. Campbell, M. C., & Tishkoff, S. A. (2008). African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet*, *9*, 403–433. https://doi.org/10.1146/annurev.genom.9.081307.164258

10. Cho, Y. S., Kim, H., Kim, H. M., Jho, S., Jun, J., Lee, Y. J., Chae, K. S., Kim, C. G., Kim, S., Eriksson, A., Edwards, J. S., Lee, S., Kim, B. C., Manica, A., Oh, T. K., Church, G. M., & Bhak, J. (2016). An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat Commun*, *7*, 13637. https://doi.org/10.1038/ncomms13637

11. Choudhury, A., Aron, S., Botigue, L. R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y. J., Ghoorah, A. W., Dareng, E., Odia, T., Falola, O., Adebiyi, E., Hazelhurst, S., Mazandu, G., Nyangiri, O. A., Mbiyavanga, M., … Hanchard, N. A. (2020). High-depth African genomes inform human migration and health. *Nature*, *586*(7831), 741–748. https://doi.org/10.1038/s41586-020-2859-7

12. Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H. C., Agarwala, R., McLaren, W. M., Ritchie, G. R., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biol*, *9*(7), e1001091. https://doi.org/10.1371/journal.pbio.1001091

13. Church, D. M., Schneider, V. A., Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C. S., Kitts, P. A., Aken, B., Marth, G. T., Hoffman, M. M., Herrero, J., Mendoza, M. L., Durbin, R., & Flicek, P. (2015). Extending reference assembly models. *Genome Biol*, *16*, 13. https://doi.org/10.1186/s13059-015-0587-3

14. Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alfoldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

15. Dewey, F. E., Chen, R., Cordero, S. P., Ormond, K. E., Caleshu, C., Karczewski, K. J., Whirl-Carrillo, M., Wheeler, M. T., Dudley, J. T., Byrnes, J. K., Cornejo, O. E., Knowles, J. W., Woon, M., Sangkuhl, K., Gong, L., Thorn, C. F., Hebert, J. M., Capriotti, E., David, S. P., … Ashley, E. A. (2011). Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet*, *7*(9), e1002280. https://doi.org/10.1371/journal.pgen.1002280

16. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

17. Duan, Z., Qiao, Y., Lu, J., Lu, H., Zhang, W., Yan, F., Sun, C., Hu, Z., Zhang, Z., Li, G., Chen, H., Xiang, Z., Zhu, Z., Zhao, H., Yu, Y., & Wei, C. (2019). HUPAN: a pan-genome analysis pipeline for human genomes. *Genome Biol*, *20*(1), 149. https://doi.org/10.1186/s13059-019-1751-y

18. Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2022). *Ethnologue: Languages of the World* (Twenty-fifth). SIL International.

19. Ewels, P., Magnusson, M., Lundin, S., & Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

20. Faber-Hammond, J. J., & Brown, K. H. (2016). Anchored pseudo-de novo assembly of human genomes identifies extensive sequence variation from unmapped sequence reads. *Hum Genet*, *135*(7), 727–740. https://doi.org/10.1007/s00439-016-1667-5

21. Fan, S., Kelly, D. E., Beltrame, M. H., Hansen, M. E. B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., Omar, S. A., Meskel, D. W., Belay, G., Froment, A., Patterson, N., Reich, D., & Tishkoff, S. A. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol*, *20*(1), 82. https://doi.org/10.1186/s13059-019-1679-2

22. Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nat Rev Genet*, *7*(2), 85–97. https://doi.org/10.1038/nrg1767

23. Foster, I. (2011). Globus Online: Accelerating and Democratizing Science through Cloud-Based Services. *IEEE Internet Computing*, *15*(3), 70–73. https://doi.org/10.1109/MIC.2011.64

24. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

25. Frith, M. C. (2011). Gentle masking of low-complexity sequences improves homology search. *PLoS One*, *6*(12), e28819. https://doi.org/10.1371/journal.pone.0028819

26. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

27. Gonzaga-Jauregui, C., Lupski, J. R., & Gibbs, R. A. (2012). Human Genome Sequencing in Health and Disease. *Annual Review of Medicine*, *63*, 35–61. https://doi.org/10.1146/annurev-med-051010-162644

28. Granger, B., & Perez, F. (2021). *Jupyter: Thinking and Storytelling with Code and Data*. 10.

29. Groza, C., Kwan, T., Soranzo, N., Pastinen, T., & Bourque, G. (2020). Personalized and graph genomes reveal missing signal in epigenomic data. *Genome Biol*, *21*(1), 124. https://doi.org/10.1186/s13059-020-02038-8

30. Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, *15*(7), e1008302. https://doi.org/10.1371/journal.pgen.1008302

31. Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, *109*(2), 83–90. https://doi.org/10.1016/j.ygeno.2017.01.005

32. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., … Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, *517*(7534), 327–332. https://doi.org/10.1038/nature13997

33. Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

34. Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

35. Hoffman-Andrews, L. (2017). The known unknown: The challenges of genetic variants of uncertain significance in clinical practice. *J Law Biosci*, *4*(3), 648–657. https://doi.org/10.1093/jlb/lsx038

36. Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491.

37. Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

38. Hu, Z., Sun, C., Lu, K. C., Chu, X., Zhao, Y., Lu, J., Shi, J., & Wei, C. (2017). EUPAN enables pan-genome studies of a large number of eukaryotic genomes. *Bioinformatics*, *33*(15), 2408–2409. https://doi.org/10.1093/bioinformatics/btx170

39. Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., Peluso, P., Boitano, M., Chin, C. S., Korlach, J., Wilson, R. K., & Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res*, *27*(5), 677–685. https://doi.org/10.1101/gr.214007.116

40. Ilboudo, H., Noyes, H., Mulindwa, J., Kimuda, M. P., Koffi, M., Kaboré, J. W., Ahouty, B., Ngoyi, D. M., Fataki, O., Simo, G., Ofon, E., Enyaru, J., Chisi, J., Kamoto, K., Simuunza, M., Alibu, V. P., Lejon, V., Jamonneau, V., Macleod, A., … for the TrypanoGEN Research Group as members of The H3Africa Consortium. (2017). Introducing the TrypanoGEN biobank: A valuable resource for the elimination of human African trypanosomiasis. *PLOS Neglected Tropical Diseases*, *11*(6), e0005438. https://doi.org/10.1371/journal.pntd.0005438

41. Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., & Antonacci, F. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, *453*(7191), 56–64.

42. Kidd, J. M., Sampas, N., Antonacci, F., Graves, T., Fulton, R., Hayden, H. S., Alkan, C., Malig, M., Ventura, M., & Giannuzzi, G. (2010). Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods*, *7*(5), 365–371.

43. Kodama, Y., Shumway, M., Leinonen, R., & on behalf of the International Nucleotide Sequence Database Collaboration. (2012). The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, *40*(D1), D54–D56. https://doi.org/10.1093/nar/gkr854

44. Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., … Snyder, M. (2007). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science (New York, N.Y.)*, *318*(5849), 420–426. https://doi.org/10.1126/science.1149504

45. Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*(1), 59.

46. Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*, *20*(1), 117. https://doi.org/10.1186/s13059-019-1720-5

47. Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), R12.

48. Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLOS ONE*, *12*(5), e0177459. https://doi.org/10.1371/journal.pone.0177459

49. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., & FitzHugh, W. (2001). *Initial sequencing and analysis of the human genome*.

50. Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health Aff (Millwood)*, *37*(5), 780–785. https://doi.org/10.1377/hlthaff.2017.1595

51. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

52. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., … Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Research*, *39*(Database issue), D28–D31. https://doi.org/10.1093/nar/gkq967

53. Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

54. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

55. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

56. Li, Q., Tian, S., Yan, B., Liu, C. M., Lam, T.-W., Li, R., & Luo, R. (2021). Building a Chinese pan-genome of 486 individuals. *Communications Biology*, *4*(1), 1–14. https://doi.org/10.1038/s42003-021-02556-6

57. Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., Zhou, G., Zhu, X., Wu, H., Qin, J., Jin, X., Li, D., Cao, H., Hu, X., Blanche, H., … Wang, J. (2010). Building the sequence map of the human pan-genome. *Nat Biotechnol*, *28*(1), 57–63. https://doi.org/10.1038/nbt.1596

58. Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research*, *21*(6), 940–951. https://doi.org/10.1101/gr.117259.110

59. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

60. Magi, A., D'Aurizio, R., Palombo, F., Cifola, I., Tattini, L., Semeraro, R., Pippucci, T., Giusti, B., Romeo, G., Abbate, R., & Gensini, G. F. (2015). Characterization and identification of hidden rare variants in the human genome. *BMC Genomics*, *16*, 340. https://doi.org/10.1186/s12864-015-1481-9

61. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

62. Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., & Kohane, I. S. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine*, *375*(7), 655–665. https://doi.org/10.1056/NEJMsa1507092

63. Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, *12*(2), 213–218. https://doi.org/10.1038/nprot.2016.182

64. Maretty, L., Jensen, J. M., Petersen, B., Sibbesen, J. A., Liu, S., Villesen, P., Skov, L., Belling, K., Theil Have, C., Izarzugaza, J. M. G., Grosjean, M., Bork-Jensen, J., Grove, J., Als, T. D., Huang, S., Chang, Y., Xu, R., Ye, W., Rao, J., … Schierup, M. H. (2017). Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*, *548*(7665), 87–91. https://doi.org/10.1038/nature23264

65. Marroni, F., Pinosio, S., & Morgante, M. (2014). Structural variation and genome complexity: Is dispensable really dispensable? *Curr Opin Plant Biol*, *18*, 31–36. https://doi.org/10.1016/j.pbi.2014.01.003

66. Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). *Hidden 'risk' in polygenic scores: Clinical use today could exacerbate health disparities*. https://doi.org/10.1101/441261

67. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

68. Miga, K. H., & Wang, T. (2021). The Need for a Human Pangenome Reference Sequence. *Annual Review of Genomics and Human Genetics*, *22*, 81–102. https://doi.org/10.1146/annurev-genom-120120-081921

69. Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., & Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, *24*(24), 2818–2824. https://doi.org/10.1093/bioinformatics/btn548

70. Need, A. C., & Goldstein, D. B. (2009). Next generation disparities in human genomics: Concerns and remedies. *Trends Genet*, *25*(11), 489–494. https://doi.org/10.1016/j.tig.2009.09.012

71. Nishimura, D. (2000). RepeatMasker. *Biotech Software & Internet Report*, *1*(1–2), 36–39. https://doi.org/10.1089/152791600319259

72. Novak, A. M., Hickey, G., Garrison, E., Blum, S., Connelly, A., Dilthey, A., Eizenga, J., Elmohamed, M. A. S., Guthrie, S., Kahles, A., Keenan, S., Kelleher, J., Kural, D., Li, H., Lin, M. F., Miga, K., Ouyang, N., Rakocevic, G., Smuga-Otto, M., … Paten, B. (2017). Genome Graphs. *BioRxiv*, 101378. https://doi.org/10.1101/101378

73. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Cheng, H., … Phillippy, A. M. (2021). The complete sequence of a human genome. *BioRxiv*, 32. https://doi.org/10.1101/2021.05.26.445798

74. Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., Oljira, T., Mekonnen, E., Luiselli, D., Bradman, N., Bekele, E., Zalloua, P., Durbin, R., Kivisild, T., & Tyler-Smith, C. (2015). Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am J Hum Genet*, *96*(6), 986–991. https://doi.org/10.1016/j.ajhg.2015.04.019

75. Paszkiewicz, K., & Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in Bioinformatics*, *11*(5), 457–472. https://doi.org/10.1093/bib/bbq020

76. Paten, B., Novak, A., & Haussler, D. (2014). Mapping to a reference genome structure. *ArXiv Preprint ArXiv:1404.5010*.

77. Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Res*, *27*(5), 665–676. https://doi.org/10.1101/gr.214155.116

78. Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G., Barreiro, L., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J., Fernandes, V., Pereira, L., & Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, *356*, 543–546. https://doi.org/10.1126/science.aal1988

79. Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

80. R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

81. Rand, K. D., Grytten, I., Nederbragt, A. J., Storvik, G. O., Glad, I. K., & Sandve, G. K. (2017). Coordinates and intervals in graph-based reference genomes. *BMC Bioinformatics*, *18*(1). https://doi.org/10.1186/s12859-017-1678-9

82. Reed, F. A., & Tishkoff, S. A. (2006). African human diversity, origins and migrations. *Curr Opin Genet Dev*, *16*(6), 597–605. https://doi.org/10.1016/j.gde.2006.10.008

83. Rosenfeld, J. A., Mason, C. E., & Smith, T. M. (2012). Limitations of the human reference genome for personalized genomics. *PLoS One*, *7*(7), e40294. https://doi.org/10.1371/journal.pone.0040294

84. Ros-Freixedes, R., Battagin, M., Johnsson, M., Gorjanc, G., Mileham, A. J., Rounsley, S. D., & Hickey, J. M. (2018). Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genet Sel Evol*, *50*(1), 64. https://doi.org/10.1186/s12711-018-0436-4

85. Rotimi, C. N., Bentley, A. R., Doumatey, A. P., Chen, G., Shriner, D., & Adeyemo, A. (2017). The genomic landscape of African populations in health and disease. *Human Molecular Genetics*, *26*(R2), R225–R236. https://doi.org/10.1093/hmg/ddx253

86. Satya, R. V., Zavaljevski, N., & Reifman, J. (2012). A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res*, *40*(16), e127. https://doi.org/10.1093/nar/gks425

87. Schlebusch, C. (2010). Issues raised by use of ethnic-group names in genome study. *Nature*, *464*(7288), 487–487. https://doi.org/10.1038/464487a

88. Schlebusch, C. M., Sjodin, P., Breton, G., Gunther, T., Naidoo, T., Hollfelder, N., Sjostrand, A. E., Xu, J., Gattepaille, L. M., Vicente, M., Scofield, D. G., Malmstrom, H., de Jongh, M., Lombard, M., Soodyall, H., & Jakobsson, M. (2020). Khoe-San Genomes Reveal Unique Variation and Confirm the Deepest Population Divergence in Homo sapiens. *Mol Biol Evol*, *37*(10), 2944–2954. https://doi.org/10.1093/molbev/msaa140

89. Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*, *27*(5), 849–864. https://doi.org/10.1101/gr.213611.116

90. Shapiro, J. A., & Sternberg, R. von. (2005). Why repetitive DNA is essential to genome function. *Biological Reviews*, *80*(2), 227–250. https://doi.org/10.1017/S1464793104006657

91. Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., Levin, A. M., Eng, C., Yazdanbakhsh, M., Wilson, J. G., Marrugo, J., Lange, L. A., Williams, L. K., Watson, H., Ware, L. B., … Salzberg, S. L. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet*, *51*(1), 30–35. https://doi.org/10.1038/s41588-018-0273-y

92. Sherman, R. M., & Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nat Rev Genet*, *21*(4), 243–254. https://doi.org/10.1038/s41576-020-0210-7

93. Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). DbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

94. Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., Lintner, K. E., Ding, Q., Wang, Z., Hu, J., Wang, D., Wang, F., Wang, L., Lyon, G. J., Guan, Y., … Wang, K. (2016). Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun*, *7*, 12065. https://doi.org/10.1038/ncomms12065

95. Shumate, A., Zimin, A. V., Sherman, R. M., Puiu, D., Wagner, J. M., Olson, N. D., Pertea, M., Salit, M. L., Zook, J. M., & Salzberg, S. L. (2020). Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol*, *21*(1), 129. https://doi.org/10.1186/s13059-020-02047-7

96. Simpson, J. T., & Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, *22*(3), 549–556. https://doi.org/10.1101/gr.126953.111

97. Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

98. Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048

99. Slater, G. S., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31. https://doi.org/10.1186/1471-2105-6-31

100. Sloan-Heggen, C. M., Bierer, A. O., Shearer, A. E., Kolbe, D. L., Nishimura, C. J., Frees, K. L., Ephraim, S. S., Shibata, S. B., Booth, K. T., Campbell, C. A., Ranum, P. T., Weaver, A. E., Black-Ziegelbein, E. A., Wang, D., Azaiez, H., & Smith, R. J. H. (2016). Comprehensive genetic testing in the clinical evaluation of 1119 patients with hearing loss. *Human Genetics*, *135*(4), 441–450. https://doi.org/10.1007/s00439-016-1648-8

101. Slotkin, R. K. (2018). The case for not masking away repetitive DNA. *Mob DNA*, *9*, 15. https://doi.org/10.1186/s13100-018-0120-9

102. Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res*, *32*(Web Server issue), W309-12. https://doi.org/10.1093/nar/gkh379

103. Stevenson, K. R., Coolon, J. D., & Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, *14*(1), 536. https://doi.org/10.1186/1471-2164-14-536

104. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H., Konkel, M. K., Malhotra, A., Stutz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

105. The 100,000 Genomes Project Pilot Investigators. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care—Preliminary Report. *New England Journal of Medicine*, *385*(20), 1868–1880. https://doi.org/10.1056/NEJMoa2035790

106. The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

107. The Genome Reference Consortium. (n.d.). *Frequently Asked Questions*. Retrieved June 16, 2020, from https://www.ncbi.nlm.nih.gov/grc/help/faq/

108. The H3Africa Consortium, Matovu, E., Bucheton, B., Chisi, J., Enyaru, J., Hertz-Fowler, C., Koffi, M., Macleod, A., Mumba, D., Sidibe, I., Simo, G., Simuunza, M., Mayosi, B., Ramesar, R., Mulder, N., Ogendo, S., Mocumbi, A. O., Hugo-Hamman, C., Ogah, O., … Rotimi, C. (2014). Enabling the genomic revolution in Africa. *Science*, *344*(6190), 1346–1348. https://doi.org/10.1126/science.1251546

109. The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258

110. Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., … Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science*, *324*(5930), 1035–1044. https://doi.org/10.1126/science.1172257

111. Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews. Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117

112. U.S. Department of Energy Office of Science. (2008). *Genomics and Its Impact on Science and Society: The Human Genome Project and Beyond*. https://web.ornl.gov/sci/techresources/Human_Genome/publicat/primer2001/primer11.pdf

113. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

114. Wang, M., & Kong, L. (2019). pblat: A multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinformatics*, *20*(1), 28. https://doi.org/10.1186/s12859-019-2597-8

115. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., … Haussler, D. (2022). The Human Pangenome Project: A global resource to map genomic diversity. *Nature*, *604*(7906), 437–446. https://doi.org/10.1038/s41586-022-04601-8

116. Wong, K. H. Y., Ma, W., Wei, C.-Y., Yeh, E.-C., Lin, W.-J., Wang, E. H. F., Su, J.-P., Hsieh, F.-J., Kao, H.-J., Chen, H.-H., Chow, S. K., Young, E., Chu, C., Poon, A., Yang, C.-F., Lin, D.-S., Hu, Y.-F., Wu, J.-Y., Lee, N.-C., … Kwok, P.-Y. (2020). Towards a reference genome that captures global genetic diversity. *Nature Communications*, *11*(1), 5482. https://doi.org/10.1038/s41467-020-19311-w

117. Yan, D., Tekin, D., Bademci, G., Foster, J., Cengiz, F. B., Kannan-Sundhari, A., Guo, S., Mittal, R., Zou, B., Grati, M., Kabahuma, R. I., Kameswaran, M., Lasisi, T. J., Adedeji, W. A., Lasisi, A. O., Menendez, I., Herrera, M., Carranza, C., Maroofian, R., … Tekin, M. (2016). Spectrum of DNA variants for nonsyndromic deafness in a large cohort from multiple continents. *Human Genetics*, *135*(8), 953–961. https://doi.org/10.1007/s00439-016-1697-z

118. Ye, C., Ma, Z. S., Cannon, C. H., Pop, M., & Douglas, W. Y. (2012). *Exploiting sparseness in de novo genome assembly*. *13*, S1.

119. Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: Improvements for better sequence analysis. *Nucleic Acids Research*, *34*(Web Server), W6–W9. https://doi.org/10.1093/nar/gkl164

120. Yoo, A. B., Jette, M. A., & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. In D. Feitelson, L. Rudolph, & U. Schwiegelshohn (Eds.), *Job Scheduling Strategies for Parallel Processing* (pp. 44–60). Springer. https://doi.org/10.1007/10968987_3

121. Yu, Y., & Wei, C. (2020). A powerful HUPAN on a pan-genome study: Significance and perspectives. *Cancer Biol Med*, *17*(1), 1–5. https://doi.org/10.20892/j.issn.2095-3941.2019.0317

122. Yuan, S., & Qin, Z. (2012). *Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression*. 718–724.

123. Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

# Appendix A

**Table A1. The 100 samples making up the 1000 Genomes Project dataset used to test the HUPAN pipeline.** Sample IDs are those used by the International Genome Sample Resource.

| Sample ID | Population | Population code | Assembled |
|---|---|---|---|
| HG02947 | Esan in Nigeria | ESN | Yes |
| HG02971 | Esan in Nigeria | ESN | Yes |
| HG02974 | Esan in Nigeria | ESN | Yes |
| HG03099 | Esan in Nigeria | ESN | Yes |
| HG03100 | Esan in Nigeria | ESN | Yes |
| HG03109 | Esan in Nigeria | ESN | Yes |
| HG03124 | Esan in Nigeria | ESN | Yes |
| HG03159 | Esan in Nigeria | ESN | Yes |
| HG03163 | Esan in Nigeria | ESN | Yes |
| HG03189 | Esan in Nigeria | ESN | Yes |
| HG03202 | Esan in Nigeria | ESN | Yes |
| HG03265 | Esan in Nigeria | ESN | Yes |
| HG03271 | Esan in Nigeria | ESN | Yes |
| HG03298 | Esan in Nigeria | ESN | Yes |
| HG03301 | Esan in Nigeria | ESN | Yes |
| HG03352 | Esan in Nigeria | ESN | Yes |
| HG03370 | Esan in Nigeria | ESN | Yes |
| HG03515 | Esan in Nigeria | ESN | Yes |
| HG03517 | Esan in Nigeria | ESN | Yes |
| HG03518 | Esan in Nigeria | ESN | Yes |
| HG02461 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02465 | Gambia in the Western Division (Mandinka) | GWD | No |
| HG02561 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02573 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02583 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02588 | Gambia in the Western Division (Mandinka) | GWD | No |
| HG02679 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02769 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02771 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02807 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02810 | Gambia in the Western Division (Mandinka) | GWD | Yes |

| | | | |
|---|---|---|---|
| HG02840 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02851 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02860 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02890 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG02891 | Gambia in the Western Division (Mandinka) | GWD | No |
| HG03049 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG03241 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG03258 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| HG03259 | Gambia in the Western Division (Mandinka) | GWD | Yes |
| NA19028 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19031 | Luhya in Webuye, Kenya | LWK | No |
| NA19038 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19309 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19317 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19318 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19328 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19332 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19350 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19372 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19380 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19385 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19434 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19445 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19448 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19461 | Luhya in Webuye, Kenya | LWK | No |
| NA19463 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19467 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19471 | Luhya in Webuye, Kenya | LWK | Yes |
| NA19473 | Luhya in Webuye, Kenya | LWK | Yes |
| HG03054 | Mende in Sierra Leone | MSL | Yes |
| HG03079 | Mende in Sierra Leone | MSL | Yes |
| HG03086 | Mende in Sierra Leone | MSL | Yes |
| HG03097 | Mende in Sierra Leone | MSL | Yes |
| HG03224 | Mende in Sierra Leone | MSL | Yes |
| HG03225 | Mende in Sierra Leone | MSL | Yes |
| HG03380 | Mende in Sierra Leone | MSL | Yes |
| HG03385 | Mende in Sierra Leone | MSL | Yes |
| HG03388 | Mende in Sierra Leone | MSL | Yes |
| HG03391 | Mende in Sierra Leone | MSL | Yes |
| HG03439 | Mende in Sierra Leone | MSL | Yes |
| HG03449 | Mende in Sierra Leone | MSL | Yes |

| | | | |
|---|---|---|---|
| HG03457 | Mende in Sierra Leone | MSL | Yes |
| HG03464 | Mende in Sierra Leone | MSL | Yes |
| HG03470 | Mende in Sierra Leone | MSL | Yes |
| HG03473 | Mende in Sierra Leone | MSL | Yes |
| HG03478 | Mende in Sierra Leone | MSL | Yes |
| HG03485 | Mende in Sierra Leone | MSL | Yes |
| HG03559 | Mende in Sierra Leone | MSL | Yes |
| HG03563 | Mende in Sierra Leone | MSL | Yes |
| NA18498 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA18499 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA18501 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA18504 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA18505 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA18871 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA18878 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA18879 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA18933 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19099 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19107 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19118 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19129 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19131 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19141 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19149 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19171 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19190 | Yoruba in Ibadan, Nigeria | YRI | Yes |
| NA19209 | Yoruba in Ibadan, Nigeria | YRI | Yes |

# Appendix B

**Table B1. The 285 samples making up the initial pan-African dataset**. Sample IDs are those used by the respective datasets of origin. Sample details, such as ethnolinguistic groups and countries of origin, were obtained from the metadata contained with the datasets. If samples were not included in the final pan-African dataset, the reason is given. SGDP: Simons Genome Diversity Project dataset, 1kGP: 1000 Genomes Project dataset, HGDP: Human Genome Diversity Project dataset, TrypanoGEN: Trypanosomiasis Genomics Network dataset, H3Africa: H3Africa Consortium dataset, EHC: Ethiopian high-coverage dataset, ELC: Ethiopian low-coverage dataset, KSP: Khoe-San project dataset.

| Sample ID | Ethnolinguistic group | Country of origin | Regional ancestral group | Included in African pan-genome | Dataset of origin |
|---|---|---|---|---|---|
| LP6005441-DNA_B02 | BantuKenya | Kenya | BantuNC | No - failed QC | SGDP |
| LP6005441-DNA_F01 | BantuHerero | Botswana/Namibia | BantuNC | Yes | SGDP |
| LP6005443-DNA_A01 | BantuKenya | Kenya | BantuNC | Yes | SGDP |
| LP6005443-DNA_E02 | BantuHerero | Botswana/Namibia | BantuNC | Yes | SGDP |
| LP6005443-DNA_F02 | BantuTswana | Botswana/Namibia | BantuNC | No - failed QC | SGDP |
| LP6005443-DNA_G02 | BantuTswana | Botswana/Namibia | BantuNC | Yes | SGDP |
| NA19028 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19038 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19309 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19317 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19318 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19328 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19332 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19350 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19372 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19380 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19385 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19434 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19448 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19461 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19463 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19467 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19471 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| NA19473 | Luhya | Kenya | BantuNC | Yes | 1kGP |
| SAMEA2580844 | Bantu | South Africa | BantuNC | Yes | HGDP |
| SAMEA2580845 | Bantu | South Africa | BantuNC | Yes | HGDP |
| SAMEA2580848 | Bantu | South Africa | BantuNC | Yes | HGDP |
| SAMEA2580849 | Bantu | South Africa | BantuNC | Yes | HGDP |
| SAMEA2580852 | Bantu | Kenya | BantuNC | Yes | HGDP |
| SAMEA2580853 | Bantu | Kenya | BantuNC | Yes | HGDP |
| SAMEA2580854 | Bantu | Kenya | BantuNC | Yes | HGDP |
| SAMEA2580855 | Bantu | Kenya | BantuNC | Yes | HGDP |
| SAMEA2580856 | Bantu | Kenya | BantuNC | Yes | HGDP |
| SAMEA2580858 | Bantu | Kenya | BantuNC | Yes | HGDP |

| | | | | | |
|---|---|---|---|---|---|
| SAMEA2580859 | Bantu | Kenya | BantuNC | Yes | HGDP |
| SAMEA2580861 | Bantu | Kenya | BantuNC | Yes | HGDP |
| SAMEA2580862 | Bantu | Kenya | BantuNC | Yes | HGDP |
| LP6005441-DNA_A08 | Mbuti | Congo | CARF | No - failed QC | SGDP |
| LP6005441-DNA_B08 | Mbuti | Congo | CARF | No - failed QC | SGDP |
| LP6005441-DNA_G02 | Biaka | Central African Republic | CARF | Yes | SGDP |
| LP6005441-DNA_H02 | Biaka | Central African Republic | CARF | Yes | SGDP |
| LP6005592-DNA_C03 | Mbuti | Congo | CARF | Yes | SGDP |
| SAMEA2580802 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580803 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580807 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580808 | Biaka | Central African Republic | CARF | No - contamination | HGDP |
| SAMEA2580809 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580813 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580814 | Biaka | Central African Republic | CARF | No - contamination | HGDP |
| SAMEA2580815 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580816 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580817 | Biaka | Central African Republic | CARF | No - unable to assemble | HGDP |
| SAMEA2580818 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580819 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580820 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580821 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580822 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580823 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA2580824 | Biaka | Central African Republic | CARF | No - unable to assemble | HGDP |
| SAMEA2580826 | Mbuti | DRC | CARF | No - unable to assemble | HGDP |
| SAMEA2580828 | Mbuti | DRC | CARF | No - unable to assemble | HGDP |
| SAMEA2580829 | Mbuti | DRC | CARF | No - unable to assemble | HGDP |

| | | | | | |
|---|---|---|---|---|---|
| SAMEA2580830 | Mbuti | DRC | CARF | Yes | HGDP |
| SAMEA2580831 | Mbuti | DRC | CARF | Yes | HGDP |
| SAMEA2580834 | Mbuti | DRC | CARF | Yes | HGDP |
| SAMEA2580836 | Mbuti | DRC | CARF | Yes | HGDP |
| SAMEA2580837 | Mbuti | DRC | CARF | No - unable to assemble | HGDP |
| SAMEA3506072 | Biaka | Central African Republic | CARF | Yes | HGDP |
| SAMEA3506073 | Biaka | Central African Republic | CARF | Yes | HGDP |
| 10_150304_L001 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |
| 11_150302_L002 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 12_150304_L004 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |
| 13-RB103T-C_150611_L007 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |
| 14_150304_L001 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 16-RB106T-C_150611_L008 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 17_150304_L002 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 18-RB108T-C_150611_L008 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |
| 20_150302_L001 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 22_150304_L004 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |
| 24-RB114T_C_150616_L001 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 24-RB114T_C_150616_L002 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 25-RB115T_C_150616_L001 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 25-RB115T_C_150616_L002 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 26-RB116T_C_150616_L001 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 26-RB116T_C_150616_L002 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 27_150304_L003 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |
| 28-RB118T_C_150616_L003 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |
| 29-RB119T_C_150616_L003 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 30-RB122T_C_150616_L004 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |
| 31_150304_L003 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 32-RB124T_C_150616_L004 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 37-RB142T_C_150616_L005 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 38-RB151T_C_150616_L005 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 39_150304_L002 | Ngongo/Mbala | DRC | CWNC | Yes | TrypanoGEN |

| | | | | | |
|---|---|---|---|---|---|
| 4-RB091T-C_150611_L006 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 5_150302_L001 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 6_150302_L002 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| 7-RB094T-C_150611_L006 | Ngongo/Mbala | DRC | CWNC | No - failed QC | TrypanoGEN |
| CB12 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB14 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB15 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB16 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB17 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB24 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB29 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB31 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB32 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB33 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CB7 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| CF24 | Mundani | Cameroon | CWNC | Yes | H3Africa |
| CF25 | Mundani | Cameroon | CWNC | Yes | H3Africa |
| CF26 | Mundani | Cameroon | CWNC | Yes | H3Africa |
| CF27 | Mundani | Cameroon | CWNC | Yes | H3Africa |
| CF30 | Mundani | Cameroon | CWNC | Yes | H3Africa |
| CF37 | Mundani | Cameroon | CWNC | Yes | H3Africa |
| CF46 | Mundani | Cameroon | CWNC | Yes | H3Africa |
| CF49 | Mundani | Cameroon | CWNC | Yes | H3Africa |
| CP17 | Bamileke | Cameroon | CWNC | Yes | H3Africa |
| CP28 | Bamileke | Cameroon | CWNC | Yes | H3Africa |
| CP36 | Bamileke | Cameroon | CWNC | Yes | H3Africa |
| CP38 | Bamileke | Cameroon | CWNC | Yes | H3Africa |
| CP40 | Bamileke | Cameroon | CWNC | Yes | H3Africa |
| CP48 | Ngoumba | Cameroon | CWNC | Yes | H3Africa |
| egpg5305762 | Amhara | Ethiopia | EAAA | Yes | EHC |
| egpg5305763 | Amhara | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305765 | Amhara | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305770 | Amhara | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305771 | Amhara | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305773 | Amhara | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305778 | Amhara | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305781 | Amhara | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305904 | Oromo | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305910 | Oromo | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305911 | Oromo | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305918 | Oromo | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305919 | Oromo | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305920 | Somali | Ethiopia | EAAA | Yes | EHC |
| egpg5305936 | Somali | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305937 | Somali | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305944 | Somali | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305945 | Somali | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305949 | Oromo | Ethiopia | EAAA | Yes | EHC |

| | | | | | |
|---|---|---|---|---|---|
| egpg5305952 | Somali | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305953 | Somali | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305959 | Wolayta | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305960 | Wolayta | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305961 | Wolayta | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305967 | Wolayta | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305968 | Wolayta | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305969 | Wolayta | Ethiopia | EAAA | No - failed QC | ELC |
| egpg5305974 | Wolayta | Ethiopia | EAAA | Yes | EHC |
| LP6005442-DNA_D09 | Somali | Kenya | EAAA | Yes | SGDP |
| 10-UO_113T_150731_L003 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 12-UO_125T_150731_L004 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 13-UO_134T_150731_L005 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 14-UO_136T_150731_L005 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 15-UO_140T_150731_L006 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 16-UO_142T_150731_L006 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 17-UO_144T_150731_L007 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 18-UO_148T_150731_L007 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 19-UO_153T_150813_L001 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 1-UO_005T_150730_L001 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 20-UO_157T_150813_L001 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 21-UO_006C_150813_L002 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 22-UO_045C_150813_L002 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 23-UO_064C_150813_L003 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 24-UO_065C_150813_L003 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 25-UO_068C_150813_L004 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 26-UO_070C_150813_L004 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 27-UO_072C_150813_L001 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 28-UO_073C_150813_L001 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 29-UO_075C_150813_L002 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |

| | | | | | |
|---|---|---|---|---|---|
| 2-UO_044T_150730_L001 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 30-UO_077C_150813_L002 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 31-UO_079C_150813_L003 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 32-UO_080C_150813_L003 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 33-UO_082C_150813_L004 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 34-UO_088C_150813_L004 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 35-UO_090C_150813_L005 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 36-UO_091C_150813_L005 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 37-UO_092C_150902_L001 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 38-UO_093C_150902_L001 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 39-UO_098C_150902_L002 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 3-UO_046T_150730_L002 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 40-UO_099C_150902_L002 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 41-UO_100C_150902_L003 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 42-UO_105C_150902_L003 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 43-UO_106C_150902_L004 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 44-UO_107C_150902_L004 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 45-UO_109C_150902_L005 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 46-UO_114C_150902_L005 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 47-UO_115C_150902_L006 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 48-UO_127C_150902_L006 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 49-UO_137C_150902_L007 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 4-UO_048T_150730_L002 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 50-UO_185C_150902_L007 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| 5-UO_063T_150731_L001 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |

| | | | | | |
|---|---|---|---|---|---|
| 6-UO_069T_150731_L001 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 7-UO_071T_150731_L002 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 8-UO_074T_150731_L002 | Lugbara | Uganda | EANS | Yes | TrypanoGEN |
| 9-UO_076T_150731_L003 | Lugbara | Uganda | EANS | No - failed QC | TrypanoGEN |
| egpg5305767 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305775 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305776 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305783 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305791 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305792 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305793 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305798 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305799 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| egpg5305814 | Gumuz | Ethiopia | EANS | Yes | EHC |
| egpg5305831 | Gumuz | Ethiopia | EANS | No - failed QC | ELC |
| LP6005442-DNA_F09 | Luo | Kenya | EANS | No - failed QC | SGDP |
| LP6005443-DNA_B09 | Dinka | Sudan | EANS | No - contamination | SGDP |
| LP6005443-DNA_E06 | Masai | Kenya | EANS | No - failed QC | SGDP |
| LP6005443-DNA_F06 | Masai | Kenya | EANS | Yes | SGDP |
| LP6005443-DNA_H08 | Dinka | Sudan | EANS | No - contamination | SGDP |
| LP6005677-DNA_G01 | Luo | Kenya | EANS | No - failed QC | SGDP |
| HG03225 | Mende | Sierra Leone | FWNC | Yes | 1kGP |
| HG03380 | Mende | Sierra Leone | FWNC | Yes | 1kGP |
| HG03385 | Mende | Sierra Leone | FWNC | Yes | 1kGP |
| HG03388 | Mende | Sierra Leone | FWNC | Yes | 1kGP |
| HG03449 | Mende | Sierra Leone | FWNC | Yes | 1kGP |
| IAJ0K | Mossi | Burkina Faso | FWNC | Yes | H3Africa |
| IAM0N | Mossi | Burkina Faso | FWNC | Yes | H3Africa |
| IAP0Q | Mossi | Burkina Faso | FWNC | Yes | H3Africa |
| IBJ0M | Mossi | Burkina Faso | FWNC | Yes | H3Africa |
| IBQ0T | Mossi | Burkina Faso | FWNC | Yes | H3Africa |
| LP6005441-DNA_F07 | Mandenka | Senegal | FWNC | Yes | SGDP |
| LZP0B | Kassena | Ghana | FWNC | Yes | H3Africa |
| NA18498 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA18499 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA18501 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA18504 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA18505 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA19099 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA19107 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA19118 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA19129 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| NA19131 | Yoruba | Nigeria | FWNC | Yes | 1kGP |
| PAV0Q | Kassena | Ghana | FWNC | Yes | H3Africa |

| | | | | | |
|---|---|---|---|---|---|
| PAZ0V | Kassena | Ghana | FWNC | Yes | H3Africa |
| PBA0X | Kassena | Ghana | FWNC | Yes | H3Africa |
| PBC0A | Kassena | Ghana | FWNC | Yes | H3Africa |
| SAMEA2580885 | Mandenka | Senegal | FWNC | Yes | HGDP |
| SAMEA2580886 | Mandenka | Senegal | FWNC | Yes | HGDP |
| SAMEA2580889 | Mandenka | Senegal | FWNC | Yes | HGDP |
| SAMEA2580891 | Mandenka | Senegal | FWNC | Yes | HGDP |
| SAMEA2580897 | Mandenka | Senegal | FWNC | Yes | HGDP |
| SAMEA3506075 | Mandenka | Senegal | FWNC | Yes | HGDP |
| SAMEA3506076 | Mandenka | Senegal | FWNC | Yes | HGDP |
| SAMEA3506077 | Mandenka | Senegal | FWNC | Yes | HGDP |
| SAMEA3506078 | Mandenka | Senegal | FWNC | Yes | HGDP |
| KSP062 | Karretjie People | South Africa | KhoeSan | Yes | KSP |
| KSP063 | Karretjie People | South Africa | KhoeSan | No - contamination | KSP |
| KSP065 | Karretjie People | South Africa | KhoeSan | Yes | KSP |
| KSP067 | Karretjie People | South Africa | KhoeSan | No - contamination | KSP |
| KSP069 | Karretjie People | South Africa | KhoeSan | Yes | KSP |
| KSP092 | \|Gui and \|\|Gana | Botswana | KhoeSan | Yes | KSP |
| KSP096 | \|Gui and \|\|Gana | Botswana | KhoeSan | Yes | KSP |
| KSP103 | Ju\|'hoansi | Namibia | KhoeSan | Yes | KSP |
| KSP105 | Ju\|'hoansi | Namibia | KhoeSan | No - contamination | KSP |
| KSP106 | Ju\|'hoansi | Namibia | KhoeSan | Yes | KSP |
| KSP111 | Ju\|'hoansi | Namibia | KhoeSan | No - contamination | KSP |
| KSP116 | Ju\|'hoansi | Namibia | KhoeSan | Yes | KSP |
| KSP124 | Nama | Namibia | KhoeSan | No - contamination | KSP |
| KSP134 | Nama | Namibia | KhoeSan | Yes | KSP |
| KSP137 | Nama | Namibia | KhoeSan | Yes | KSP |
| KSP139 | Nama | Namibia | KhoeSan | Yes | KSP |
| KSP140 | Nama | Namibia | KhoeSan | Yes | KSP |
| KSP146 | !Xun | Angola | KhoeSan | Yes | KSP |
| KSP150 | !Xun | Angola | KhoeSan | No - contamination | KSP |
| KSP152 | !Xun | Angola | KhoeSan | Yes | KSP |
| KSP154 | !Xun | Angola | KhoeSan | No - contamination | KSP |
| KSP155 | !Xun | Angola | KhoeSan | Yes | KSP |
| KSP224 | \|Gui and \|\|Gana | Botswana | KhoeSan | Yes | KSP |
| KSP225 | \|Gui and \|\|Gana | Botswana | KhoeSan | No - contamination | KSP |
| KSP228 | \|Gui and \|\|Gana | Botswana | KhoeSan | Yes | KSP |
| LP6005441-DNA_A11 | Ju\|'hoanNorth | Namibia | KhoeSan | No - failed QC | SGDP |

| | | | | | |
|---|---|---|---|---|---|
| LP6005441-DNA_B11 | Ju\|'hoanNorth | Namibia | KhoeSan | No - failed QC | SGDP |
| LP6005443-DNA_G08 | Ju\|'hoanNorth | Namibia | KhoeSan | Yes | SGDP |
| LP6005592-DNA_C05 | KhomaniSan | SouthAfrica | KhoeSan | No - contamination | SGDP |
| LP6005677-DNA_D03 | KhomaniSan | SouthAfrica | KhoeSan | No - contamination | SGDP |
| SAMEA2580840 | San | Namibia | KhoeSan | Yes | HGDP |
| SAMEA2580841 | San | Namibia | KhoeSan | No - contamination | HGDP |

# Appendix C

The full MultiQC (Ewels et al., 2016) reports obtained from the QUAST analysis of the starting assembled pan-African dataset of 279 samples can be accessed at https://github.com/BournSupremacy/BioinformaticsTools/tree/main/QuastMultiQC/startingDataset.

The full MultiQC reports obtained from the QUAST analysis of the final pan-African dataset of 168 samples can be accessed at https://github.com/BournSupremacy/BioinformaticsTools/tree/main/QuastMultiQC/finalDataset.

The scripts for plotting and comparing the MultiQC data for each dataset can be accessed at https://github.com/BournSupremacy/BioinformaticsTools/tree/main/QuastMultiQC.

# Appendix D

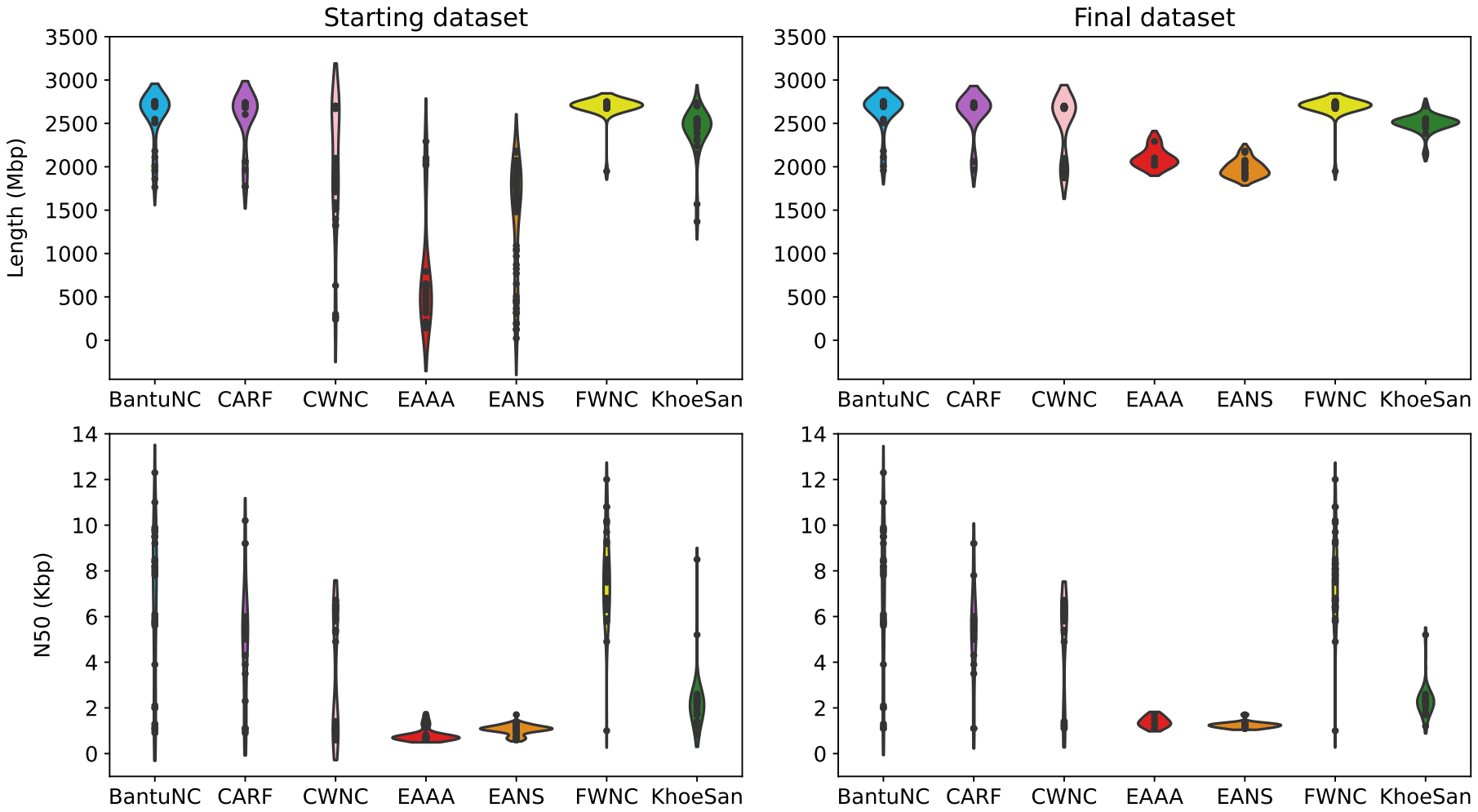


**Figure D1. Comparison of the distribution profiles of the total length (Mbp) and N50 values (kbp) before and after removing low-quality samples from all seven regional ancestral groups.** The width of each violin plot indicates the density of samples at those values. Each point represents a sample. Each violin plot shows density distributions extending past both the lowest and highest datapoints due to the nature of the plots.

# Appendix E

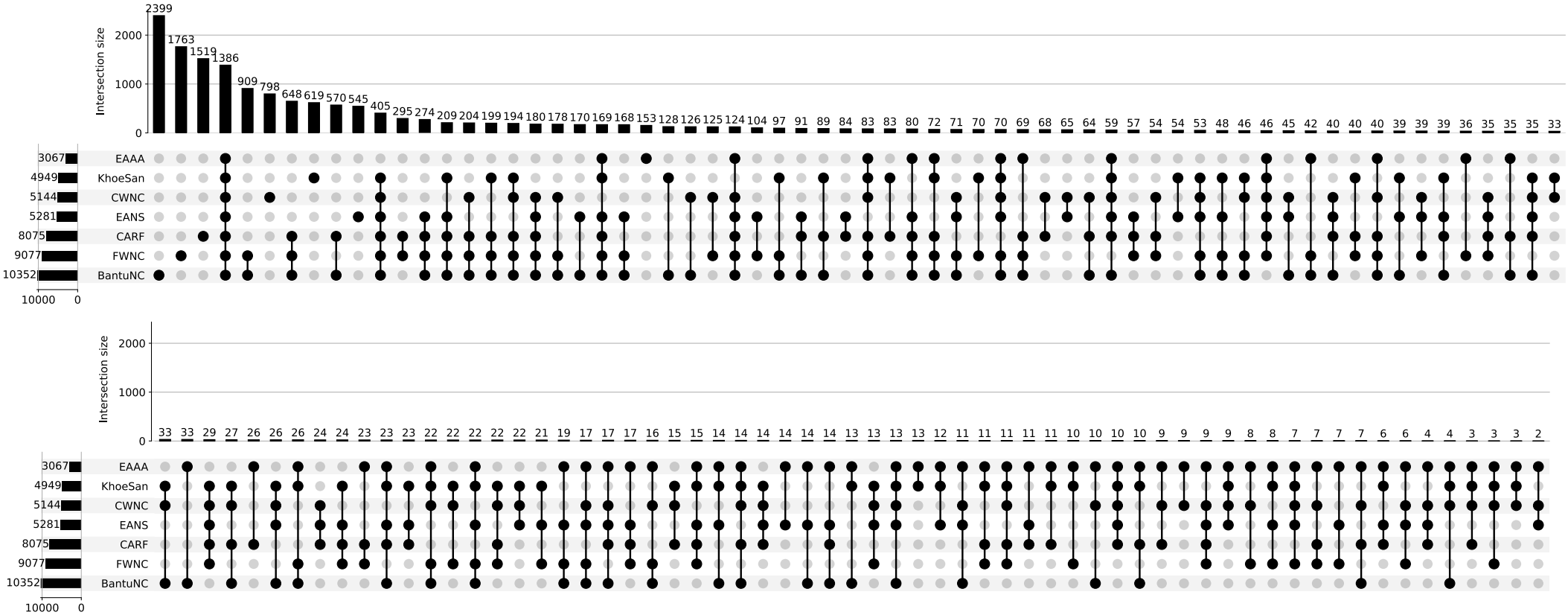


**Figure E1. Upset plot showing the sample membership of the 17 550 nucleotide clusters produced by CD-HIT following the merging and redundancy removal of the pan-African non-reference sequences.** Intersection size indicates the number of times that union of populations is found within a cluster. All intersection size values add up to 17 550. Values next to the population names indicate how many clusters that population contributed towards.

# Appendix F

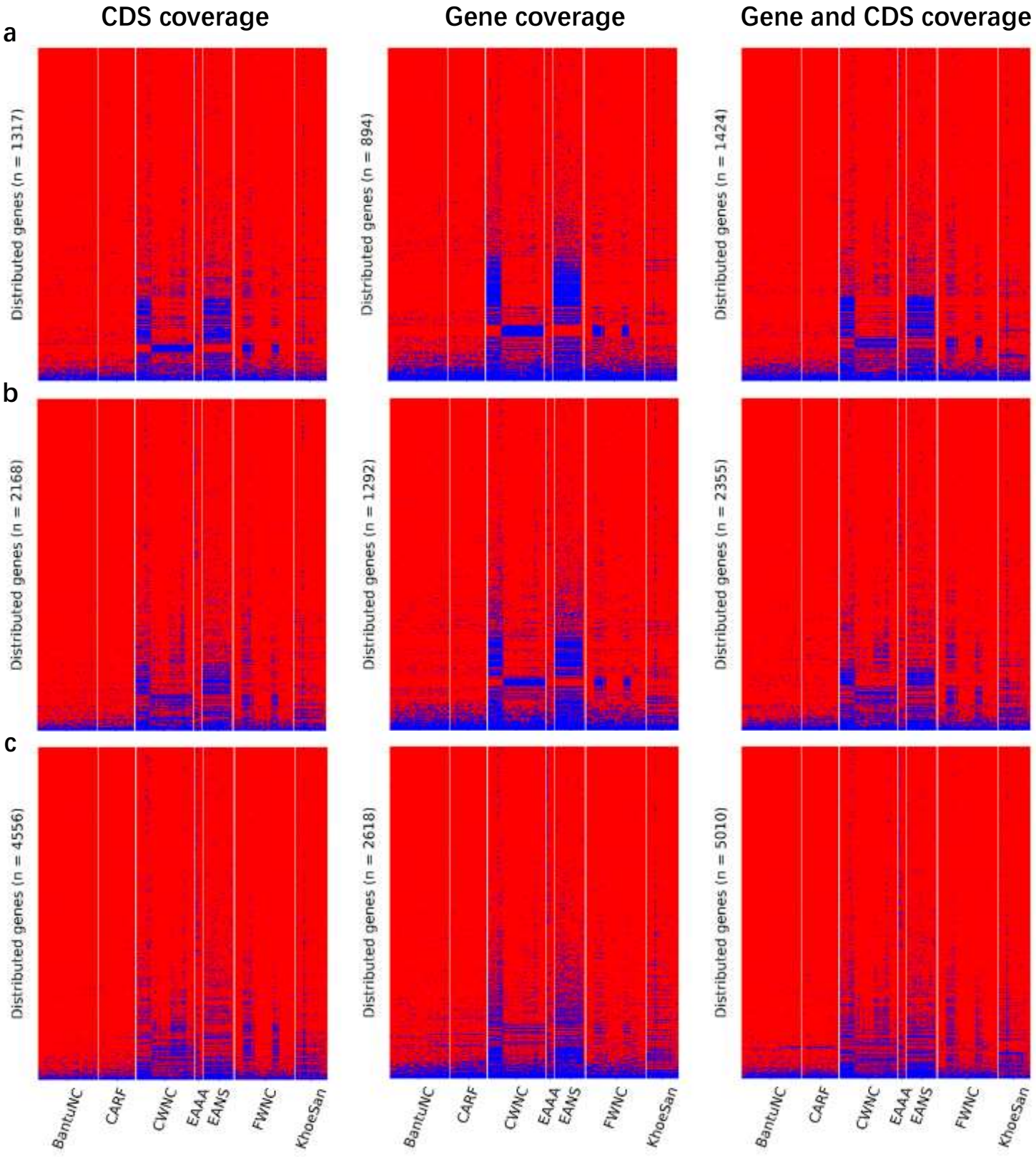


**Figure F1. Gene presence-absence variation plots at different thresholds for the 168 samples used to assemble the African pan-genome.** Red signifies presence and blue signifies absence. The first column shows when only CDS coverage of genes is required, the second column shows when whole gene coverage is required, and the third column shows when both are required. The number of distributed genes for each threshold is shown on the y-axis. The distributed genes are ordered from those present within the most samples to those present within the least samples. Regional ancestral groups are separated by white lines. The samples are ordered alphabetically by sample ID within each group. For each row of plots, the CDS and/or gene region was defined as present if **(a)** ⩾ 85%, **(b)** ⩾ 90%, or **(c)** ⩾ 95% of the sequence was covered by raw sequenced reads from each sample.