Admixture History of the South African "Coloured" Populations

Ryan Joseph Daniels

St. Edmund Hall
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity Term 2019

Abstract

Admixed people formed from colonial era migrations are a major feature of present-day diversity. In Southern Africa, establishment of the Dutch Cape Colony instigated the so-called "Coloured" communities, associated, essentialistically, with 'mixed descent'. This multi-way admixture involved diverse regional groups producing a fascinating genetic story. In this thesis I develop further the history of the Cape Colony admixture using genome-wide SNP chip data.

In chapter 4 I characterise the ancestry of 733 Cape Town "Coloured" individuals, identifying specific contributions using haplotype coalescence. I find no substructure despite identifying 11 ancestral contributions. Results concur with historic accounts of South Asian, European, African KhoeSan and Bantu-speakers, and South-East Asian contributions from slaves, indigenous people and settlers. Signals from known slaving regions are detected as is a previously unknown Central European and Iberian contribution.

In chapter 5 I explore evidence of admixture events and sources using linkage-disequilibrium decay curve fitting for the above dataset. Results indicate recent, continuous admixture pre-dating the Cape Colony. Europeans, South-East Asians and the KhoeSan are likely sources for the pre-Colonial signals possibly reflecting shipwrecks along the coast. I find no clear evidence for Southern Bantu and Malagasy contributions to the South African Coloured (SAC).

In chapter 6, I evaluate geographic expansion and admixture in the "Coloured" sub-identity: Cape Malay, Griekwa and Baster. I genotype 116 new individuals. Geography and ethnicity correlate with genetics but inconsistently. Griekwa and Cape Malay show evidence of gene flow with geographic neighbouring communities while the Namibian Baster are the clearest example of endogamy, associated with isolation and inbreeding.

Novel and challenging aspects are added to our understanding of the "Coloured" communities. This work highlights that admixture pre-dates the arrival of settlers and how social identities are formed along more dynamic processes than simple admixture proportions.

Admixture History of the South African "Coloured" Populations

Ryan Joseph Daniels

St. Edmund Hall

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2019

Dedicated to those generations who endured
persecution, discrimination and victimisation.
Yet survived to contribute, culturally and genetically,
to those alive today.
To my parents.
To their parents.

Declaration

This thesis comprises three unpublished data chapters. All work presented here is my own, and where I have benefited from other people's research, I have declared this in the text. In addition, in Table 7.1, I have listed all individuals that have helped, either by contributing DNA samples, through technical help in the laboratory or with analyses.

Abstract

Admixed people formed from colonial era migrations are a major feature of present-day diversity. In Southern Africa, establishment of the Dutch Cape Colony instigated the so-called "Coloured" communities, associated, essentialistically, with 'mixed descent'. This multi-way admixture involved diverse regional groups producing a fascinating genetic story. In this thesis I develop further the history of the Cape Colony admixture using genome-wide SNP chip data.

In chapter 4 I characterise the ancestry of 733 Cape Town "Coloured" individuals, identifying specific contributions using haplotype coalescence. I find no substructure despite identifying 11 ancestral contributions. Results concur with historic accounts of South Asian, European, African KhoeSan and Bantu-speakers, and South-East Asian contributions from slaves, indigenous people and settlers. Signals from known slaving regions are detected as is a previously unknown Central European and Iberian contribution.

In chapter 5 I explore evidence of admixture events and sources using linkage-disequilibrium decay curve fitting for the above dataset. Results indicate recent, continuous admixture pre-dating the Cape Colony. Europeans, South-East Asians and the KhoeSan are likely sources for the pre-Colonial signals possibly reflecting shipwrecks along the coast. I find no clear evidence for Southern Bantu and Malagasy contributions to the South African Coloured (SAC).

In chapter 6, I evaluate geographic expansion and admixture in the "Coloured" sub-identity: Cape Malay, Griekwa and Baster. I genotype 116 new individuals. Geography and ethnicity correlate with genetics but inconsistently. Griekwa and Cape Malay show evidence of gene flow with geographic neighbouring communities while the Namibian Baster are the clearest example of endogamy, associated with isolation and inbreeding.

Novel and challenging aspects are added to our understanding of the "Coloured" communities. This work highlights that admixture pre-dates the arrival of settlers and how social identities are formed along more dynamic processes than simple admixture proportions.

Contents

List of Figures

List of Tables

List of Abbreviations and Special Characters

1KGP . . . . . 1000 Genome Project

α . . . . . The proportion of ancestry contributed to an admixed population by a source during admixture

θ . . . . . Per site mutation rate used in CHROMOPAINTER

μ . . . . . Mutation rate

BCE . . . . . Before Common Era (=BC)

CE . . . . . Common Era (=AD)

cM . . . . . CentiMorgan

CP . . . . . CHROMOPAINTER

DNA . . . . . Deoxyribose Nucleic Acid

f_3 . . . . . f_3 indices for testing tree-like relationships/admixture as per Patterson, Moorjani, Luo, *et al.* [1] .

fS . . . . . fineSTRUCTURE

GWAS . . . . . Genome-wide Association Study

HapMap . . . International HapMap Project

HGDP . . . . . Human Genome Diversity Panel

IBS . . . . . Identity by state

IBD . . . . . Identity by descent

Kb . . . . . Kilobase

Km . . . . . Kilometre

Kya . . . . . Thousand years ago

LD . . . . . Linkage disequilibrium

LGM . . . . . Last Glacial Maximum

MALDER . . Multiple Admixture Linkage Disequilibrium for Evolutionary Relationships (software)

MANOVA . . Multivariate analysis of variance

MCMC . . . . Markov Chain Monte Carlo

Mb . . . . . Megabase

mtDNA . . . . mitochondrial DNA

Mya . . . . . Million years ago

NNLS . . . . . Non-negative-least-squares regression

NRY . . . . . Non-recombining sections of the Y-chromosome

PC(A) . . . . . Principal Component (Analysis)

POPRES . . . Population Resource Database

SAC . . . . . South African (Cape) "Coloured" and associated ethno-racial affinities

SAHGP . . . . Southern African Human Genome Programme

SNP . . . . . Single Nucleotide Polymorphism

STR . . . . . Short Tandem Repeat

TMRCA . . . Time to the most recent common ancestor

TVD . . . . . Total variance distance

Glossary

1KGP . . . . . 1000 Genome Project; Collaborative project run between 2008 - 2015 to characterise the geographic and functional spectra of human genetic variation specifically for understanding disease. The project initially included 14 populations and 1 092 individuals but now includes 2 504 individuals from 24 populations [2], [3]. See Internationalgenome.org

AGVP . . . . . African Genome Variation Project; SNP chip dense genotypes from 1,481 individuals and whole-genome sequences from 320 individuals across sub-Saharan Africa [4].

Copying vector A vector of values representing the proportion ancestry received from each of a possible number of donors. Produced by applying a mixture model such as Non-negative least squares regression to the total genome length copied or the number of haplotype chunks copied as output by CHROMOPAINTER.

Donor . . . . . An individual or population which is the nearest genetic proxy for an ancestral contribution identified in a recipient.

HapMap . . . International HapMap Project; Collaborative project to develop a haplotype map of the human genome. The project includes 11 global populations and a total of 1 397 individuals and was released in three phases [5]–[7].

HGDP . . . . . Human Genome Diversity Panel; Collaborative project to map human genetic diversity for disease discovery, to encourage the study of human genetics and to understand how genetic variation is formed. The project included 52 indigenous populations from across the world and 1 062 individuals [8]

Painting profile A vector of values representing the total genome length copied or the number of haplotype chunks copied from each of a possible list of donors. Values output by CHROMOPAINTER.

POPRES . . . Publicly available resource genotyped with a commercially available genome-wide 500,000 single-nucleotide polymorphism panel (Affymetrix 500K). This project includes nearly 6,000 subjects from various disease control groups of African-American, East Asian, South Asian, Mexican, and European origin. [9].

Race . . . . . "Each of the major divisions of humankind, having distinct physical characteristics." (Oxford Dictionary Online). More specifically, a sub-population of a species with a distinct set of phenotypic and genotypic frequencies which differ from other sub-populations [10], [11]. The most regular use in this thesis: A population grouped together by their perceived 'otherness' and assumed similarity.

Recipient . . . An individual or population which has received an ancestral contribution from a donor.

SAHGP . . . . Southern African Human Genome Programme; Pilot study of deep whole-genome sequencing of 24 Southern African individuals [12].

SGDP . . . . . Simon's Genome Diversity Project; high-quality genome data from 300 individuals from 142 diverse populations spanning much of human genetic, linguistic, and cultural variation [13].

SGVP . . . . . Singapore Genome Variation Project; Publicly available resource of 1.6 million single nucleotide polymorphisms (SNPs) genotyped in 268 individuals from the Chinese, Malay, and Indian population groups in South-East Asia [14].

Source . . . . . An individual or population which has contributed an identified ancestry in a recipient.

1 Introduction

In this thesis I set out to detail the admixture dynamics which have affected the genetic history of the people of mixed descent in South Africa, the so-called "Coloured" people.

Gene flow among human populations is no longer viewed as an anomalous consequence of the age of discovery but is now known to have been common in human evolution. A thorough understanding of the socio-genetic consequences of admixture is important for developing a holistic understanding of our history, and for developing analytical and societal tools to alleviate public health burdens.

In the following section, Chapter 2, I introduce the background concepts necessary for understanding the current South African "Coloured" community and the admixture dynamics in the region.

I first discuss how ancient and recent developments in trade and civilisation have influenced our collective views on human diversity and ethnicity. Then, in Chapter 3, I provide an account of the developments in the molecular bioinformatic tools commonly applied in genetic anthropology and specifically those taking advantage of recombination and linkage disequilibrium. This is followed up with an introduction to the populations of the Southern African region, a simple account of the pre-history and the most recent political history. I discuss the Indian Ocean Slave trade and

the transition of the indigenous hunter-gatherer KhoeSan communities of Southern Africa to their present-day creolised identity as "Coloured" people which began with the arrival of the European settlers and their slaves to Southern Africa in the 1600s.

In Chapter 4 I use single nucleotide polymorphism data from a large set of South African Cape "Coloured" (SAC) people to characterise the possible genetic contributing sources. I employ linkage disequilibrium information between SNPs to identify sources for ancestry and to cluster individuals. I identified the genetic sources for the haplotypes within the SAC population. I further explore for possible cryptic structure within the SAC from the Cape and find little support, suggesting panmixia within the dataset.

In Chapter 5 I continue with the dataset from Chapter 4, employing linkage disequilibrium decay curves to date the admixture events that have brought together the identified sources. The temporal resolution allowed further discussion of the possible source populations.

In Chapter 6 I investigate the influence of geographic expansion and endogamy on a comprehensive dataset of the urban "Coloured" populations of South Africa. I characterise the trends in ancestry contributions beyond the Cape Colony. Specifically, I discuss the relationship of the Cape Town Cape Malay and the Griqua to that of the broader "Coloured" and Baster communities. Despite their shared ethno-genetic affinity as "Coloured" the new genetic data from the Cape Malay and Griqua highlight the influence of geography and socio-cultural norms on the admixture history across the South African "Coloured" communities.

Lastly in the concluding chapter 7, I discuss the results in context of what was established in previous work and the difficulty in characterising recent admixtures which have yet to 'settle' to a simple signal. I further highlight areas of research which are of interest for understanding the 'Coloured' community and the genetics of admixture more generally.

2
Literature Review

2.1 Notes on Identity Terminology

The terms used to refer to various people and communities globally and in Southern Africa have changed continuously, and sometimes abruptly, over the past 350 years. What has resulted is that sometimes terms are used inconsistently between sources. This is reflected in the literature on ethno-racial identities, in particular of the KhoeSan of Southern Africa. I have tried to use the most commonly used terminology to avoid confusion for the reader, but I have also tried to balance this with the acceptance of new developments in what are considered more respectful ways to address and identify communities. Unless specified differently I used the following definitions.

'Black' : I use the word in the South African political sense in that it would identify anyone not considered 'white', including Indian, "Coloured", Chinese and Bantu-speaking Africans.

'African' : While the word is often used to refer to indigenous Bantu-speaking communities in Southern Africa, I follow the lead of Adhikari [15] here to explicitly include the KhoeSan communities and their descendants (including the "Coloured" people of Southern Africa).

'Coloured' : A "Coloured" person is here considered to be a person who identifies as a "Coloured" person. In the South African context this is usually thought of as someone with mixed ancestry associated with the development of the Cape Colony circa 1652 [16] though the definition is lax and extremely subjective and variable. With young people of mixed ancestry, the terms bi-racial or multi-ethnic may be preferred over "Coloured". I use 'bi-racial', 'multi-ethnic' or 'of mixed descent' to refer to people of mixed descent who are not necessarily culturally or historically associated with the South African "Coloured" community. "Coloured" people active during the fall of Apartheid may prefer the term 'So-called Coloured' [17].

'KhoeSan' : The term is used to jointly refer to the indigenous Khoikhoi who were pastoralists traditionally and indigenous San communities who were Hunter-gatherers traditionally. KhoeSan revivalism has thrown in many new questions regarding the way in which we do book-keeping of Southern Africa's diversity. Many people of the "Coloured" community, particularly among the young adults, have now taken to identifying as KhoeSan to emphasise their shared indigenous ancestry with the KhoeKhoe and San people of Southern Africa[15], [18]. I use the term specifically for those non-urban communities of predominant KhoeSan descent [19].

'White' : White people are here understood to be people of predominantly European descent though most often this identity is associated with 'the appearance of' predominantly European ancestry as understood by Southern African 'blacks'. For example, Arabs, Jews, British and the Afrikaners are all typically considered 'White', Indians, Sri-Lankans are not. This differs from other authors. For example, Howells [20] based 'white' on skull morphology, thus including East Africans while Brues [10] equated 'white' with western Eurasian.

'Afrikaner' : "Afrikaner" is commonly understood to mean the white descendants of the earliest Dutch settlers at the Cape Colony [21], [22]. Today approximately half of the white population of Southern Africa identifies as "Afrikaner" [23]. Afrikaner-dom is culturally characterised by the use of Afrikaans (Dutch-derived language) and membership to the Dutch Reformed Church (Nederduitse Gereformeerde Kerk; NGK) however this is not strictly observed [21]. The term "Afrikaanse" is sometimes

used to refer more broadly to the Afrikaans-speaking communities (i.e. including "Coloured" people) [16].

I recognise that all of these categories are subjective social constructs and do not have essentialist characteristics, including of a genetic nature.

2.2 Anthropology, Identity, Genomics

2.2.1 Understanding Human Diversity

The earliest human societies would have been small groups of closely related individuals. This is the case for contemporary hunter-gatherer communities who are often used to infer characteristics of past communities [10], [24]. Daily interactions, reproduction and family rearing would have been among culturally and biologically related people. These populations would have been sparsely spread across large areas. Communities would rarely venture further than what is necessary to forage and hunt, as exploration is costly and dangerous without a local 'push-factor' such as resource depletion or conflict [10]. Even when an adjacent area was unoccupied and with abundant food, a group's culture may not have prepared them to make use of the resources. For example, fishing communities on the coast may not have the knowledge of edible plants in the inland forests, making such habitats appear barren.

The neighbouring communities would then be related as an area is occupied by a gradual spill over from adjacent areas and the local knowledge is updated over several generations. On the whole, communities would rarely have encountered people who were greatly different from themselves and the understanding of differences among humans would be localised [10], [25].

This may have been true for recent hunter-gatherer communities and sedentary societies, but considering that there appears to have been multiple waves of range expansion and contraction in hominin evolution which resulted in repeated events of species range overlap [10], [20], [26], we should expect that sharing an environment was rather common. Human diversity may not have been so well graded that people should seldom have encountered a person starkly different from themselves.

There is growing evidence that early modern humans (EMH) and anatomically modern humans (AMH) would have co-existed alongside several other forms of humans for thousands of years [25], [27]–[29]. As recently as 9 - 12 Kya Nigerians in West Africa and Pygmies in Central Africa may have lived alongside and interbred with archaic forms [27], [30]. Beyond Africa, AMH, Neanderthals, Denisovans would have known of each other and possibly other species as well [13], [20], [31], [32]. Going further back to the early days of Hominin evolution during the "Hey-day" of great ape diversity in the Miocene-Pliocene, co-existence may have been the norm as there would have been a greater diversity of representatives on any lineage.

Without detailed written history of these early encounters, our collective experience of hominin diversity would have been lost to historic amnesia after a few generations or possibly distorted into legends and fables.

It was with the rise of transport innovations, trade, raiding and slavery that moving greater distances more often would have brought people into contact with the perception of 'otherness' more regularly [10], [25], [33]. Transport innovations such as the domestication of horses and cattle, the development of the sleighs and carts, and the invention of rafts would have allowed faster overland and over-water transport of people and goods. With these innovations people could trade, raid and enslave over greater distances [10].

With the development of written history in its various forms, records of encounters became available to subsequent generations such that people could be aware of communities beyond the immediate neighbours. Literature from Homer of 9th century BCE Greece describes Ethiopians as a mystical people to the South, showing that Africans were rarely encountered in Europe at the time. Images and pottery from 5 - 6th century BCE Egypt and Greece depict sub-Saharan Africans, indicating some movement between the Mediterranean and below the Sahara [10], [26], while Herodotus (435 BCE) discusses Indians with dark skin like Ethiopians indicating knowledge of South Asia [10]. In 6th century BCE, the Persian empire would have already encountered a range of ethnicities as they conquered from Egypt to the Himalayas [33]. Eastern and Central Europe encountered the Huns in 450s

CE, and a diverse collection of communities were responsible for Attila's final defeat [33]. With the rise of Islam in the 7th century, slaving and trade brought people from across Southern Europe, Arabia, Central Asia and Northern and Eastern Africa into the same cities [33]–[35]. With the development of the T'ang Dynasty in China, trade allowed Persian, Syrian, Arab and Mongolian merchants to inhabit Chinese cities in the 10th century CE [10], [33], and movement between East Asia and South Asia was already well established [33].

Over the many iterations, empires expanded further afield and more frequently, bringing home foreign people and "exporting" equally foreign people. Thus, over the global shifts in empires - Persians, Greeks, Romans, Huns, Mongols, Chinese, Islam, Spanish, French, Portuguese, Dutch, and British - we've collectively developed an understanding of the diversity of people across the world.

Studying Human Diversity

The earliest attempts to understand human diversity were based on the characteristics most accessible; the outward physical appearance and variations in culture and behaviour such as religion and language [10], [20], [26].

This is reflected in the accounts of travellers and traders and in the art and literature during the rise of sedentary colonial societies such as Mesopotamia, Egypt, Athens and the Arab Caliphates during the last 5 Ky [10], [26], [33], [34], [36]. While the earliest writers seem to show limited interest in ethnic differences compared to the fascination that arose in the age of enlightenment [10], [26], some attempts were made to explain the various forms of man. The Greek name *Ethiopian* implies that the Greeks considered Ethiopians to have been 'scorched' by the sun [10]. Near Easterners had a similar view; Northern Europeans were pale because the temperature at which they carried children in the womb was too low, while for Africans the temperature was too hot [10], [34].

There was an upturn in the understanding of human variation with the shift of learning and enquiry among elites in the Mediterranean and Middle East toward western Europe [33]. A plethora of information about peoples from around the world

was made available during the 'Age of Exploration' (1400 - 1700s) as European powers began traversing the world for trade and conquest [10], [20], [33]. As Western Europe was moving from relative isolation and stagnation, news of the new peoples encountered shocked those back home [33]. Unlike the Mediterranean where people were already acquainted with both Africans and Eastern Asians [10], [33], [34], North and Western Europe had had far less exposure to people beyond the region. Roman occupation had done little to introduce foreign figures to the West as compared to the Mediterranean and shorter-lived events such as the Huns and Viking's encounter with 'Eskimos' flickered out of memory [10]. Philosophers and the church frantically tried to align the new observations of human diversity with the Holy Scriptures. A plethora of theories were produced to explain inconsistencies, for example, to explain how there could be such apparent long history for various communities if the global floods of Noah occurred less than 3,000 years ago [10].

Importantly, the shift of commercial and military power to Europe coincided with the beginnings of a formalised scientific method which emerged in the 18th century during the 'Age of Enlightenment' [33]. This led to many areas of study to be systematised, including that of human origins [10].

The works on the classification of humans by Linnaeus (1756) and Blumenbach (1776) and the acceptance of language families (1786) [10], [26] are considered major milestones in the scientific study of man. These early classifications continued to use easily observable characteristics such as colour variation in eyes, hair and skin, differences in hair and skin texture, skull dimensions, body shape and stature, language, religion and customs. Human diversity and origin was a topic initially riddled with speculation and often met with disinterest [10] but soon became a popular research focus for naturalists but remained quite full of speculation [20]. Blumenbach's proposed 'colour' system generalised five major races of man based on outward characteristics. The work gained notoriety among anthropologists and later among white-supremacy advocates, a connection for which the classification is still known [10]. The five groups recognised were 'popularly' dubbed "White", "Black", "Yellow", "Red" and "Brown" and corresponds more or

less to the later iterations of classification: "Caucasoid", "Negroid", "Mongoloid", "American Mongoloid/Amerindian", with "Brown" being divided into a collection of other groups which were not readily grouped (see Figure 2.1 and [10], [20], [37]).

Subsequent classifications provided labels for isolated communities who were distinct from the major groups identified. For example, the KhoeSan language communities were grouped as "Capoid", the Australian and Papuan communities as "Australoid", and Negrito and Pygmies were considered separately [20]. Still today many anthropologists argue that human variation can be divided into no fewer than three groups (~Mongoloid, Caucasoid and Negroid) but this has been significantly challenged by recent work, particularly genetic work, demonstrating the absence of any clear discontinuities in defining features and thus the arbitrary nature of these divisions [38]–[40].

The plasticity of many of the features used in early classifications has come under scrutiny. Even non-cultural traits such as limb-length, skin tone and immunological response (e.g. adaptations to malaria) may be recent enough not to be good indicators of long-standing divergences [20].

There was evidently a need for multiple lines of evidence for our origins and within each of these lines, a means of separating those plastic traits, disposed to recent influences, from those traits which suggest some meaningful historic or ancient connection between populations.

Molecular Genetic Anthropology

The study of cryptic variation as found in protein polymorphisms, immunological responses and behavioural traits is rather young and still younger is the study of the genetic code which underlies these many differences [26]. The earliest works in molecular genetic anthropology included documenting simple differences between populations based on easily detectable markers; the inheritance system of which was at the time poorly understood. Landsteiner in 1900 discovered the first genetic polymorphisms in the ABO blood group system [41]. Subsequently differences in protein polymorphism for the ABO system, Rh factor, Duffy factor and NMSU

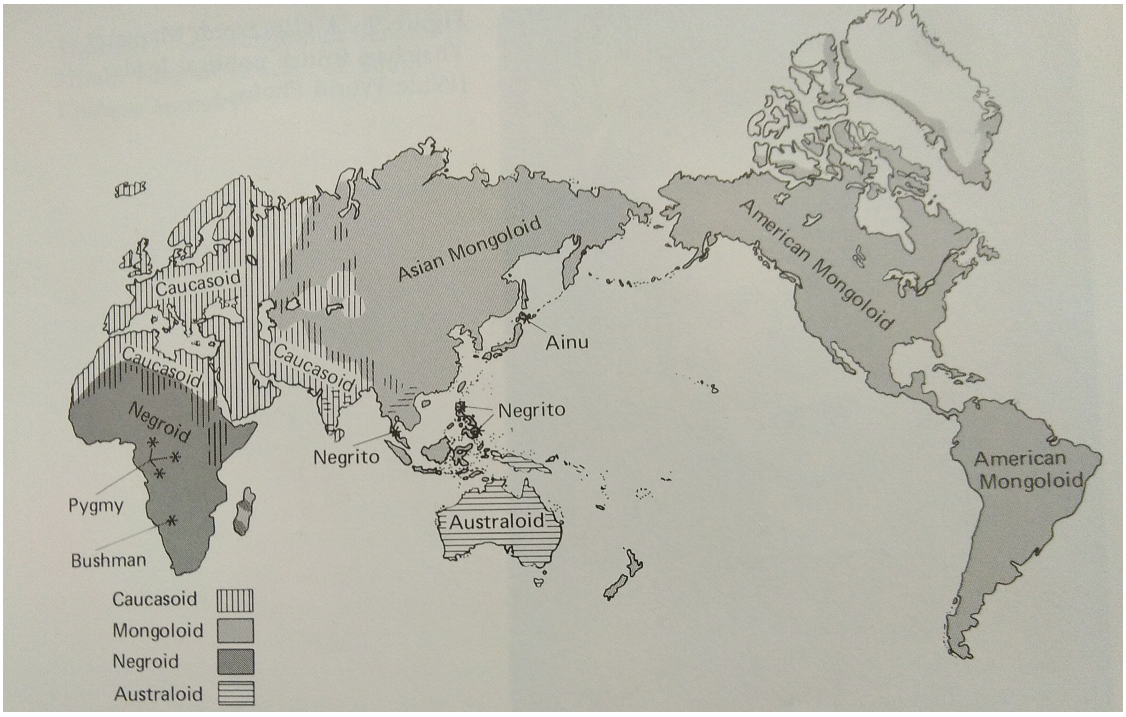


Figure 2.1: Anthropological classification of human 'super-group' diversity circa 1960 . Showing the 'major races' distribution across the world as of 1500s, taken from [10]

system became widely used to characterise the 'races of man' [10], [37], [41]. Work on these 'classic markers', their heritability and link to human ancestry sparked an interest in human polymorphism variation and global substructure.

The frequencies of different blood groups vary between populations and have allowed anthropologists to refine race classification using combinations of blood features. For example, the 'Caucasoids' have particularly low frequencies of Rh positive compared to other groups, in particular "Mongoloids" [42]. The Duffy factor, discovered in the 1950s, has the lowest frequency among Sub-Saharan Africans (~0 - 6%) and notably higher proportions in other groups, particularly Eastern Asians [42]. Combinations of these characteristics have been used to describe racial and geographic relationships between populations [10], [42]. Phylogenetic analyses have produced proposed relationships between human populations which have been remarkably congruent with subsequent investigations [43]. Despite their usefulness, they only provide some indication of relationship between groups but are not more forthcoming in explaining how the patterns came to be.

Direct analyses of DNA were only possible after theoretical advances on the structure of DNA and biochemistry of DNA replication were made. Such advances included the description of the double helix structure of DNA in 1953 [44], and the discovery of the DNA polymerase enzyme in 1955 [45].

Early genetic research was heavily restricted by the available technology and the cost of finding and profiling relevant gene regions for a sufficiently large number of samples to make the results generalizable [46], [47]. A series of landmark advances massively improved the efficiency of such work and, with some time, reduced the cost of performing genetic research.

The development of the Sanger sequencing (chain-termination) process in 1977 and the subsequent modifications to include florescent-labelled di-deoxynucleotide-triphosphates instead of radioisotopes, allowed for the detection of nucleotide-level sequence variation in a high-throughput and automated system [26]. This led the way to the production of whole genome sequences for various species, including the first human genome sequence [48].

The invention of Polymerase Chain Reaction (PCR) in 1985 provided a massive improvement in the efficiency and reliability of DNA detection [46]. It allowed hundreds of samples to be processed in a reasonable space of time and from tiny amounts of DNA (less than ~0.1μg) [26]. This allowed DNA analysis from material with low DNA quantity thus extending the commercial and scientific interest in DNA amplification (e.g. use in crime scenes, use on fossil material, profiling single-cell variation), driving cost reduction for such techniques.

It became possible to measure variation at specific loci by designing oligonucleotide primers for the PCR process to amplify these loci, thus targeting genes of interest.

The PCR process allowed DNA analyses to take on new forms of quantification too, including measuring template amplification in real-time, allowing, for example, the quantification of the frequency of different variants in a sample or of transcription products from an experiment [26]. In an effort to identify variants of interest for medical research, methods were developed to profile specific single nucleotide

polymorphisms (SNPs) as identified from previous genome sequences. Initially, methods using restriction enzymes to identify SNPs, such as PCR-RFLP (restriction fragment length polymorphism), were very limiting in their low through-put nature, allowing only a few SNPs to be examined. Today, profiling with SNPs has been notably more cost effective than sequencing when testing known variants and the most recent technology using micro-array based "SNP-chips" has allowed the profiling of comprehensive datasets involving millions of SNPs. Here DNA undergoes whole genome amplification, the products are fixed to an array of silica beads annealing to a specific oligonucleotide, targeting a specific SNP. Dye-terminated amplification and florescence detections are used to detect the alleles present [26].

Recent iterations of technological advances include the development of next generation sequencing which involves the nebulisation of DNA (shearing into ~200bp fragments) and adhesion of fragments to a solid surface, amplification of fragments to produce clusters of identical strands and finally massively paralleled simultaneous shotgun sequencing [26]. Third generation sequencing by-passes the need for library preparation entirely, reducing ascertainment bias [26].

Continual technological advancements and commercial interest in genetics has produced an unprecedented amount of data, much of which is available freely and publicly. Research teams across the globe are making continual use of this and undoubtedly the geneticists toolset will become an increasingly powerful tool in anthropology and archaeology.

2.2.2 Using Molecular Tools to Characterise Ancestry

Mendelian Inheritance

The work of the Augustinian friar, Gregor Mendel, is a key tier to the modern synthesis of evolutionary theory [49], [50]. His work describes the passing of traits from parent pea plants to their offspring in units of inheritance [11], [49], [51]. The discussion of units of inheritance was a needed complement to the work on the evolution of organisms as formulated by Charles Darwin and Alfred Wallace of the same period (c. 1858) [50]. The "units of inheritance" Gregor identified are now

recognised as "genes" and are seen as being the material (partially) responsible for phenotypic variation [11], [50], [52]. The identification of the units of inheritance allowed for the mathematisation of evolutionary theory in the 20$^{\text{th}}$century under the modern synthesis of evolutionary theory, and this has been arguably the driving force behind the relatively rapid uptake within the biological science community [50], [52].

Important aspects of Mendel's work included the recognition that only half of the hereditary factors were passed on to the offspring such that during gamete formation, gametes are equaly likely to be formed from either of the genotypes (alleles) present in the parent organism [49]. This principle of segregation describes how the organism's ploidy level, i.e. the number of genomes per individual in subsequent generations, is maintained. Furthermore, Mendel's principle of independent assortment postulates that alleles at different loci segregate independently of each other [49]. Secondly, traits are passed intact and not blended, hence referred to as 'units' of inheritance, and thirdly, some traits are dominant over others. These were postulated in Mendel's principle of dominance [49]. Based on these, Mendel could identify an expected ratio of phenotypes among the offspring, given the genotypes of the parents [11].

The work and principles provided the fundamental framework needed for the later founding and development of genomic analyses including the establishment of the Hardy-Weinberg equilibrium - that genotype frequencies are maintained in a population in the absence of evolutionary forces, and the predictions for the effects of such forces; drift, natural selection, mutation, founder effects, bottlenecks etc.

Tracing Ancestry through Lineages

The unit nature of gene inheritance provides a way with which we can measure differences between populations. Trait variation is common across organisms and populations [11]. Describing similarity or differences between populations based on shared traits allows us to develop a sense of shared ancestry. Given that all living organisms should share a common ancestor and that the sole evolutionary force generating novel alleles is random mutation, shared genetic traits will be an indication of shared ancestry. Homoplasy, or shared traits evolved independently,

will have a relatively small impact on the inferences when sufficient traits are examined. This can be said because mutations are infrequent and thus mutations resulting in co-incidental shared traits are even more unlikely [11]. Markers such as sequence data, single nucleotide polymorphisms, small tandem repeats etc. can be measured and the differences quantified. For example, by considering the proportion of sites across sequences which are variable (segregating sites) or the frequency profile for alleles at a set of loci in populations [11], [26]. Statistics such as F_{ST}, which uses heterozygosity measures, provide an index for considering the apportionment of variation across a set of populations and levels of organisation [11], [26], providing an indication of relative differences among populations [11], [39].

Added resolutions can be gained by considering the frequencies of haplotypes instead of individual variants and their respective allele frequencies. A haplotype is a permutation of the available alleles for a set of loci. The expected haplotype frequencies can be estimated from the product of the allele frequencies [11]. However, there are a greater number of possible gametes for a multilocus haplotype than a system of loci considered independently. For example, with n possible alleles at two loci, there are $2(n-1)$ independent allele frequencies with which one could base relatedness estimates [11]. However, $n^2 - 1$ independent gamete frequencies exists when considering multilocus haplotypes for the same system [11].

When loci are physically linked, i.e. they are in relatively close physical proximity on the same chromosome, there is a correlation in the transmission probabilities (discussed further below). When the linkage is particularly high, haplotypes can serve to track lineages of descent. This is possible because mutation is the predominant force introducing variation [11] and it becomes much less likely that a shared haplotype is shared by chance rather than descent. Such haplotypes are grouped by genetic similarity based on marker characteristics which indicate shared ancestry assuming a relatively recent divergence of the human species [11], [26]. By this we can measure genetic diversity, distinguish relative age of lineages and we can begin to describe scenarios of splits between groups and bottlenecks.

Human populations can be grouped by shared collections of haplotypes and relatedness interpreted in conjuction to patterns across groups in culture, geography and languages [46]. Uniparental markers, and haplotypes in particular, are a common tool in genetic research as they are relatively simple systems to model and understand [11], [53]–[56]. Uniparental markers are inherited solely through the paternal or maternal line, respectively, allowing one to trace sex-specific events in evolution [53]. Importantly, the high linkage between loci means that variation is primarily introduced by mutation and haplotypes reflect lineages with datable divergence times.

The mitochondrial DNA (mtDNA) is most often used to trace the female lineage [49], [54] and is a particularly useful tool as it is housed in the mitochondrial organelle which is universally present in eukaryotes and replicates independently of chromosomal DNA [49]. The mtDNA is inherited solely through the maternal lineage and has high levels of polymorphism, no recombination and a high copy number within some tissues, such as muscle and liver tissue [26], [54], [57]. It is thus easy to investigate and is often the first locus considered.

The Y-chromosome provides a relatively simple system for reconstructing an evolutionary phylogeny of males and is complementary to the mtDNA system [53], [58]. In mammals the X-chromosome is heterogametic and the presence of the reduced form, i.e. the Y-chromosome, is male-determining. Thus, the Y-chromosome is haploid in males and absent in females. It has few genes, does not recombine (mostly) and is not essential for survival as females don't have a copy [53].

These markers have been used to support the 'Out-of-Africa' hypothesis for a recent origin of man as a challenge to the independent, regional origins from *Homo erectus* [2], [20], [53], [59], [60]. Teams have used these markers to study genealogy, kinship and paternity [22], [61], possible climatic refugia, historic migration and recent dispersals, and biases in gene flow [55], [62]–[66].

The use of uniparental markers however limits our understanding to non-recombinant ancestry and thus a bifurcating relationship between haplotypes. These markers can only reflect a single lineage for each individual included and thus a single

ancestor for any number of generations back [11], [53], [67]. Furthermore, being sex-linked, such markers are heavily influenced by demographic and genetic events which have different consequences for each sex [e.g. 53], [63], [65]. Over a number of generations and following several different scenarios of demographic change, it can become unclear what exactly a proportion of haplogroups in a population reflects.

Recombination and Linkage Disequilibrium

An important aspect of inheritance not covered by Mendel's laws of segregation and independent assortment was that not all traits show independence.

In the system which Mendel had studied, loci were far enough apart in the genome to be inherited independently of each other giving the impression that traits are unlinked.

When loci are on separate chromosomes or physically distant on a single chromosome, at each generation there is a 0.5 probability of the two loci being co-inherited, this is due to independent assortment of chromosomes into gametes during anaphase of cell division and due to homologous recombination during meiosis [11], [68].

Between loci on a chromosome, independence is introduced via recombination [49]. Following double stranded breaks in the DNA helix, either through enzymatic action or radiation, the DNA repair process is initiated and this involves two homologous chromosomes forming a junction for recombination [51]. During mitosis this process functions primarily as a means to repair double strand breaks in DNA, contributing to the preservation of the genome integrity [51], [69]. During meiosis, homologous chromosomes exchange large tracts of genetic material (everything past the break point), creating recombinant chromosomes or non-parental chromosomes [68]. The breaking up of haplotypes introduces independence between loci as a function of the physical distance between them [11], [68], [70].

The term gametic phase disequilibrium refers to the state in which alleles at loci of interest approach non-independence in their transmission probabilities, for example, some correlation in transmission of an allele at locus A with an allele at

locus B [11]. Linkage disequilibrium (LD) is often used to mean gametic phase disequilibrium but it is misleading as physical linkage between loci is not a pre-requisite for LD [11]. Gene conversion, for example, can introduce LD between alleles at two physically unlinked loci [51].Gene conversion results in the transformation of small segments (~300bp) of one haplotype into a copy of that present on the homologous chromosome, typically near the recombination point [51], [68]. While such conversion are prevalent, they are difficult to detect because of the size of the effected fragments [68]. Generally the amount of LD observed between loci is a consequence of the interactions of selection, mutation, drift, non-random mating, gene flow and recombination while in closely linked loci the primary forces are recombination and genetic drift [11], [69].

Formally, LD is measured as the deviation of the haplotype/gamete frequency from the expected frequency [11], [26]. Considering a two loci (letters A and B), two allele system (subscripts 1 and 2), the deviation is estimated as $D = x_{11} - q_1 p_1$, where x_{11} is the observed haplotye frequency for A_1B_1 and q_1 and p_1 are the allele frequencies for A_1 and B_1, respectively.

The value of D ranges from 0.25 to -0.25 when all alleles have equal frequencies but the values are notably smaller when this is not the case [11]. As such, the value is normalised to D' by dividing by the maximum possible D value for a given set of allele frequencies, thus setting the range of possible values to -1 to 1 [11].

The rate of recombination is measured as a genetic map distance (e.g. for closely linked loci, percentage recombination expected between two loci as estimated through pedigree studies) [11]. The recombination rates are often species- and sex-specific but may also vary with chromosome size and the position within the chromosome. In humans, females have a 60% higher autosomal recombination rate [71], shorter chromosomes have higher recombination rates than longer ones and as much as 95% of recombination is observed at hotspots with exons having low rates [72]. Some chromosomes, particularly 21 and 22 in humans, can have recombination rates twice that of other chromosomes [71].

The importance of recombination in evolution is that during meiosis homologous recombination produces genetic variation using existing genetic material by creating new allele combinations [11], [26]. This production of novel phenotypes is done at a greater rate than mutations could produce [11]. Another benefit of this process is that, in contrast to sexually reproducing organisms, asexual organisms or non-recombining regions of a genome accumulate deleterious mutations, a process referred to as Muller's ratchet [11], [26]. Without recombination, purifying selection on deleterious mutations would necessarily purge all linked genes as well even when they are beneficial. Recombination allows genes to 'escape' selection on neighbouring loci and thus the collective gene pool of a population can be purged of bad mutations without necessitating the loss of useful, otherwise linked, genes. Similarly, recombination allows adaptations which have evolved independently to "meet" in a single genome. Without recombination, this would require the independent evolution of one trait followed by the other making it increasingly unlikely [11], [26] and slowing the rate of adaptation.

Autosomal Markers and Admixture Linkage Disequilibrium

In human population genomics, the value of understanding recombination events has come to the fore in its use to identify genomic fragments identical-by-descent within populations and thus quantifying shared ancestry, inbreeding and for use in local ancestry characterisation [73], [74].

The use of autosomal markers (i.e. non sex-linked chromosomes) in evolutionary studies has the advantage of included haplotypes from more than a single ancestor as was introduced onto recombinant chromosomes by recombination. As such, autosomal markers are less susceptible to genetic drift due to the larger effective population size [11] and they do a better job at 'recording' past encounters between populations. For example, the absence of Neanderthal mitochondria in Humans supported the argument against admixture, but prevalence of autosomal ancestry shows the contrary [57], [75]. By contrasting across populations the observed autosomal and uni-parental markers, we can begin to understand how sex-biased

demographic processes can affect the genome and how the sole use of uni-parental markers has biased our understanding of population genetics [e.g. 76], [77].

As with uni-parental markers, there is added value in investigating haplotypes with physical linkage. Linkage disequilibrium is thought to be negligible beyond a few hundred kilobases distance between loci due to the drift-recombination equilibrium [5]. However, for most populations longer range LD can be detected related to ancestral mixing between previously separated groups and this LD is refered to as admixture LD (ALD) [78], [79]. Within each ancestral population there may indeed be independence between most pairs of loci due to the drift-recombination equilibrium but, due to allele frequency divergences between the admixing sources, an admixed group, even when well mixed, will have detectable LD between loci extending beyond a few hundred kilobases [78]. With generations of recombination breaking down ancestral haplotypes, each chromosome thus becomes a patchwork of haplotypes, each haplotype representing a stretch of DNA inherited from an ancestor [80], [81].

This is also used most recently in characterising the proportion of the genome derived from specific populations and estimating a date for admixture events [78], [82]–[84]. The use of this information has produced intriguing details on the evolution of populations across the globe by providing information directly comparable to archaeological information (i.e. who moved and when) [e.g. 82], [85]–[87].

With the reduced cost of genotyping autosomal DNA, several ambitious projects have successfully contributed to the collection of data from thousands of humans across the world, representing several hundred ethno-linguistic groups with coverage ranging from several thousand SNPs to whole genome sequences [4], [6], [14], [88]–[92]. This growth has fostered better representation of the variation within an individual's genome and allowed datasets to grow to include diverse populations. Ultimately this has improved robustness and precision of research and the growing data sizes has created a drive for the next generation of analyses [47].

Developments have prompted the re-evaluations of past results which were based on far fewer markers. The new data has corroborated preceding archaeological,

linguistic and genetic work, including confirming trends in genetic variation and linkage disequilibrium (LD) across the planet [6], [68], and further supported the out-of-Africa bottleneck [3], [47], [90].

With the added resolution, investigations have gone several steps further in providing support for multiple migrations out of Africa and for a "Back-to-Africa" migration [26], [93]. Most impressive is the successful amplification of ancient DNA from humans and extinct hominins which has provided for the first time clear evidence for cross-species admixture and accounts who are the closest living relatives to the earliest inhabitants of some regions [32], [57], [94]. Creative analyses have also shown admixture between the earliest hominins and the Pan-hominin ancestor [95] and identified ghost ancestors [27]. The resolution has even allowed the discovery of previously unknown ancestral source populations for which there was previously only suggestive evidence e.g. the ancestral South and North Indian populations [85], [96] and the identification of extinct ancient human populations where fossils are near absent such as with the Denisovans [32], or non-existent such as in Central Africa [27].

With this data, researchers have traced the source populations for people with known genetic admixture but for which details are poorly recorded [97]–[101], providing some information on the historic scenarios of social structure, marital taboos, gene-flow biases and migration routes [19], [63], [100], [101]. What remains of many indigenous populations today is sometimes solely represented by people of mixed descent. Genomics allows us to retrieve information from the ancestral components inherited from the indigenous people at high enough resolution to provide a glimpse of earlier human history that would otherwise have been lost. For example, with the /Xam-speaking people of Southern Africa [19], [99] and the Andamanese of South Asia [85].

Genomics has become an increasingly important tool in reconstructing human past, both recent and ancient. This is particularly so for pre-literate societies and in the instances where little or degraded archaeological artefacts or fossils hinder understanding [26], [32]. In conjunction with other fields such as physical

anthropology, linguistics and archaeology, genomics and genetics have provided corroborative and sometimes unexpected insight into human history beyond what has been possible before.

2.3 Peopling of Southern Africa

In this thesis I specifically focus on the historic admixture dynamics for the slave and serf community of the Cape Colony as these people would have contributed to the heritage of the present-day "Coloured" community. As such it is necessary to review the history of the region in the run up to the establishment of the Cape Colony and the subsequent politics which shaped the people and identity as encountered today.

To develop the backdrop to understanding the admixture dynamics of the Cape "Coloured" community I, below, introduce the populations of the region briefly and then provide a simple account of the pre-history, followed by an account of the Indian Ocean Slave trade and the most recent historic developments in South African politics.

2.3.1 Contemporary People of South Africa

Written history for Southern Africa extends back ~550 years to the arrival of the first European explorers at the Cape of Good Hope, but preliterate history for anatomically modern humans (AMH) extends back tens of thousands of years at least and for the genus *Homo*, millions of years [20], [46], [102], [103].

The Republic of South Africa today is a multi-ethnic country of 52 million people and an area of 1.22 million km^2 [23], [103], [104].

There are approximately 48 million Bantu-speaking Black Africans of various cultural affiliation. The most populous of whom are the amaZulu, amaXhosa, Bapedi and Batswana [105]. 'White' people are the second largest racial demographic at two million. The largest component (~50%) is the Afrikaner people who are purportedly descendants of earliest settlers at the Cape Colony. The second largest is the descendants of the English 1820 settlers [21], [103].

There are two million people of mixed ancestry who are referred to colloquially and formally, and who often self-identify as 'Coloured people' [16], [99], [105]. National Survey data does not accommodate for the explicit documenting of the earliest occupants of Southern Africa, the KhoeSan people. These are a collection of culturally, economically and linguistically diverse people [102], [106]–[108]. Presently most KhoeSan communities are small and fragmented or otherwise have been absorbed into the Bantu-speaking or "Coloured" communities or have adopted the "Coloured" identity [109]–[111].

The "Asian" population reaches just over one million and includes both South, South-East and East Asians [21], [103], [112].

Today there is on-going large-scale immigration to South Africa as the region receives refugees, labour migrants and opportunity-seekers from across the world, simultaneously the European population has been in decline due to emigration since the 1980's [23].

2.3.2 Pre-Colonial Southern Africa

The KhoeSan People

Southern African has been one of the suggested regions of origin for AMH [20], [113] and the indigenous populations who were well established circa 350 before present (BP) included some of the earliest divergences of contemporary AMH [28], [93], [113] supporting the importance of the region for the history of humanity.

"KhoeSan" refers to a diverse set of languages characterised by the use of several click consonants and the people who speak those languages. The KhoeSan languages are broadly divided into three major families; the Kx'a (geographically Northern), Taa (geographically Southern) and Khoe-Kwadi of which the members are debated [19], [108].

I hereafter refer to the groups of people by their respective language and use "KhoeSan" to refer to the collective people who speak the languages but to the exclusion of the Bantu-related groups who have adopted the language, e.g. Damara [102], [108].

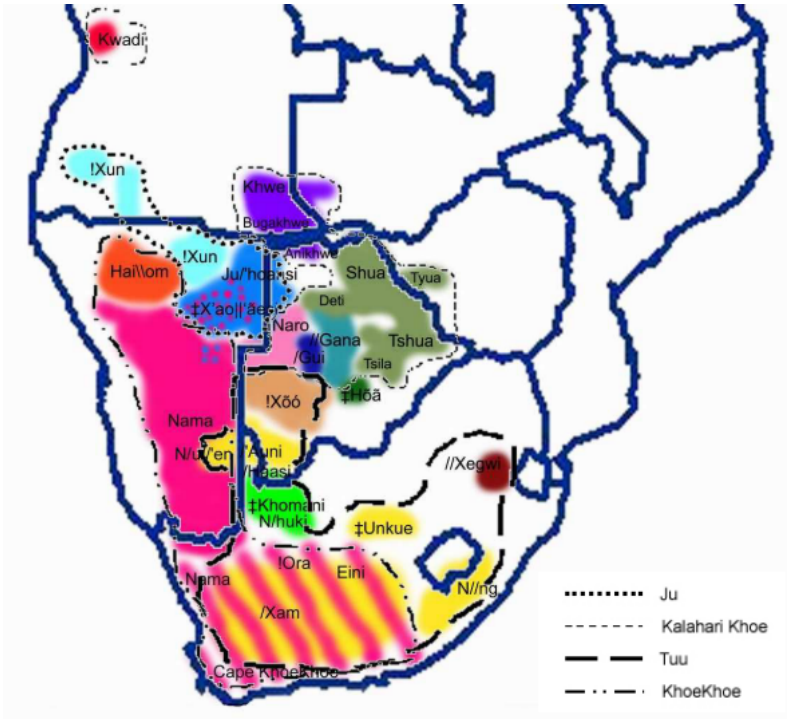


Figure 2.2: Historic distribution of the KhoeSan communities as summarised by Schlebusch [19]

Genetic diversity in the KhoeSan people is larger than in any other group of people despite their small current population sizes and recently reduced geographic occurrence. This indicates that they would have had the largest populations for the longest time in recent history [107], [114]. Of the contemporary humans, the KhoeSan possess some of the earliest diverging Y and mtDNA lineages with the split estimated around 100 - 260 Kya [25], [28], [94], coinciding with the earliest evidence of anatomically human fossils. Recent ancient DNA evidence however, suggest the oldest divergence may be between non-KhoeSan West Africans and East Africans [94] in line with a multi-regional origin for AMH. The forest hunter-gatherer pygmies of central Africa are the most closely related population to the Southern African KhoeSan both genetically and linguistically [27], [113] and fossil evidence suggest that there once was a continuous distribution along the eastern African region which was later interrupted by the Bantu expansion, as observed in the ancestry of ancient samples from Malawi circa 700 BCE [19], [20], [27], [94].

Overall, the high diversity and earliest divergences observed for these populations

suggest that an important stage in the earliest steps of modern human evolution are recorded in their genomes.

Between the KhoeSan, some population splits have been estimated to 20 - 45 Kya [93], [108], [114], [115] coinciding with glacial - inter-glacial migration events for various human populations [116]. The traditional KhoeSan societies would have been hunter-gatherers (HG) and pastoralists, consisting of small family groups of a few tens of people [19], [24], [102], [114], [117]. However, pastoralism was likely introduced by the Khoe-Kwadi speaking ancestors to the Nama and Khoikhoi of Southern Africa and this is associated with a fairly young (~1.9 - 2.2 Kya) migration from Eastern Africa [108], [118]–[120].

The genetic distinctions between KhoeSan groups do not parallel linguistic and subsistence distinctions. Current geographical distributions, however, correlates well with ancestral components, with evidence for a Northern ancestral component predominant in the Juu|'hoansi and !Xun, a Central genetic component associated with the ≠Hoan, G|ui and G||ana and a Southern component represented in the Nama, Karretjie and ≠Khomani [108], [115], [117].

Several populations show a geographic-genetic mismatch. The Kwe (Khoe-Kwadi speakers) and Naro show mixed cluster assignment which Montinaro, Busby, Gonzalez-santos, *et al.* [108] found to suggest ancestral allele sharing rather than recent admixture. The divergence between the Kwe and other KhoeSan based on KhoeSan specific ancestry did not suggest mixed ancestry despite allele sharing with all Northern, Central and Southern genetic groups.

The recent decline of the KhoeSan populations is associated with the arrival of what are today sedentary societies. Firstly, with the arrival of the Bantu agro-pastoralists in the east of the country ~2,000 BP [64], [102] and more markedly with the European settlers ~350 BP, the KhoeSan populations were drastically reduced in number over a relatively short time. What we observe today are remnants of these communities [99], [102], [121].

While the traditional lifestyle has 'vanished' for many KhoeSan groups, the genetic legacy of the KhoeSan ancestry has persisted in the "Coloured" people of Namibia and South Africa who are their descendants [110].

The East African Pastoralists and the Bantu Expansion

Beginning ~5,000 BP the proto-Bantu language communities within the Nigeria - Cameroon region began expanding east and southward into the rest of sub-Saharan Africa, a moving front which arrived in Southern Africa 2,000 - 1,200 BP [20], [64], [102], [103], [122]. This was one of the largest and most rapid demographic movement in the pre-history of the human species and resulted in the spread of genes, language, and culture across sub-Saharan Africa [20], [102], [123], [124]. The expansion was likely related to the development of agriculture and iron technology in the region of origin [20], [102]. Today the Bantu-language group is the largest of the African languages both geographically and by number of speakers [56], [102].

This demographic and cultural event resulted in what was essentially the loss of ancient Y-chromosome diversity from the indigenous hunter-gatherers in many regions and the replacement with younger lineages associated with the Bantu expansion [123]. In contrast many mtDNA lineages with deep divergences from the HG people have persisted [123], [124], reflecting the sex bias in assimilation and reproduction.

After the first phase involving a southward diffusion through the central African rainforests, there were at least two major dispersal routes. Broadly speaking, one route was along the east and one along the west of Africa [123], [125]. The expansion which moved along the east coast of Africa was relatively rapid, such that the amount of gene flow with indigenous people at the time was low as is seen in the absence of a unique HG signal from the SE-Bantu populations of Mozambique [64], [94], [125].

Contemporary genetic patterns of the Bantu-speakers in the southern parts of the continent do show signs of ancient admixture with populations of KhoeSan speakers from the south, including the expected sex-biased in gene flow [123], [124], [126]. The change from a moving to a static frontier would have allowed higher

rates of admixture and assimilation of the KhoeSan [64], [102]. As much as 20% of the genotyped SNPs in South African Bantu-speakers (e.g. Zulu, Xhosa, Sotho) reflects KhoeSan ancestry [4], [12], [99], [108].

With respects to East Africa, there was another important ancient event. Sometime between 2.7 - 3.3 Kya there was an admixture event in East African populations which introduced a genetic component most closely related to that of West Eurasians (European and Middle Eastern), possibly from southern Arabia [93], [102], [118]. The contribution is seen in several Ethiopian populations and can represent as much as 50% of the genetic components present [122], [127]. A similar admixture is detected between 900 - 1 800 ya in the Southern African Khoe-Kwadi language speakers, which is also associated with the presence of lactase persistence in the KhoeSan pastoralists [118], [120]. This contribution was likely the result of a pastoralist migration south from East Africa with gene flow between the arrivals and the autochthonous groups coupled with the spread of pastoralism. This event predates the arrival of the Bantu-speakers to Eastern and Southern Africa and thus not all groups in these areas show evidence of both genetic contributions [4], [94], [122].

2.3.3 The Indian Ocean Slave Trade

Slavery

Historian Francois Valentijn referred to slavery as "den oudsten handel in de wereld" (The oldest trade in the world) [35], [128]. It is not possible to definitively say when slavery would have first appeared in human society, but it is likely an early tool in the human repertoire, as the forerunners to slavery, i.e. raiding and abduction are early activities.

Slavery is often thought of as the opposite of 'freedom' though it is really a point along a gradient of coerced labour, being characterised in particular by violent force [129]. Cultural, ideological and economic pressures still play a major role in forcing slaves to work, just as they do today among 'free labourers' [128]. Slaves lacked autonomy and control over their lives. This included physical reproduction, sexuality, social ties, kinship and inheritance [35]. Patterson [130, pg 13] identified three

major characteristics of enslavement, the first two of which are truly characteristic of slavery [35]. Firstly, there is a perpetual domination which is enforcement by violence. Secondly, there is a natal alienation. That is to say the loss of all claims at birth, such as family ties, property, identity and religion. All protection and privilege provided by ancestors cannot be claimed, likewise, the enslaved cannot provide protection and privilege to their descendants.

These two characteristics distinguish slavery from other forms of involuntary servitude such as serfdom, pawns, indentured labourers etc. [130].

The agricultural revolution was clearly an important upturn in the history of slavery as the need to cheap labour increased. The Nieboer-Domar hypothesis links abundant free land to the imposition of slavery, as landowners need to coerce labourers to work when the option exists for labours to take up land independently [35], [131]. Similarly, large-scale labour systems necessitate the development of state-systems to protect the interest of slave owners by providing the force and coordination (physical, religious and legal) needed to acquire, discipline and coerce slaves [35]. Slavery has many non-commercial or minor-commercial incentives such as use for domestic, administrative, military, sacrificial and sexual services [33], [35], [128], [132], these of course don't require large sedentary societies.

The Dutch Slave Trade

The Dutch came to dominate the Indian Ocean slave trade for nearly 200 years during the 17-18th century [35], [133], [134]. Dutch business in the Indian Ocean basin was largely orchestrated by the *Vereenigde Oost-Indische Compagnie* (*VOC*) which had been granted authority to act on behalf of the Dutch Republic.

Unlike the Atlantic Ocean slave trade, the 17th century Indian Ocean slave trade was built on the back of an established systems of slavery and coerced labour in the region established before 3,000 BP [33], [35], [129].

As long-distance trade in luxury goods became established along the silk route and the coastal networks, the demand for slavery picked up, both to mass produce the luxury goods cheaply and as a luxury item of its own right [34], [35]. Coinciding

with the rise of Islam from the Arabian Peninsula in the 7th century, the slave trade became increasingly long-distance, systematised and better recorded.

As Islam prohibited enslavement within the Islamic kingdom it became necessary to seek out infidels beyond [34], [35] thus the majority of the Islamic slaves were collected from Southern and Eastern Europe, Central Asia and Sub-Saharan Africa [34]. As many as 6.4 million African slaves alone may have left for Islamic societies between 650 CE and 1500 CE [pg 24 135]). Islamic enslavement of groups further east was infrequent prior to the 1500s [34] and the islamisation of South East Asia was not a colonial conquest but a more voluntary conversion [36].

Slavery under Dutch domination and under the preceding Portuguese experienced an upturn tied to the increased over-ocean transport efficiency [35]. This allowed the newly arriving Europeans to create a network of direct and extensive connections between the far ends of the pacific rim. There was thus a massive increase in the numbers of people transported, the distances covered, and the demand for slaves.

The Enslaved

The focus of research on the Indian Ocean slave trade has almost systematically shifted away from the slave trade itself and toward urban and general trade history, economic governance or the more recent forms of coerced labour. This is in part due to the multi-player nature of the Indian Ocean trade and that a large portion of the trade was private and thus poorly documented [35], [133], making research cumbersome. Furthermore, Slavery contributed a tiny fraction of the *VOC*'s annual income even at its height of importance during the 18th century, reducing the incentive for research [35]. Lastly, a good deal of the early research was written by sympathisers to the trade who held that the trade was a domestic, 'benign' and 'paternalistic' form compared to the Atlantic Ocean trade [35]. This view dampened the focus on slavery, the enslaved and its social consequences.

Over the past 10 years, a body of work focused on the enslaved has begun to emerge. The summary of the influence of the *VOC* below is largely based on the work by Vink [35], [136].

The *VOC* controlled three interlocking circuits within the Indian Ocean Basin [see 35]. The East African circuit included slave transport between Mozambique, Madagascar, and the Mascarene islands (Reunion, Mauritius). The Southern Asian circuit included the major sources from Bengal, Arakan, Coromandel and Malabar. The South East Asian circuit covered much of Malaysia, Indonesia, New Guinea and the Southern Philippines.

Slaves were exported to centres of demand which included colonial *VOC* cities such as the "Chinese colonial city of Batavia" (present-day Jakarta) which was the *VOC* headquarters, and the regional centre for the Western District at Ceylon (present-day Sri-Lanka) [35]. Other centres of demand included emporia (e.g. Malacca and Makassar), agricultural estates (e.g. Cape Colony) and plantation economies in eastern Indonesia, e.g. Maluku, Ambon, and Banda Island [35].

Slaving cities were built on already established major emporia. These were true slave communities as 20 - 40 % of the population were slaves and the enslaved contributed to the creolisation and emerging cultures [35], [132], [137]. The cities acted as urban centres and the *VOC* excised power to peripheral towns and villages from which slaves were collected. In South Asia, the enslaved were predominantly from micro-state or stateless people beyond the "House of Islam" as Muslim raiders contributed to the collection [35].

For the *VOC*, South Asia was a major source of slaves until 1660s [35]. Between 1626 - 1662 the coast from Bengal (present day Bangladesh) to Arakan (present day Rakhine province of Burma) provided the *VOC* with a regular supply of 150 - 400 slaves annually. For example, between 1646 - 1649 211 slaves were manumitted in Batavia, of these 126 (60%) came from South Asia and 86 (41%) from Bengal. The peak export periods were in 1647 and 1655 with 1 046 and 1 803 slaves exported, respectively. In 1666 the Eastward expanding Mughal empire captured Chittagong (present day Islamabad), cutting off traditional supplies from the region [35].

The West coast of India continued to provide slaves up to the capture of the Portuguese Malabar cities by the Dutch (1658 - 1663). After 1666 slaves arriving

at Batavia and Ceylon show that exports from Cochin were reduced to 50 - 100, 80 - 120 slaves collected annually, respectively.

The East Coast (Coromandel coast), in contrast to the West, remained a spasmodic source of slaves throughout the 17th century. A series of droughts and political incursions instigated famines and produced displaced and desperate people. This created vulnerable people, the economic incentives for people to enslave others and the incentive for families to sell their children [35]. European, Arab and Indian raiders and slavers capitalised from the sales. Vink [35] has identified at least six peaks in the exports from South Asia in relation to these calamities.

The first large-scale export was between 1618 - 1620 following prolonged drought. Around 1 900 slaves were exported between 1622 -1623 and collectively thousands are estimated to have been exported throughout the period. In 1645 the Nayaka revolt of Hindu rulers against Vijayanagara overlordship picked up at Thanjavur, Senji, Madurai in Tamil Nadu [35]. The Bijapur (Vijaypur) armies devastated the Thanjavur country side and famine set in. Muslim armies captured ~150 000 people and took them to Bijapur (~Vijayapura in Karnataka) and Golconda (Telangana, Hyderabad) [35]. Some slaves were collected from further south at Tondi, Adirampatnam, and Kayalpatnam as well [35].

Again between 1659 - 1661 a series of Bijapuri raids created 'famine-slave' cycles in the Thanjavur area. The *VOC* benefited from the purchase of eight to ten thousand slaves most of whom were sent to Ceylon, some to Batavia and Malacca.

Drought in Madurai and the ongoing struggle between Madurai - Maratha for Thanjavur between 1673 - 1677 again provided the *VOC* with 1 839 slave exports.

In 1688 the Mughal advancement into Karnatak resulted in thousands of Thanjavur slaves sold by traders from Nagapattinam to areas such as Aceh, Jahor and other markets.

A sixth boom followed further conflict in the region [136]. An estimated 3 895 slaves were exported by private individuals.

South East Asia grew in importance from 1660 due to 'endemic war and raiding', and particularly after 1667 - 1669 with the collapse of the Sultanate of Makassar (Goa,

Southwestern Sulawesi) [35]. At the time Makassar was the main transport point for slaves from Buton (Butung), Sulawesi, Borneo (Kalimantan), the north eastern islands and the eastern Tenggara islands (Lombok, Sumbawa, Bima, Manggarai, and Solor). The Kingdom of Bali was an independent slave exporter at this time too.

As many as 10 000 Indonesian slaves were exported to Batavia by Asian traders between 1653 - 1682 of which 42% (4 086) came from South Sulawesi, 24% (2 352) from Bali, 12% (1 184) from Buton, 7% (679) from the Tenggara islands, and 7% (646) from Maluku (Ambon and Banda) [see 35], [138]. Between 1620 - 1830 Hindu Bali exported over 100 000 members of their own and neighbouring Lombok, Sumbawa, Sumba, and elsewhere as slaves.

East African slavery took off due to plantation slavery on the Swahili coast and the Mascarene Islands (Mauritius and Reunion) ending in the late eighteenth century. As Africa forms the south western periphery of the Indian Ocean basin, it was typically avoided for sourcing slaves due to its distance [35]. The Cape Colony was not an important destination by comparison to others, but most of the slaves exported by the *VOC* from Africa were destined there [35], [133].

Records indicated 502 Malagasy exported from 1641 - 1647 [35], [139]. Later between 1652 - 1699 there were 12 *VOC* sponsored voyages to Madagascar which returned with 1 069 slaves and one voyage to Dahomany, bringing back 226 slaves. Malagasy were only 24% of all slave imports to the Cape during this period [35].

The Dutch trade of Malagasy pales in comparison to the British, Arab and French trade. For example, 70% (31 076 of 44 394) of slaves reaching French Mauritius and Reunion between 1670 - 1769 were Malagasy while the rest came from Mozambique and the Swahili Coast (19% or 8,435), India (9% or 3,995), and West Africa (2% or 888) [35], [134], [140].

2.3.4 Dutch Settlement and Cape Colony Slavery

The first European settlers to arrive in Southern Africa were employed by the *Vereenigde Oost-Indische Compagnie* (*VOC*) to establish a stop-over station for ships *en-route* to the East Indies [21], [103]. The initial group of 92 arrived at

the Cape of Good Hope in 1652 and consisted of Dutch settlers lead by Jan van Riebeek. Following the revocation of the Edict of Nante in 1685, French Huguenots supplemented the settler colony, later still, from the 1700s onward, unskilled German labourers seeking new opportunities immigrated [21], [61], [103].

In 1659 the *VOC* began importing large numbers of slaves to fill the need for labour. The earliest large imports were predominantly from Angola and Dahomeny (present day Benin) [16], [21], but in later years people of Madagascar, Mozambique, South Asia and South-East Asia were also imported [21].

When the *VOC* began releasing Europeans from contract and allowing land ownership in 1657, it led to the gradual expansion of the Cape colony from Table bay to the Kei River [21], [141]. Released labourers, referred to as *Vrijburghers* or free burghers (citizens), farmed in the vicinity of *De Kaap* (the central town, present day Cape Town), leading a sedentary lifestyle with small families on very extensive farms [21]. An expansive pastoralist life-style was assumed because the sandy Cape region had limited productivity [121].

It was only after Commander Simon van der Stel extended the Dutch land ownership west of the first mountain ranges (present-day Paarl, Stellenbosch) that wheat and wine become profitable and productivity increased [121].

Another, semi-nomadic form of farming emerged from cultural influences from the nomadic Khoikhoi communities which were subjugated and assimilated [16], [142]. These farmers, the *trekboeren* (loosely translated to 'moving farmers'), were of mostly European descent and lived at the frontier of the expanding settlement [21]. This life was isolated from main society and most often *trekboeren* lived in close quarters with their servants and slaves and likely produced children with them [21], [141], [142].

Between 1700 and 1795 the white settler population (whites and *VOC* employees) grew from ~500 to ~12 000 with relatively little European immigration (Figure 2.3) [121]. This is thus a "textbook example" of a founder population [61].

At its earliest days, the settler population was predominantly male (e.g. in 1717, only 350 of ~2 000 free Europeans were women [103]. The sex skew encouraged interracial coupling between settler men and women slaves [16], [21]. Assimilation of

the offspring was sex-biased as mixed-descent males were more readily assimilated into the 'black' population and girls more likely to join their white families [16], [21]. From genealogical estimates, six - eight percent of the Afrikaner ancestry is non-European, likely South-East Asian or South Asian origin though with notable contributions from KhoeSan people [16], [21], [61], [143].

At the time of European arrival somewhere between 20 000 - 50 000 KhoeSan may have lived in the Cape region and were possibly half the Cape population in 1780 [145], [146]. The initial interactions of the settlers with the indigenous KhoeSan was peaceful and included much trading for livestock using trinkets, metal, tobacco and alcohol [21], [103], [104], [142]. However, as the settlers expanded and began using more land, conflicts broke out over land ownership [141], [142]. Raiding became frequent on both sides and the KhoeSan children were at times captured as entourages and intermediates for negotiations with KhoeSan communities, despite prohibition of enslaving locals [21], [103], [104]. In 1660 a commando of 150 men send out by Van Riebeek expelled the Khoikhoi from the area surrounding the settlement. Some Khoikhoi returned and asked to remain in their birth place [142], perhaps taking on a serf status. There are several references to the KhoeSan as labourers as late as 1688 and 1695 and court records indicate that KhoeSan were often permanently 'attached' to arable farms [147], [148], thus they inherited a serf status among settlers.

Despite the number of KhoeSan, the growth of the Cape Colony was largely through reproduction among the slaves and settlers as the KhoeSan population suffered a demographic collapse in response to their expulsion from the region and disease outbreak [16], [21], [99], [111]. There were two smallpox outbreaks in 1713 and again in 1755, killing as much as 90% of the population [21], [103], [134].

The number and presence of the KhoeSan on the outlying pastoralist farms is less contentious as here they clearly outnumbered the slaves [21], [111], [121]. In both regions, *De Kaap* and beyond, sound census information is not available and many (often unsound) assumptions are needed to produce population estimates [see 121].

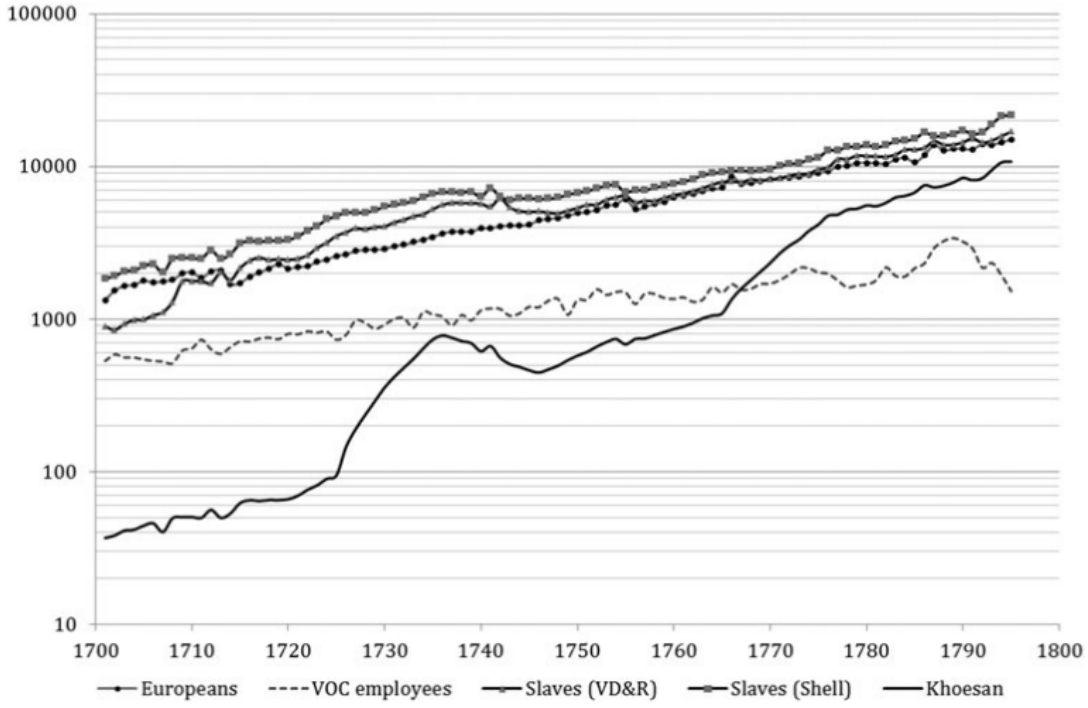


Figure 2.3: Growth of administrative groups at the Cape between 1701 - 1795. Figure taken from [121]. Note that the y-axis is in logarithmic scale. Groups include Europeans (settlers and servants), VOC employees, and slaves. Two estimates are provided for slaves; from [144] and from [140].

The cost of hiring white servants, (*knechten*), on farms was prohibitive because produce price was predetermined and set low by the *VOC* [121]. Enslavement of the indigenous people was not considered due to the 'low' population sizes and the risk of retaliation [21]. Slave labour became an important resource for the Colony. From 1657 to emancipation in 1838, over 60 000 slaves were imported to the Cape Colony from the Indian Ocean slaving networking [16], [121], [132]. The slave arrivals included privately owned individuals and *VOC* funded purchases [35], [133]. The bulk of the enslaved from the *VOC*'s African circuit were sent to the Cape Colony and these constituted the largest and earliest of slave imports [35], [149]. The importance of South-East and South Asian slaves grew in later years. Between 1680 - 1731, of 666 company slaves imported to the Cape, 30% (201) were from Indonesia, 25% (165) from India, and 22% (147) from Madagascar and other parts of Africa [35]. Privately owned slaves were often as young as

children to early adolescents, the great majority were from South Asia (65%), the remainder came from Indonesia (33%) [133], [149].

The geographic distribution of slaves was uneven across the country. The farmers near the city had more slaves than serfs compared to the farmers further out (9:1 KhoeSan to slaves in the interior in 1806 and ~6:1 at Swellendam [see 121]), reflecting the relative abundance of each [21], [111], [121]. Most farmers had only a few slaves and serfs (<10) [16], [21].

After threat of French capture, the British briefly took over the Cape Colony in 1795 [36]. Then again on the backdrop of the anti-slavery movement, the British captured the Cape Colony permanently in 1806 and slavery was banned in 1807 [103], [121] but persisted in practice much longer [103], [109].

Forms of 'new slavery' [see 129] were implemented in its place, allowing imported Mozambican and other slaves captured from rival fleets to be assigned to extended labour contracts [16]. In 1838 slaves were emancipated resulting in a gradual influx of former slaves to various town across the Cape Colony [21], [103].

British Imperialism, Apartheid and Post-Apartheid South Africa

By 1843 over 5,000 British immigrated to South Africa taking up residence at several towns along the Southern and Eastern coast [103]. British governance and the abolition of slavery was heavily contested by the white Afrikaner people who were economically dependent on slave labour and in many ways had established their identity against that of the slaves and 'blacks' [21], [103], [141]. A mass exodus from the Colony of over 15,000 Afrikaners and their servants followed in response between 1834 - 1838 [21], [103]. This historic event became known as the 'Great Trek' and was a key point in the history of the Afrikaners, "Coloured" and many African communities. As the *Voortrekkers* migrated, the settler and "Coloured" community increasingly came into contact and conflict with already present communities [103]

African communities similarly had a dynamic history during this period linked to the expansion of the Zulu Kingdom and the formation of new identities over

existing and displaced communities, for example, giving rise to the present-day amaSwazi, Ndebele and Basotho kingdoms [103], [104], [142].

Sugarcane plantations became a lucrative business in KwaZulu-Natal, increasing the demand for labour. Over 150,000 indentured Indian labourers were imported between 1860 - 1911 [103], [112]. Most were likely lower caste Indians and tribal though many upper castes came voluntarily to start businesses [103], [104].

The mineral revolution starting in 1867 and the imperialist 'scramble for Africa' led to complete loss of independence for African states as the need for unskilled labour and simultaneous fight against slavery necessitated 'encouraging' Africans to work by imposing various living costs [103], [104], [129]. The need for skilled labourers lead to influxes of white Europeans from Australia, Britain and the U.S.A [103].

The growing political unease between the Afrikaners and the British developed into the South African War of 1881 (Anglo-Boer war) and again 1899 during which the Afrikaner people were effectively defeated [21], [103]. The Union of South Africa was formed under British rule but with strong ideological interests in the welfare of the both Afrikaner and British citizens over the indigenous Africans. This ushered in the next century of formal, institutionalised white-supremacy and legalised race-based discrimination [15], [103], [150].

The Apartheid era began from 1948 after the Afrikaner Nationalist party won the election [15], [21], [103], [111]. Laws were imposed based on race implementing restrictions on access to facilities, rights to land ownership, forced removals from communities and large-scale relocations, particularly in the cities and surrounding areas. 'Blacks' were further subjected to the migrant-labour economy with near complete disregard for the societal consequences.

After over 46 years of liberation struggle, the United Democratic Front composed of various organisations successfully negotiated the first non-racial free election for 1994 [15], [103]. Since then the country has had a non-racial and liberal constitution with ongoing efforts to redress the myriad of ongoing societal troubles.

2.4 The So-called "Coloured" People of South Africa

Circumscribing a study population within the "Coloured" identity is tricky because of the inherently blended and ambiguous nature of the identity and the recency of the events which have instigated the identity. Below I outline some of the aspects of the people's history and the developments in the discussion of the identity to provide the reader with some context as to how populations were chosen.

Despite geographic and historic diversity, the "Coloured" people are often introduced via the Cape "Coloured" community and described as the product of miscegenation between settlers and their slaves [17]. This reflects the overarching essentialist narrative which understands race and racial distinction to be a self-evident concept, making 'mixedness' the hallmark of "Coloured" and specifically a biological 'mixedness' [15], [109], [151].

Some authors have denied the existence of a unified "Coloured" identity entirely. Ross [132] said there was no mutually embracing identity beyond the sporadic ideas of some authors. Others [15], [17], [109], have considered it to be a real phenomenon with elements of both a self-evident phenomenon and something artificial.

At present, people who identify as "Coloured" represent ~10% of the South African population, while the proportion of the Namibian population may be larger. The greatest proportion in South Africa resides in the Western Cape (~50%), Northern Cape (40%) and Eastern Cape (8%) [105]. This reflects the longest history in Colonial settlement and the historic distribution of the KhoeSan. Beyond the former Cape Colony, the "Coloured" communities represent below 4% of the population and their history is patchy and detached from the Colony [109], [143].

The majority of the "Coloured" population speaks Afrikaans, a derived Dutch language, and are members of one of the Dutch Reform Churches [152], again reflecting their tie to with the *VOC* settlement and developments in the Dutch society. As the history, language and religion are shared with many white Afrikaans-speaking communities there is a familiarity between the two which is not seen with other groups [152]. At the same time, "Coloured" people often shun their

KhoeSan heritage, and the non-KhoeSan African heritage is seldom acknowledged. This is not the case for all communities, as is shown by the Griekwa identity and KhoeSan revivalism [15], [18]

2.4.1 The "Coloured" Identity

The meaning of "Coloured" has shifted over time [111] and the way in which the term is used today administratively ("miscellaneous mixedness") [23] has dissociated it from the Cape history.

The early colony included distinct social ranks; "free blacks", white servile class (*knetchen*), the free and serf indigenous groups and the slaves of various backgrounds. The "Coloured" community didn't exist as we identify it today and non-Europeans were collectively referred to as 'black' while more specific terms existed for various ethnic groups, such as Bengali, Madras Khuli, Malay, Hottentot [16]. Free Blacks had civil rights similar to European settlers in that they were free citizens, property owners (including land and slaves), and had the right to marry and defend themselves in a court of law [16], [21]. This distinguished them from the *knetchen*, slaves and servants [153], while *knetchen* enjoyed greater social mobility than many non-whites.

A developing culture, including the common language of Creole-Dutch (proto-Afrikaans) and shared religious outlooks, would have facilitated communication and co-operations between people with such diverse backgrounds [15], [109]. Some have argued that in the years after the emancipation of the KhoeSan (1828) and the slaves (1838) integration was more fluid and a shared identity based on a common socio-economic status and culture could develop [15], [154], [155].

There is also evidence for a top-down assignment of the "Coloured" identity. The first administrative use of "Coloured" in reference to non-whites was in the Stellenbosch census of 1841 following the change to British governance and this was repeated in the 1849 general census [16], [111]. British census taking was concerned with the size of the native populations whereas the Dutch censuses were not [111]. Conceivably, the incoming authorities were unfamiliar with the dynamics of the lower rungs of society and produced a blanket term for 'blacks'. The casual

discourse wouldn't have changed for more than two decades after the introduction of the administrative term. Only in the 20th century did the earliest words such as "Hottentot" fall into disuse, being replaced by more general terms such as *Braunen* (Brown people) or *Kapenaars* (Cape people) [16].

From 1870 onwards, the mineral revolution drew in immigrants prompting the colonial "blacks" to hold themselves separate from the immigrants, mostly amaXhosa [15]. The "Coloured" identity may have come to symbolise their higher economic status and distance from 'Africaness' in emphasis of European descent [15].

The intensified segregation including the removal of the "Coloured" vote lead to further political mobilisation under the shared identity, however rather late (early 20th century) [15]. Under Apartheid's policies of separationalist civil life, development, and city planning, the identity was galvanised due to both regular administrative use and by "Coloured" people taking action for and against the imposition [15].

For the longest part of their history, the primary desire for many "Coloured" people was white-ward assimilation [15], [17] but with the rise of the Black Consciousness ideology of the later 1970s, a black solidarity arose. With this came the rejection of the "Coloured" identity as a Marxist style false-consciousness view imposed by the minority rule [15]. This was predominantly among the middle class, politically active and educated [17] as the majority of "Coloured" people are in favour of their shared identity. The anti-racist lobby groups and KhoeSan revivalists hold strong to denouncing the "Coloured" identity [17], [109].

Under recent Social Constructivism and Creolisationism views, "Coloured" people are not seen as a product of racial mixing but of cultural creativity in a political environment of marginalisation. The creolisation happens out of elements of the ruling class and the cultures of the 'blacks' and is made by the people themselves to give meaning to everyday life [15], [109].

The above history of the "Coloured" discourse strongly focuses on the Cape "Coloured" communities. Undoubtedly there are regional variations and possibly very different understandings to the "Coloured" identity across the country. Even during

the earlier days of the "Coloured" identity, many communities were in rural settings in the countryside [16], [21], [141]. These people would have had clear kinship and occupational ties within the community and cultural forms that distinguished them from the focal history of the Cape "Coloured" [109]. Unfortunately, the dynamics of their identities are less well covered.

Many identity factions have existed under the "KhoeSan-Coloured" spectra historically and some of these persist to present. Often identities are localised, with few individuals but have had a large role in the early settler politics. For example, the Koranna and Oorlam raiding bands were a prominent feature of the frontier [18], [142], but are not prominent in contemporary society.

In this thesis I focus on factions which have some support or suggestions of genealogical continuity and have the largest populations. Narratives of genealogical continuity could act as a signal for genetic distinctions between identities and these signals may relate to historic differences in admixture dynamics which are of interest in this thesis. Below I discuss the three main factions investigated in the thesis.

2.4.2 The Griqua

The Griqua or Griekwa are a faction of the "Coloured" population though not all Griqua identify as "Coloured". The identity of the Griqua are distinct from the "Coloured" in being strongly self-determined and highly politicised [142]. Griqua, along with the Basters, have a long history of migration across Southern Africa in response to political developments [142]. They are predominantly Christian under the Dutch Reform Church and speak Afrikaans [21].

Their journey preceded the emancipation of the slaves. The founder, Adam Kok, was a manumitted slave of a Dutch governor who was provided with a stretch of land outside North-West Stellenbosch by his former owner [21], [142], [156]. A collection of free blacks, servants and runaway slaves and white labourers joined his community. In later years they were encouraged by a resident missionary to adopt the name of previous aboriginals of the Cape, the Chariguriqua or ≠Karixurikwa [142],

[156]. Another story goes that after Kok married the daughter of the Khoikhoi Chariguriqua Chief, he adopted the name [21].

The expanding colony pushed them further north to Piketberg then little Namaqua, and through a series of further expansions of the colony and conflicts, they moved further east [21], [142]. On moving north, they lived alongside, assimilated and fought against local Khoikhoi and Baster communities, including the !Ora (Koranna), Bergenaars and Oorlam raiders [142]. Their encounters with South-central Africans, the Batlapin, Baralong and Matabele, resulted in the same [142]. Their long history reports on migrants to and from these neighbouring groups [142], but some historians report that they married mostly among themselves [21]. By the 1820s the community which included the Kok family, the Barends family (another Baster community) and various neighbouring groups was now collectively known as the Griqua [21]. They had established the community of Klaarwater (present-day Griquatown) in 1819 and Phillipolis after 1825 and finally East Griqualand and Kokstad in the 1860s following an exodus of 2 000 from Philippolis across the Drakensberg consequent to a failed land agreement with the British [21], [142]. Conflict persisted until complete British annexure in the 1870s [142].

The Griqua, like many other KhoeSan - settler creole communities, sought self-determination and self-sufficiency. They were stock farmers and hunters and, to a lesser extent, agriculturalists [142]. The Griqua of Philipolis became famous for the production of Merino wool and arms trade between the *Boere* and African communities. They established their own system of captaincy including councils and enforcers of the code of law [21], [142]. Their role in modern politics is far lighter than that of their predecessors.

2.4.3 The Basters

The word Baster is derived from the Dutch word *Bastaard* for bastard, but in the Cape Colony context it describes a person with European and non-European parents, the former usually the father and the latter usually a Khoikhoi [16], [21], [142]. In later years it was used to describe any person suspected of mixed KhoeSan heritage.

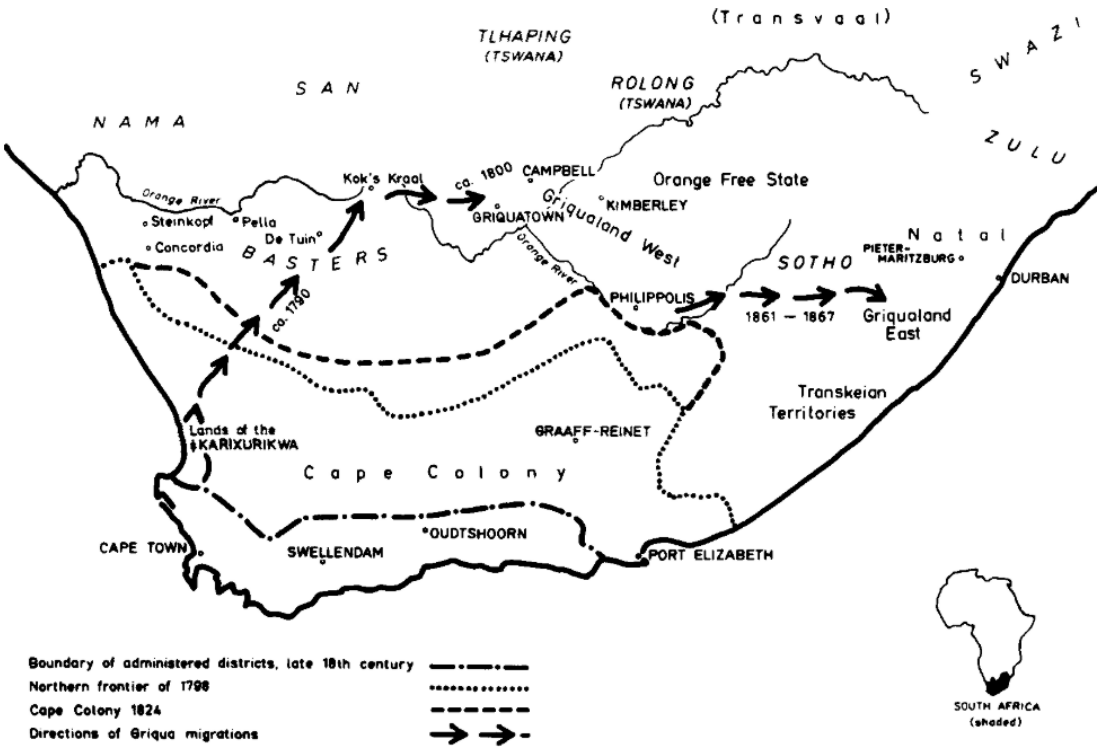


Figure 2.4: Migration of the Griqua across South Africa. Circa 18th and 19th century, taken from [156]

Basters were a predominant feature of the *trekboeren* lifestyle on the frontier where settlers lived in close quarters with their servants in isolation from other European communities [21]. The population boom in the settler society in the Cape Colony and the periodic pulses of immigration from Europe resulted in a shortage of land. Sons of settlers were thus often drawn to move into the hinterland to establish a property of their own [157]. Through necessity and familiarity, the frontier settlers developed a cultural likeness with the Khoikhoi, viewed as approaching barbarisation by some [21]. Single European farmers would then take brides from their neighbouring Khoikhoi and mixed-descendant farming communities. Small Baster communities existed throughout the northern and eastern frontier of the colony and in 1798 as much as five to ten percent of the Graaff-Reinet farming community were Basters [21]. On the frontier, the slaves were outnumbered by the KhoeSan servants and serfs often more than five to one [121]. With continued emigration of Dutch-descendants from the Cape Colony, the Baster communities were increasingly pressured into moving further north [157].

While the term has been employed most generally over the history of the Cape Colony [21], [111] it is also a term used by some to self-identify their affiliation and history apart from the Cape "Coloured" and other groups [99]. Such 'Baster' communities persist in South Africa and Namibia. The Basters of Namibia are an early offshoot of the Cape emigrants from the Colony in 1800s. Nearly 90 of 100 Basters families intended on departing from their homes at *De Tuin* near the Orange river in 1868 in search of more favourable land [158]. The Rehoboth Basters migrated northward and established a town in 1870 on the previously abandoned mission station (1864) of Rehoboth, marking their identity as distinct from the "Coloured" [99], [157]. In 1872, population estimates were 333 Baster representing ~30 families [158]. The population continued to grow as more families arrived, totalling 800 people by 1876. As with the Griekwa, the Baster consider some lineages to be associated with the founding fathers. However, the importance and legitimacy of different genealogies is disputed [157].

Under German administration the Basters or Rehoboths enjoyed a position of privilege over other groups and legal protections of their lands which was upheld after the Union of South Africa invaded in 1915 [142], though their political governing "Captain" was taken over by a South African appointed magistrate [157].

2.4.4 The Cape Malay

The Cape Malay have what is arguably a more distinct identity compared to the overarching "Coloured" identity [15], [109], [159]. Despite this, national census only included Cape Malay as an identity option, along with Griqua, in the early 20th century and it is presently no longer an option [105]. The present-day Cape Malay (~163,000 people) are predominantly Sunni Muslim and many speak Afrikaans though English is becoming more prevalent and as is conversion to Christianity [112].

The earliest founder, Shaikh Yusuf, was brother to the King of Goa (Gowa, Macassar) and was expelled to the Cape in 1694 [112]. Here he established Islam in the Cape. The slaves who adpoted Islam and contributed to the present-day Cape Malay were brought from Dutch South-East Asia, largely in the Indonesian

archipelago, not the Malayan archipelago [35], [133], [152]. The term "Malay" may have been used colloquially for people from the general region ([160, pg 120]) in the same way the term "Chinese" seems to have been used for East and South-East Asians [21, see pg 2] [152]. Islam became known as the 'black' man's religion and an alternative to the religion of the oppressor [15] and grew popular among the slaves.

One could tentatively draw a connection between the earliest mentions of the Malay in the Cape to the present-day Cape Malay. Davenport [155] even suggested that the Malay may be an exception to the "miscegenation among ethnic groups" which produced the "Coloured", however closer scrutiny shows diverse contributions to their heritage [112].

The Cape Malay are mentioned fairly early in the history of the Cape (e.g. 1772 by Thunburg [161]). From the earliest mentions, the Indonesians and subsequently the Cape Malay are described as artisans and fisherman [112], [133], [153]. Scott [162] describe them as the aristocracy of the Cape "Coloured". In 1904 the Malay were again recognised to be proportionally more involved in skilled labour than other "Coloureds"; in commerce (10% vs 6.5% of their respective population) and industry (33% vs 22%) [153].

The distinction was reflected geographically as well. In 1857 the population of the fishing region of Simonstown included 20% Cape Malay, i.e. they were enumerated separately from the other "Coloured" people [163]. In 1875 census indicated Claremont, Wynberg and the Malay quarter in central Cape Town as places of predominantly Malay residence [153]. The former two were subsequently converted to White-Only communities under Apartheid.

The Cape Malay thus have the makings of a sub-group within the "Coloured" identity in having a spatial territory and social coherence in the form of religion, education and language [159]. Unlike the Griqua and Baster who are separated by geography from the Cape Coloured, the Cape Malay may have been separated through some culture-based endogamy.

3
Methods Description

In the data chapters of this thesis I make use of a set of analyses which include those often applied in genetics and those which are specific to population genomics. I discuss here some details about the data analysis process employed and the theory behind these analyses.

3.1 Population Assignment and Ancestry Characterisation

Assigning individuals to a population or meta-population based on shared genetics signals, i.e. delimiting the population, is an important first step in trying to unravel the history and patterns of relatedness of a population. Some of the most commonly employed methods include dimensionality reduction algorithms such as Principal Component Analysis and model-based clustering algorithms as implemented in STRUCTURE and ADMIXTURE.

3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that summarises a set of variables into a few independent 'eigenvectors' which are linear combinations of the correlated variables within the original data [164].

Each eigenvector captures as much of the variation between data points as possible, with each subsequent eigenvector capturing less than the previous one and being uncorrelated to previous eigenvectors [26], [164]. This analysis is applicable to a wide variety of data from fields as different as image software development, ecology, medicine and astronomy [164], but was first used for genetic data more than 40 years ago [165] and has since been employed routinely for evolutionary and population genetics [26], [46], [166]. It requires no *a priori* knowledge of the populations of interest, can be applied to nearly any dataset and is often performed in a matter of minutes with contemporary software, making it an extremely useful tool. It is common practise to visually or analytically investigate a PCA of genetic data ahead of other analyses to identify outlying individuals or populations, detect population substructure and more generally to pick up on any unexpected or undesired artefacts of data preparation [e.g. 4], [12], [96], [118], [166], [167].

The use of the PCA for interpretation of demographic related processes is not simple. Cavalli-Sforza et al. [46], [58] used a PCA to explore the relationship between protein polymorphisms and geography in Europe and interpreted the results as evidence of the spread of Neolithic farmers from Levant 6 - 9 Kya. Although the demic diffusion of Neolithic farmers has since been supported by other investigations [168], the interpretation of the PCA synthetic maps could have been incorrect. Correlations of PCA plots with geography more often reflect isolation-by-distance even in the absence of a demographic event such as range expansion or migration, moreover some patterns can be a complete artifact of how the PCA groups variables into eigenvectors [169].

Interpretations of PCAs can be very useful and fairly straight-forward when the demographic cause is not of interest, otherwise the results should be considered in conjunction with other analyses and data [166]. The PCA does not try to group individuals into a population or a linear combination of populations and this may be advantageous when population history cannot be constructed as simply [165].

3.1.2 Model-Based Population Clustering

Explicitly modelling genetic processes allow us to more accurately examine genetic patterns in ancestry and potentially tease apart more complex scenarios of population history. A step toward this was the development of clustering algorithms which attempt to identify the best clustering of samples to reduce within-group genetic distances compared to between-group distances.

The program STRUCTURE, introduced by Pritchard et al. [170], is a population assignment analysis which uses Bayesian modelling to assign population membership and estimate allele frequencies jointly. A number of populations are assumed, represented by K and iteratively the parameters are updated and iterations compared for improved fit. Much like the PCA, these modelling processes take advantage of allele frequency differences between populations. Groups are formed to maximise the F_{ST} distances between populations.

These types of analyses are often referred to as STRUCTURE-like analyses and have been used for many species to successfully contribute to our understanding of population structure, including identifying global population structure in humans [38] and to interrogate our understanding of recent and ancient admixture history related to debates of demic diffusion or replacement in archaeology [13], [97], [100], [110], [171].

Subsequently, modifications of this process have been developed in programs like ADMXITURE [172] and FRAPPE [173] to extend functionality. In particular, ADMIXTURE maximises the likelihood of a parameter set instead of sampling the posterior distribution following MCMC as done in STRUCTURE, thus speeding up the computations and increasing the size of the datasets that can be analysed. FRAPPE uses the same model but estimates parameters using maximum likelihood with an EM algorithm.

While these analyses have proven extremely useful, there are distinct drawbacks which influence the use. The program was developed to assign genetic diversity among individuals most parsimoniously. Human history is however not (entirely)

parsimonious. When all individuals are admixed the cluster assignments are made by shifting the allele frequencies of inferred ancestral populations to reflect shared diversity across all individuals [174]. Thus, the algorithm is influenced by the interplay of demographic history and sampling structure.

The selection of a best K value is difficult as K values above 10 often encounter issues of convergence due to computational costs and the presence of distinct local optima [83]. The criterion used to decide the optimum number for of K is in itself a difficult decision to make and numerous indices have been proposed [see 175].

The exact interpretation of the identified structure is not entirely clear either. Many possible demographic scenarios could result in the same STRUCTURE results [174]. More specifically, group assignments from STRUCTURE-like analyses are understood as a high assignment to a single cluster. This distinction can be a consequence of a bottleneck or divergence, and the analyses can be strongly influenced by sample selection, sample sizes and the chosen K value. The identified 'K components' thus do not necessarily represent an 'ancestral component' as it is often interpreted to mean. For example, a population can have >90% assignment to a single K cluster but this does not mean they are an ancestral proxy to other groups that share this component at lower proportions [e.g. 94], [127] and conversely a population with high assignment probabilities to >2 clusters is not necessarily an admixed population; this assignment pattern can be a result of sampling a sister population with notably lower diversity or a by-product of the chosen K and sampling (e.g. global STRUCTURE at K=2 under this interpretation would suggest a large part of the world is a mix of African and East Asian [38], [174]).

The limitation of STRUCTURE-like analyses affect how precisely I can interpret the output into a demographic and gene flow history. Despite the distinct limitations to the use of STRUCTURE-like analyses, its application in combination with other data (geographic, linguistic and haplotype-based data) has been very insightful [e.g. 97], [125], [174].

3.1.3 Haplotype Coalescence

Many analyses were developed with the assumption of independent inheritance and independent fitness consequences for each locus of concern. This was done in part to simplify the models used and in part because the sparsity of the genetic markers at the time made this assumption functionally valid [11].

With the development of improved genotyping techniques and the reduced costs associated, issues of correlation between markers became tangible and techniques for removing such linkage disequilibrium (LD) were developed.

The growth in availability of high density data has exceeded the development of statistical methods with which researchers could exploit the available LD information. Many methods employed still require the removal of loci in linkage disequilibrium. While necessary for analyses using PCA [166], STRUCTURE [83], [170] and ADMIXTURE [172], trimming the data means losing information. Some developments have tried to make use of this information [78], [82], [83], [176]. By identifying loci in LD, I can reconstruct the genome of an individual as a series of haplotypes. Each haplotype would contain at least as much information as the component SNPs but likely more as it also then allows one to construct a relationship between haplotypes and produce haplotypes more variable than the composite SNPs.

Thus, I have the option of reducing the number of variants considered by clustering SNPs into haplotypes and increasing their informativeness for identifying relationships between individuals [83]. One of the most commonly implemented methods for this process is the CHROMOPAINTER (CP) and fineSTRUCTURE (fS) pipeline [83] which I discuss below.

3.1.4 CHROMOPAINTER and FineSTRUCTURE

Haplotype-based analyses, such as CHROMOPAINTER, utilise the discontinuities in linkage disequilibrium between SNPs which result from recombination events in a population's history [83]. CHROMOPAINTER identifies sources for the haplotypes in another donor population [83]. It uses a Hidden Markov Model (HMM) to form each recipient chromosome as a mosaic of donor chunks and

allows for emission and transition probabilities for each donor to be adjusted. The emission probabilities are the probability of a mutation assuming the SNP is copied from a donor and the transition probability is the probability of moving from one haplotype chunk to another.

3.1.5 Production of the Co-Ancestry Matrix

At each locus within a chromosome, the sample history can be represented as a genealogical tree based on the allelic state at a set of positions from the SNP data. The structure of the tree changes along the genome reflecting ancestral recombination and admixture history [83]. Haplotypes in this context are considered to be segments on a chromosome bounded by recombination sites, i.e. where the tree is similar across the segment. Haplotypes of individuals within a sample can be matched to a nearest neighbour haplotype on other individuals. In this case I describe the focal individuals as 'recipient' individuals having received a haplotype from a 'donor' individual. Nearest neighbour segments correspond to distinct coalescence events and each segment is assumed to provide reasonably independent information on the ancestry of the individual. The reconstruction of coalescence for each haplotype and the estimation of the number and length of chunks produces a Co-Ancestry Matrix (CoAM) upon which estimates of shared ancestry are based [83].

The identified donor of each chunk represents the nearest neighbour of the recipient haplotype for that stretch. The likelihood of a donor being the nearest neighbour is assessed by the similarity in the chunks, for example when using unlinked (binary) SNP data, each locus is independent such that all possible donors are equally likely, and all non-donors are equally unlikely [83].

A co-ancestry matrix (CoAM) is constructed as a x_{ij} matrix. Here x is the expected number of haplotype chunks copied from donor j at a haplotype for individual i, summed over chromosomes [83]. The intuitive interpretation of this is that the CoAM counts the number of observed recombination events leading to individual i being most closely related to individual j. Similarly, a matrix of chunk lengths l_{ij} and of mutation in the chunks m_{ij} is produced.

CHROMOPAINTER (CP) requires no *a priori* information on donor importance, but any available prior information can be included as a probability of copying from a donor compared to other donors.

CHROMOPAINTER Output

CHROMOPAINTER provides several output files. The most commonly used are the chunk count files, providing the number of chunks copied from a donor, and the chunk length files, providing the length of genetic material copied from a donor. While both data have found use in population genetics, each has its own suite of advantages and drawbacks.

Chunk counts

A high chunk count will not necessarily reflect a high proportion of ancestry should the chunks be small compared to other ancestries. Differences in the size of contiguous chunks/haplotypes can be the result of several influences. Firstly, haplotypes inherited from an older admixture event will have underwent a greater number of recombination events and thus be broken into smaller chunks. Secondly, when possible source populations have different levels of genetic variation at the SNPs considered, haplotypes inherited from population with low SNP diversity will be broken up less often than haplotypes inherited from populations with high SNP diversity. This data is necessary for the construction of the fineSTRUCTURE (fS) tree.

Chunk lengths

This data consists of the expected total length (cM) of genetic material copied by the recipient from the donor across all SNPs.

This data is extremely useful as it allows one to calculate the average length of material copied from a donor $\frac{chunklength}{chunkcounts}$. It also has an inherent property of having a consistent upper bound across individuals [88], [177] which is particularly valuable in understanding relative contribution to ancestral proportions.

Mutation rates

This output consists of the expected number of SNPs copied from donors with error. This is referred to as emissions and is understood as mutations necessary for the recipient's haplotype chunk to best match the donor's. This data has been far less utilised in population genetic papers to date.

3.1.6 Population Clustering with Haplotypes

The local ancestry information, specifically the chunk counts, can be processed to identify genetically similar clusters by using the fineSTRUCTURE (fS) clustering program [83]. The fS program uses the Markov Chain Monte Carlo (MCMC) algorithm progression related to that implemented in the STRUCTURAMA [178] to explore sample partitioning space. The best number of clusters is estimated by Random Jump-MCMC. The Random Jump-MCMC algorithm proposes new configurations from the previous step by allowing the MCMC to move between different models with different types of information. The new configuration is accepted with a probability depending on the ratio between the two respective Likelihoods. From this best partition of samples, groups are merged successively, and the most probable merge identified at each step to create the bifurcating tree [83].

Individual samples are assigned a cluster based on an identical recipient and donor distribution. An individual from population a with n individuals in it should share the same underlying fraction of chunks received from other individuals in that population. Furthermore, individuals in a should share identical relationships with other groups such that they receive the same fraction of their chunks from, say, population b and donate the same fraction of chunks to population b [83].

The likelihood that an individual i in population q_i receives a single chunk from individual j of population q_j is $\frac{P_{q_i q_j}}{\hat{n}_{q_j}}$. Where $\hat{n}_{q_j}$ is the number of individuals in population q_j and $\hat{n}_{q_j} = n_{q_j}$ if $q_i \neq q_j$ otherwise $\hat{n}_{q_j} = n_{q_j} - 1$. The overall likelihood is calculated by multiplying across chunks, assuming chunks are independent.

The expected number of chunks x_{ij} of the co-ancestry matrix is divided by a correction factor, c, to produce an 'effective number of independent chunks' to account for non-independence between chunks in practice and the use of the expected

rather than the observed number of chunks copied in the likelihood estimation [83]. The cause for deviations between the expected number of chunks and the observed number is due to several influences. Some influences discussed by Lawson, Hellenthal, Myers, and Falush [83] include the double counting of a chunk when two individuals act as donors to each other, non-independence of adjacent chunks on the same haplotype due to limitations in modelling recombining genealogies, the non-Markovian nature of genealogical relationships and inaccuracies in data leading to misleading chunk boundaries. The c parameter is estimated empirically from non-overlapping chromosomal regions large enough to be approximately independent.

Tree construction for fS differs from what is often practised in interpreting ADMIXTURE results in that successively reducing the value of K does not perform well [83]. Instead relationships are inferred at the 'natural' value of K. In this process the maximum *a posteriori* state is attained by taking the MCMC iteration with the highest observed posterior likelihood and performing additional hill-climbing steps to identify merges or splits that further improve the posterior probability.

3.1.7 Cluster Assignment Re-evaluation

The final clustering produced by fS may not be the sole best clustering. The algorithm does not explore the entire possible space of parameters or reconfigurations and as such the choice of priors can influence the outcome should there be a number of possible local maxima in a parameter space. Multiple configurations may be equally good, or the final outcome could reflect the algorithm getting stuck at a local maximum. A few techniques are available to further evaluate the configuration.

Maximum Marginal Posterior

The final tree states are dependent on a single MCMC sample observation with the maximum posterior probability. This probability is calculated assuming fixed values for a large number of parameters (including an individual's placement and the number of clusters). There is a chance that the posterior distribution will be relatively flat across an extensive state space so that a fairly divergent

set of parameter values may result in similar posterior probabilities [88]. Leslie, Winney, Hellenthal, *et al.* [88] proposed a process in which the marginal posterior distribution for each individual's assignment to a cluster across MCMC runs can be used to construct the final tree.

Following the work by [88], we begin with N individuals, M MCMC samples to be considered and K clusters at the fineSTRUCTURE 'final inferred state', i.e. maximum posterior probability state.

For each individual i we find x_i^m which is the number of individuals (including i) which cluster with i at MCMC sample m; for $i \in 1 \dots N$ and $m \in 1 \dots M$. we identify the number of individuals y_{ik}^m that cluster with i at m and who are in cluster k at the final inferred state. That is, $y_{ik}^m \leq x_i^m$ for $k \in 1 \dots K$. Individual i is re-assigned to a cluster k where k has the maximum marginal y_{ik}^m, i.e. $\sum_{m=1}^{M} [y_{ik}^m / x_i^m]$ for $k \in 1 \dots K$.

we can use the value $\sum_{m=1}^{M} [y_{ik}^m / x_i^m]$ as an indication of our confidence in the assignment of individual i to each cluster $k \in 1 \dots K$. The value from the final iteration is normalised across k to sum to 1, producing a $P_{K,i}$ vector of length K.

Following the re-assignment of individuals, we construct a tree. Lower level clusters are successively merged, accepting the merge with the highest probability at each step. This process produces a bifurcating tree which empirically performs well in capturing approximate similarity between clusters as an indication of relationships [83].

To reduce the number of branches on the tree we performed a branch collapsing step. Here we consider a vector $P_{J,i}$ which is analogous to $P_{K,i}$ above. $P_{J,i}$ is produced at level L_J, where $J \in 2 \dots K$ and each L_J represents a height on the tree. At each L_J , each cluster (or node, in tree terminology) j is composed of all clusters below it (or sub-trees, in tree terminology). Here $P_{J,i}$ is the sum of all $P_{K,i}$ values for the branches which form node j .

A threshold value, typically $t > 0.7$, is used to collapse branches. Here t is the minimum allowed value for $P_{J,i}$ at L_J before a pair of branches are collapsed.

Total Variation Distance Cutting

The number of clusters retrieved by the fS tree is not always useful for interpreting population histories as there may be a rather large number of clusters with marginal differences. To reduce the number of clusters Dr Francesco Montinaro suggested a Total Variance Distance (TVD) cut procedure (pers. comm.). Here the TVD is calculated according to [88] and clusters sharing a node on the final maximum concordance tree are merged when the TVD is below a decided value, a $TVDmin$. The process is iterated until the lowest TVD value is above the $TVDmin$. The TVDcuts are performed based on the genomic length matrix as genome lengths share an upper value across for all individuals and are not influenced by the time since admixture in the same way as chunk count values would be (i.e. decay in average chunk length would correspond to an increase in chunk count).

Non-Negative Least Squares Regression

Because donor populations themselves may share haplotypes and even have qualitatively similar chunk count copying profiles, the use of mixture models has been suggested to select the most likely donor populations for a recipient individual or cluster. The non-negative least-squares (NNLS) regression uses the recipient population's copying profile as a response based on the donor population's copying profile which is the predictor. The NNLS coefficients $\in [0, 1]$ and summing to 1 are used to describe the proportion ancestry from each potential donor population. Specifically, I create a copying vector Y_P of length G where G is the number of donor clusters identified and P is a recipient cluster or individual. Each element $g \in Y_P$ is an indication of shared ancestry with cluster g. This measure of shared ancestry can be the sum of the length of genome (cM) copied from all individuals in cluster g (which by definition has the same upper limit for all individuals) or it can be the sum of chunks copied from all individuals in g (which then needs to be adjusted appropriately to account for the potential relationship of chunk size and chunk length) (see section 3.1.5 for more details). In a similar fashion, I then define X_g to be the copying vector of a donor population g from other donor populations of length G.

I set up the regression equation ; $Y_P = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_G X_G$

Where $\sum_{g=1}^{G} \beta_g = 1 \; and \; \beta_g \geq 0$

The vector of β_g values for each cluster in P is referred to here as an ancestry profile and I use this as an indication of ancestral contributions to individual or cluster P.

A jack-knife estimation of error can be performed, jack-knifing over chromosomes, following the procedure by [97], [179]. Here chromosomes are weighted as unequal groups based on SNP count.

3.2 Admixture Dating

One can make further use of recombination to characterise the ages of admixture events. A background level of linkage disequilibrium is maintained in any population through a combination of processes which create LD (drift, selection and population structure) and remove LD (recombination) [78]. In humans, LD becomes negligible beyond a few hundred kilobases [5] but in populations with a recent admixture there can be longer-range Admixture LD (ALD). There is thus a heightened LD between SNPs co-inherited from an ancestral source. This LD breaks down with time since admixture [79].

The relationship of LD between SNPs and the distances between SNPs can be used to characterise this breakdown in LD. The use of a weighted LD decay curve to examine admixture history was first proposed by Moorjani, Patterson, Hirschhorn, *et al.* [180]. Considering diploid genomes from a test population P_C and two reference populations P_A and P_B and for a pair of SNPs ($x \; and \; y$) at a distance d, I can estimate a weighted measure of linkage disequilibrium. Weighting is based on allele frequency differences between reference populations.

Where, $w(x)$ is the allele frequency divergence of P_A from P_B and $D_2(x, y)$ is the sample covariance between genotypes at x and y for P_C. If P_C is derived as an admixture of P_A and P_B then LD (averaged over SNPs at distance d) should show an exponential decay with increasing values of d as the probability of a recombination having happened increases with increasing d. The software ROLLOFF, developed to estimate this [180], infers parameters of admixture based on the decay constant

(λ) and the amplitude of the curve. These parameters correspond to, respectively, the date of admixture and proportion contribution.

MALDER (Multiple Admixture Linkage Disequilibrium for Evolutionary Relationships) [78] offers an extension of the original ROLLOFF software in that it uses a new weighted LD statistic which is more robust, and it allows us to interpret the amplitude of the curve. It further provides a statistical test for the admixture and as it uses a fast Fourier Transformation algorithm for computing weighted LD, the calculation times are reduced compared to ROLLOFF.

To filter out possible confounding signals due to other demographic events such as shared bottlenecks between the admixed population and the reference populations, MALDER follows a multi-step procedure.

1. The LD correlation between the admixed and reference populations is determined. This is used to decide a minimum distance cut-off for the calculations to prevent the confounding influence of background LD and to determine if the admixed population and reference population share a bottleneck. Curve fitting beyond the LD correlation threshold fails when the two populations have a shared bottleneck which has caused the shared LD pattern.

2. A two-reference weighted LD is calculated and the amplitude and decay constant estimated for each pair of reference populations.

3. A one-reference weighted LD is calculated. This is done as a test for admixture because should Population P_C be truly admixed, a decay curve should be detected even if the second reference population is P_C.

4. Comparison of the two-reference test and the two one-reference tests. Standard errors are estimated by jack-knifing over chromosomes and a p-value estimated by dividing the mean by the estimated standard error to retrieve the Z-score.

3.2.1 Drift

MALDER results may be influenced by drift since admixture in the reference populations because the reference populations are likely to have had allele frequency changes since the event. The level of drift changes the noise - signal ratio and with higher drift there would be higher noise - signal ratio, producing a weaker fit to the curve.

MALDER uses an unbiased sample covariance in place of the sample correlation coefficient and $w(x)w(y)$ used in ROLLOFF [78], [180] making the results more robust to the influence of post-admixture bottlenecks.

MALDER is conservative in identifying a population as admixed when there is a confound from other demographic events, and in particular a shared bottleneck with one of the reference populations. If the reference populations share a bottleneck history with the admixed population, MALDER will identify long-range LD (step 1. of the process above) and indicate that the reference is not suitable.

3.2.2 Multiple Sources of Admixture

MALDER can be used to test for multiple contributors to admixture in a population in a similar fashion to that described above. A further multiple-hypothesis correction is made to account for multiple tests. To account for tests not being independent when populations are similar, an effective number of populations is estimated based on a PCA of the allele frequency matrix of the reference populations.

4

Fine Characterisation of Cape Town "Coloured" Ancestry

4.1 Introduction

The age of exploration beginning in the 15th century and the subsequent globalisation has led to the voluntary and coerced mass movement of people across global regions in a relatively short period of time. Admixture among migrant and indigenous communities have become an established feature of many post-colonial societies [97], [181], [182], creating 'new' genetic and cultural creolised ethno-racial identities [18], [183].

The Cape "Coloured" communities (hereafter SAC) of Southern Africa are one of such a set of communities, associated with the arrival of the European settlers and their slaves and servants in the 17th century [16], [21]. Today this ethno-racial group is recognised as a distinct 'racial' group in the Southern African identity landscape. Despite their importance in shaping the identity landscape and the historic political developments, there is a stark gap in the available information on their predecessors. Records are few, often scant, and biased in representation toward higher socio-economic status individuals. This leaves open many questions on the influences of the slaves and servants on genetic and cultural creolisation at the Cape.

4.1.1 The Missing Slave Identities

The First European settlement at the Cape of Good Hope, South Africa was established in 1652 by ~90 people, of which most were European peasants, and slaves and servants. Settlers were employed by the *Vereenigde Oostindische Company* (*VOC*) to establish a refreshment station for ships *en route* to the East Indies [184]. While the *VOC* had relatively little engagement in the slave trade in comparison to other commercial interests, mainly spices and, later, fabrics, tea and coffee, slaving expeditions were undertaken to meet labour demands [133], [149]. As employees at the Cape were later released from their contracts and allowed to take up land to promote farming, such imports were necessary [16], [121].

The private slave trade far exceeded the company sponsored trade [133]. Private traders and mid- to senior-level *VOC* employees could profit from small-scale trading by exploiting the *VOC*'s' transportation networks. This allowed them to move slaves to areas of higher demand [133]. Unregistered slaves were often transported, and ships frequently rerouted illegally to follow demand [133]. In contrast to the triangular shipping circuit of the Atlantic Ocean Slave Trade, the Indian Ocean Slave Trade was more complex as movements were multi-directional and influenced by multiple players in the trade [35], [149].

The chaotic nature of the trade has hindered systematic studies [35], limiting information on the ethnic background of the arrivals. What was reported often depended on the interest of the traders involved. For example, Asian traders recorded more ethnically detailed notes compared to European traders, names were changed, and little was recorded of the geographic origin except for the port of departure or sale [133]. In addition, the procedures and purpose of population census through the Cape Colony's history has shifted and ethnic and 'racial' identifiers were not used consistently over time [111]. This hinders systematic studies to back-track changes in demographics as it is unwise to assume that a term for ethnicity is either precise or consistent [111], [158]. The many discontinuities, biases and imprecisions in available records limits our understanding of the origins of the arriving slaves and servants.

4.1.2 Slave, Servant and Settler Arrivals

The multi-regional nature of the Indian Ocean Slave Trade is reflected in the genetics of the present-day SAC. STRUCTURE-like analyses [12], [99], [110] show that the largest and most consistent contributions were from the KhoeSan speaking indigenous communities (~25% on average). This is followed by the Western Eurasian component (second in prevalence, 40 - 50% when present), a non-KhoeSan African component (~15%) and an infrequent Asian component (17 - 20%).

For the first three decades after 1652, the European settler population consisted of ~30% Dutch, 30% German and 20 - 30% French [61]. The French arrived in 1688 as 180 Huguenots following the revocation of Edict of Nantes, while prior to and after this, European immigration was largely from unskilled Dutch and Germans [21]. Over 5,000 British settlers arrived after the capture of the Cape Colony in 1806 [103].

Throughout the history of the Cape Colony there was a heavy male bias in the European population which 'encouraged' reproduction between European men and the slaves and servants [16], [21]. Inter-ethnic coupling and the assimilation of children was, however, strongly sex-biased. For example, non-white men who had sex with a white woman were severely punished and male 'basters' of European men were not assimilated into settler society as readily as their female siblings [21]. The complex socio-economic dynamics and stratification are reflected in uniparental genetic markers [63]. There is a near absence of European mtDNA among the SAC and a near absence of South and East Asian Y-chromosomes. By contrast, Bantu, KhoeSan and Asian mtDNA were predominant as were Bantu and European Y-chromosomes, reflecting the societal power asymmetry and numbers of arrivals.

There were several large-scale *VOC* sponsored slave imports. The first arrived in 1658 from Angola and Dahomey (Benin), West Africa, and later from Madagascar and Mozambique [16], [132]. Between 1680 and 1731 half of the arriving slaves were Malagasy, while the proportion of Indian and Indonesian slaves from private trading grew to a third [16], [35], [133]. The importation of African slaves continued in the 18th century, but with still increasing imports from South Asia and the Indonesian

archipelago [133]. Half of the slaves imported from the Asian circuits and arriving through Batavia were from the Indonesian archipelago while the largest part of the other half was from Bengal and the southern coast of India [133] (Figure 4.1).

Most non-African slaves arrived from *VOC* outposts such as the Sunda Islands, Moluccas, Ceylon, India and Bengal (Bangladesh) [133] while some East Asians may have arrived as private servants (possibly 140 - 230) [152]. The degree of variation within the South-East Asian ancestry, which appears to be mostly Indonesian in origin, is low, possibly indicating low diversity among the founders or a subsequent bottleneck after arrival at the Cape and there is no evidence so far for an East Asian contribution [185]. The South Asian component appears most like Bengali ancestry [12].

While the *VOC* did not permit the enslavement of local communities, the local semi-nomadic Khoikhoi were assimilated as serfs. Some of the present day KhoeSan communities have a demonstrated genetic connection with Khoikhoi occupants of the Cape from 2 Kya [28], [94] and were likely the earliest contacts for European settlers. Initially, trade was common between settlers and the Khoikhoi but as *VOC* employees took up land for farming, the Khoikhoi were increasingly forced out of grazing land [103]. Conflict became frequent and the indigenous communities were pushed north- and eastward and into conflict with other groups [142]. Many KhoeSan entered servitude among settlers [121].

Despite the recent work on the "Coloured" ancestry [12], [63], [87], [99], [108], [110], [115], [152], [185]–[188] there is still a clear lack of a detailed characterisation of the source populations. Current identified sources are tentative as the choice of representatives has always been *a priori* the most likely sources based on the historical records and tests have included few alternatives and low SNP density, potentially lowering resolution for closely related populations [12], [99], [107]. I here produce a fine-scale characterisation of the ancestry of the SAC employing mixture models to identify proximal rather than distal sources. I test a substantially broader collection of possible sources from South Asian, European and Eastern Asian populations, making this the first fine-scale assessment of specific European

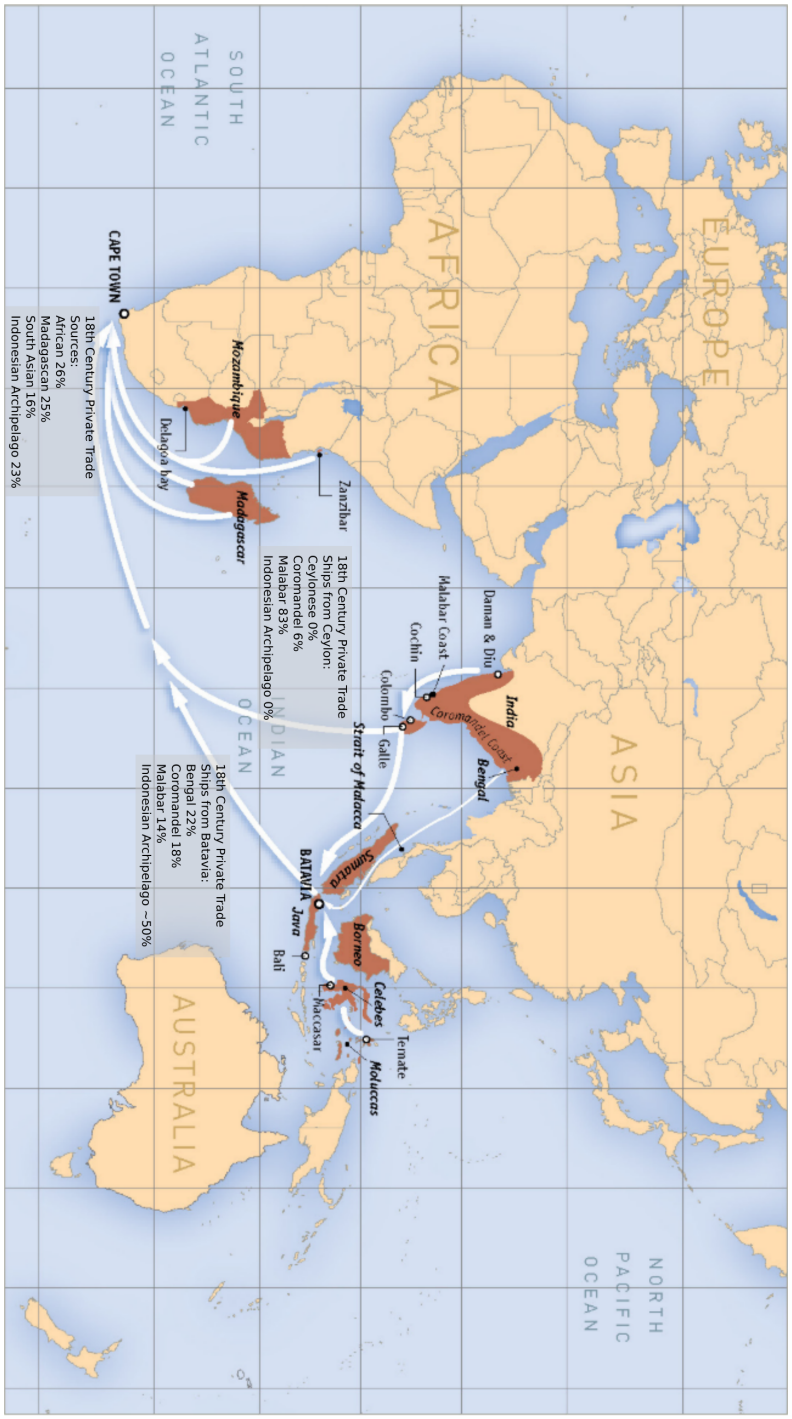


Figure 4.1: Slave trade routes across the Indian Ocean. Both private and *VOC* sponsored imports are shown. Port of origin for slave ships from Ceylon and Batavia are indicated. Figure modified from [149].

contributions to the SAC genome. Finally, I use the available data to examine for possible uncharacterised population structure within the Cape Town SAC.

4.2 Methods and Data Analysis

All supplementary material for this chapter is available in Appendix A.

4.2.1 The Community

Uitsig and Ravensmead are neighbouring low socio-economic standing communities in the City of Cape Town. The area is only 3.4 km^2 but hosts over 36,000 people (as of 2001) [189], [190]. The area was established on a flattened rubbish dump in the 1970s as part of a low-cost housing development under Apartheid. The communities today still face marginalisation which affects their living conditions and consequently their health. The communities were 95% Coloured-identifying in 1996 but there has been a possible decline (85% in 2011 for the Elsie's Rivier Suburb Area[1]). No data exist on the 'within-Coloured' identity variability.

Table 4.1: The demographic profile of Uitsig from the 1996 census.

Group	Male	%	Female	%	Total	%
Black(African)	165	1.35	144	1.18	309	2.54
Coloured	5539	45.45	6081	49.89	11620	95.34
Indian/Asian	39	0.32	49	0.40	88	0.72
White	4	0.03	4	0.03	8	0.07
Unspecified	81	0.66	82	0.67	163	1.34
Total	5828	47.82	6360	52.18	12188	100.00

4.2.2 Sample Collection

Data were made available from a pervious gene association study [110]. A total of 959 individuals were collected for a Tuberculosis case-control study and genotyped on the Affymetrix 500K SNP chip. The variant calling process is described by [110] and further details on the sample quality control can be found in [187]. Briefly, the initial

[1]Includes the Uitsig community but not the Ravensmead. Ravensmead is included under the Parow Suburb which includes some White-majority neighbourhoods with disjunct socio-economic standings making the Suburb-averaged demographics misleading.

Table 4.2: The demographic profile of Elsie's Rivier Suburb from the 2011 census. Uitisg but not Ravensmead is included in this area.

Group	Male	%	Female	%	Total	%
Black(African)	3303	3.65	2771	3.06	6074	6.71
Coloured	38182	42.16	42121	46.50	80303	88.66
Indian/Asian	1468	1.62	1556	1.72	3024	3.34
White	112	0.12	78	0.09	190	0.21
Other	625	0.69	358	0.40	983	1.09
Total	43690	48.24	46884	51.76	90574	100.00

Data taken from www.capetown.gov.za

sample size was reduced to 733 individuals with 390 887 SNPs (381 558 autosomal SNPs, 642 cases and 91 controls; 406 males of which 361 are cases and 45 controls) through checks for relatedness, minimum allele frequency (1%) and missingness of data (0.05). The samples were collected under a project approved by the Institutional Review Board of Stellenbosch University, Tygerberg, South Africa and provided for population analysis by Dr Eileen Hoal [110]. The median age of birth was roughly the same for males (1967, n=208) and females (1969, n=203) (Figure 4.2).

4.2.3 Datasets, Merging and Quality Control

I follow the analysis procedure outlined in Figure 4.3 for the work in this chapter and Chapter 5.

SNP-chip Data

I merged genomic data from several papers which covered much of the old-world populations thought to be relevant potential sources (Figure 4.4). I placed emphasis on reference populations from South Asia and Europe. All data were genotyped on an Affymetrix 6.0 or 500K chip. Data curation was performed using PLINK [191]. An outline of the change in SNP and sample numbers can be found in Table 4.3.

Merging with Genome Sequences

I included whole genome sequence data from the Simon's Genome Diversity Project [13] (SGDP). Sites homozygous for the reference allele were then added back to the VCF files by merging SGDP individuals using bcftools `merge -0`. Data were

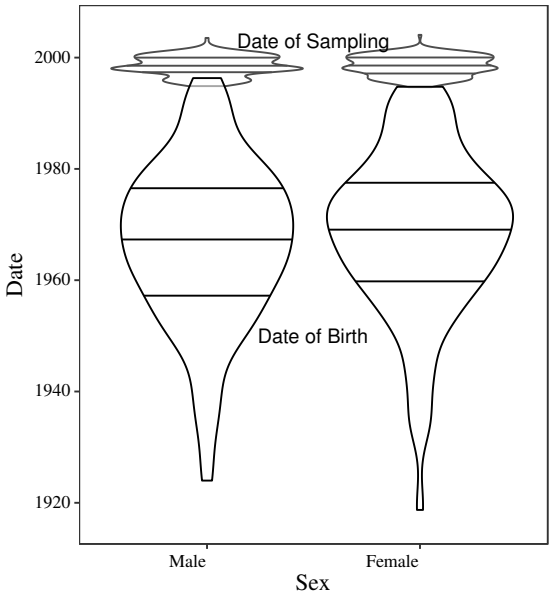


Figure 4.2: Frequency distribution of sampling and birth dates of the SAC individuals Dates (sampling; n=416, median=1998 and birth; n=411, median=1968) grouped by sex. Width of the violin plot indicates relative proportion of the total sample set. Black lines indicated median and 25/75% quartiles for the dates.

converted from VCF to PLINK binary files (`plink -vcf FILE.vcf -make-bed -out PLINK.out`). Data were sub-setted for SNPs present in the Affymetrix 6.0 chips using the `-extract` option in PLINK. Variants missing from the SGDP but present on the Affymetrix chip were then added back as a reference using a custom script. The modified VCF data were then compared to the 1000 Genome Project (1KGP, [2]).

An ADMIXTURE run and PCA were performed on the SGDP data to confirm the absence of a batch or chip effect on the identified population structure. A batch or chip effect refers to variant calls which are inconsistent across different genotyping chips or across sample batches on the same chip [for an example see 4]. Loci were removed when incongruent across more than two individuals (total 2 271 SNPs) using the `-exclude` option in PLINK.

Data Curation

Data were lifted to UCSC build37 using the UCSC liftOver tool [192] and the relevant liftover file with the command : `liftOverPlink.py -map file.map -out`

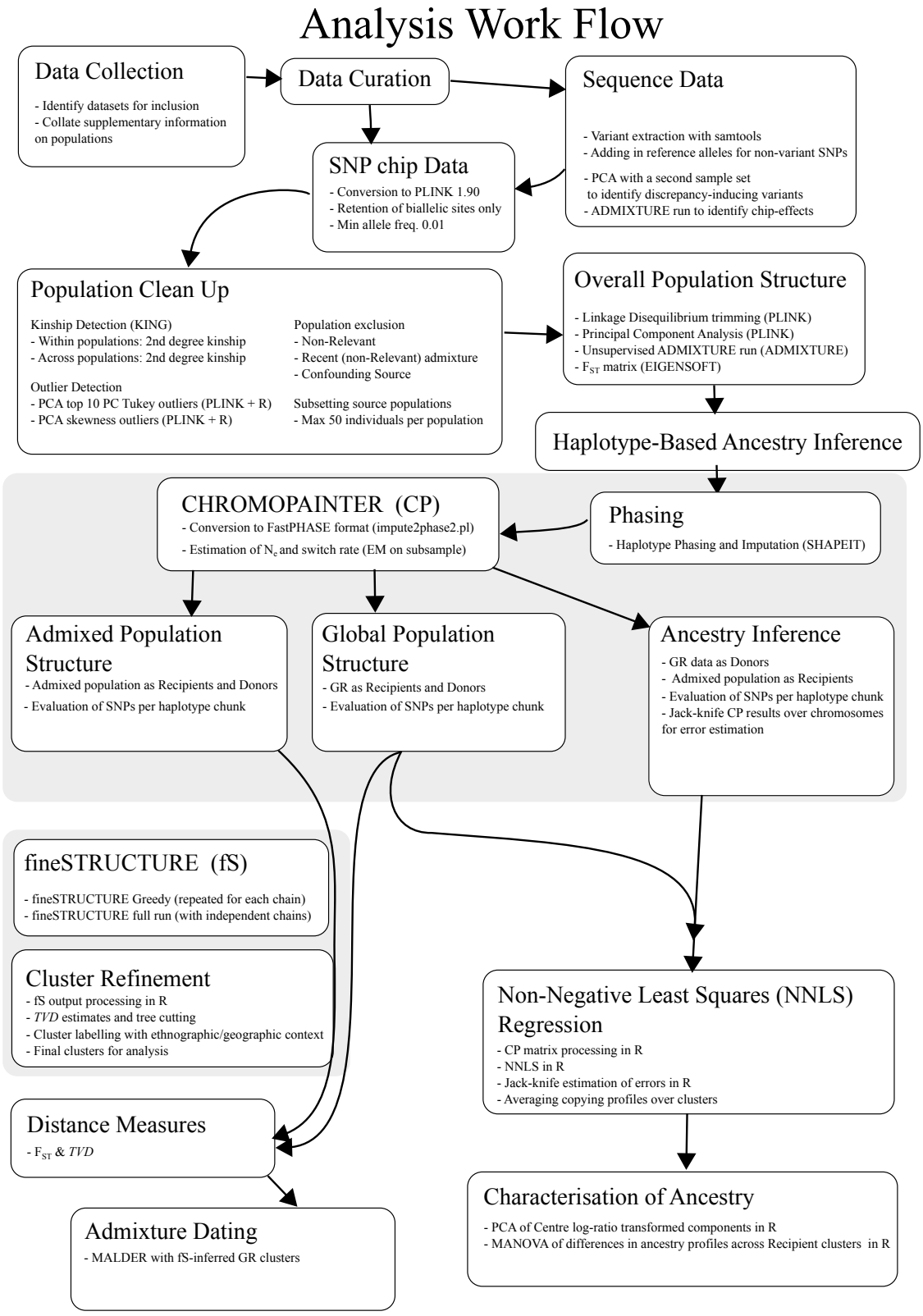


Figure 4.3: Analysis workflow used in Chapter 4 and 5.

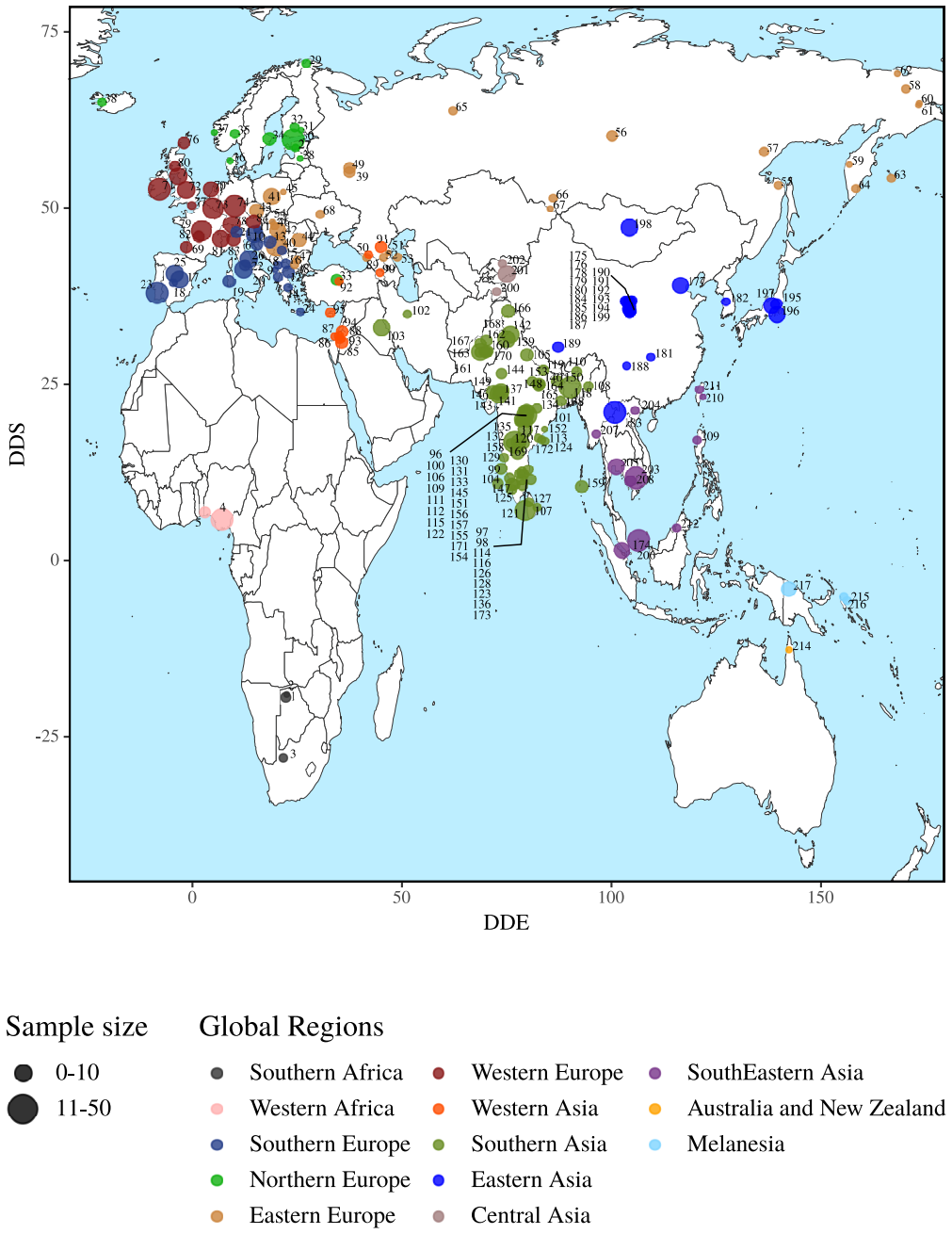


Figure 4.4: Distribution of Global Reference (GR) samples included in the analyses. Size of the plotting symbols indicate sample sizes and colour indicates regional divisions according to United Nations (UN Regions). Plotted numbers correspond to Supp. Table A.1 where information on the samples can be found. Abbr. DDS/E - Decimal Degrees South/East

`file.out -chain hg18ToHg19.over.chain.gz`. Individual datasets were then pruned for T-A/C-G to prevent flip-strand issues for such ambiguous calls. Any SNP unplaced for Chromosome or bp position were removed. Coordinates were updated from previous rsIDs to the db150 build using the PLINK `-update-map -update-chr` options. All SNPs were renamed to a Chr_position[b37] ID to ensure a match across datasets. All multi-allelic SNPs were removed. On merging datasets SNPs were flipped using PLINK `-flip` as needed.

Minimum allele frequency was set to 1% (`-maf 0.01`), removing 1 458 SNPs. I trimmed missing genotype data per individual and per locus to a maximum 5% (`-geno 0.05 -mind 0.05`). After the data curation process the dataset consisted of 168 450 SNPs and 8 610 individuals in 292 populations. Apart from the focal SAC populations, all populations were restricted to 50 individuals. Where multiple datasets included the same population (e.g. Yoruba from HGDP and YRI from 1KGP), I randomly sampled individuals from all available datasets to make up the 50 individuals.

Linkage Disequilibrium Pruning

When performing any of the Principal Component Analyses (PCA) and ADMIXTURE analyses, I trimmed the datasets for linkage disequilibrium as both programs do not account for correlation between SNPs [170]. I opted for a lenient trimming of R^2>0.7 for a 50bp frame with a 5bp sliding window [100] with the command `-indep-pairwise 50 5 0.7` in PLINK. This resulted in a final SNP count of 127 481.

Kinship assessment

I assessed the level of relatedness between samples by estimating the kinship coefficient using the software KING v1.4 [193]. The algorithm is very efficient and stable when testing for kinship in very large datasets with potential substructure. Only one individual from each pair with second degree kinship was retained (kinship coefficient >0.087). In each case, the individual with the most SNPs was retained.

The `-kinship -ibs` flags were used to generate kinship and identity-by-descent estimates of relatedness to compare.

Kinship detection resulted in 1 398 individuals from 51 populations being removed based on within population kinship (Supp. Table A.2); additional individuals from 21 populations were removed due to kinship across populations. Most pairs of related individuals were from known trios or matched those identified in the original publications.

Outlier detection

Individuals identified as outliers in the original paper were removed from the analysis from the onset. Additional outliers were identified by any of three means.

I performed a Principal Component Analysis in PLINK using all GR data. The results were investigated both visually and by using the interquartile range of the first six PCs adjusted for skewness to identify Tukey outliers [194]. This was done with the `Robustbase` package in R using the function `adjOutlyingness` [195]. Outlyingness is a generalization of the Donoho-Stahel outlyingness measure using the medcouple [194]. Only populations with n>5 were included, repeating the process 150 times each with default parameters and `ndir=250`. Individuals identified in at least 30% of the runs were considered outliers. Full Rank on the input matrix was enforced with the `fullRank()` function.

A third method was implemented which included identifying individuals with outlying identity-by-state (IBS) values as employed by [167]. This allows us to identify alien individuals with lower genetic similarity to other individuals within the population. I followed the procedure of [196] in which individuals with at least 60% of their pairwise IBS values below the $median - 3 \times IQR$ (the 'Tukey outlier criterion') were excluded.

The results of the outlier detection were in little agreement, with only 3 individuals identified consistently across methods (Supp. Table A.3). All individuals identified by any of the methods were excluded resulting in 102 individuals from 44 populations being removed.

Population exclusion

Populations were excluded on several grounds. In brief, I excluded populations which demonstrated non-homogeneous PCA results, populations which are known to have a very recent admixture history (~600 years old) and populations with ancestry profiles I considered a risk for downstream analysis (Supp. Table A.4). For example, citizens of Malaysia who are ethnically Chinese/Indian - Malay admixed may supplant a genuine Chinese or Indian copying, thus downplay the proportion of Eastern Asian and Southern Asian ancestry. I removed all African populations with the exception of Yoruba, Ju|'hoansi and ≠Khomani; individuals in these populations were selected as they are known to have little historic admixture.

Population homogeneity was assessed by estimating variance of the multivariate Euclidean distance between individuals of the same ethnic group. I estimated the Euclidean distance d_{ij} based on the first ten Principal Components from the GR PCA for each individual $i \in 1 \ldots n$ and $j \in 1 \ldots n$ where n is the total individuals for that population and $i \neq j$. The population average distance was estimated as $\sum_{i=1}^{N} d_{ij}$. Outlying populations were identified based on their average distance, where inter-individual distances were $\mu \pm 3 \times \sigma$ above the dataset average. The process was performed on a global dataset and repeated with data split into UNESCO global regions. Across the project the results of this comparison were stable. I again trimmed for missing genotype data (`-geno 0.05`). A total of 82 populations were removed leaving a final dataset of 2 884 individuals, 150 populations and 166 529 SNPs.

Table 4.3: Changes in the Affymetrix dataset through quality control.

Stage	No. Individuals	No. Populations	No. SNPs
Merged data	8610	292	168 450
Post-cleanup (PCA, ADMIXTURE)	3086	232	166 529
"CP-fS[1] (PCA, ADMIXTURE, FST)"	2884 (2151 + 733 SAC)	150	166 529

[1]CP-fS values from after the removal of excess populations.

Global Population Structure

I investigated patterns in global reference (GR) population structure within our data as a sanity check for the data QC process. This information was further used to discuss the relationship of the South African "Coloured" (SAC) to the GR data. Prior to analysis, I removed SNPs in linkage disequilibrium. A summary of population structure was assessed using a Principal Component Analysis including the SAC as implemented in PLINK v1.9 [191] using the `-pca` option.

I further evaluated population structure using the clustering algorithm implemented in ADMIXTURE v1.3.0 [172]. The analysis was run for 12 replicates of each K value between 2 - 17 where K is the number of assumed clusters, with a random seed, 5-fold CV estimation with 100 bootstraps for estimating standard errors (options `-s time -B100 -cv INPUTFILE.bed 2..17`). The optimum K-values were selected based on the lowest CV error [172]. Any K-value in the proximity of the elbow of the curve were considered equally well suited. Replicates of the ADMIXTURE runs were processed with CLUMPPAK [197] to identify common modes. I used default settings including LargeKGreedy algorithm with 2,000 random permutations. The software uses DISTRUCT to coordinate membership colours. Results were visualised using `ggplot2` [198] in R v.3.5.1 [199].

I investigated further the specific KhoeSan contributions to the SAC using a larger set of populations of KhoeSan-related ancestry [from 93] (reducing the SNP density to 48,940) following the same procedure above.

I estimated the F-indices of pairwise distances between populations as implemented in Eigensoft 7.2.1 [165].

4.2.4 CHROMOPAINTER, fineSTRUCTURE and NNLS

I employed haplotype phasing and local ancestry determination utilising linkage disequilibrium from the dense SNP data to produce informative haplotype chunks for ancestry characterisation as described in Section 3.1.4.

Phasing

The CHROMOPAINTER-fineSTRUCTURE (CP-fS) pipeline requires phased data. The process of phasing estimates haplotypes from genotyped data thus re-creating diploid information for individuals [200], [201]. Data were phased using the improved Hidden Markov Model implemented with Segmentation Haplotype Estimation and Imputation Tool, ShapeIT v.2 [201]. I used the HapMap Human genome build 37 recombination maps from the ShapeIT website. The effective population size, N_e, was set to 1500 and the main number iterations was set to 50, the burn-in stage set to 10 and the prune stage was set to 10. Default settings were retained for other parameters (`-effective-size 1500 -burnin 10 -main 50 -prune 10 -states 100 -window 2`).

Haplotype Coalescence

Phased data were processed to coalesce haplotypes as implemented in CHROMOPAINTER. I estimated the number of haplotype-chunks copied and total genomic length copied by each recipient from each possible donor individual.

The chromosome painting was performed using three different set-ups:

1. SAC individuals copied from other SAC, to reconstruct the pattern of within population ancestry relationships;

2. GR individuals were allowed to copy only from other GR individuals; by doing so I characterize the pattern of shared ancestry among GR individuals to identify appropriate clusters used to discuss populations in this chapter;

3. SAC individuals copied from GR individuals, to identify the potential source populations.

The matrix of haplotype-chunks copied referred to as the co-ancestry matrix (CoAM) was used to identify clusters within the SAC and GR dataset to discuss possible structure. This is based on similar copying profiles and is implemented in the clustering algorithm fineSTRUCTURE (fS) [83]. The process is described in section 3.1.6.

GR - GR CP-fS

Haplotype Coalescence and Sample Clustering

To identify GR source-populations with low heterogeneity, I performed the CP-fS pipeline on the GR dataset. This was done as the data may include genetically equivalent sample sets from different sampling events labelled differently or include sub-structure within the *a priori* population groups.

I employed the default settings in CHROMOPAINTER unless otherwise specified. Estimation of chunk count and lengths for the CoAM requires two scaling parameters; recombination switch rate and the mutation rate [83], [84]. The switch rate accounts for necessary switching between donor populations when identifying ancestry along a chromosome in a recipient. The mutation rate is an emission rate which accounts for the necessary number of errors needed to best match a recipient's haplotype to that of the available donors. To simplify the analysis these parameters were estimated from a diverse subset of six donor populations; Ju_Hoansi_North (4), YRI (Yoruba, Africa) (50), UK (United Kingdom, Europe) (25), CDX (Dai Chinese, China) (50), BEB (Bengali, Bangladesh) (50), MAS (Malay, Singapore) (50), and a randomly selected set of five chromosomes; 1, 6, 15, 20 and 22. I supplied the initial values (100 N_e and 0.5 μ) and used 15 iterations of the expectation-maximization algorithm in CHROMOPAINTER (options specified as `-a 0 0 -n 100 -m 0.5 -i 15`)). Previous work with CHROMOPAINTER has demonstrated that results are stable even with a 10-fold change in the switch rate [88]. Parameters were visually inspected for convergence and estimated as an N_e of 330.34 and μ of 0.0016.

When combining results across chromosomes, I estimated the 'c' correction parameter (see section 3.1.6) to account for non-independence between chunks in practice and the use of the expected rather than the observed number of chunks copied in the likelihood estimation [83]. I retrieved a c-value of 0.0919. I ran fS on the chunk count CoAM output from CP. The default settings in fS were employed unless specified otherwise. I used 4 million iterations for the Hidden Markov Chain, I remove half for burnin and retain 500 samples (`fs fs -m oMCMC -x 4000000`

`-y 2000000 -z 4000 -t 100000`). The number of hill-climbing iterations were set to 100,000 for the selection of the best tree.

Cluster Evaluation

To improve upon the final assignment of individuals to clusters as output by fS, I used the marginal posterior distribution across MCMC runs for each individual's assignment to a cluster. This was done for every 10 000 iterations using the built-in option in fS (`-T 1 -K 2`).

I compared the results across trees from independent runs using the average and standard deviation of pairwise coincidence of individuals across trees.

Further refinement of groups was performed based on a randomly chosen tree. This is permissible because trees were compared before proceeding and are known to be topologically consistent (see section 4.3.2), at least at higher levels.

I reduced the number of groups/clusters on the tree by performing pairwise Total Variance Distance (TVD) estimation and merging leaves on a branch for which the TVD values were below a threshold value of 0.021 as guided by our evaluation of the change in minimum TVD at different heights (Supp. Figure A.1).

While the TVD estimation has been performed before and used to regroup fS-inferred clusters, the evaluation of the minimum TVD threshold value is novel. I describe the decision-making process below.

The TVD values are estimated at varying heights on a tree and using the minimum TVD value at a height, I calculated the difference from the $minTVD$ at the preceding height. An increase in $minTVD$ is considered an increase exceeding $3 \times \frac{1}{H} \sum_1^H |d_h|$ where d_h is the change in $minTVD$ from heights h to $h-1$ and H is the final height. These changes in clustering structure reflect a large jump in the minimum TVD and thus loss of substructure. I wish to minimise the number of significant losses of sub-structure.

Simultaneously, I tried to minimise the number of clusters with few individuals (e.g. five individuals), resulting in a trade-off between retaining the lowest $minTVD$ and the fewest number of clusters with few individuals.

I evaluated how well this cut performed at grouping co-occurring individuals using the pairwise coincidence averaged across independent chains within each cluster. I validated that clustered individuals had meaningful groupings with regards to regional origin and/or linguistic characteristics.

The fS-clusters were named to reduce the number of assumptions needed when discussing a cluster's identity or representatives. The names consist of three parts as follows 00_[Area]_[linguistic group]_[*a priori population*]. Where "00" is a sequential index and "Area" is a description of the region in which the samples are distributed, defaulting to the country of sampling when samples are not widespread. "Linguistic group" is a description of the linguistic families present in the cluster. The languages are identified by available information where possible, otherwise the majority languages for the country of sampling was used. "*A priori population*" is used when the cluster has three or fewer *a priori* subgroups. A detailed breakdown of the geographic distribution of each fS-cluster can be found in Figures A.9 -A.18.

I estimated the genetic distances between clusters using F-indices implemented in Eigensoft [165] and total variance distance (TVD) as implemented by Busby, Band, Si Le, *et al.* [177].

The fS clustering produced a dendrogram which did not have Africans as an outgroup. To confirm that this distortion is a consequence of under-representation of African populations, I performed a linear correlation of the F_{ST} values against the average length of genome copied as output by CP. Here I am expecting to find that genome length copied is least informative for the African groups such that there is no relationship of the F_{ST} distance and genome length copied, thus the position in the fS dendrogram would be poorly informed. I included the Africans and a subset of non-African populations as recipients, respectively, and include only the non-African GR populations as donors.

SAC CP-fS

I investigated the substructure within the SAC population by performing the CP-fS pipeline on the SAC dataset without other populations, allowing individuals to copy from each other. This was done as described for the GR dataset above and similar to that of [88]. The CoAM was estimated using the two scaling parameters estimated for 12 EM iterations on 30% of the individuals and all of the chromosomes (`-a 0 0 -n 100 -m 0.5 -i 12`). Parameters N_e and μ were visually inspected for convergence and estimated as 139.83 and 0.0012, respectively. I inferred a 'c' value of 0.468. A maximum concordance tree was produced and interrogated as described above. I reduced the number of groups by performing the pairwise *TVD* estimation with a threshold value of 0.032.

As pointed out by previous researchers using this dataset, there is no information on the specific ethnic identifiers preferred by the participants thus there is a possible confound of the "Coloured" identity being used as a blanket term for several ethnicities each of which may have different genetic ancestries (Griekwa, Nama, \Xam, Cape Malay). The fS run on the dataset may highlight these 'outliers', if present. Further to this there may be structuring within the Cape "Coloured" community as a consequence of the history in the Cape. I estimated the genetic distances between clusters using F-indices implemented in Eigensoft [165] and total variance distance (*TVD*) as implemented by [177].

Identifying Ancestral Contributions to the SAC

Mixture Model

The SAC populations were painted by the GR samples and considering the clustering affiliations obtained for both sets. The copying vector of both the SAC individuals copying from the GR data and the copying vectors of the GR individuals copying from each other were used to perform a Non-Negative Least Squares Regression [97]. The NNLS analysis summarises the copying vector of each SAC recipient as a set of contributions from the GR clusters identified by accounting for the fact that donor clusters are themselves compositions of ancestries

from other donor clusters. I estimate jack-knife standard errors over chromosomes using a weighted block jack-knife [179], weighting by chromosome SNP counts. The NNLS and jack-knife analyses were implemented in R [199].

Relationship between Genetic and Genealogical Ancestors and the Exclusion of Signals

Typically for CHROMOPAINTER analyses no cut-off criteria are applied beyond a jack-knife or bootstrap estimation of error (e.g. [88], [97], [177] from which all error ranges including zero can be used to eliminated some ancestries and the copying profile rescaled to sum to one. This may be a viable procedure when considering simple admixture scenarios with few sources but in the case of a complicated, recent multi-way admixture this may result in many spurious source populations being identified.

In the case of this thesis, I focus on admixture within the last 3 - 15 generations (~90 - 450 years ago or 1510 - 1870 CE) so my interest is in identifying recent ancestral contributions. Recent incidental contributions are less likely to have been lost through drift and thus a stricter criterion may be justifiable to overlook such signals. I consider ignoring ancestral contributions which are not relatively prevalent in the population and contributions which are at low proportions in any individual when present.

To define the criteria for an 'incidental' contribution, I consider records of slave arrivals to South Africa [21], [133], [152]. From this I identify three sets of populations based on their representation, this includes groups with "few arrivals" (1-20 individuals; e.g. South Asians Nairo, Oellada), groups with a moderate number of arrivals, "moderate arrivals" (21-99 individuals; e.g. Polia, Chego) and a groups with "many arrivals" (100 and upward; e.g. Madagascan, Dutch, French).

The "few arrivals" will likely have drifted out or will be ruled out as noise due to their low prevalence.

The "many arrivals" group are undoubtedly represented in the genomic data of the present South African "Coloured" (SAC) as they would constitute the majority

of ancestry. I here discuss the lower values of the "moderate arrivals" group to decide where to draw the line for exclusion.

I build the arguements below based on the explanation of genealogical and genetic ancestry from Graham Coop's research group (UC Davies) webpage [202]. Firstly, I should relate the number of genealogical ancestors to the likely number of genetic ancestors. Genealogical ancestors do not all contribute to the genetic ancestry of an individual as the recombination and independent assortment process results in genetic drift, the stochastic loss of genetic material with subsequent generations. To quantify contributions from ancestors I follow the model described by Donnelly [73]. The work estimates the number of haplotype chunks identifiable to a specific ancestor in the genealogy. The size of the chunk is not regarded and I assume a single recombination event per chromosome per generation and no limits on detectability are considered for simplicity.

Given a set number of generations in the past, k, a diploid individual is expected to have 2^{k-1} ancestors per haploid genome (Figure 4.5). The number of genealogical ancestors thus increases exponentially with the time. The number of genealogical ancestors can be estimated from the equation below (from [73]).

$$2^{k-1} \times 1 - exp(-\frac{22 + 33(k-1)}{2^{k-1}})$$

Here I assume 22 autosomal chromosomes of interest and that each recombination event introduces new haplotypes chunks which are either a chunk of a chromosome or an entire chromosome. The formula $22 + 33(k-1)$ gives the number of haplotype chunks present in the chromosome. The probability of contribution from an ancestor can then be estimated from a Poisson distribution considering a mean number of haplotype chunks contributed at generation k to be as below:

$$\frac{22 + 33(k-1)}{2^{k-1}}$$

From this, the expected genealogical and genetic ancestors correspond for the first seven generations (128 ancestors) after which there is an increasing discrepancy as genetic ancestors increase only linearly for large k values.

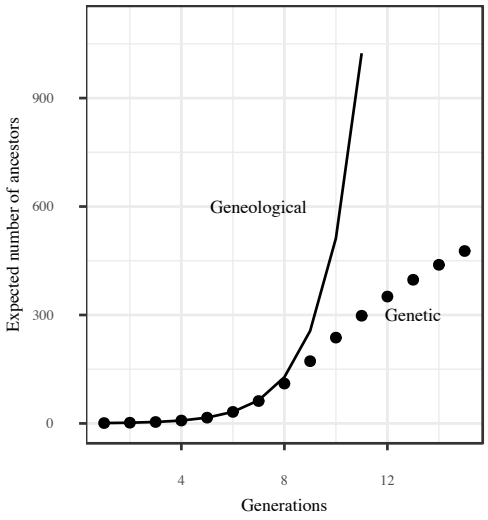


Figure 4.5: Increase in the number of genealogical and genetic ancestors with generations in the past.

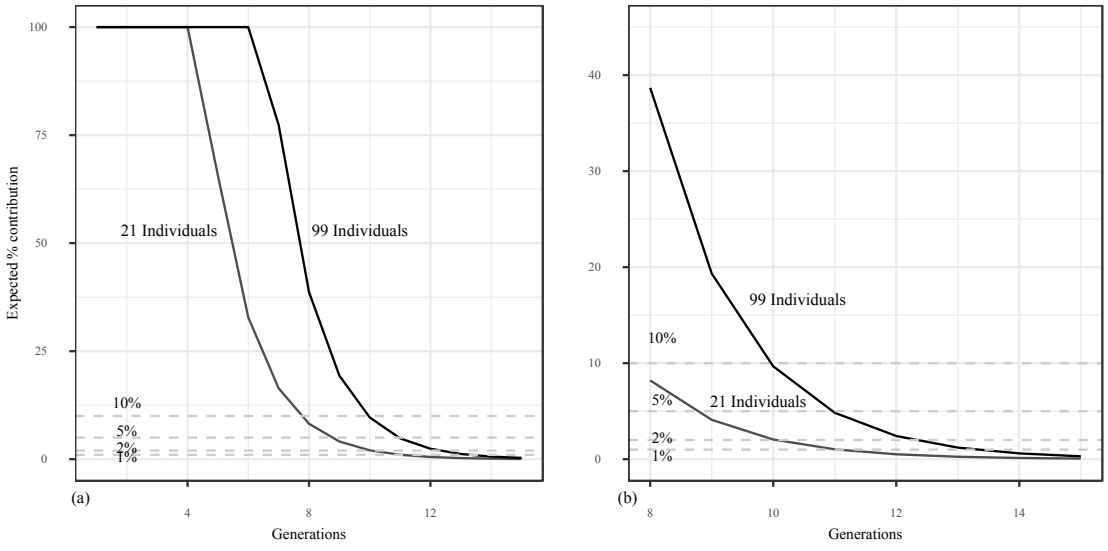


Figure 4.6: The change in the expected genetic contribution from a set of ancestors (21 or 99 individuals) at varying dates of arrival. (a) 1-15 generations ago and (b) a close-up of 8-15 generations ago . Estimates are of a diploid genome. Estimates assume as many of the arrivals as possible are *de facto* ancestors.

For the sake of developing the argument, let us assume all genetic ancestors contribute equally. Thus, the genomic contributions can be divided into vector A of length $2^k \times 1 - exp(-\frac{22+33(k-1)}{2^{k-1}})$ with each element representing the contribution from any particular ancestor. More generally these contributions will be a fraction of the genome and with equal contributions, each contribution is $\frac{100}{A}\%$.

Using the above information, I can predict the expected genetic contribution from those individuals in the "moderate arrivals" group. I do this by calculating the number of expected genetic ancestors given the observed genealogical arrivals and then sum up their contribution for the population ancestry detectable.

For the upper bound of the "moderate arrivals", 99 individuals, the expected contribution drops below 10% from generation 9 and below 1% from generation 14. For the lower bound, 21 individuals, at generation eight the expected contribution is below 10% and from generation 11 (~1630 CE) it drops below 1% (Figure 4.6). Thus, to exclude the possible signals from contributions from the 'few arrivals' category arising prior to the Dutch settlement in 1652 CE a lower limit of 1% may be necessary. This estimation assumes that all the individuals in the group reproduce. If some members don't reproduce, the overall contribution of the group would be lowered. Thus, these estimations are a maximum expected contribution from a "moderate arrivals" group.

A further complication lies in redundant ancestors. The early Cape Colony had a rapid increase in population size largely through reproduction despite an initial small founder population [21]. Thus, many genealogical ancestors are repeated even as recently as 10 generations ago [see 61]. As such the number of genetic ancestors too is lower and the relative contribution of each genetic ancestor increases. Inbreeding may result in higher contributions than expected. Overall, the signal from this would be reduced genetic diversity related to this ancestral group and overestimation of the number of arrivals. This is not accounted for in my analyses.

For admixture events around the period of interest for our study, the maximum expected contribution from the "moderate arrivals" groups lies between 1-10% of

the total genome. I therefore proposed the use of a lower bound of 1% on the NNLS jack-knife errors jointly with a prevalence in the population of ~5%.

This should eliminate very recent (and thus low prevalence) admixtures as well as possible noise. I acknowledge here that there may be actual signals which too are eliminated through this process.

Compositional Analysis of NNLS-inferred Ancestry

I explored the influence of various ancestries on the substructure of the SAC dataset by performing a principal component analysis on the ancestral composition of each SAC cluster.

The ancestral components identified above are a form of compositional data (CoDa) as they are restricted to sum to 100%. I need to account for the nature of the resultant compositional data because of the inherent risk of spurious correlations between components when investigating the characteristics of ancestral composition. I therefore follow the suggested best practises [203]–[205]. The primary adjustment is a log-ratio transformation which removes the constraints on the CoDa allowing the values to be used with a slightly broader range of analyses [203].

I used a centred log-ratio (CLR) transformation on the ancestry composition values after the imputation of zero values to a low non-zero value using an EM-based parametric imputation as implemented in `robCompositions` R package [206]. I used the `impRZilr` function setting the lowest detectable ancestry at 1×10^{-5} and 0.01 as the convergency criteria under a linear regression model (options `dl=(1e-05*number of columns), eps=0.01, method="lm"`). The CLR transformations are based on the division of each component by the geometric mean of the components in an observation-by-observation basis, implemented with the `pcaCoDa` function. The PCA was performed using the `robpca`.

To test for a significant difference between SAC clusters in terms of their composition, I performed a multivariate analysis of variance using the `stats` package [199] in R. To identify which principal components, and thus ancestries, influenced differences between clusters, I performed a post-hoc Tukey Honest

Significant Difference test to identify which Principal Components were different between SAC clusters.

4.3 Results

4.3.1 Global Contributions to the South African "Coloured" Summarised by Principal Component Analysis and ADMIXTURE Proportions

I evaluate the global contributions to the SAC by examining the genetic structure of the GR and SAC samples jointly. This also validated the merging and quality control processes. The genetic structure within the GR dataset as detailed through PCA and ADMIXTURE analyses concurred with what has been previously reported in the original papers for each dataset. The first five principal components (PCs) account for more than 91% of the genetic variation (Supp. Figure A.2). I found that the SAC individuals show evidence of genetic influences from several world regions as reported by earlier work [99], [110] (Figure 4.7 and GR plots in Supp. Figures A.3 - A.8).

The SAC forms a cloud stretching between global populations, the greatest variation was seen between Eurasians and Africans along PC1 and the spread between KhoeSan and Non-KhoeSan Africans along PC4 (Figure 4.7). There was less spread along PC2, which described variation differentiating Eastern and Western Eurasians and notably less variation along PC5 and 6 which described Eastern Asian and South East Asian ancestry variation. The latter result concurs with earlier reports of lower variation in contributions from these regions [185]. The position of the SAC on PC5 and PC6 overlaps with European, Near Eastern and South Asian populations suggesting a significant influence from these contributions.

The positioning along PC3, describing variation along a South Asian - Central Asian cline, showed a slight skew toward the positions of other South Asian populations (see Figure 4.7).

When discussing ADMIXTURE Bayesian cluster assignments, I refer to assignment probabilities explicitly as assignment probabilities and as 'ancestral components' to reflect that shared group assignment may reflect shared ancestry.

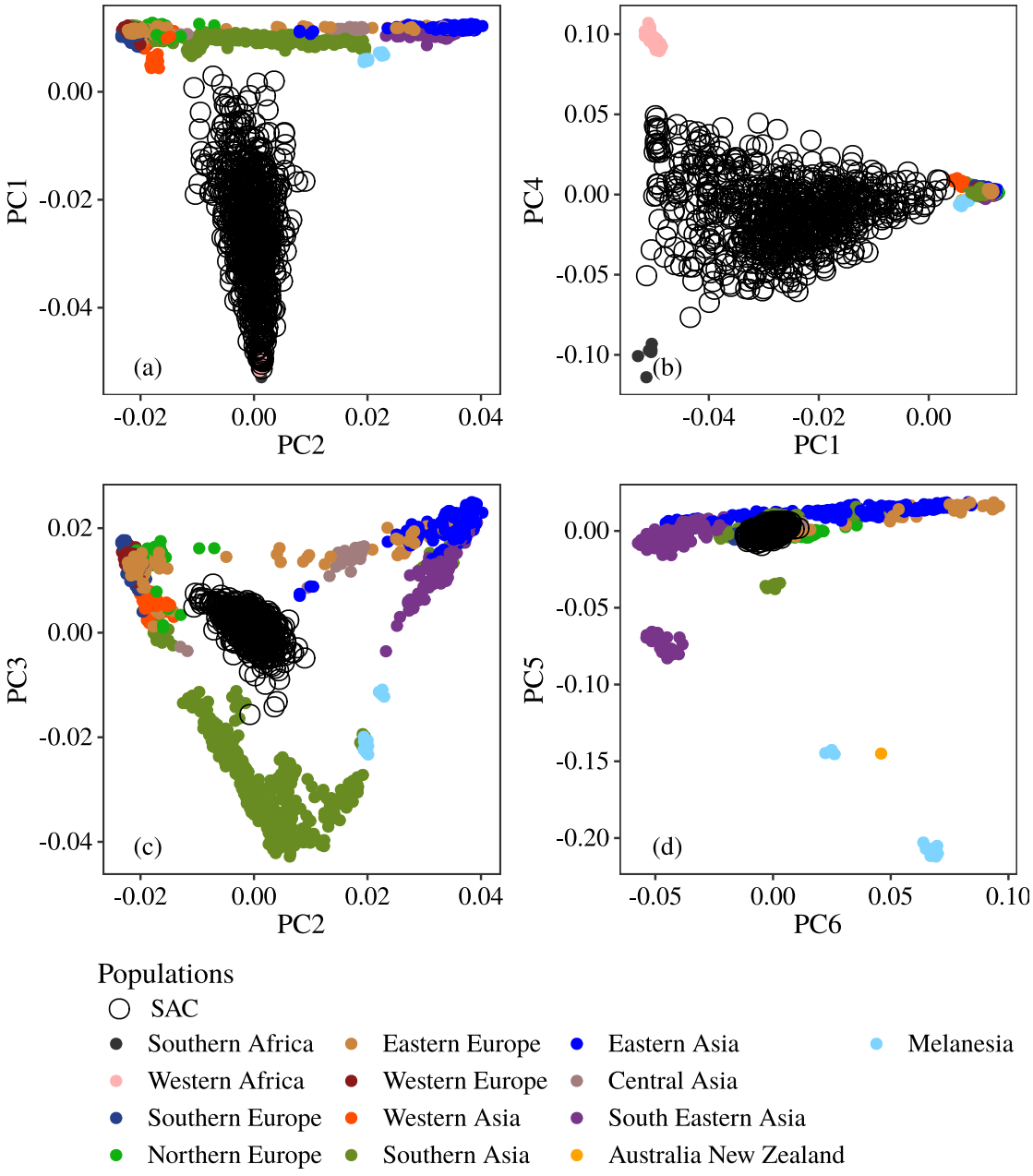


Figure 4.7: Global scale PCA focused on the South African "Coloured" (SAC) individuals. Shown (a-c) are the first 6 PCs with the SAC in black and the remaining populations coloured by global regions.

The decline in cross validation errors appears minimal past K = 10, suggesting the best K value in that proximity, however K = 15 produced the lowest value (Figure 4.8). The exact sequence of component separation seen differed from that published in previous work, which is expected since the samples included here differed (e.g. [38], [97], [99], [100], [115]. However, the division between global regions appeared consistent with published work.

The earliest split occurs between African and non-African populations (K = 2) (Figure 4.9), which is followed with the Eastern Eurasian-Western Eurasian split (K = 3) and then a Southern, South-East and Central Asian component from the Eastern Asian component (K = 4). At K = 5, the KhoeSan related ancestry in the ≠Khomani, Ju|'hoan separates from non-KhoeSan African ancestry. This approximately mirrors the first five principal components.

Within the SAC data, as early as K = 5 there is evidence of a shared component with East Asia (shown in dark blue), South Asia (green), Europe (black), and two groups of Africans; non-KhoeSan (peach) and KhoeSan + pygmies (teak) (Figure 4.10). At K = 15, many of the GR populations were characterised by several ADMIXTURE components which makes identification of specific proximal ancestry in the SAC unclear (e.g. contributions from BEB (Bengali) or MAS (Malay) could produce similar ADMIXTURE profiles). To facilitate discussion, I describe each component by the GR population in which the component is best represented. The major components observed in the SAC at K = 15 are in common with Juu|'hoan (KhoeSan) (mean 36% in the SAC), YRI (non-KhoeSan African) (20%), Ireland (European) (12.5%), Paniyas (Dravidian Indian) (6.6%) while several other components contribute lower proportions; Dai Chinese (4.5%), Bedouin (Levant) (3.4%), BedouinB (Levant) (3.8%), and a component present in the Kalash, Makrani, Balochi and Brahui (Pakistan) (3.3%). An additional four components contributed slightly over 1%.

The top four components account for 75% of the total ancestral assignment within he SAC despite notable variation in ancestry. In particular the KhoeSan-related ancestry was ubiquitous (range 1% - 84%). I considered a broader KhoeSan

dataset which included Taa, Khoe-Kwadi and K'xa langauage groups from [93] (48 940 SNPs). Here I found the largest component in the SAC was shared with Taa and Khoe-Kwadi speakers from Namibia and North-Western South Africa as compared to the K'xa from Northern Botswana and Namibia (Figure 4.11).

Neither the PCA nor the ADMIXTURE analyses suggested any obvious discontinuities in ancestry within the SAC, indicative of the high variability in ancestry proportions.

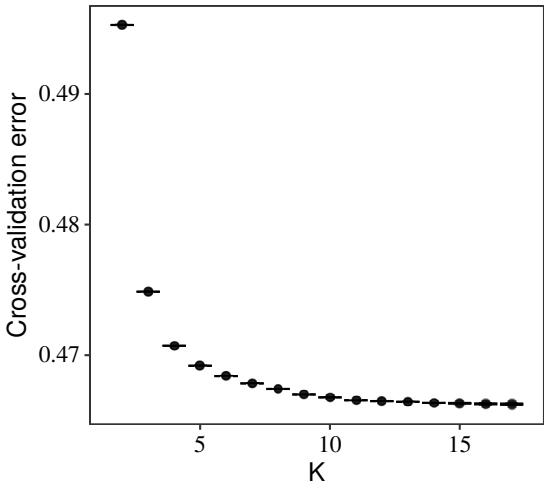


Figure 4.8: Cross validations errors for ADMIXTURE runs for the GR-SAC dataset at K = 2...17

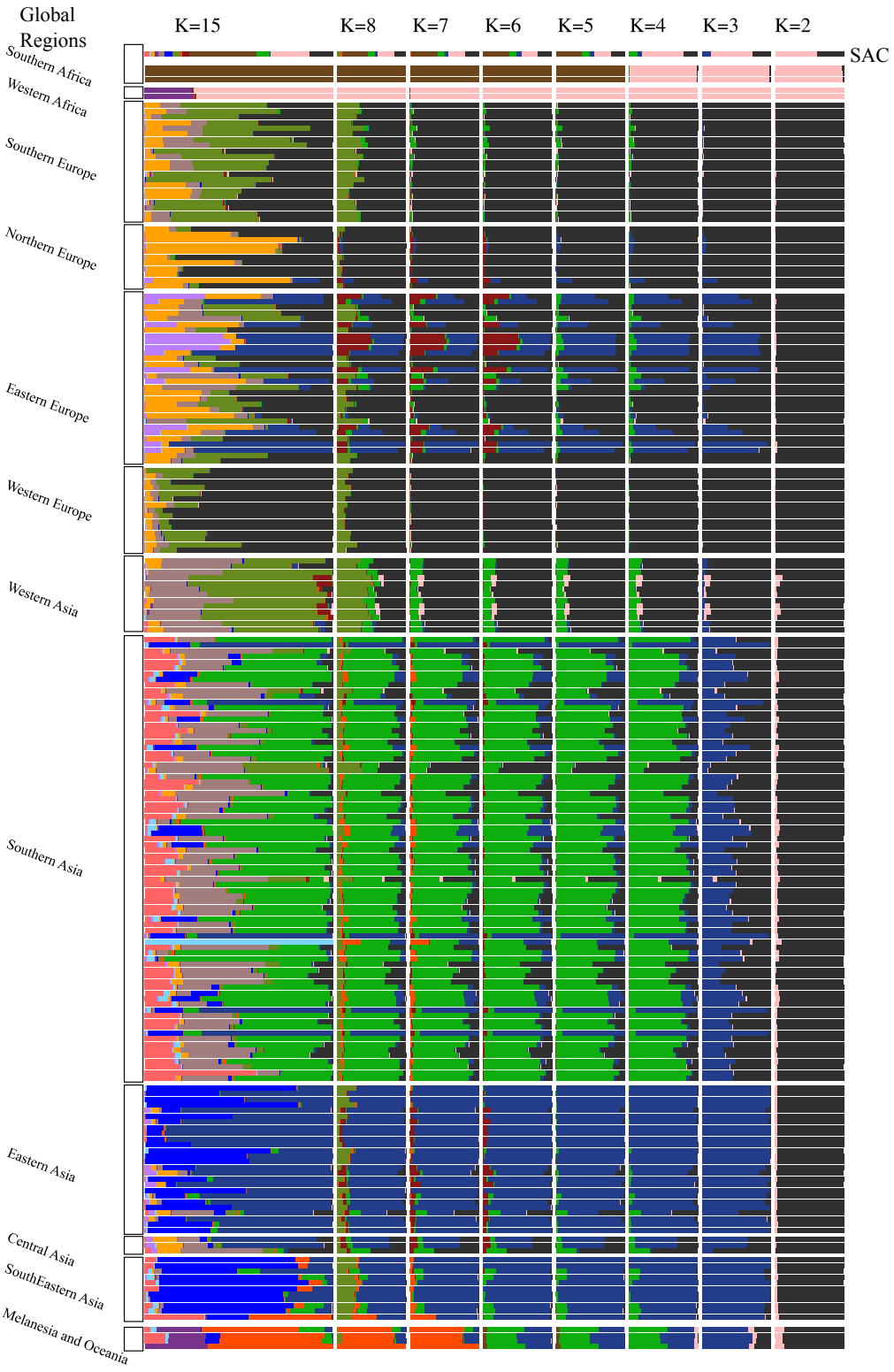


Figure 4.9: ADMIXTURE run for the GR dataset at K = 2...8 and 15. Bars indicate population averages. Labeled by UN global regions.

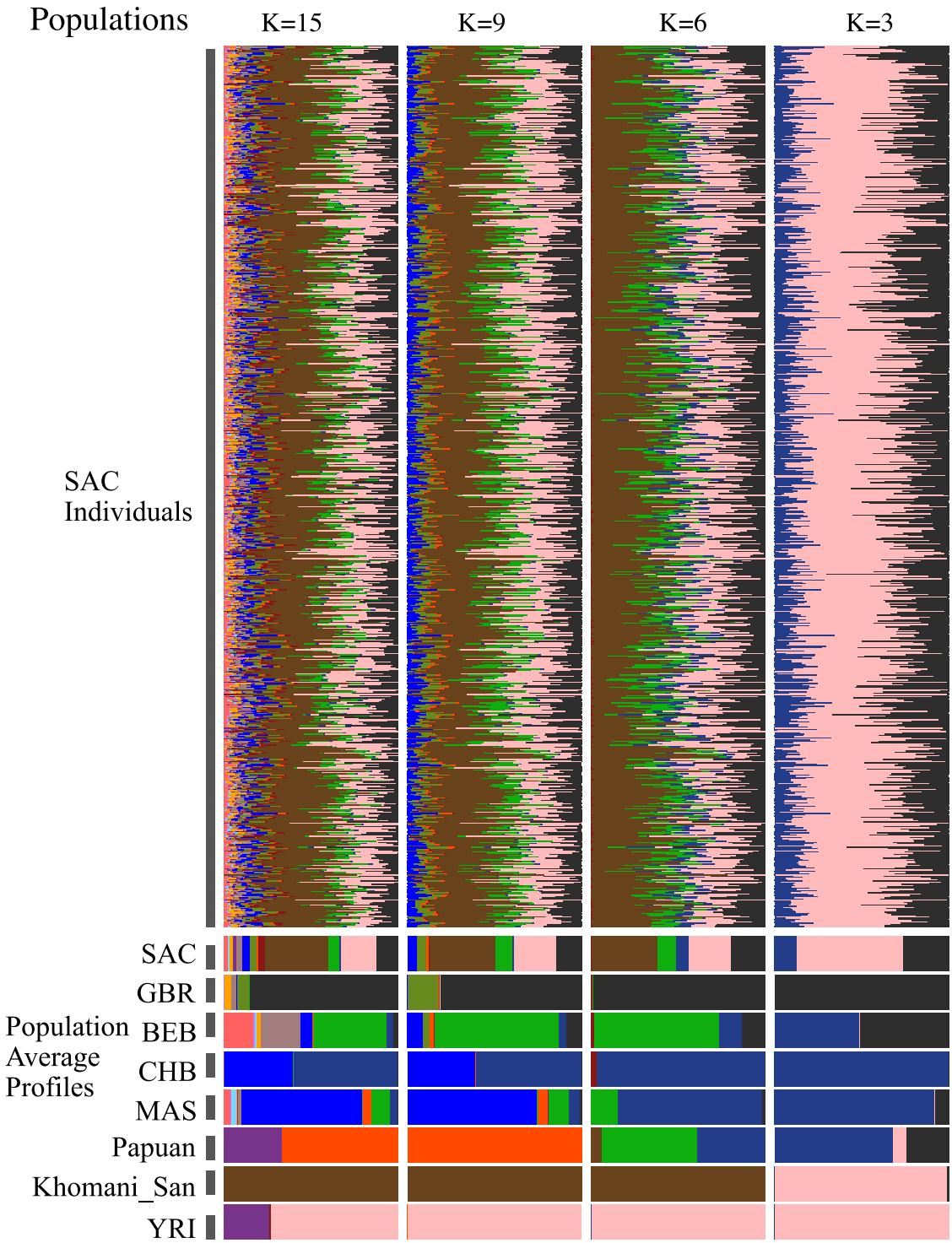


Figure 4.10: ADMIXTURE profiles (K = 3, 6, 9 and 15) for the SAC + GR dataset. SAC profiles plotted individually above a set of *a priori* populations from the GR averaged over samples. Abbr. SAC - South African Cape Coloured, BEB - Bengali from Bangladesh, CHB - Chinese Han from Beijing, China, GBR - British from Great Britain, Khomani_San - ≠Khomani, MAS - Malay from Singapore, YRI - Yoruba form Ibadan, Nigeria.

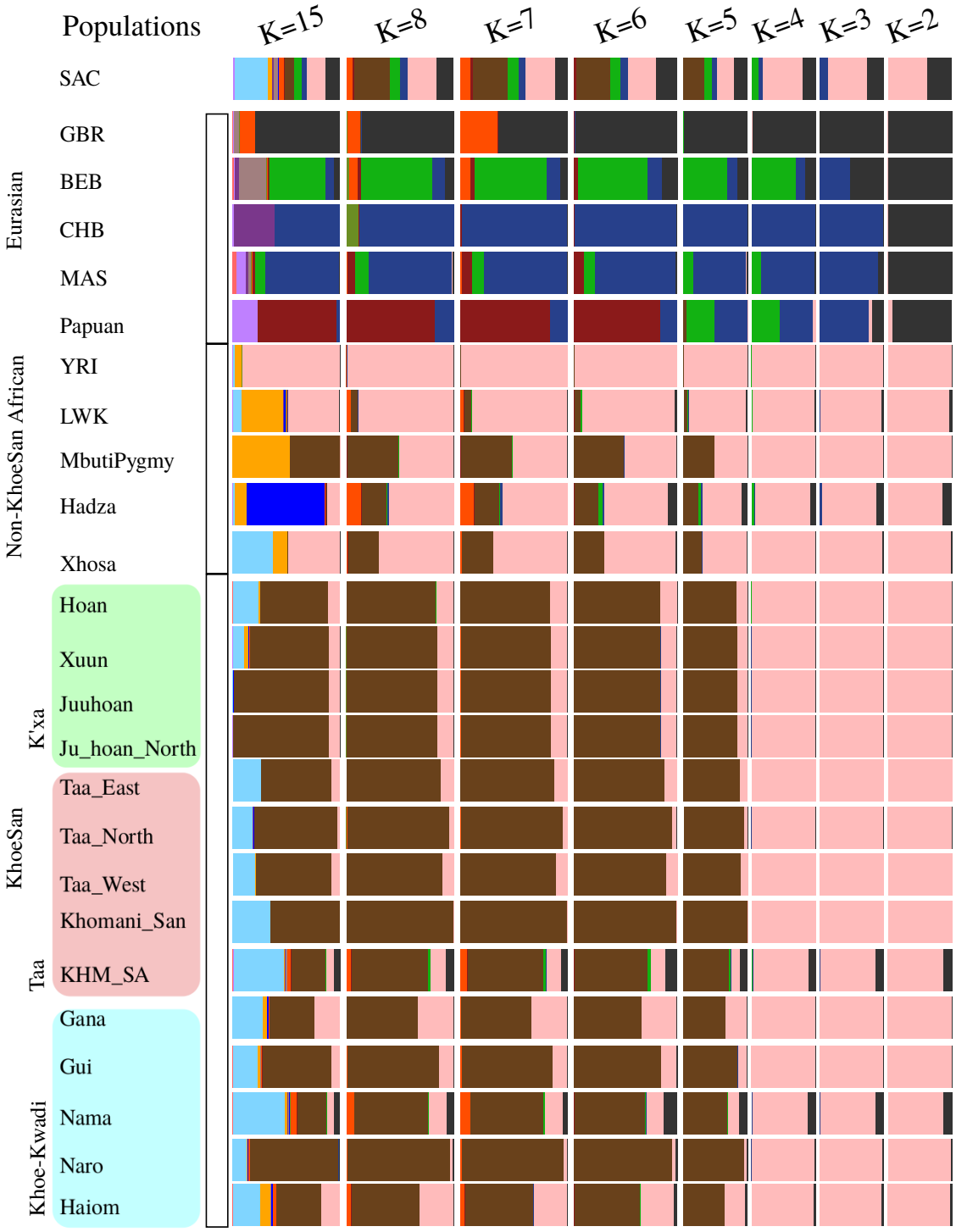


Figure 4.11: ADMIXTURE profiles (K = 2 . . . 8, 15) from GR + SAC dataset merged with a broader KhoeSan dataset. KhoeSan data from [93]. KhoeSan samples have language groups indicated. Results averaged across *a priori* labels and a subset of the populations are shown. Abbr. SAC - South African Cape Coloured, BEB - Bengali from Bangladesh, CHB - Chinese Han from Beijing, China, GBR - British from Great Britain, MAS - Malay from Singapore, YRI - Yoruba form Ibadan, Nigeria, LWK - Luhya from Kenya, KHM_SA - ≠Khomani from South Africa.

4.3.2 Haplotype-based Ancestry Inference from Chromosome Painting

Evaluating SNP Density per Haplotype for Ancestry Estimation

I used CHROMOPAINTER to characterise local ancestry and thus identify the most closely related sources of ancestry for each recipient individual. The process was repeated in three 'recipient - donor' arrangements; GR - GR, SAC - SAC and SAC - GR.

The per individual chunk counts produced were largest for the GR - GR and SAC - GR run, indicating smaller haplotypes than produced by the SAC - SAC run (Supp. Table A.5 and A.6 vs. A.7). The GR - GR values were comparable to those found for investigations considering many, differentiated populations (e.g. for admixture across continental Africa [177]). The average chunk lengths copied in the GR - GR and SAC - GR run (~0.4 and ~0.3 cM, respectively) were in line with previous work (average 0.3-0.4 cM chunk sizes in [82], while the SAC - SAC values were double these (~0.75 cM). The average haplotype-chunk length copied from any SAC individual and the largest chunk copied were both larger than those observed in the GR - GR and SAC - GR runs.

From these values I could show that all CP arrangements produced SNPs per haplotype chunk (SNPs haplotype-chunk^{-1}) greater than seven, approaching values used in other work [e.g. 13 - 32 SNPs in 97]. The number of SNPs haplotype-chunk^{-1} varied with the chromosome size and the available SNPs (Figure 4.12). Lowest SNPs haplotype-chunk^{-1} were observed on shorter chromosomes and chromosomes with lower SNPs counts per bp. Chromosome two and three were clear outliers in the SAC - SAC run, (median 3.6 and 13 SNPs haplotype-chunk^{-1}, respectively) but this was not seen in the SAC - GR run nor the GR - GR run (Figure 4.12).

The lowest SNP densities were found in the SAC - GR run (~5.6 - 9.2 SNPs haplotype-chunk^{-1}) and the greatest in the SAC - SAC run (~20). Chromosome 19 is an outlier having consistently low values, ~3.8 SNPs haplotype-chunk^{-1}. This chromosome is known to have the highest gene-density [7] and open proportion of

chromatin [207]. Recombination hotspots are less concentrated locally compared to other chromosomes [68]. This means that the distribution of recombination events are more spread across the chromosome, and segments copied from a GR population may be shorter on average as IBD tracts are broken up more rapidly. The SNP density per haplotype is not, however, unexpected compared to the available SNPs on Chromosome 19.

The SAC - SAC run values are expected to be large as IBD tracts should be longer between individuals with shared recent admixture, indicative that the SAC - SAC run recovers more recent recombination breaks than the SAC - GR run.

Forming Regional Population Clusters to Characterise Global Sub-structure

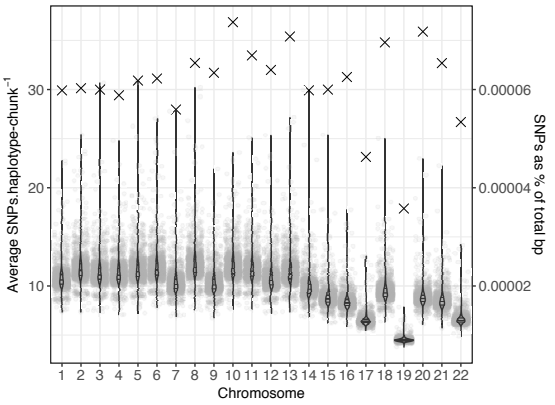
The fS-inferred clustering of GR data from three independent chains were convergent (Figure 4.13).

I recovered 175 clusters with an average of 12 individuals per cluster (range 1-152). To make clusters more interpretable, I performed a Total Variance Distance (TVD) cut with a threshold value of 0.021. This involved estimating a pairwise Total Variance Distance (TVD) between clusters, merging pairs of clusters on a branch for which the TVD values were below the chosen threshold. The threshold was determined by evaluation of the change in minimum TVD at different cluster resolutions (i.e. heights at which the clustering tree was cut) with the intention to maximise the number of individuals per cluster and maximise the number of clusters retained (Supp. Figure A.1).

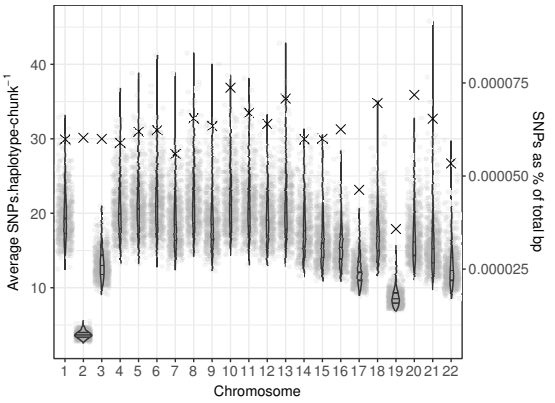
I retained 56 clusters without removing any individuals (Figure 4.14). Information on the sample composition of each cluster can be found in Supp. Table A.8.

The resultant tree had three major branches which reflected global regions, corresponding to Eastern Eurasia, South-Central Asia, and Western Eurasia.

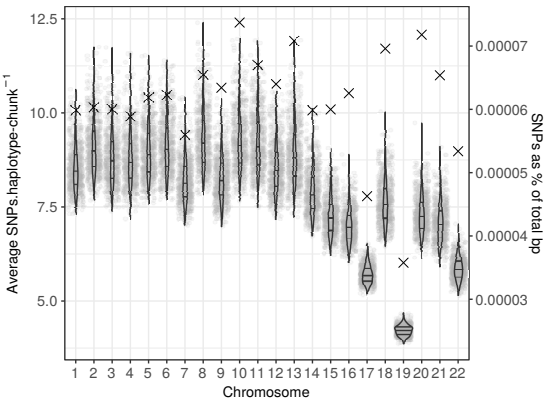
Clustered individuals had meaningful groupings with regards to regional origin and/or linguistic characteristics (Supp. Figure A.9 - A.22). Within each global region, major patterns of diversity structuring were recovered (Figure 4.15). I present an overview of the patterns below to provide context for later discussion. The



(a) GR - GR



(b) SAC - SAC



(c) SAC - GR

Figure 4.12: Average SNPs haplotype-chunk^{-1} recovered from CP-fS analysis. SNPs haplotype-chunk^{-1} for all recipient individuals per chromosome, averaged across copying sources, shown. Subplots (a-c) refer to the three different painting set-ups. The grey cloud of points are individual values. Median and IQR values indicated by black horizontal lines and range of values indicated by vertical extent of violin plots. The total SNPs per chromosome are indicated (X) as a percentage of the total bp positions currently sequenced [208].

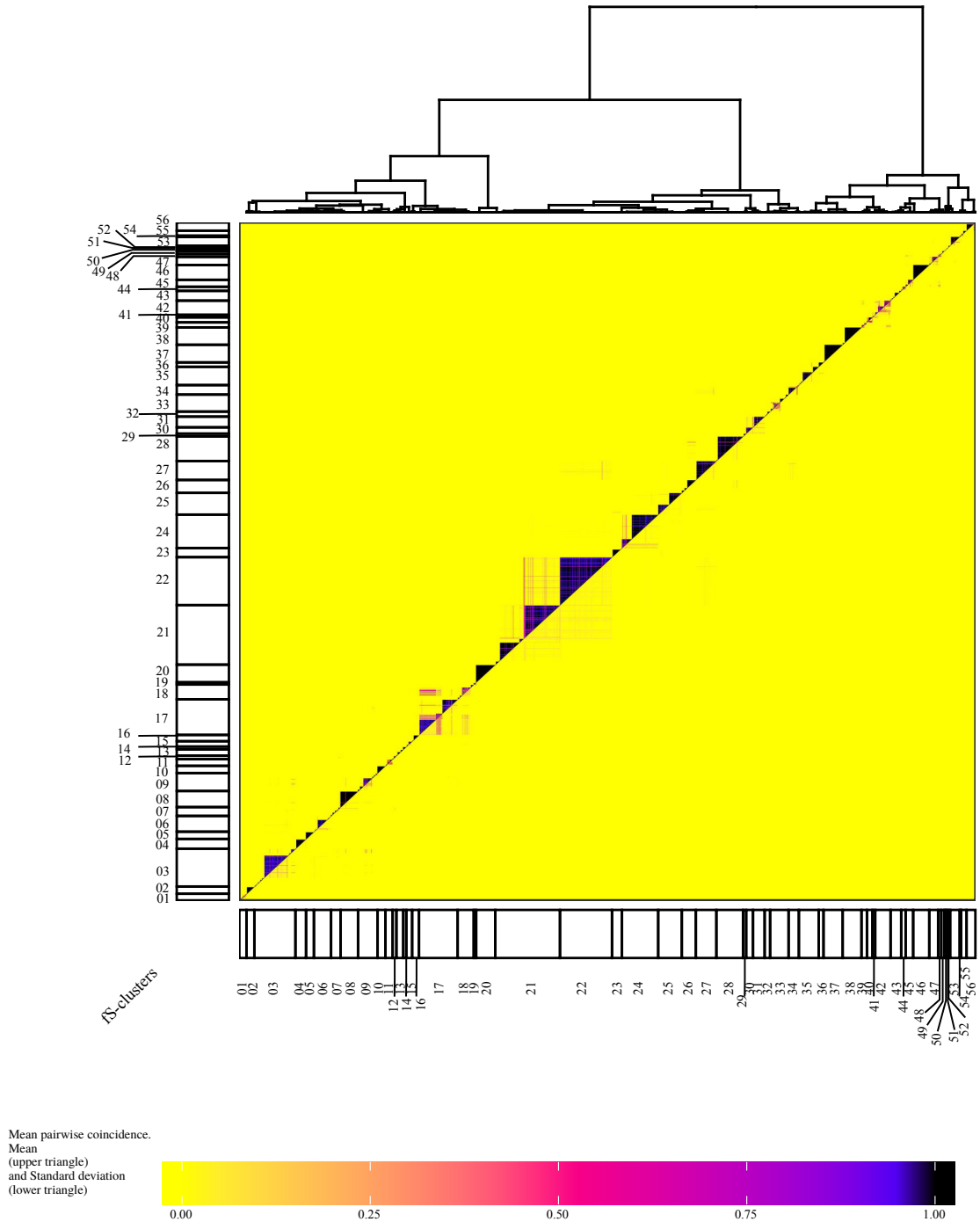


Figure 4.13: Tree structure convergence shown by pairwise coincidence values for the GR - GR CP-fS chains. Mean (upper triangle) and standard deviation (lower triangle) values from three independent chains indicated. Matrices sorted by the same randomly chosen tree (chain 0). Bars below and left of the matrix indicate the *TVD*-based clusters created. Pairwise coincidence plots of independent chains available in Supp. Figures A.23 - A.25

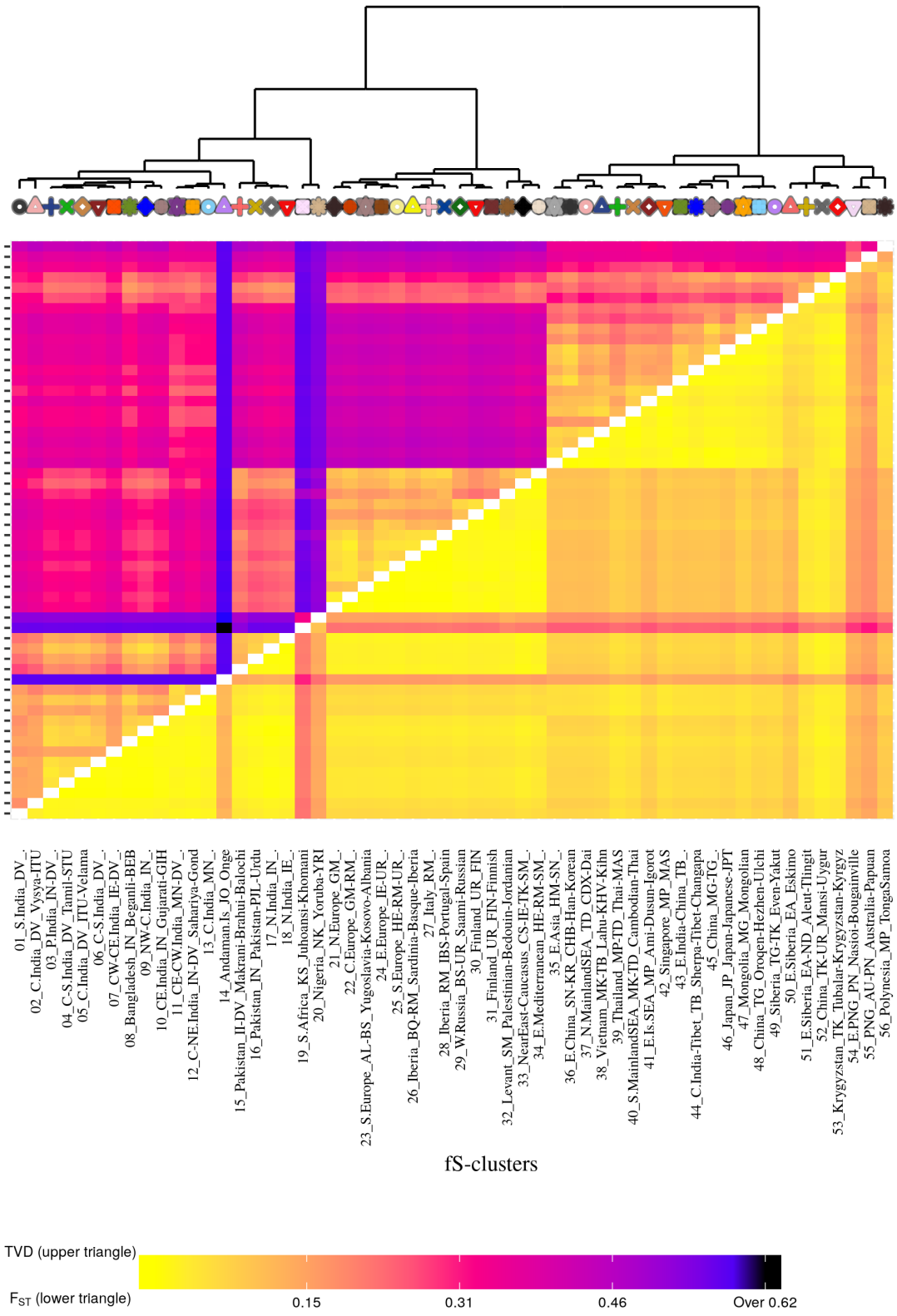


Figure 4.14: FineSTRUCTURE-inferred clusters, F_{ST} and *TVD* from GR data following a *TVD*-based cut. Pairwise distances estimated as Total variance distance (*TVD*) (top left) and F_{ST} (bottom right). The *TVD* cut threshold was 0.021 retaining 56 clusters. Cluster names indicated at the bottom and fS dendrogram above. Symbols correspond to labels Figure 4.15. Diagonal values set to white.

fS-clusters were named with three parts as follows: 00_[Area]_[linguistic group]_[*a priori population*]. Where "00" is a sequential index and *a priori* population labels are listed when fewer than four (see section 4.2.4) and when necessary for clarification.

The only two African fS clusters included in the GR dataset clustered within South and Central Asia (Figure 4.14). This is possibly a consequence of the under-representation of African samples [83]. To test if there was support for the usual clustering of African groups as an outgroup to Eurasians, I performed a linear correlation of the pairwise F_{ST} between GR clusters as estimated with SNP allele frequencies against the average length of genome copied from each fS cluster. I found that correlation coefficients estimated with African populations as recipients in the CP analysis had less explanatory power. While correlations for most fS clusters were significant, the adjusted r^2 was ~0.1 for 20_Nigeria_NK and 0.3 for 19_S.Africa_KS compared to 0.09 - 0.73 for the remaining groups, with 0.3 being the 25th percentile (Supp. Figure A.26 - A.28). Thus, for African samples, the copied lengths of genome are more uniform when copying from non-Africans and consequently less informative for fS. A similar distortion was found by Busby [171] when effectively reducing the representation populations for some regions to form "super individuals". Such "super individuals" act as "populations" by representing an average copying vector for the individuals included within, and thus reducing the computation time by lowering the number of individuals included. FineSTRUCTURE output is not meant to reproduce a phylogeny [83] so this deviation is not a concern. Based on both *TVD* and F_{ST} values, the African data are the most divergent from the rest of the dataset, supporting the traditional clustering of Africa as an outgroup to other populations.

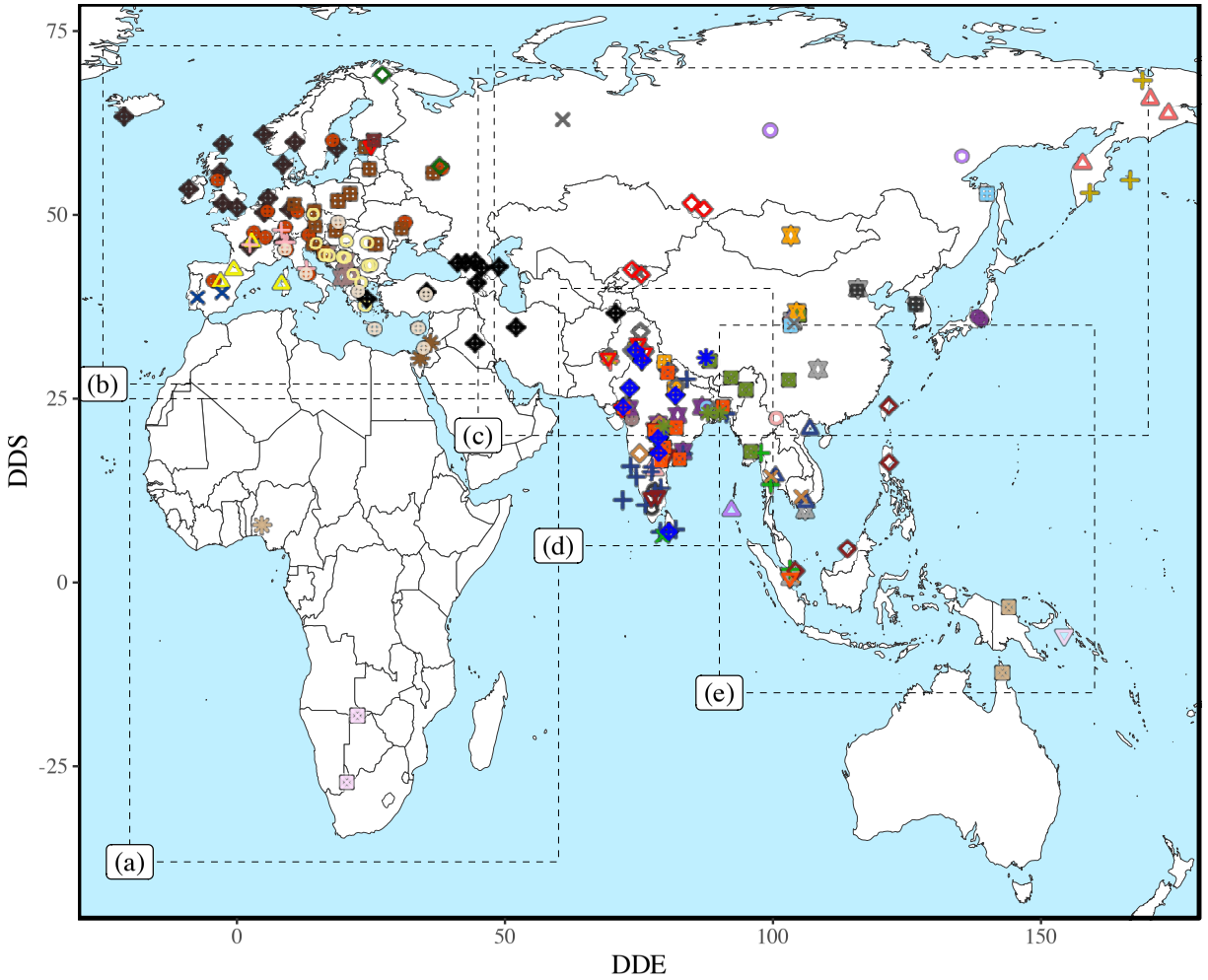
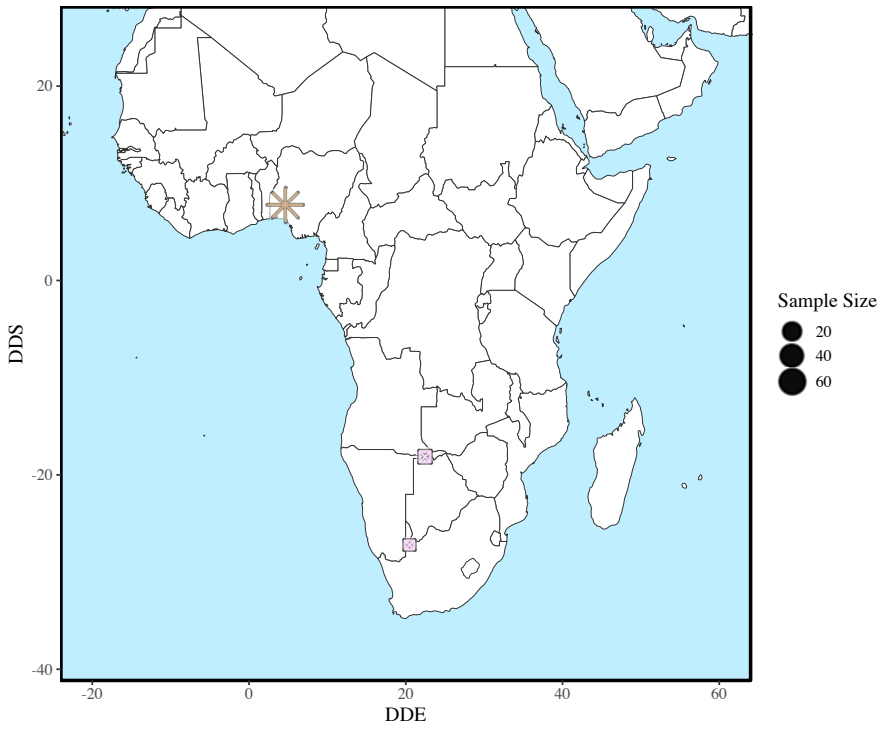
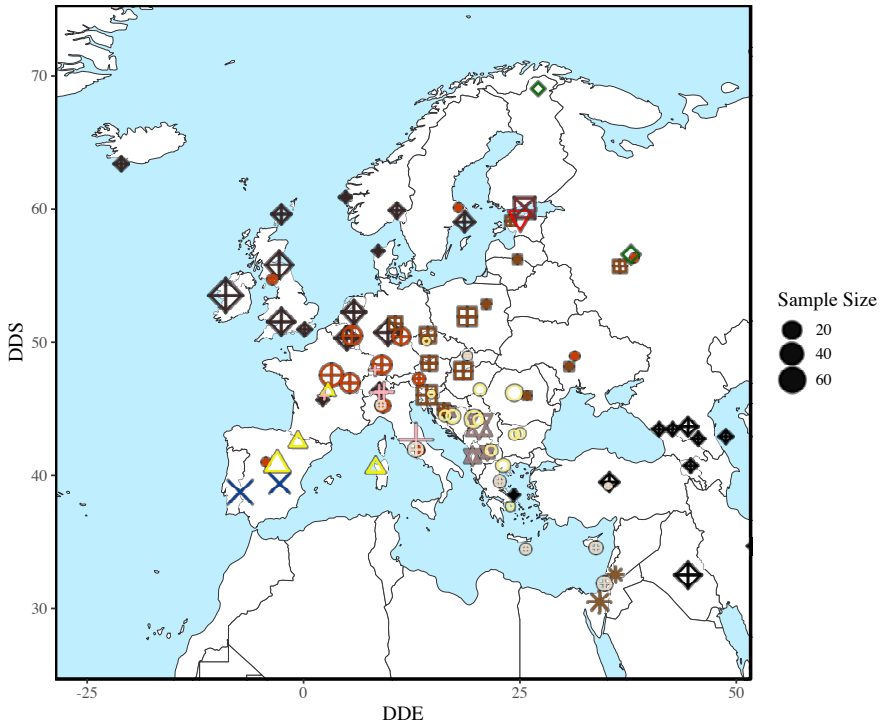


Figure 4.15: All FineSTRUCTURE inferred GR clusters mapped. Blocked regions (a-e) correspond to figures 4.16 - 4.18 . Points have been jittered to aid visualisation. Symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

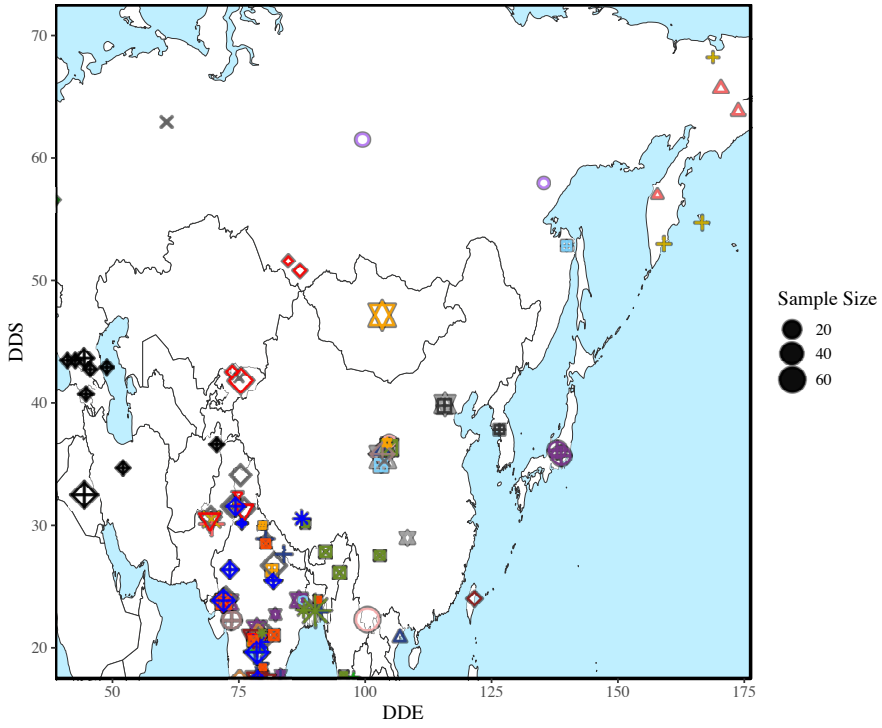


(a) Sub-Saharan Africa

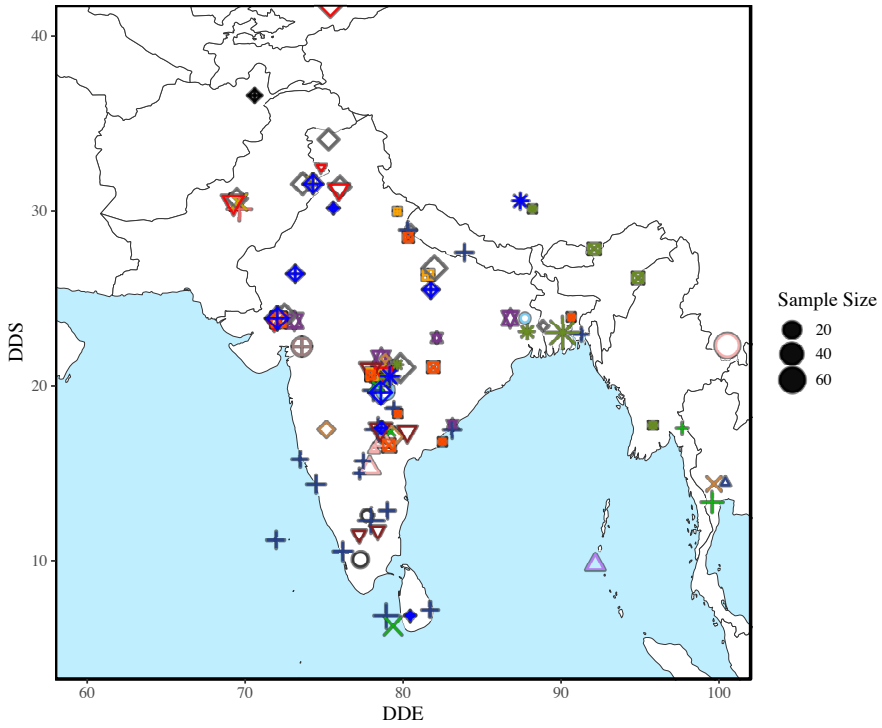


(b) Europe and Western Asia

Figure 4.16: FineSTRUCTURE inferred GR clusters mapped by global regions. Plot sub-headings correspond to inset blocks in figure 4.15. Points have been jittered to aid visualisation. Symbols match Figure 4.14. Size of the plot symbol indicates sample size from that location. Abbr. DDS/E - Decimal Degrees South/East



(c) Central and Eastern Asia



(d) Indian Subcontinent

Figure 4.17: FineSTRUCTURE inferred GR clusters mapped by global regions. Plot sub-headings correspond to inset blocks in figure 4.15. Points have been jittered to aid visualisation. Symbols match Figure 4.14. Size of the plot symbol indicates sample size from that location. Abbr. DDS/E - Decimal Degrees South/East

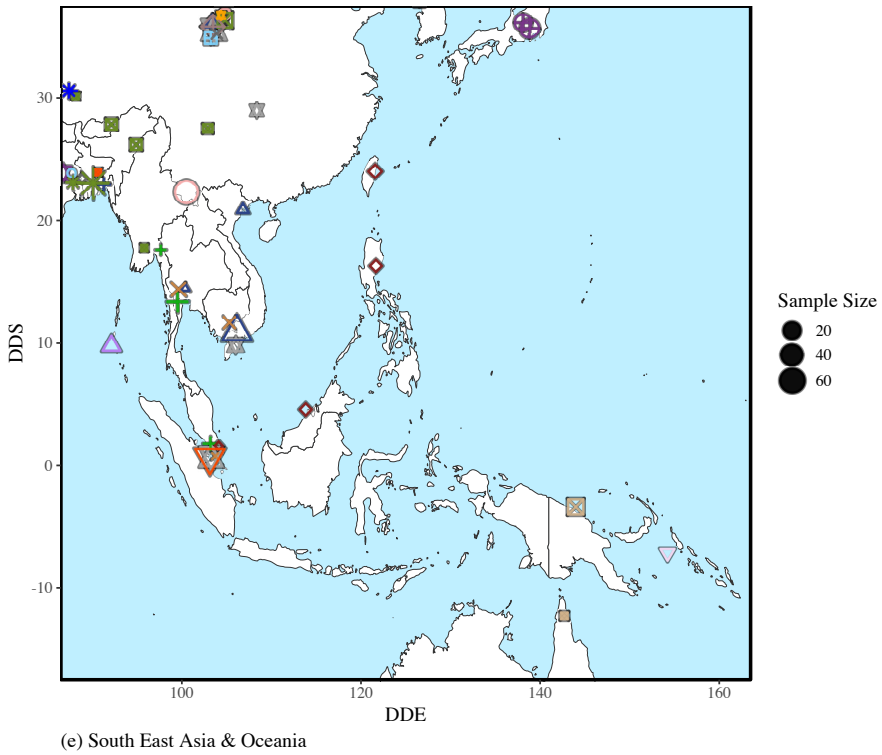


Figure 4.18: FineSTRUCTURE inferred GR clusters mapped by global regions. Plot sub-headings correspond to inset blocks in figure 4.15. Points have been jittered to aid visualisation. Symbols match Figure 4.14. Size of the plot symbol indicates sample size from that location. Abbr. DDS/E - Decimal Degrees South/East

Figure 4.19: Legend of symbols for FineSTRUCTURE GR cluster maps. Plot symbols used for figures 4.15 - 4.18. Symbols match Figure 4.14.

The African samples formed two distinct groups with the Yoruba samples and the KhoeSan samples (≠Khomani and Juu|'hoan) clustering separately. Within Europe *TVD* and F_{ST} values were on average the lowest (Figure 4.14) but major regional structuring was easily identified from the fS clusters (Figure 4.15 b). This included clusters for the Northern, Central, Eastern, two Southern regions as well as two Finnish, an Iberian and an Italian cluster. I further identified a cluster for two known genetic outliers represented by language isolates (Basques and Sardinians). The Near-East and Eastern Mediterranean were divided into three clusters. The patterns of clustering in South-East and Eastern Asia (Figure 4.15 c & e) mostly reflected the discontinuity of sample distribution geographically and the ethnographic nature of the sources publications [see 38], [86], [209], [210]. This was reflected in the estimates of *TVD* and F_{ST} within regions which were marginally elevated compared to Europe and South Asia (Figure 4.14). Structuring within the Indian subcontinent was more layered than the other regional clusters (Figure 4.15 d). I identified 18 geographically overlapping clusters distinguishing groups of Dravidians, Indo-European, Austro-Asiatic and Tibeto-Burmese language speakers as well as geographic structuring within each language group. With few exceptions, these clusters were ethnically diverse based on their *a priori* population labels (see Supp. Figure A.9 - A.13 a-r & A.18 ar). Intra-regional *TVD* and F_{ST} were comparable to that of the South-East and Eastern Asian clusters.

To evaluate if the available linkage disequilibrium for informing haplotype coalescence varied between the inferred clusters, I examined the number of SNPs haplotype-chunk^{-1} within each cluster (Figure 4.20). Again values are slightly below earlier research [97], [177]. The lowest values, ~7.5 (African groups), are likely the result of higher genetic diversity and poorer regional representation for these populations which may cause a higher frequency of switching between different haplotypes while characterising local ancestry along the chromosome. The highest number of SNPs haplotype-chunk^{-1} were reported for 14_Andaman.Is_JO (>16 SNPs haplotype-chunk^{-1}) which reflects the high degree of inbreeding as the population has fewer than 200 individuals remaining [85]. Similarly, the

56_Polynesia_MP, 55_PNG_AU-PN, 30- and 31_Finland_UR clusters have elevated SNPs haplotype-chunk^{-1} which mirrors the elevated average chunk length copied (Figure 4.20), reflecting relatively recent bottleneck events or founder effects [209], [211], [212].

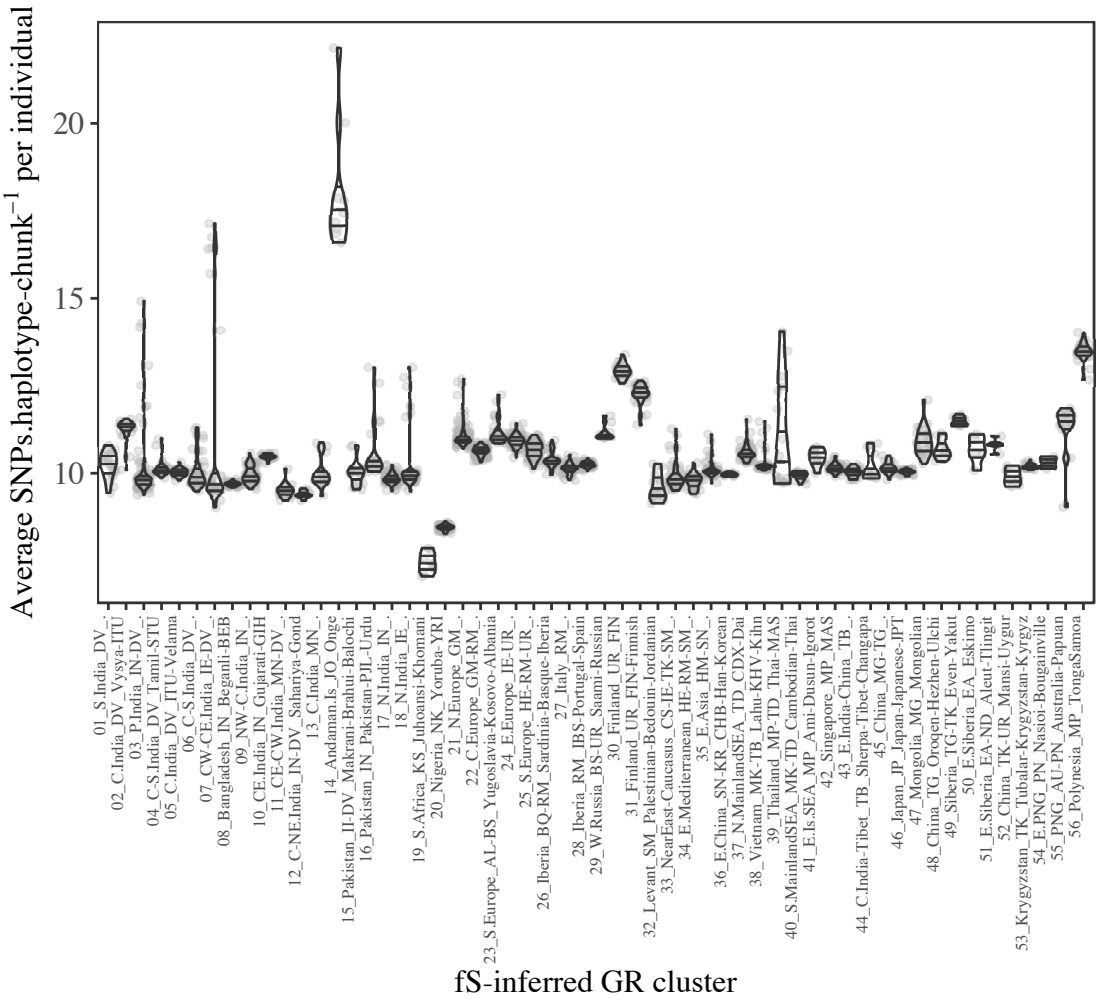


Figure 4.20: Average SNPs haplotype-chunk^{-1} per individual from GR - GR CP-fS analysis arranged by fS-inferred clusters. SNPs haplotype-chunk^{-1} averaged across chromosomes and donors. The grey cloud of points are individual values. Median and IQR values indicated by black horizontal lines. Range of values indicated by vertical extent of violin plots.

Identifying Possible Cryptic Sub-structure within the SAC Dataset using Haplotype-based Clustering

I explored for the possibility of genetic structure within the SAC which may have not been detected with PCA or ADMIXTURE. Some degree of structuring within the SAC community may exist because of recent admixture history including assortative mating by geography, social status etc. and therefore be reflected in the genetic relationship among SAC.

The procedure followed mirrors that of the GR - GR CP run above. In this case

copying vectors do not necessarily reflect only the underlying distal source ancestry (i.e. slave and settler origins) but are potentially more influenced by very recent ancestry (i.e. haplotypes shared between SAC individuals which are constructed out of the recombined slave and settler haplotypes). The SNPs haplotype-chunk^{-1} for the SAC - SAC run were elevated above that of the SAC - GR run (Figure 4.12 and Supp. Table A.6 & A.7) which suggests that haplotypes constructed in the SAC - SAC run are indeed more influenced by recently formed haplotypes.

The SAC individuals were clustered according to similar copying profiles using fS. The assignments based on the maximum concordance trees showed convergence across five independent chains, though individuals were more often assigned to multiple clusters than seen in the GR - GR run (Figure 4.21 and Supp. Figure A.29 - A.33). In total 40 clusters were identified with 2-55 individuals each (median 17).

I applied a *TVD* cut threshold of 0.032, following the procedure described above, retaining 14 clusters (Supp. Figure A.34). Clusters were named with a running index, followed by the sample size .e.g. 14_17SAC indicates cluster 14 with 17 SAC individuals.

Three major branches can be seen in the fS topology (a-c) (Figure 4.22) and one of the branches (c) had notably higher self-copying values for total genomic length (Supp. Figure A.36 & A.37) but not chunk counts or mutations (Supp. Figure A.35 & A.38) and correspondingly had higher *TVD* values when compared to the other two branches.

In general, the F_{ST} (~0 - 0.07) and *TVD* (~0 - 0.33) values were low between clusters (Figure 4.22) suggesting little differentiation. Higher F_{ST} values seen between branch (c) and branch (b) are not seen among comparisons of clusters in branches (a) and (b) indicating a particularly large distinction here. In particular, cluster 14_17SAC has high F_{ST} and *TVD* values in comparisons with most other clusters. Based on F_{ST}, it is most different from cluster 10_67SAC, while based on *TVD*, it is most different from cluster 08_25SAC.

More than 80% of the F_{ST} values fell below the 25$^{\text{th}}$percentile of the GR pairwise F_{ST}. The top 90$^{\text{th}}$ percentile of the SAC dataset (F_{ST}>0.033) all included

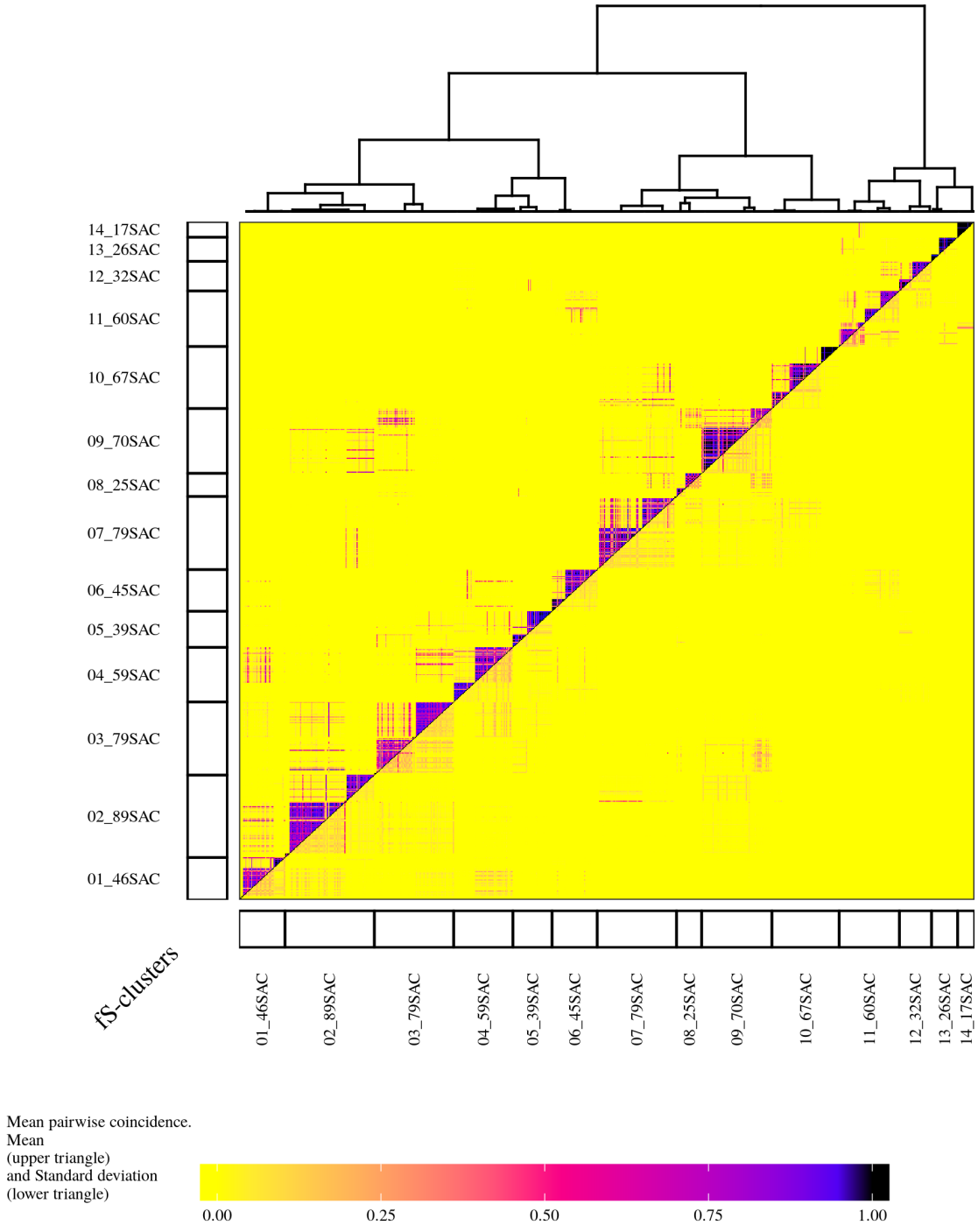


Figure 4.21: Tree structure convergence shown by pairwise coincidence values for the SAC - SAC CP-fS run. Mean (upper) and standard deviation (lower) values from five independent chains per pair of individuals indicated. Matrices sorted by the same randomly chosen tree (chain 0). Bars below and left of the matrix indicate the *TVD*-based clusters created.

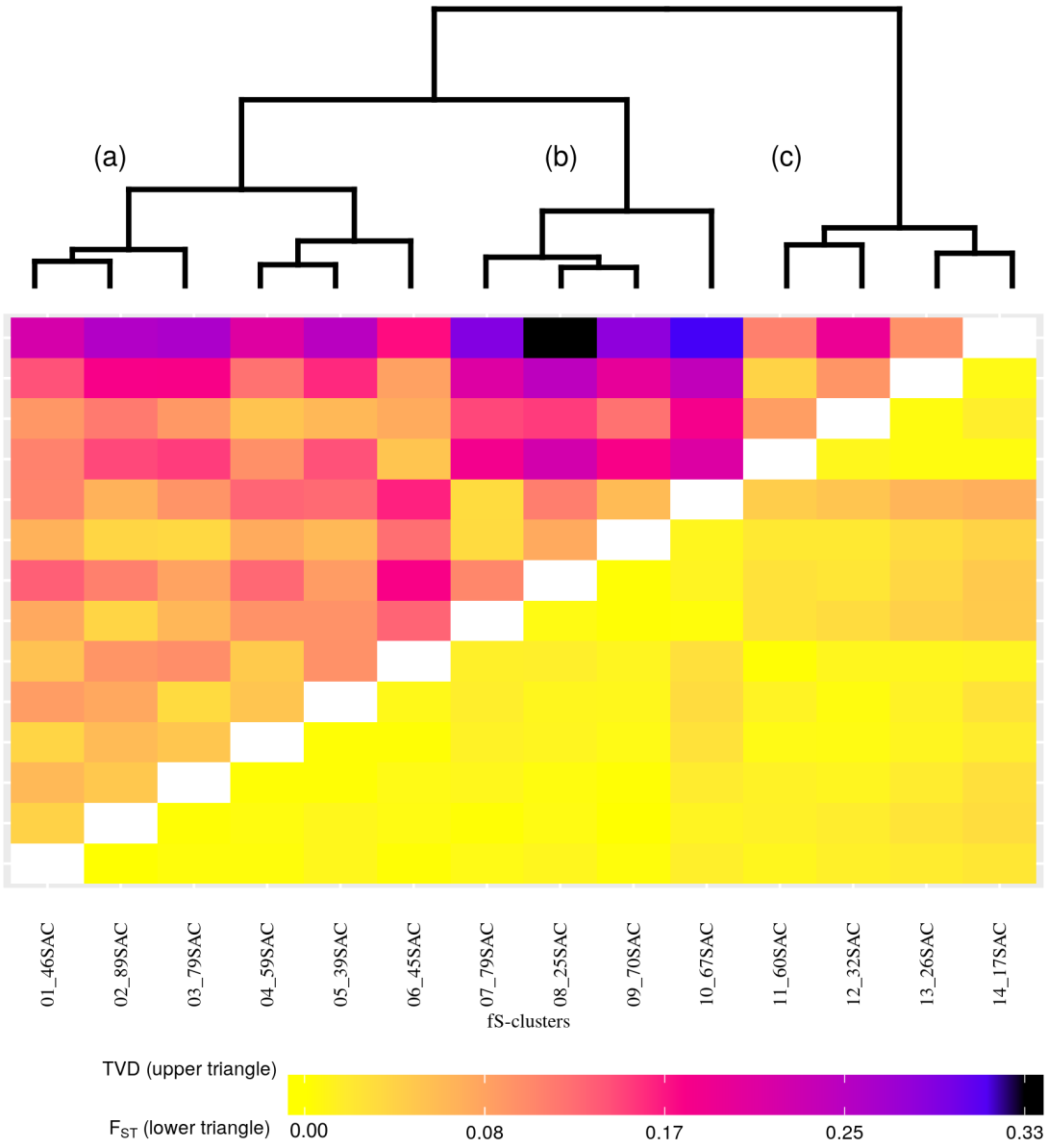
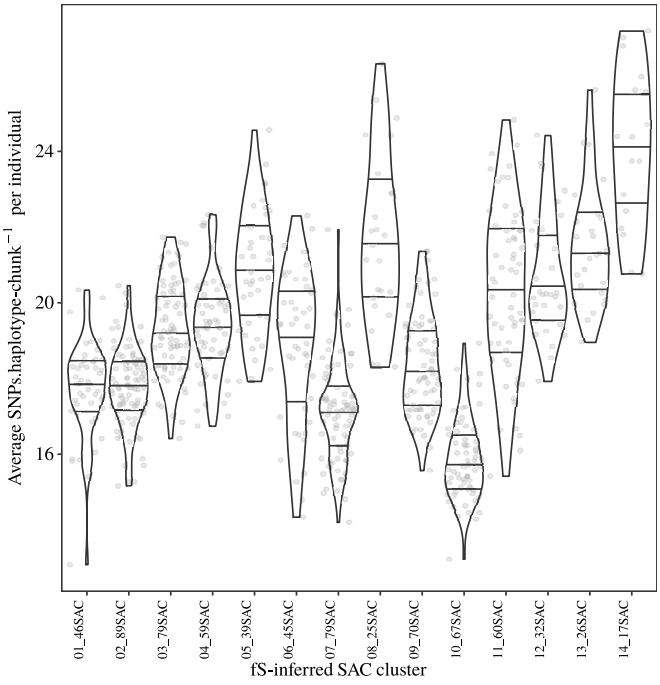
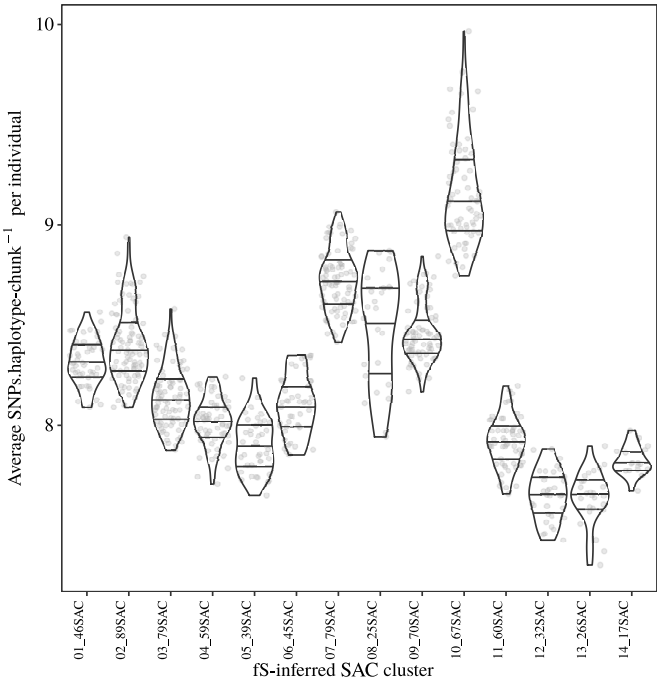


Figure 4.22: FineSTRUCTURE-inferred clusters, F_{ST} and TVD from SAC - SAC CP data following a TVD-based cut. Pairwise distances estimated as Total variance distance (TVD) (top left) and F_{ST} (bottom right). The TVD cut threshold was 0.032 retaining 14 clusters. Cluster names indicated at the bottom and fS dendrogram above. Diagonal values set to white.



(a) SAC - SAC



(b) SAC - GR

Figure 4.23: Average SNPs haplotype-chunk^{-1} per individual from SAC CP-fS analysis organised by fS-inferred clusters. Shown are (a) the SAC copying from SAC individuals and (b) SAC copying from GR individuals. The grey cloud of points are individual SAC averages. Median and interquartile values indicated by black horizontal lines. The full range of values are indicated by the vertical extent of the violin plots. Here the SAC-SAC fS analysis was used, informing clustering of SAC based on copying only from other SAC.

comparisons of branch (c) to (b). The F_{ST} values equal to or larger than the median of the GR dataset (~0.05) all involved a comparison of 10_67SAC to a cluster from branch (c). The *TVD* values showed less support for structure as 90% of the SAC values fell below the 25th percentile for the GR dataset. The top 10% from the SAC all included comparisons of 13_26SAC and 14_17SAC to clusters from branch (a) or (b).

The estimated SNPs haplotype-chunk^{-1} varied across clusters suggesting possible differences in the recency of admixture or different rates of switching between haplotypes (Figure 4.23). Clusters 11 - 14 had the largest haplotype-chunks (>20 SNPs haplotype-chunk $^{-1}$). Cluster 08 appeared as an upper outlier within its branch, which on average had the lowest values.

ADMIXTURE components for the fS-inferred SAC clusters (Figure 4.24) show a notable distinction between 14_17SAC (and to a lesser extent 13_26SAC) compared to the other clusters in that this cluster has fewer ADMIXTURE components. The components were predominantly shared with African fS-inferred GR clusters (19_S.Africa_KS_Juhoansi-Khomani and 20_Nigeria_NK_Yoruba-YRI) with an additional component which is shared with Levantine, Pakistani and Mediterranean GR clusters (e.g. GR cluster 28, 32, 34, 15 but also 20_Nigeria_NK_Yoruba-YRI).

In summary, I recover clustering among the SAC which may reflect shared history between individuals. Cluster 13_26SAC and 14_17SAC show some suggestion of differentiation from the remaining clusters by F_{ST}, *TVD* and ADMIXTURE proportions but for the remaining clusters there is no clear evidence.

4.3.3 Characterising Specific Global Ancestral Contributions to the SAC

When allowing SAC individuals to copy only from the GR data, I can reconstruct the copying vector of each SAC sample as a linear combination of contributions from each GR cluster, identifying the best GR combination to produce the SAC profiles. This was done implementing a mixture model in Non-Negative Least

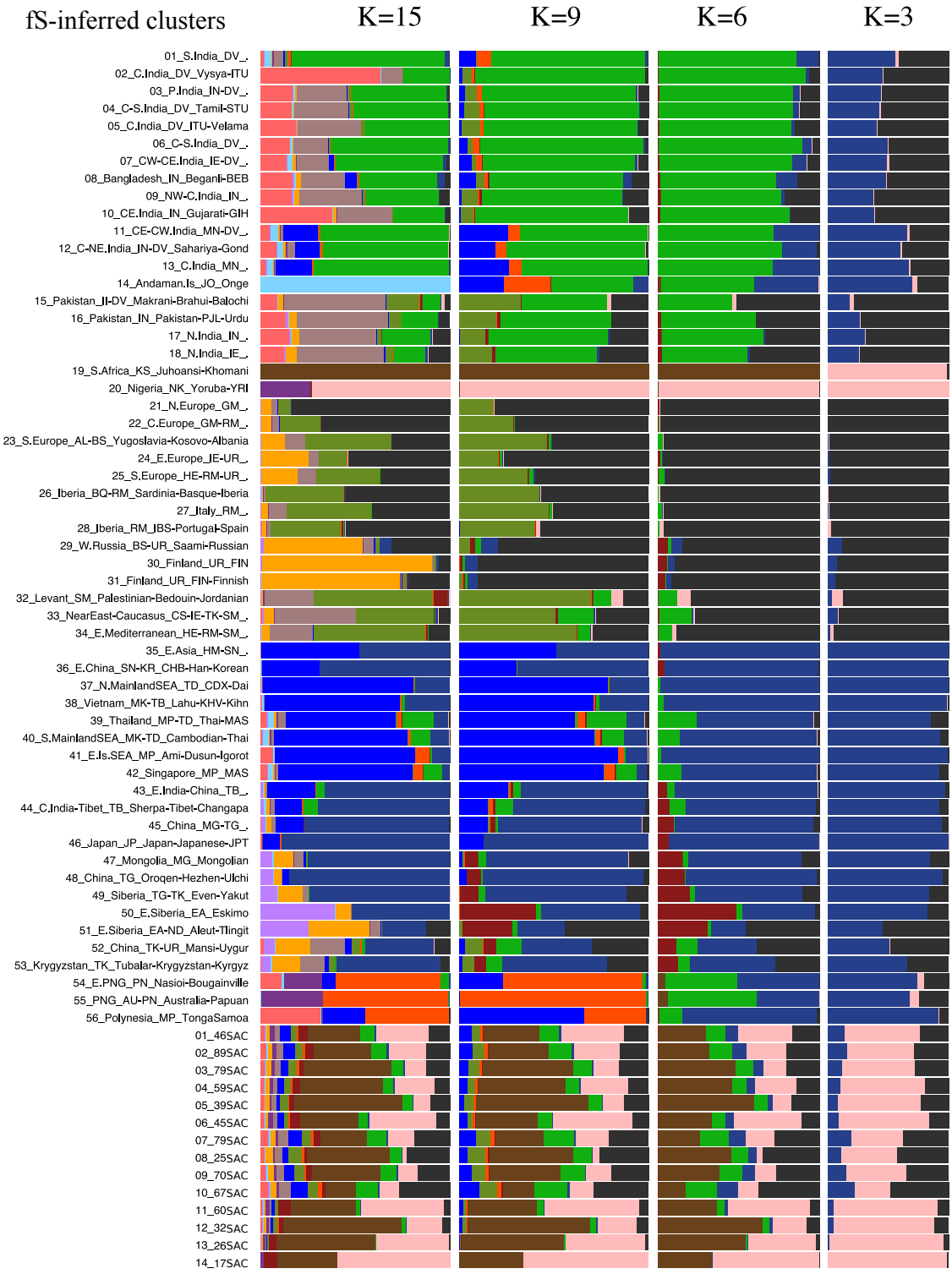


Figure 4.24: ADMIXTURE profiles (K = 3,6,9,15) for GR + SAC dataset averaged across fS-inferred clusters. The SAC clusters are located at the bottom.

Squares (NNLS) regression [97]. The GR copying profiles are the output from the GR - GR run (see above Section 4.3.2).

Following the NNLS, a contribution >0 (mean - jack-knife error) was identified from 53 fS-inferred GR clusters (of 56). The three clusters which did not contribute were 36_E.China_SN-KR_CHB-Han-Korean, 40_S.MainlandSEA_MK-TD_Cambodian-Thai and 52_China_TK-UR_Mansi-Uygur. I applied a cut-off criterion to find contributions which were of interest. The criteria employed emphasised ancestry relevant to most of the dataset and minimised idiosyncratic/individual-specific contributions. I excluded signals with <5% prevalence in the population and with a lower threshold of 1% in any individual based on the lower jack-knife error estimate. Details for the justification can be found in Section 4.2.4.

Only 11 GR clusters were identified as relevant sources post-criteria (Figure 4.25). Except for two sources (32_Levant_SM and 06_C-S.India_DV), the identified GR clusters were resilient to an increased threshold of 2% on the exclusion criteria for minimum ancestry per SAC individual.

I further examined an alternative cut-off criteria procedure which is to perform the NNLS after averaging copying profiles within each of the SAC fS-inferred clusters. The results were overall similar (see Supp. Mat. A.1.1) however an additional five contributing sources were identified. A theoretical issue with averaging prior to the NNLS is that in a population of diverse ancestry, one may inadvertently average two copying profiles which are relatively different and reflect different contributions, but when averaged resemble a new ancestry source. Considering both the greater number of possible sources identified and the possibility of finding spurious sources, the conservative option is to perform the NNLS prior to fS-cluster averaging. I thus discuss the 11 sources identified by the cut-off criteria. For the rest of the chapter I refer to the identified 11 GR clusters as the 'sources' or 'donors'.

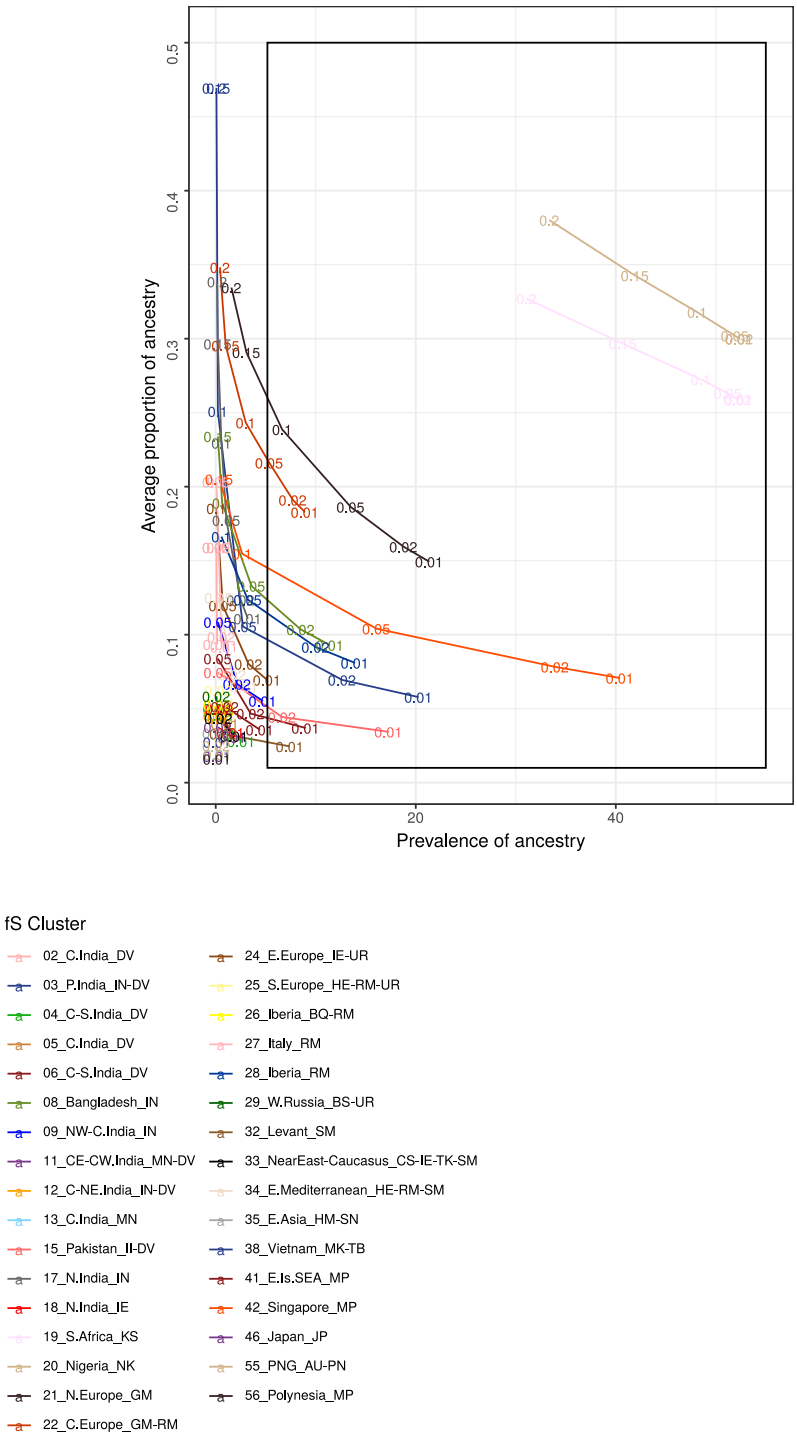


Figure 4.25: Change in NNLS ancestry prevalence and proportion under different exclusion criteria. Values shown are based on $mean \pm jack-knife\ error$ of the remaining samples after the cut-off criteria was applied. Exclusion criteria for recognising ancestry in any individual varied from 0.01 - 0.2, indicated beside the line for each GR source. Box indicates the decided exclusion; 5% prevalence and 0.01 proportion ancestry.

The African GR sources, 20_Nigeria_NK and 19_S.Africa_KS, contributed the largest proportions on average, 0.28 ± 0.15 and 0.24 ± 0.10 respectively (Figure 4.26 and 4.27). Their contribution was the most prevalent, present in 100% of the sample in both cases. The next most prevalent ancestry was 42_Singapore_MP (~77%) followed by 21_N.Europe_GM (40%) and 03_P.India_IN-DV (38%). All the other sources had 14 - 33% prevalence.

Following the two African sources, the next largest contributors by average proportions were Northern and Central Europe (~0.08 each) followed by 42_Singapore_MP (~0.05), then 08_Bangladesh_IN and 28_Iberia_RM (0.04). The remaining sources contributed ~0.03 or less on average (Supp. Table A.11 & A.12). The two African sources, and Northern and Central Europe had the largest variability in the average proportion of ancestry present (Figure 4.26).

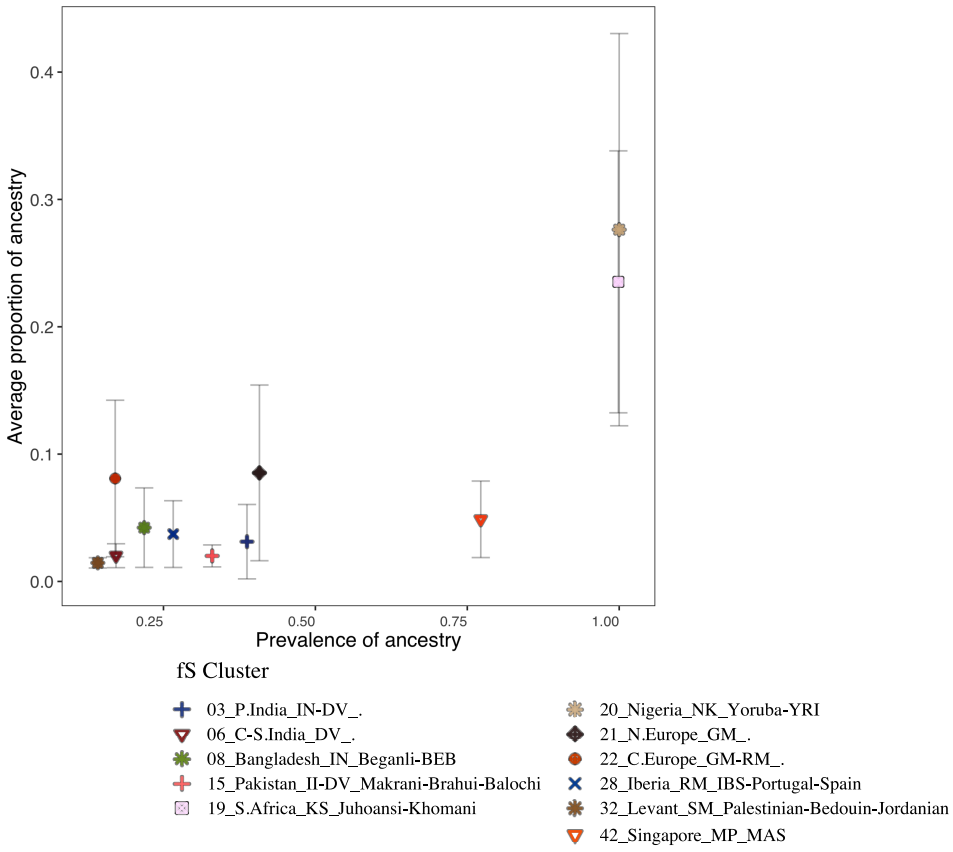


Figure 4.26: Prevalence and average± *s.d.* proportions of the eleven identified NNLS components. The average proportion ancestry was calculated after excluding individuals with <1% of the relevant ancestry. Prevalence refers to the prevalence in the population.

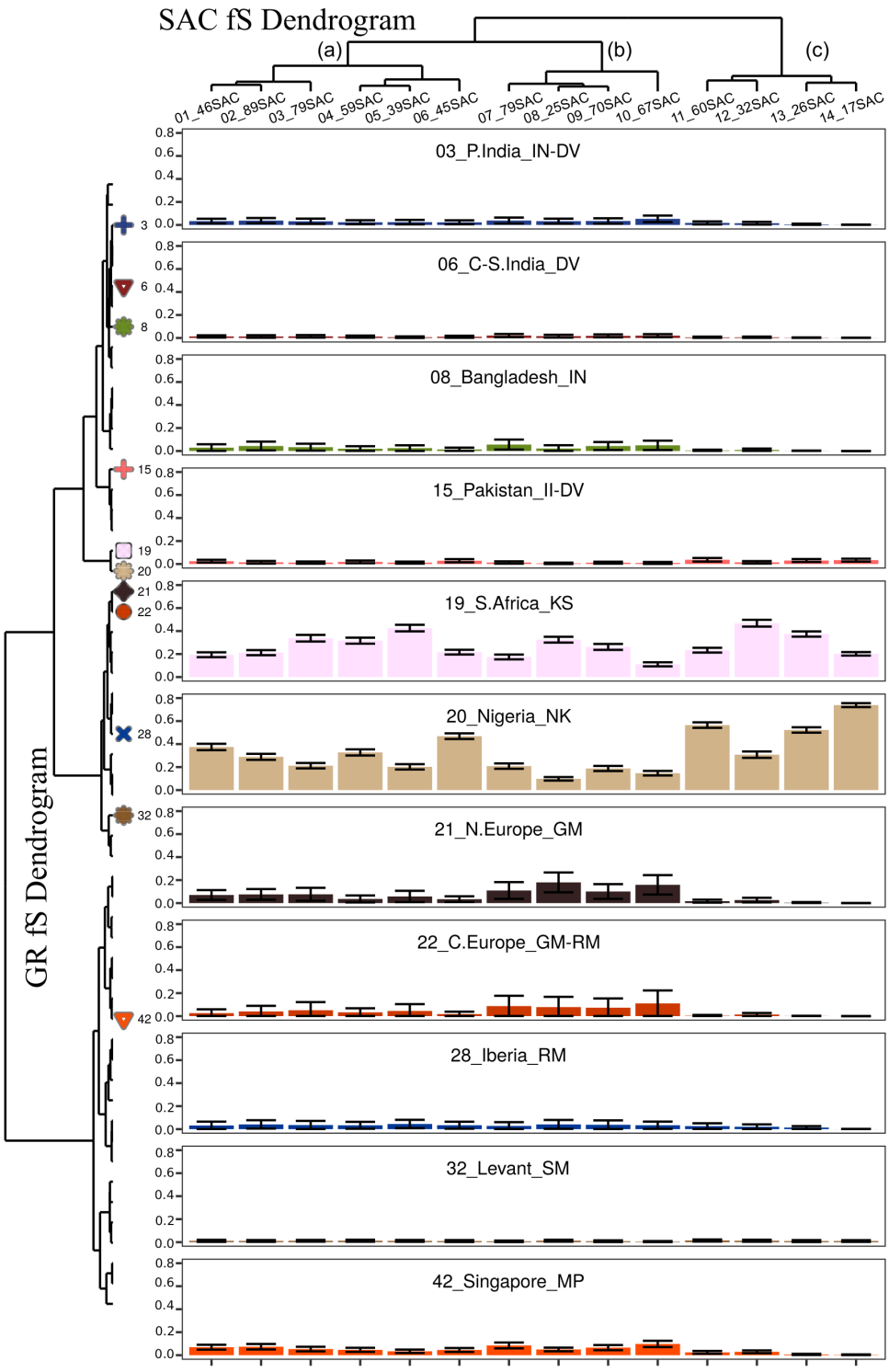


Figure 4.27: Ancestral NNLS contributions to the SAC arranged by fS-inferred clusters. Identified clusters contributed at least 0.01 ancestry to at least 5% of individuals in the SAC dataset. Mean proportion genome length copied, and mean jack-knife error indicated within each cluster (including those individuals with <0.01 ancestry). Branch and cluster labels on SAC fS-tree indicated. Symbols, cluster numbers and colours correspond to previous figures (see Figure 4.26)

4.3.4 Testing for Ancestry-based Sub-structure to the SAC fS-inferred Clusters

To formally test for possible structure within the SAC dataset based on these ancestries identified, I performed a principal component analysis on the centred log-ratio transformed values of the proportion ancestry after imputing zero values to a near-zero value [204] (Figure 4.28 - 4.30). The first three principal components explained ~76% of the variance and the first five explain 98% of the variance (Figure 4.31). The primarily drivers for the first three components were 08_Bangladesh_IN, 21_N.Europe_GM, 22_C.Europe_GM-RM and 28_Iberia_RM, which are relatively prevalent ancestries with moderate representation in SAC individuals (Figure 4.28 - 4.29).

The 22_C.Europe_GM-RM and 28_Iberia_RM components were directed in near opposite directions on PC1, and both directed away from 08_Bangladesh_IN on PC2 (Figure 4.28). 08_Bangladesh_IN and 21_N.Europe_GM were directed in opposite directions along PC3 while the remaining ancestries were directed away from the above four along PC4 (Figure 4.29). Along PC5 06_C-S.India_DV was strongly directed away from almost all other ancestries (Figure 4.30). Except for 14_17SAC and 13_26SAC, all the remaining clusters overlapped extensively in PC space. This indicates little evidence of structure within the SAC based on fS-inferred clusters.

I performed a multivariate analysis of variance on the top five principal components for the fS-inferred clusters (Table 4.4 and 4.5). There were significant differences among clusters for all PCs ($p<0.00001$). The results of a Tukey post-hoc HSD test with sequential Bonferroni correction ($\alpha = 0.01$) broadly conform to the results of the F_{ST} and TVD distances (Figure 4.32). The clusters on branch (c) of the SAC fS-inferred tree produced the greatest number of significant comparisons with clusters on branch (b). Cluster 14_17SAC had at least one significant PC difference with each cluster except those within its own branch.

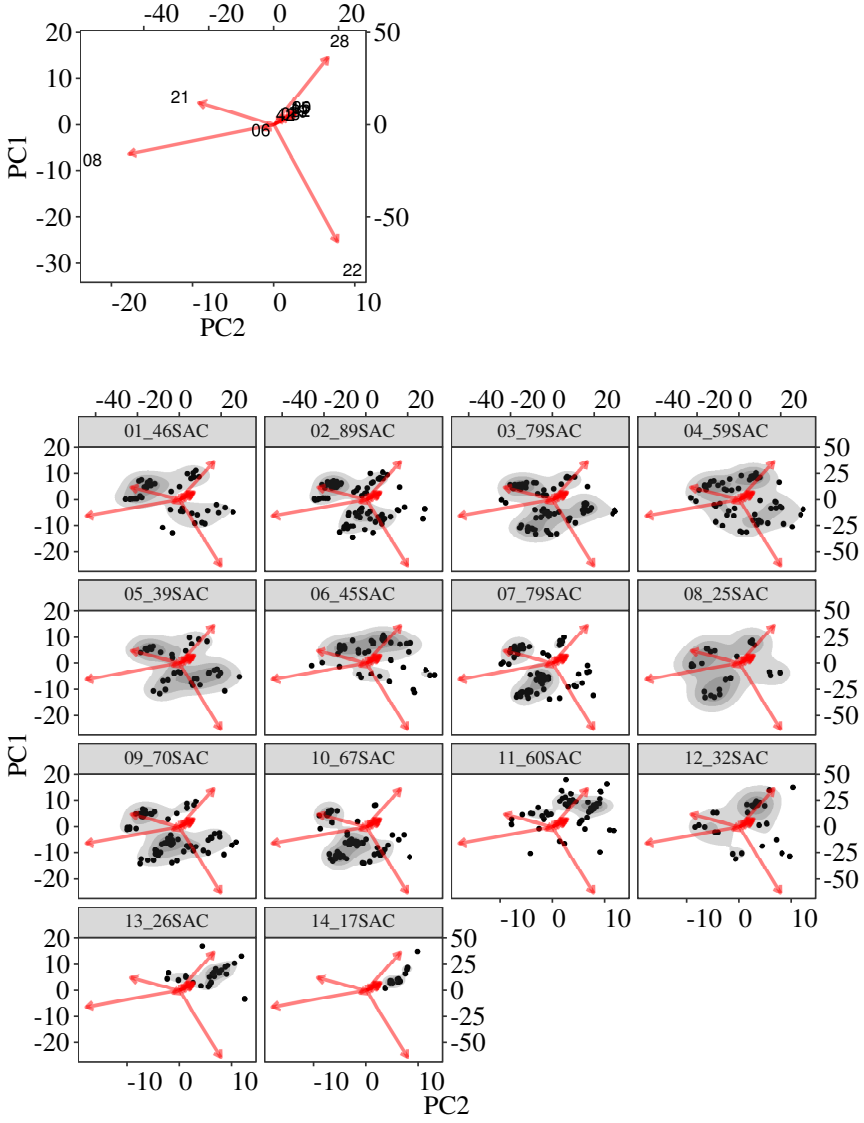


Figure 4.28: Principal component analysis on centred log-ratio (CLR) transformed ancestral components showing PC 1-2. The top inset plot shows the loading vectors for the 11 NNLS ancestral components. Top and right axes are scaled by loading vector. Plotted numbers correspond to the GR numeric prefix. e.g. '28' corresponds to '28_Iberia_RM'. Shown below are eigenvectors and labels and the distribution of samples from each fS-inferred SAC clusters. Zero-ancestry values imputed to a near-zero low value.

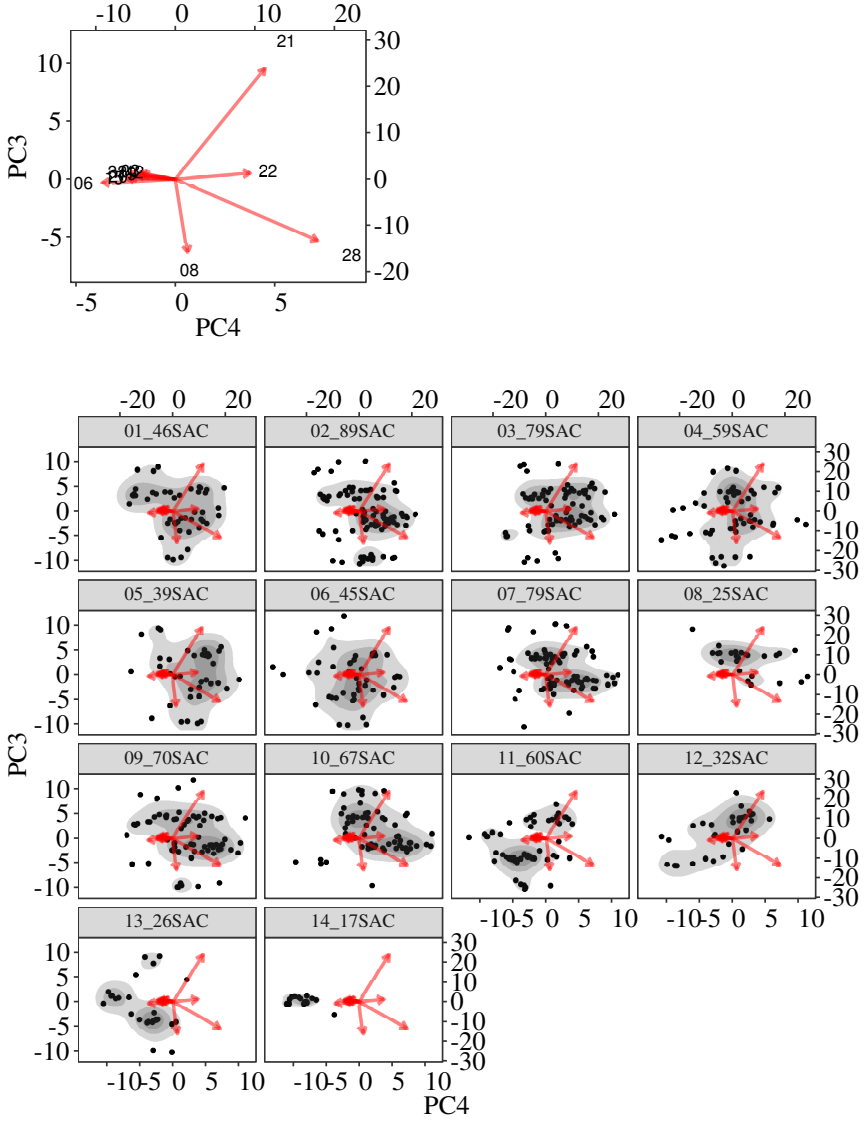


Figure 4.29: Principal component analysis on centred log-ratio (CLR) transformed ancestral components showing PC 3-4. The top inset plot shows the loading vectors for the 11 NNLS ancestral components. Top and right axes are scaled by loading vector. Plotted numbers correspond to the GR numeric prefix. e.g. '28' corresponds to '28_Iberia_RM'. Shown below are eigenvectors and labels and the distribution of samples from each fS-inferred SAC clusters. Zero-ancestry values imputed to a near-zero low value.

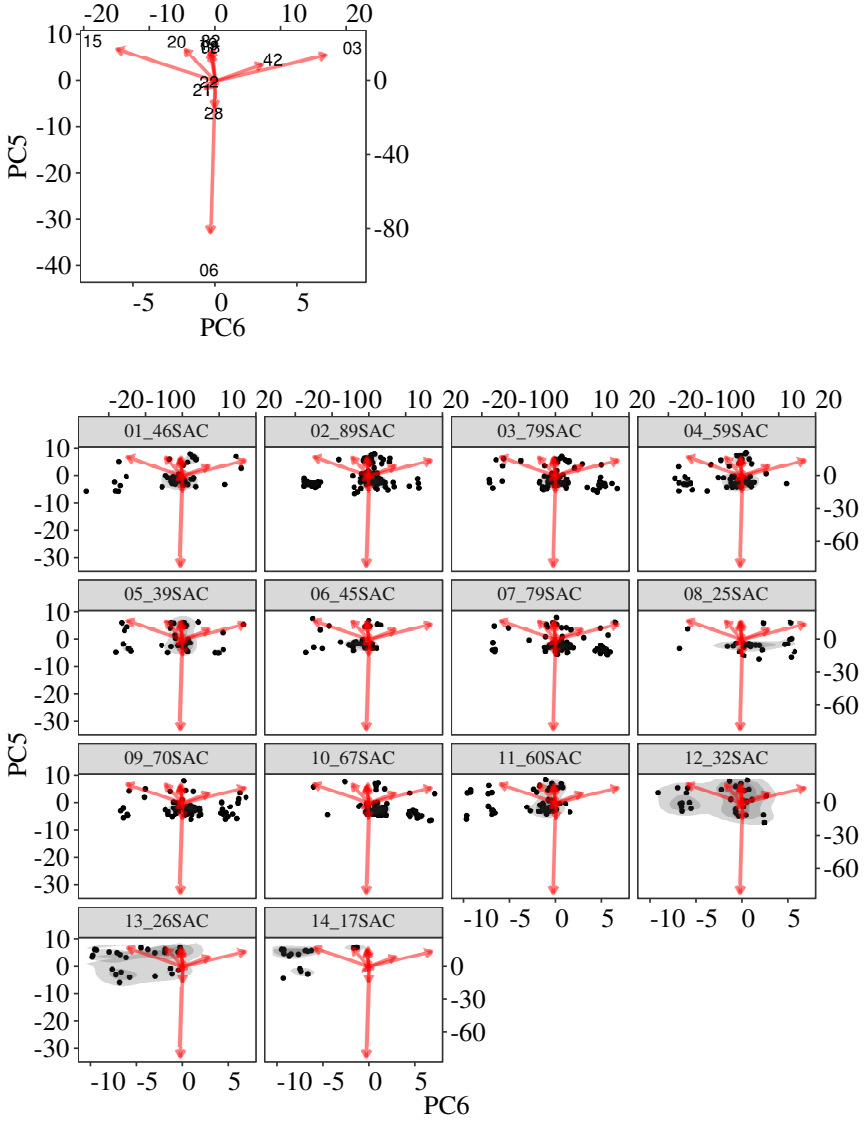


Figure 4.30: Principal component analysis on centred log-ratio (CLR) transformed ancestral components showing PC 5-6. The top inset plot shows the loading vectors for the 11 NNLS ancestral components. Top and right axes are scaled by loading vector. Plotted numbers correspond to the GR numeric prefix. e.g. '28' corresponds to '28_Iberia_RM'. Shown below are eigenvectors and labels and the distribution of samples from each fS-inferred SAC clusters. Zero-ancestry values imputed to a near-zero low value.

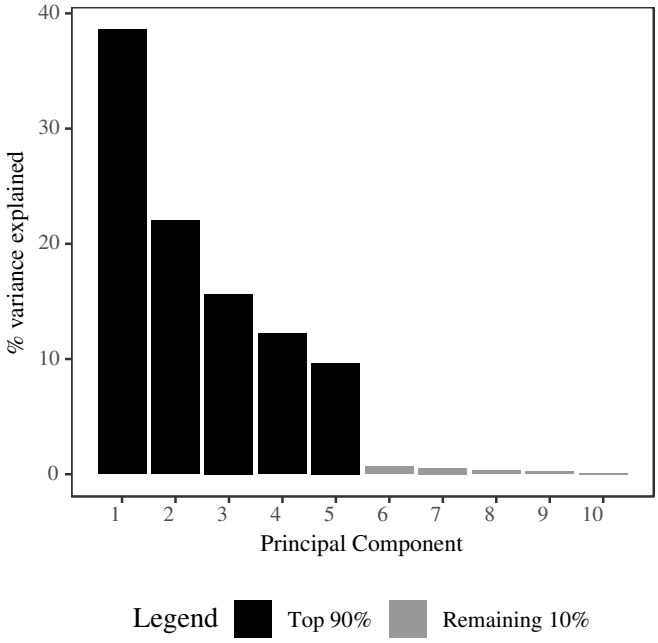


Figure 4.31: Variance explained by each principal component formed on the CLR-transformed NNLS ancestral proportions. Centred log-ratio (CLR) transformed ancestral components were used to inform the eigenvectors. The eigenvectors contributing to the top 98% are coloured in black.

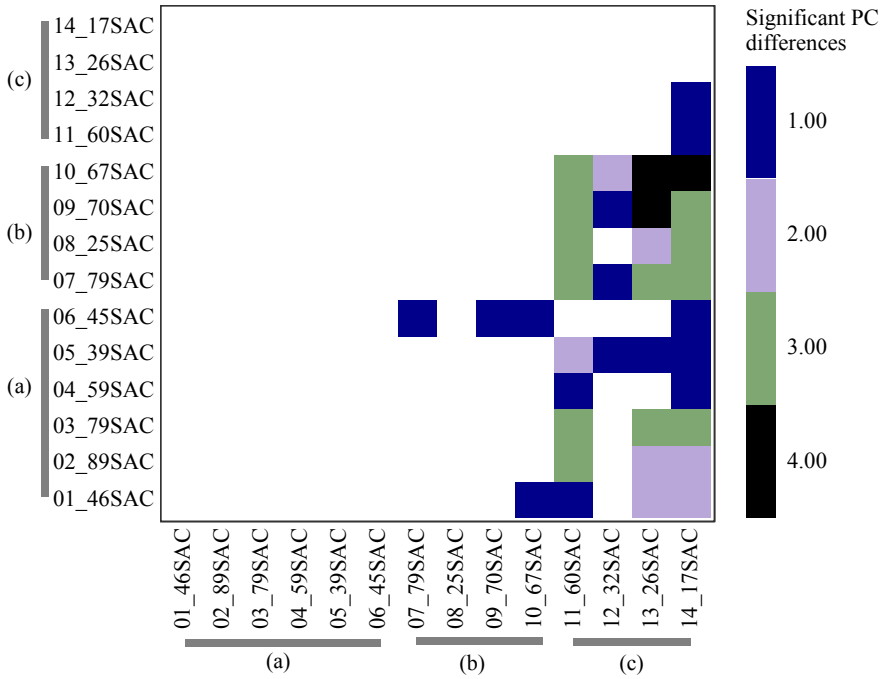


Figure 4.32: Number of significantly different pairwise comparisons between fS-inferred SAC clusters based on top five principal components. Principal components summarise the centre log-ratio transformed NNLS ancestry proportions. Comparisons made with Tukey Honest Significant Difference post-hoc test with sequential Bonferroni corrections, $\alpha = 0.01$. Letters on the axes indicate the fS branches.

Table 4.4: Multivariate analysis of variance of the first five principal components for the SAC fS-inferred clusters. Values based on centre log-ratio transformed NNLS ancestries as input data. Abbreviations: Df - Degrees of Freedom, Pillai - Pillai's trace statistic, approx. - approximate

	Df	Pillai	approx. F	Pr(>F)
Intercept	13	0.711	9.171	$< 1 \times 10^{-8}$
Residuals	719			

Table 4.5: Analysis of variance results for each of the first five principal components for the SAC fS-inferred clusters. Values based on the centre log-ratio transformed NNLS ancestries. Abbreviations: Df - Degrees of Freedom, Sum Sq. - Sum of Squares, Mean Sq. - Mean of Squares.

Principal Component	Df	Sum Sq.	Mean Sq.	F value	Pr(>F)
PC1	13	8373	644.1	15.07	2.2e-30
Residuals	719	30740	42.8		
PC2	13	4538	349.1	14.79	8.4e-30
Residuals	719	16969	23.6		
PC3	13	746	57.4	2.83	5.5e-04
Residuals	719	14558	20.2		
PC4	13	4098	315.2	19.86	3.6e-40
Residuals	719	11413	15.9		
PC5	13	1107	85.1	6.60	5.2e-12
Residuals	719	9270	12.9		

4.4 Discussion

Global mass migration events have been a prominent feature of historic socio-political developments. Consequently, geneflow between divergent lineages which were previously separated by continental distances is now common. The South African Cape Coloured, as with many other post-colonial admixed communities, are characterised by their creolised genetics and culture [97], [99], [181], [182], [213]. The community is often described as a five-way admixture [12], [110], [187] and indeed I show that the dynamics of the Indian Ocean slave trade has brought together genetic components from several Old-world regions.

Limited Sub-structure within the SAC Dataset

I found little substructure within the SAC dataset which is most clearly shown from the F_{ST}, *TVD* and ADMIXTURE results (Figure 4.24) and further by the MANOVA

of PC components based on the NNLS-inferred ancestral contributions (Figure 4.32). While the fS procedure did cluster individuals with similar profiles, the clustering produced limited genetic differentiation. The lack of substructure reflects the recency of the neighbourhood (~1970s) and the small geographic extend of sampling.

The most notable differences by F_{ST} and TVD included branch (c) of the fS dendrogram (Figure 4.22), and more specifically 13_26SAC and 14_17SAC. It appears that the predominant Yoruba (~Bantu) and KhoeSan ancestry in branch (c) drives the difference estimates when compared to clusters with greater Eurasian contributions (Figure 4.27).

Only SAC cluster 14 provides enough genetic evidence of differentiation to warrant further discussion. The F_{ST} and TVD values suggest elevated differences compared with the remaining clusters. There are only 17 individuals in this cluster, and all appear to have entirely African (non-KhoeSan) and KhoeSan ancestry.

The non-KhoeSan African to KhoeSan ancestry ratio based on ADMIXTURE proportions would suggest a possible Southern Bantu-speaking contribution, (~75%:25% or 3:1 in the amaXhosa) (see Figure 4.24, Chapter 6 and [108], [125], [177]). I found 14_17SAC has the highest ratio among the SAC clusters (1.95; other clusters have average 0.64 and 0.07 - 1.95 range) but high values were not restricted to branch (c). Cluster 14 may be Bantu-speaking migrants to the region who have recently adopted the "Coloured" identity or individuals at the tail of the distribution of Bantu - KhoeSan ancestry, reflecting drifted genomes. The Southern Bantu amaXhosa make up most of the neighbouring populations in Cape Town and may contribute to ongoing admixture. This cluster amounts to 2% of the entire dataset, figures in agreement with the census proportions for 'Black' South Africans in Ravensmead and Uitsig (~2 - 7%) [214].

Predominant African Ancestry Reflects Large Indigenous Communities and African Slave Imports

The Cape Colony was the western most destination for slaves on the Indian Ocean slavery circuit [35], [133] and this is reflected in the genomes of the SAC. Cape Town had the largest African slave intake of all the *VOC* cities, taking in the slaves from

VOC sponsored campaigns to African nations [35]. The largest genetic contributions in the SAC reflected KhoeSan and Bantu-speaking African ancestry [12], [63], [110], both to be 100% prevalent by CP-NNLS and ADMIXTURE proportions.

The Cape KhoeSan and non-KhoeSan African slaves are a well-established presence in historic accounts [16], [132]. The most genetically similar KhoeSan to the SAC are the relatively un-admixed contemporary ≠Khomani (Taa speakers) from the 'Southern' KhoeSan group which consists of Taa and Khoe-Kwadi representatives [93], [108] (Figure 4.11). Though the direct ancestors could be the /Xam who are now ethnographically extinct but may have genetically resembled the contemporary Karretjie (Khoe-Kwadi speakers) of the Northern Cape, South Africa [108], [215] and ≠Khomani, both of whom belong to the Southern genetic cluster [108].

An additional source of KhoeSan and Bantu-speaking African ancestry could be contributed by Southern Bantu populations who admixed with KhoeSan following their arrival in Southern Africa [64], [125]. While clustering within the SAC suggests that 14_17SAC may reflect such ancestry, in the remaining clusters this contribution is tentative. Patterson, Petersen, Ross, *et al.* [185], however, suggested a Southern Bantu contribution in a smaller but higher resolution dataset.

Non-KhoeSan African ancestry may have been contributed by several populations, specifically from Western and Eastern Africa and Madagascar but it's not possible to discern different sources from this work as no alternatives were considered.

Identified Asian Contributions Reflect Known *VOC* Slaving Ports

As the earliest arriving slaves and servants were likely privately-owned eastern Eurasians [16] and the representation of South and South-East Asian dominated during the later centuries, their contributions should be evident.

There were six possible South Asian clusters (out of 18) based on geography and ethnic composition which could have been identified as sources (see Supp. Table A.13). Of these, three were identified corresponding to suggestions from archaeology and archival work [16], [133], [149] (03_P.India_IN-DV_., 06_C-S.India_DV_., 08_Bangladesh_IN_Bengali-BEB). A further three were considered

highly likely contributions but did not meet the cut-off criteria (01_S.India_DV, 04_C-S.India_DV_Tamil-STU, 07_CW-CE.India_IE-DV).

The 03_P.India_IN-DV component was the most prevalent (~38%) and 06_C-S.India_DV the least prevalent (~17%). Both Dravidian sources contributed a low average proportion which was consistently low across SAC fS-clusters. Geographically, these clusters lie on much of the Malabar and Coromandel coast of India, including the Cochin port, all areas from which slaves may have been collected at the height of Dutch and Arab slavery in the 1600s [35], [133], [149]. The Southern coasts of India contributed some 44% of the slave arrivals on the Ceylon and Batavia route to the Cape during the 18th century [149]. Ships from Ceylon carried only ~11% Ceylonese, the remainder originated from the Malabar Coast of India [133]. This may account for the absence of a unambigiuous signal from 04_C-S.India_DV_Tamil-STU.

The 08_Bangladesh_IN_Bengali-BEB cluster contributed the largest proportion ancestry of all South Asian fS-clusters, but the prevalence was low (21%). The relevance of Bengali ancestry has been alluded to before [12] and these slaves would have made up 22% of the private slaves from Batavia (present-day Jakarta) [133]. Between 1626 and 1662, Dutch exports from the Arakan - Bengal coast was regular; thus, Bengali may have been among the earliest arrivals. Cape records reflect the presence of Bengali by the frequency with which the slave name 'Van Bengal' is encountered [16].

Slave exports from the South-East Asian circuit may have increased after 1660s following political turmoil in the region [35]. From the CP-NNLS I identified no East Asian ancestry (after the exclusion criteria) but a very prevalent Malay-like ancestry from the NNLS results, thus concurring with previous work [185]. ADMIXTURE profiles indicated a possible contribution from a population which shares ancestry with 37_N.MainlandSEA_TD_CDX-Dai though this component is present in the Malay and might reflect admixture events in the history of the South-East Asian proximal ancestors. For example, Dai-related and/or Austronesian-related predecessors to the Malay [89], [211], [216].

The SAC is characterised by a high prevalence of a low proportion of Malay-related ancestry, i.e. 77% have >0.01 (0.04 ± 0.03). This prevalence is far greater than any of the other Eurasian ancestries (40% for N. European, being the next most prevalent). There are two likely sources for this ancestry based on slave import records: 1. Direct contribution from the Indonesian slaves and 2. Indirect contribution from the Malagasy slaves. Considering that German and Dutch labourers were among the earliest contributors [16], [21] and that Bengali would have arrived in similar numbers to the Indonesians [132], [149], one expectation is that both the Indonesian slave-related ancestry prevalence and average proportions should be similar to that of the European or South Asian slaves. The observed results, however, would require a substantial contribution from Indonesian slaves at the earliest onset then panmixia among the African slaves, as they made up most of the early slave community. While this is not implausible, the more likely alternative is that the ancestry was already present at low proportions in large numbers of arriving slaves. Present day Malagasy have ~40% Indonesian genetic ancestry because of the arrival of seafaring Austronesians to the Island at least 900 years BP [86], [217]. Over 500 Malagasy slaves were imported to the Cape Colony within the first decade of settlement and well before the bulk of the Indonesian slaves arrived [35], [149] and were housed with other African slaves as they were all *VOC* owned [142]. They may have integrated more readily into the broader slave community, affecting the current genomic signal. A Malagasy contribution does not preclude a direct Indonesian contribution.

I identified a 15_Pakistan_II-DV and a 32_Levant_SM ancestry from the NNLS for which there are several possible scenarios for their detection.

The 32_Levant_SM cluster produced the lowest F_{ST} values of all Eurasian clusters for comparisons with either 20_Nigeria_NK or 19_S.Africa_KS, indicating close affinity. The nearest fS cluster to 32_Levant_SM is 34_E.Mediterranean_HE-RM-SM (distance between the two; $F_{ST} = 0.005$), but this group was more distantly related to the two African clusters compared to 32_Levant_SM (F_{ST} Yoruba 0.164 & KhoeSan 0.214). Bedouin and Palestinians were among the populations in

32_Levant_SM and an African contribution to the Arab populations is expected with the Islamic slave trade [34], [82], [177] such that this signal identified in the SAC may reflect shared African admixture in the Levantine groups.

The second lowest F_{ST} values for the African clusters were observed in comparisons with the 15_Pakistan_II-DV (0.138 & 0.207, respectively), these F_{ST} values were lower than F_{ST} between other Eurasian groups and either of the African groups (e.g. Yoruba compared to South Asian clusters : 0.141- 0.178 and 0.209-0.3 for KhoeSan compared to Eurasian clusters). The cluster 15_Pakistan_II-DV consisted entirely of Brahui, Makrani and Balochi all of whom have been suggested to have African ancestry with admixture dated ~1100CE, thus possibly related to the Arab slave trade as well [82], [85], [96], [218].

Similarly, the Levant ancestry could be a surrogate for a 'Back-to-Africa' Eurasian component which is present in East African Bantu-speakers and the KhoeSan of Southern Africa, the latter as the result of a pastoralist related migration from East Africa to Southern Africa approximately 2,000 years ago [28], [93]. This component is absent from Southern Bantu groups as the Bantu expansion post-dates this admixture event in Southern Africa. In this analysis only five KhoeSan individuals from only two populations were included which might limit the representation of variation associated with the back-to-Africa migration, encouraging copying from other clusters.

The detection of 32_Levant_SM ancestry could also reflect the presence of recent African ancestry in Middle Eastern groups [38], [218] or ancestral allele sharing between the Middle Eastern populations and African groups [38], [219] which may cause haplotypes with low confidence to be assigned to the Middle East (here 32_Levant_SM).

A direct contribution is of course also possible, however I did not find reports of large-scale immigration from the region which would be necessary to account for prevalence *en par* with the Malay and South Asian sources.

Resolution on European Settler Contributions and a Novel Iberian Ancestry Detected

The two primary European contributions identified, the Northern (21_N.Europe_GM) and Central European clusters (22_C.Europe_GM-RM), are easily related to the arrival of settlers. The Northern cluster consisted of British, Dutch, German, Scandinavian, and Belgian individuals, the former three having been recorded as the major contributors to European immigration and thus the Afrikaner populations [21], from which the SAC contribution would be derived. The Central cluster consisted of French, Austrian, German and Swiss individuals. The French accounted for 20 - 30% of arrivals as French Huguenots. While the French contribution to the Afrikaners is well established, this is the first genetic evidence for a contribution to the "Coloured" population. The arrival of 180 French Huguenots in 1688 would have done little to offset the sex bias and inter-racial pairing as most women were married [21] and only when their children reached maturity could they be married to settler men. The French were fully assimilated into the Dutch society after two generations [220] and no young people spoke French by 1750, largely resulting from an active drive by the governor to restrict French traditions from taking root [21]. We can thus assume contributions from "Dutch" settlers may have included French ancestry.

As much as ~26% of the SAC individuals have >0.01 (~0.04 on average) of Iberian-related ancestry (28_Iberia_RM_IBS-Portugal-Spain) making it more prevalent than many of the signals well supported by historic texts (Central Europe, Bengali, Dravidian South Asian). It therefore seems to warrant some discussion. To-date there are no major immigration events from the Iberian Peninsula reported for the Cape Colony. The first clear reports are from 1920 and 1975 [103], the latter following the independence of former Portuguese colonies in Africa. In both cases immigration was largely inland to the Johannesburg area and such late arrivals would not produce a signal of low proportion and moderate prevalence as observed here.

I suggest two possible routes. Firstly, there may be an early contribution present in the Cape KhoeSan. The earliest records of European presence are of Portuguese sailors from 1488 at the Cape of Good Hope and Mossel Bay (~300km to the East)

[103], [142]. There are no direct accounts of reproduction but other interactions including trade, exchange of meals and conflict were regular [103], [142]. Several visits to the Khoikhoi campsites are recorded and at least two conflicts happened which reportedly involved as many as 170 Khoikhoi and 150 Portuguese. In 1509 Portuguese men were recorded to have kidnapped Khoikhoi children following a conflict [142] and we know rape is a regularly used as weapon of conflict, both in the past and present [221], [222]. It is thus plausible that there may have been an early genetic contribution from Portuguese men through rape and friendlier exchanges.

The second route of entry could come indirectly from the slaves of former Portuguese colonies. When the Dutch captured Portuguese colonies in South and South-East Asia, many of the local communities may have already had some Portuguese genetic contribution, particularly servants and slaves of Portuguese families. Mixing between people appears more common in Spanish and Portuguese colonies than British or French colonies in part due to the different autonomy allowed to women. For example, the Dutch Reformed church allowed women to divorce their husbands and to acquire his estate [21]. This allowed women in the Cape colony to discourage extra-marital affairs, in particular with slaves and servants [21]. Such power to women is not allowed under the Roman Catholic church. It is plausible that many of the South Asian slaves may have been of mixed descent at the time of their arrival. There is, unfortunately, insufficient evidence to distinguish the routes.

4.5 Conclusion

The South African Cape "Coloured" population has been shown to have a complex admixture history involving multiple sources. Our evaluation shows little evidence of substructure within the sampled community but an overwhelming congruence with historical records for South Asian, European and African ancestry. There was a striking prevalence of KhoeSan ancestry, present in 100% of individuals, suggesting a substantial contribution. A surprising lack of diversity was observed in South Asian ancestries. Dravidian ancestry was among the most prevalent, in line with accounts of slaving along the Malabar and Coromandel coast of India. The Bengali were

among the largest group imported and this is reflected in the moderate-prevalence of Bengali ancestry. The prevalence but low proportion in any individual of Malay ancestry, I propose may support an indirect acquisition through the Malagasy slaves but may include direct Indonesian contributions. For the first time there is clear evidence for contributions from distinct regional European clusters, including a central (~French/German) and Iberian (~Portuguese) contribution. There are a few possible modes of entry for the Iberian ancestry but insufficient evidence to support any particular route of entry.

5 Detecting Admixture Dynamics by Admixture Dating

5.1 Introduction

A substantial part of genomic research is understanding the link between observed patterns in genetics and the possible environmental drivers shaping these patterns [46], be they physical, social or political causes. Identifying possible causal factors behind genetic change gives predictive power, allowing the research to be used in medical, forensic and commercial applications.

Often controlled experiments and well sampled association studies are not feasible, and we rely on 'natural experiments' and more loosely set up correlation studies to estimate parameters and identify possible causes related to patterns of interest. For example, estimating generation times common to human populations required the use of data collected over several decades by several research groups [223]. In communities which are remnant and inbred, the large sample sizes necessary for studying gene - function association studies are not possible, and research relies on relating candidate genes to the population's history and culture [99], [115]. Large scale events such as pandemics, the development of polities and global climate cycles are not reproducible and research thus relies on understanding the sequence of developments in relation to changes in genomic patterns [e.g. 114], [224]–[226].

European colonialism is arguably the most important historic event in human population genomics, affecting the genetic characteristics of the majority of the populations across the Americas, Africa and Oceania [19], [63], [97], [99], [100], [181], [227]. A consequence of this colonialism was the eradication of indigenous groups, culturally and often genetically, and the establishment of populations of mixed descent [63], [98], [100], [228], [229]. Within the Indian Ocean slavery circuit, colonial cities were exceptionally diverse as slaves and traders often arrived from throughout the circuits. The diversity extended across social rank particularly in South-East Asia [35] however the societal dynamics for the lowest ranks of society are often poorly recorded making it near impossible to accurately understand the drivers of admixture dynamics. Considering that inhabitants of colonial cities in the Indian Ocean Slave Trade (IOST) curcuits were largely slave and servants, the accounts of the lives of the largest constituent of these centres are missing from the narrative. Genomic research offers an avenue to investigate some aspect of this by allowing some study of temporal political changes in relation to dates of admixture.

The South African colonial cities were no exception to the IOST diversity as reflected in genetic data and historic accounts [19], [35], [99], [133, Chapter 4]. However, the diversity has also made it difficult to interpret ancestral contributions as being either proximal (colonial era) or distal (pre-dating the colonial settlement). Recently developed statistically procedures now allow for such resolution.

From Chapter 4 several signals were easily reconciled with historic accounts, however, I detected additional signals which require further investigation. A prevalent Iberian ancestry was present in relatively high proportions across much of the SAC dataset. Europe was well represented in the dataset, suggesting that this is not an artificial signal but a genuine Iberian contribution. As the signal is novel and there are no clear historic account of direct contributions, I suggested that the signal possibly may come from undocumented admixtures preceding the *VOC* settlement. The first suggestion was of an admixture between the 1400s Portuguese sailors who stopped along the coast and the indigenous KhoeSan communities. Alternatively, the Dutch settlers may have brought over slaves of mixed descent

which they acquired from former Portuguese colonies. The results from the previous chapter did not allow further discussion of which was most likely.

I also found Malay ancestry in the SAC which could be derived from a direct Indonesian source as a variety of South-East Asians were imported according to historic records [35], [152]. However, the largest and earliest imports were of Malagasy slaves, who have ~40% ancestry Indonesian ancestry related to the South Kalimantan Dayak [86]. These Indonesian sailors would have arrived in Madagascar ~1,000 - 2,000 BP [86], [230].

As the Malagasy contributed earliest to the SAC, their ancestry should be prevalent. The Indonesian slaves were imported in similar numbers as the Malagasy toward the end slavery at the Cape, representing 25% of arrivals each [133]. However, the later arrival should lead to a lower prevalence which should be similar to the South Asian ancestries as South Asia and South-East Asia both became increasingly important sources into the 1700s [35]. The high prevalence but low proportion ancestry in any individual found in chapter 4 support a Malagasy contribution but does not rule out a direct Indonesian contribution. Thus, it remains unclear from which source the Malay ancestry is derived.

In this chapter I use linkage disequilibrium decay curves to estimate admixture dates. Dating may add some resolution allowing us to understand the relationship between these identified components and the admixture dynamics. I predict an admixture date pre-dating the colonial settlement for the Iberian ancestry. While I cannot test for a Malagasy ancestry directly in this chapter, I can predict that a second, older admixture event between the Malay and Bantu seen in the South African "Coloured" would reflect the pre-Colonial Malagasy admixture [see 55], [86]. While if the admixture is restricted to a recent event, I can conclude that most if not all of the ancestry is from a direct Indonesian source.

5.2 Methods and Data Analysis

All supplementary material for this chapter is available in Appendix B.

Admixture Dating using Linkage Disequilibrium Decay Curves

I examine the dates of the admixture events within the SAC dataset by fitting the exponential decay curves for admixture linkage disequilibrium (LD) against the increase in distance between SNP pairs using MALDER (Multiple Admixture Linkage Disequilibrium for Evolutionary Relationships) [78]. The parameters used to fit the decay curve can be related to admixture parameters, specifically the decay constant is related to time in generations and the amplitude of the curve to the relative contribution of the proposed source populations [78], [93], [118]. An inter-cultural meta-study has indicated that inter-generation times for human populations are around 28 generations [223]. I here use 29 years, in line with other researchers [see 108], [177], [225]. Dates are calculated as $1960 - 29\ years.gen.^{-1} \times no.\ of\ generations$, where 1960 is the median date of birth of the sampled SAC.

Detecting admixture events dating to several thousand years ago may complicate the interpretation of the results as signals pre-dating the arrival to South Africa are possibly still detectable. To by-pass this issue, I opted to set the d_0 value, the smallest distance bin considered for curve fitting, to 1.9 cM (`mindis: 0.019`). This would focus the results on relatively recent events by excluding linkage disequilibrium associated with the oldest events and thus shortest haplotypes. This process further alleviates an additional issue. Due to the presence of background LD, MALDER is overly conservative when admixture is younger than 20 generations ago as it detects long-range LD correlation and fails to complete (Appendix C in [78]). A large d_0 value removes the detected long-range LD allowing us to estimate admixture dates for recent events. This does not entirely prevent detection of ancient admixture, as we find dates as far back as 30 generations ago are detectable (see section 5.3 and Chapter 6). Manual inspection of the pairwise curves fit for varying d_0 values (0.5 - 2.2 cM) showed only marginal downward shift (~2 generation) in the date estimates between the extremes d_0 values, suggesting that this will not substantially bias the dates detected. Furthermore, when d_0 is not set and instead

estimated by MALDER, it often selects d_0 >1 cM for African populations [see Supplementary Information in 177].

The standard errors are estimated by jack-knifing over chromosomes and a p-value estimated by dividing the mean by the estimated standard error to retrieve the Z-score.

The analysis was run on each of the fS-inferred SAC clusters using all of the fS-inferred GR clusters as sources. The results were then sorted by amplitudes and processed in R to identify the top-ranked pairs of sources.

Bootstrap evaluation of signals

As it is possible to identify a signal of admixture from potentially any subset of the SAC dataset, I needed to evaluate to what extent the MALDER signals detected in the fS-inferred SAC clusters were different from a random subset. This allowed me to evaluate: 1. which signals are omnipresent in the SAC data set and 2. which signals found in the SAC fS-inferred clusters were statistically significant (i.e. unique to that cluster).

I performed 1,000 bootstrap resampling from the SAC data set with replacement. Sample sizes varied between 5 - 99 (Figure 5.1). Seven iterations failed to complete, potentially due to long-range LD detection, and were excluded. The overlap between bootstrap sample sets was less than 22% (99th percentile) and no pair had >55% overlap (Figure 5.2).

I tested if the signals detected in the bootstrap iterations were similar to those identified in the fS-inferred clusters due to sample overlap. I calculated what proportion of samples from the SAC fS cluster were present in the bootstrap iterations which produced analogous signals ($\frac{sample\ overlap}{total\ samples\ in\ fS-cluster}$). I performed a t-test of this proportion and a randomly selected subset of iterations in which the signal was not identified, matching for sample size to test if there were significant differences between iterations which produced analogous signals compared to those that did not.

Identifying sources

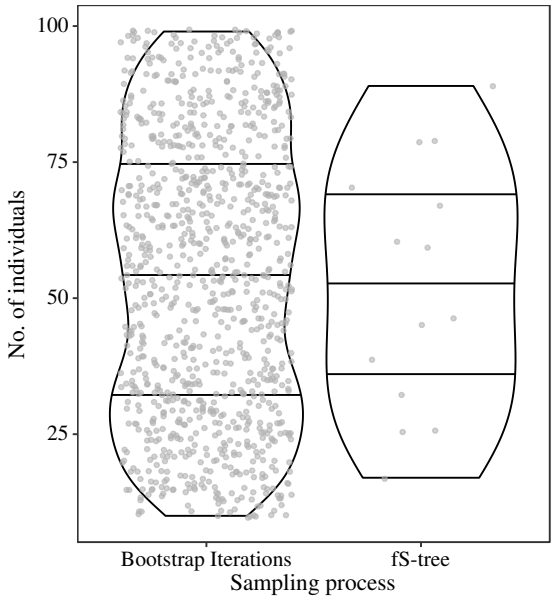


Figure 5.1: Sample size distribution for the 993 bootstrap samples compared to those retrieved for the fS-inferred SAC clusters.

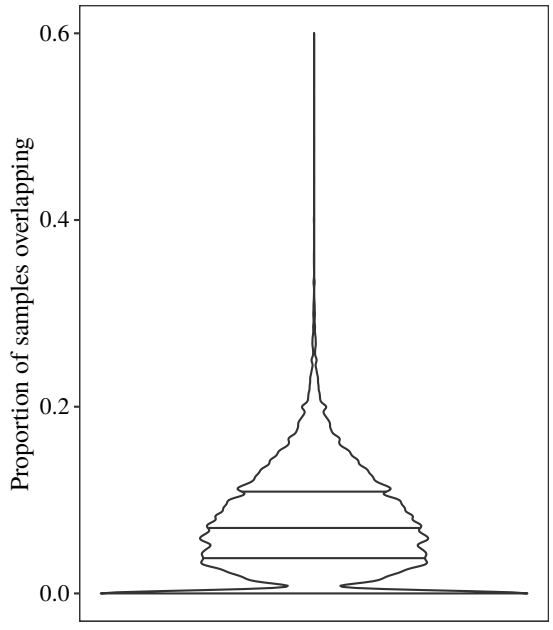


Figure 5.2: Sample overlap between MALDER bootstrap iterations. Values indicated as a proportion of the samples from iteration i found in iteration j. Black horizontal lines indicate the 25th, 50th and 75th percentile.

The results of MALDER are influenced by the tree-based relationship between source fS-clusters and as such the ranking of amplitudes which is normally employed is influenced by the number of clusters on various branches and the gaps in representation [see 78]. The amplitudes produced by MALDER are dependent on allele frequency differences between the reference populations and the LD decay observed in the admixed population [78, formula (4)] $\sum_{S(d)} z(x,y)w(x)w(y)$. It considers a bi-allelic system (Allele 1 and 0), estimates the covariance in the admixed population for this allele over varying distances between two loci, and weights the covariance estimates using allele frequency differences (at Allele 1). The amplitude is expected to be maximum when the proposed sources are exactly related to the true sources. Given two actual admixing sources, A and B, I could consider two proposed proxies for the actual sources, A' and B'. Where A' is known to be the true source, A, I could vary B' and find the source which produces the maximum amplitude to identify the best representative of true source, B [78]. However, this is not possible when A is not known nor if I wish not to assume a true source. In our case here, there are likely multiple sources of ancestry within a section of the phylogeny such that shifting B' away from B but toward a third source, D, may result in an increase in the amplitude. There may be several local maxima in amplitude across the phylogeny of populations.

Any pair of populations with large allele frequency differences can increase the amplitudes observed. A single reference test (using the admixed population as the second reference) is used to control for identifying false pairs [78]. When there are multiple contributing ancestries to an admixed population, many more populations may pass the single reference test as a consequence of shared distal or proximate ancestors with the actual ancestral populations for the admixed group. Thus, the possible number of source pairs increases rapidly, and it becomes difficult to distinguish the signals for multiple admixing sources from artefactual signals based on the inclusion of multiple 'next best' proxies which in fact did not contribute directly.

To overcome this complication, I ranked the amplitudes based on the F_{ST} values between proposed GR clusters used in the curve fitting. The comparisons were then binned into 26 ranks of approximately equal sample size (~26). Within each bin of F_{ST} values I identified the amplitudes in the top 90%. I could thus identify contributions between sources even when their genetic distances are low but simultaneously ruling out signals which may be simply a consequence of using a B' population both notably far from the actual B source and close to ancestral source D.

Results were visualised using a chord plot produced in R with the package `circilize` [231].

Three Population Test for Admixture (f_3)

I estimated the f_3 parameter as a formal test for admixture [185], [232]. The three-population test and MALDER use different information to detect admixture, complementing the interpretation of the other. The f_3 indices use shared drift between populations to measure the internal branch in an assumed tree-like relationship. When a tree-like relationship is supported, the f_3 values are positive, when admixture has occurred, the external branch of the tree have heightened lengths, resulting in negative f_3 values. Estimates were made between all GR fS clusters in the form $f_3(X, Y; SAC)$ where X and Y are GR clusters. Negative f_3 values ($Z\ score\ <\ -3$) are considered indicative of a discordant tree relationship and in support of admixture [232]. Estimates were made using Admixtools v.5.1 [1].

5.3 Results

5.3.1 Admixture Dating

The highest ranking amplitudes for LD decay curves fit by MALDER for the fS-inferred SAC clusters included an African group (typically 19_S.Africa_KS) and a European group as best source proxies (Figure 5.3). The estimated admixture dates were 5 - 8 generations ago for almost all SAC clusters except cluster 14_17SAC, where only an older pre-16th century KhoeSan - Yoruba admixture event was identified (14 - 22 generations ago; ~406 - 638 BP; ~1322 - 1554 CE). There was

variation in the dates estimated across clusters, but the error intervals across clusters 02 -13 overlapped, suggesting they reflect the same history. Cluster 01_46SAC produced two events which did not overlap with dates for 02 - 13. The earliest event for this cluster (~14 generations ago; ~1554 CE) did overlap with the single event found for 14_17SAC.

For cluster 08_25SAC, I detect an ancient 01_S.India_DV - 19_S.Africa_KS admixture dated over 30 generations ago (870 BP, 1090 CE). This is not easily explained as it may reflect admixture in the history of any of the multiple sources that have contributed to the SAC, for this reason it is not discussed further.

All of the fS-cluster date estimates were reproduced when bootstrapping over samples from the full 733 individuals and re-running MALDER (Figure 5.3), but some fS-cluster events were qualitatively unique. In particular, the late and early events of 01_46SAC (block (a) in Figure 5.3) were readily reproduced (recovered in 80% of iterations (late) and 10% (early)) (Supp. Table B.1).

These bootstrap date estimates (~9 - 18 generations ago) produced were not continuous. The distribution appears bimodal centred on 5 and 14 generations ago, respectively (1554 and 1815 CE) (Figure 5.3), indicating the possibility of two waves of admixture.

The earliest events were identified when the late event was younger than ~5 generations ago, demonstrating that the detection of two events may be artificial and most likely the signal reflects continuous admixture (Figure 5.3) at least for the most recent of the earlier dates (0 - 8 generations ago). Indeed the early events when two events are detected have an overlapping distribution with single events (Figure 5.3).

The GR fS-clusters identified as sources by the bootstrapping were similar across early and late events, with the notable exception that a Yoruba proxy was detected for younger events (Figure 5.4). Events involving 20_Nigeria_NK as a source were restricted to pairs with a South-East Asian and Northern European source, both 3 generations ago (1873 CE, n = 2) suggesting Colonial era admixture. The dates involving any South-East Asian sources (37_N.MainlandSEA_TD or 41_E.Is.SEA_MP) included the earliest event, ~31 generations ago (~1061 CE,

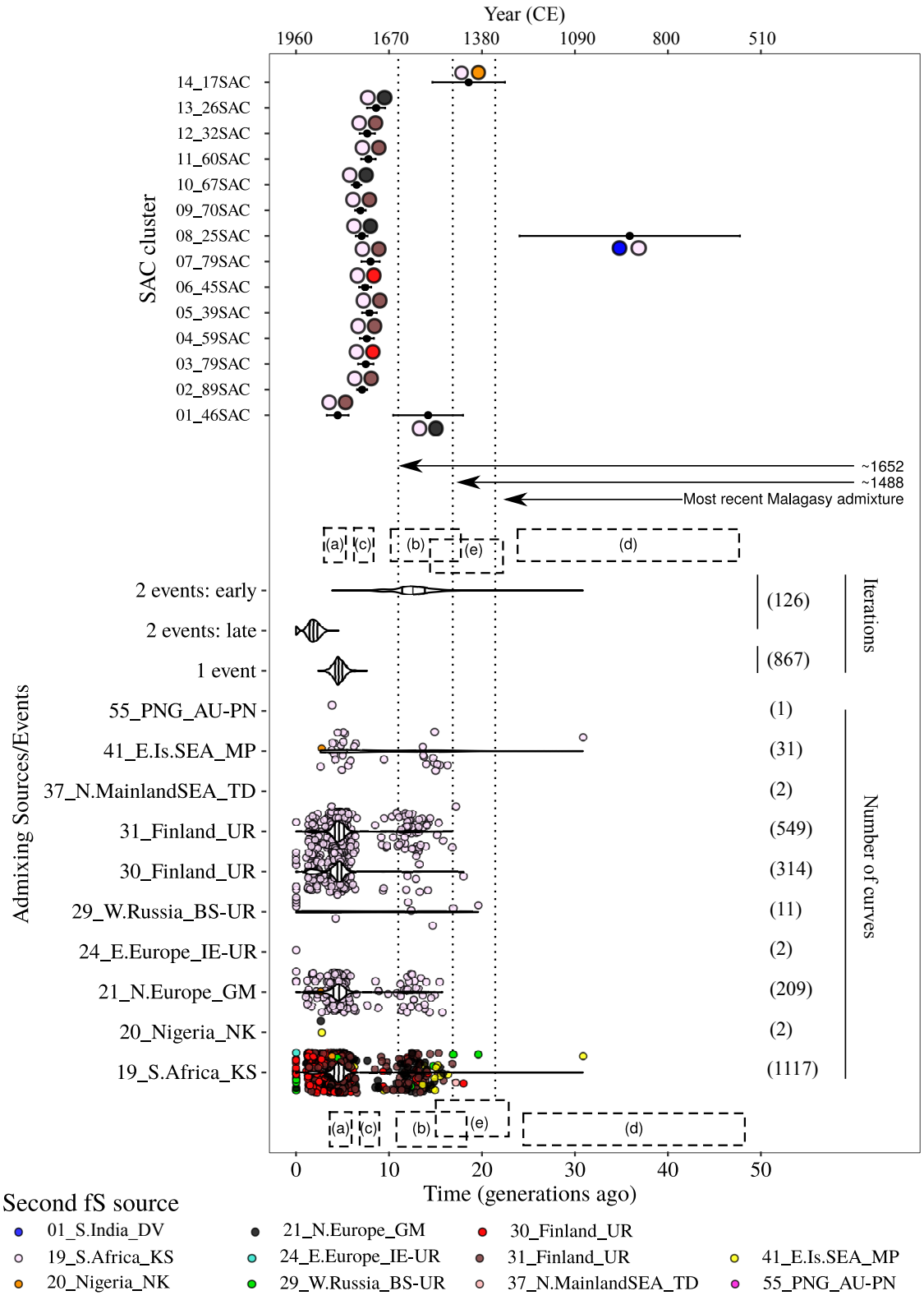


Figure 5.3: Admixture sources and date estimates from top-ranked MALDER LD decay curves. Top panel: Top result per fS-inferred SAC cluster presented as sorted by amplitude of the curves. Bottom panel: Bootstrap iterations (n=993). Violins show range of date estimates involving various sources and for single or two-event detections (early and late).

Figure 5.3: Continued. Dots indicate the second source contributing to the admixture. Number of iterations/curves for each violin shown on the right (n). Dashed-line blocks indicate the events detected by the fS-inferred clusters; (a - b) show events from 01_46SAC. (c) shows the point estimates for events from clusters 02-13. (d) shows the second event of 08_25SAC and (e) shows the event of 14_17SAC. Vertical dotted lines indicate historic events of interest; 1652 - Arrival of the Dutch, 1488 - Arrival of the Portuguese, and the most recent estimates for Bantu - Indonesian admixture on Madagascar [217].

n = 1 iteration) and later events 3 - 17 generations ago (1467 - 1873 CE; n = 98) (Figure 5.3), the latter being bimodal.

The early event from 14_17SAC (2% of iterations) and the ancient event found in 08_25SAC (0.1% of iterations) were not readily reproducible by bootstrapping. The bootstrap pre-colonial events analogous to that of 14_17SAC and 08_25SAC did not recover the same sources, indicating some qualitative difference. Dates comparable to 14_17SAC recovered KhoeSan - Europe sources and overlapped with the pre-colonial event from 01_46SAC so these iterations may better reflect the earliest European contributions and not a Bantu - KhoeSan admixture.

The majority of the SAC clusters (clusters 02 - 13), however, produced a date which was infrequently reproduced (block (c) in Figure 5.3), appearing in <3% of the bootstrap iterations. Even when considering the entire range of block (c) (Figure 5.3), this only increases to 3.6%. This suggests that the events of block (c) cannot be reproduced by chance sampling. This is peculiar as block (c) represents the bulk of the SAC data (670 individuals). These dates, which are between the oldest events (2 events: early in Figure 5.3) and the youngest events (2 events: late), may not be a consequence of the LD decay curves being averaged between the oldest and youngest dates. As the dates of cluster 02 - 13 are notably older than the most commonly recovered dates by bootstrapping, it would suggest that the fS clustering has partitioned the youngest admixture events among clusters (e.g. in 01_46SAC and 14_17SAC), allowing the detection of events at the upper edge of the 'recent' range (<10 generations ago).

To test if the inclusion of samples from specific SAC clusters influenced the age of detected events, I examined the sample overlap between iterations and the SAC

clusters. I found that detection was in general not related to the sample overlap with the respective SAC clusters. For example bootstrap detected events comparable to the early events of 01_46SAC, 08_25SAC or 14_17SAC did not come from samples with high overlap with these SAC clusters ($p>0.05$) (Supp. Table B.1). Signficiant differences in sample overlap between iterations matching versus iterations not matching specific events were restricted to SAC clusters of block (c). Counter-intuitively, the iterations producing matching events had less sample overlap than those producing non-matching events. This was the case in four instances. This can be explained by the measure of overlap. I use $\frac{No.\ individuals\ in\ SAC}{No.\ individuals\ in\ iteration}$ to measure overlap. For the events from SAC clusters 02 - 13, the MALDER dates and error bars reflect by-en-large the same event. As such, when testing for sample overlap from any one of the SAC clusters from this set (block c), there are more individuals from the other cluster available to drive the same signal. Thus the denominator increases and the overlap score declines.

I further tested the combined individuals from two suites of fS clusters; block (c) and 01_46SAC + 14_17SAC which have the youngest and older events. I find that, for block (c), there is no relationship between sample overlap and events comparable to the SAC cluster event. In contrast, samples from 01_46SAC + 14_17SAC were at higher proportions in iterations which identified these respective dates (block a,b,e in Figure 5.3). Thus the individuals from the SAC clusters which produced the youngest and some of the oldest date estimates appear to influence the date estimates for the bootstrap iterations.

While the fS-inferred clusters appear to account for specific sets of admixture events (and thus geneologies), the recovered dates are not likely the only signals present. This is evident in that fS clusters did not recover a top-ranked curve with South-East Asians while a large number of the bootstrap iterations did.

Identifying admixing sources using MALDER in populations with multiple contributors is a challenging task [233]. As expected the top results from MALDER seems to highlight the importance of allele frequency differences between proposed sources as indicated by the frequency with which unlikely sources are identified

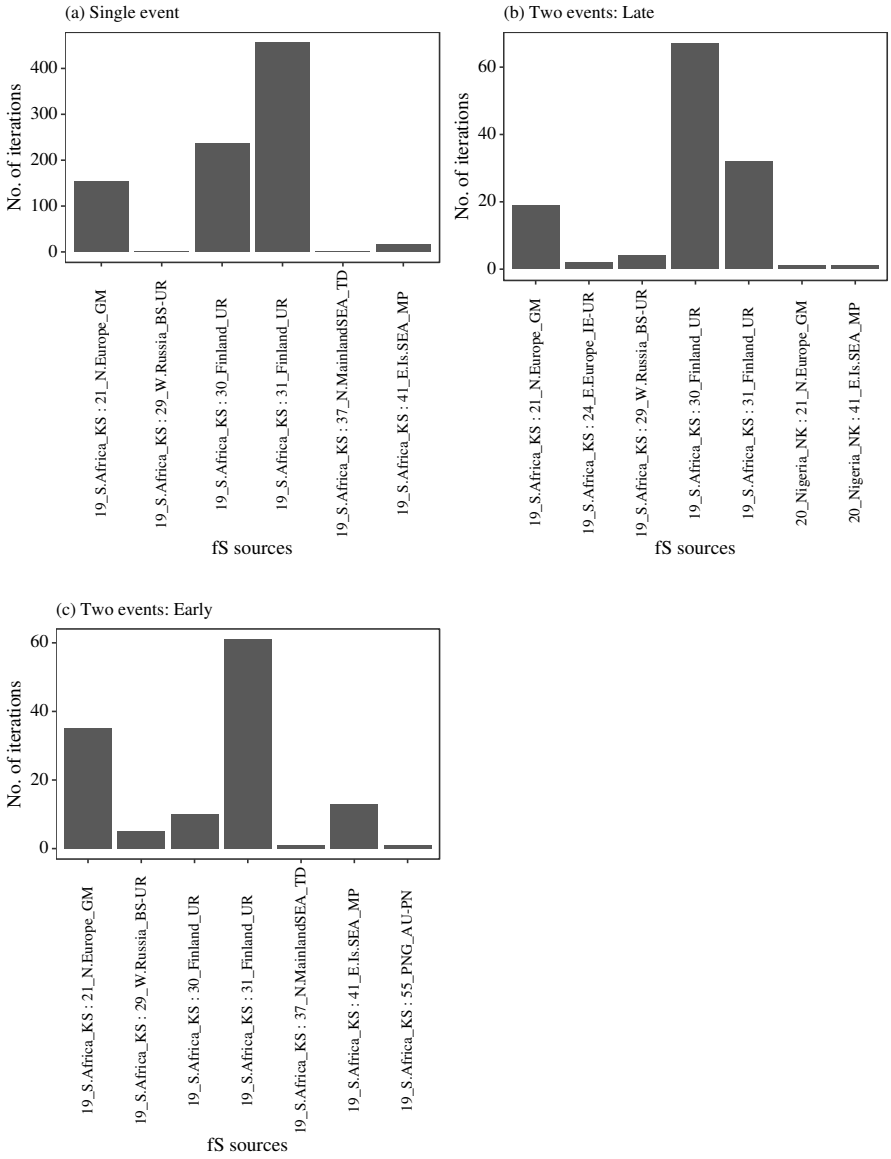


Figure 5.4: Frequency of the top pairs of sources identified from the bootstrap iterations. Indicated (a) are the sole events for iterations identifying single events , (b) the late event and (c) early event for iterations identifying two events.

in place of the most likely source, e.g. Finnish identified more frequently than 21_N.Europe_GM, and 41_E.Is.SEA_MP and 55_PNG_AU-PN are identified more frequently than more likely South-East Asian sources.

One cannot simply take the highest amplitude as representing the best sources nor can one easily rule out other sources as being redundant signals of the same ancestry. In an attempt to clear up the signal, I binned the amplitudes using the corresponding F_{ST} values for the two GR clusters considered in the curve (see Section 5.2) and identified the top 90% of curves.) These were considered to be the best signals for a given genetic distance between possible sources.

For all SAC clusters except 14_17SAC, and for both early and late events, I identified 0 - 3 top results per bin (Supp. Figure B.1 & B.2). For 14_17SAC too few curves were significant per bin for the results to be useful, I therefore examine all curves for this cluster .

Between 30 - 46 GR clusters (of 56) were identified as possible sources for the first event. I similarly identified 30 and 38 for the second event (01_46SAC and 08_25SAC, respectively). Many of these clusters were identified consistently across the SAC clusters (Figure 5.5). The overwhelming diversity of possible sources suggests that inter-regional allele frequency divergences may still be driving the signals observed, rather than a genuine multi-population admixture signal. These results do however help confirm that there are contributions from several regional sources outside of Africa.

By focusing on curves fit between proxies from the same global regions (intra-regional curves) within the top 90% within each F_{ST} bin I can evaluate the possibility of multiple contributions from that global region (Supp. Figure B.3 - B.10).

There are only a few intra-regional curves successfully fit by MALDER. This suggests that many of the clusters identified are redundant proxies for the same source. Intra-regional curves involving pairs of South Asian GR clusters are identified frequently. Typically, these were curves involving a North Indian/Pakistani cluster with a South Indian/Dravidian cluster, the 14_Andaman.Is_JO with any South Asian group, or a South Indian/Dravidian cluster with an Indo-European/Indic

cluster. In some cases, the curves fit were between clusters that were close together on the fS tree, indicating low *TVD* distances, e.g. two Dravidian clusters for 07_79SAC, two Indo-European clusters for 09_70SAC, two Pakistani clusters for 11_60SAC and two central Indian clusters for 12_32SAC and 07_79SAC, each. There is thus good support for a contribution from both of the respective sources in these cases, indicating multiple South Asian contributions to the SAC on the whole.

Three additional intra-regional curves were recovered. In SAC cluster 06_45SAC I found a 36_E.China_SN-KR - 49_Siberia_TG-TK curve. In 12_32SAC I recovered two intra-Western Eurasian curves involving 32_Levant_SM with 26_Iberia_BQ-RM and 23_S.Europe_AL-BS, respectively. This provides some support for multiple contributions from these regions (Eastern Asia and Western Eurasia).

In cluster 14_17SAC the signals were clearly driven by a few sources (Figure 5.7). The greatest amplitude is seen with KhoeSan - Yoruba comparison but both also produce curves with the rest of the GR clusters, suggesting a Eurasian contribution. The amplitudes of curves fit involving Yoruba in particular suggest a South Asian contribution corresponding to Dravidian and Indo-European language groups from North and South India. Additional curves are fit between Eurasian fS-inferred clusters. There is thus some evidence of a broad Eurasian contribution (Europe, Near East, South Asia and East Asia) though it seems marginal based on the amplitudes suggesting either ancient or incidental contributions.

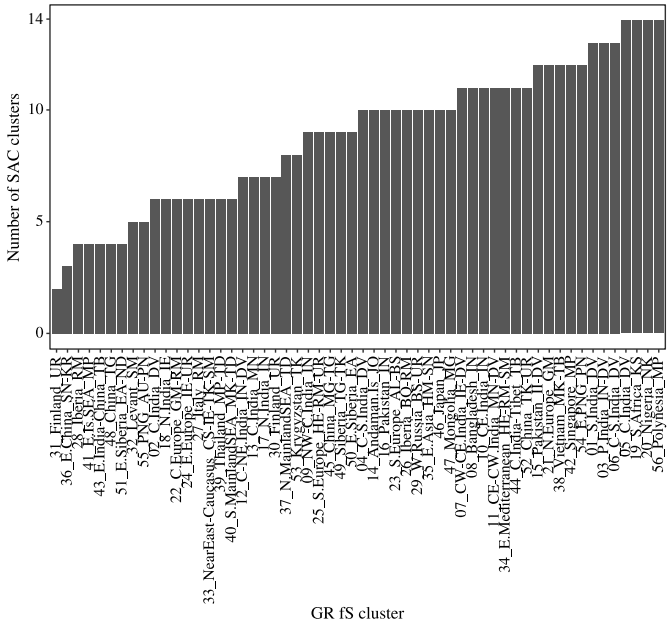


Figure 5.5: Frequency of appearance of GR fS clusters in the top 90% of curves from F_{ST}-based binning of the first events. Frequency based on occurrence across SAC clusters (max = 14).

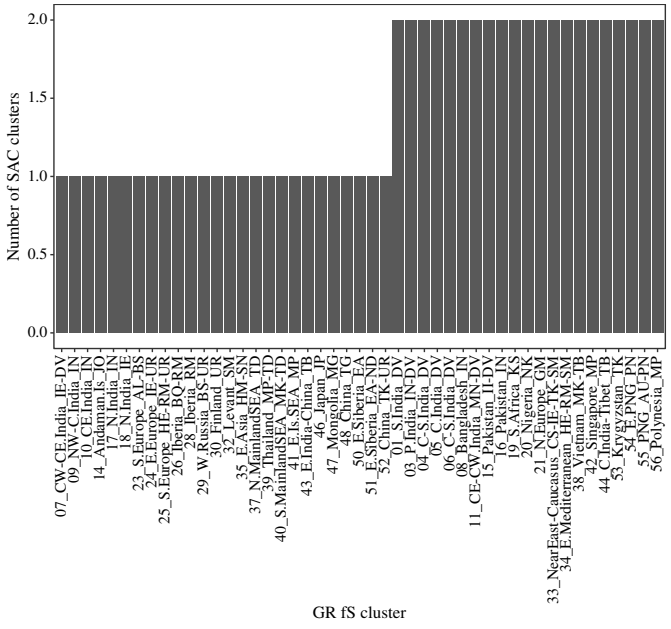


Figure 5.6: Frequency of appearance of GR fS clusters in the top 90% of curves from F_{ST}-based binning of the second events. Frequency based on occurrence across SAC clusters (max = 2).

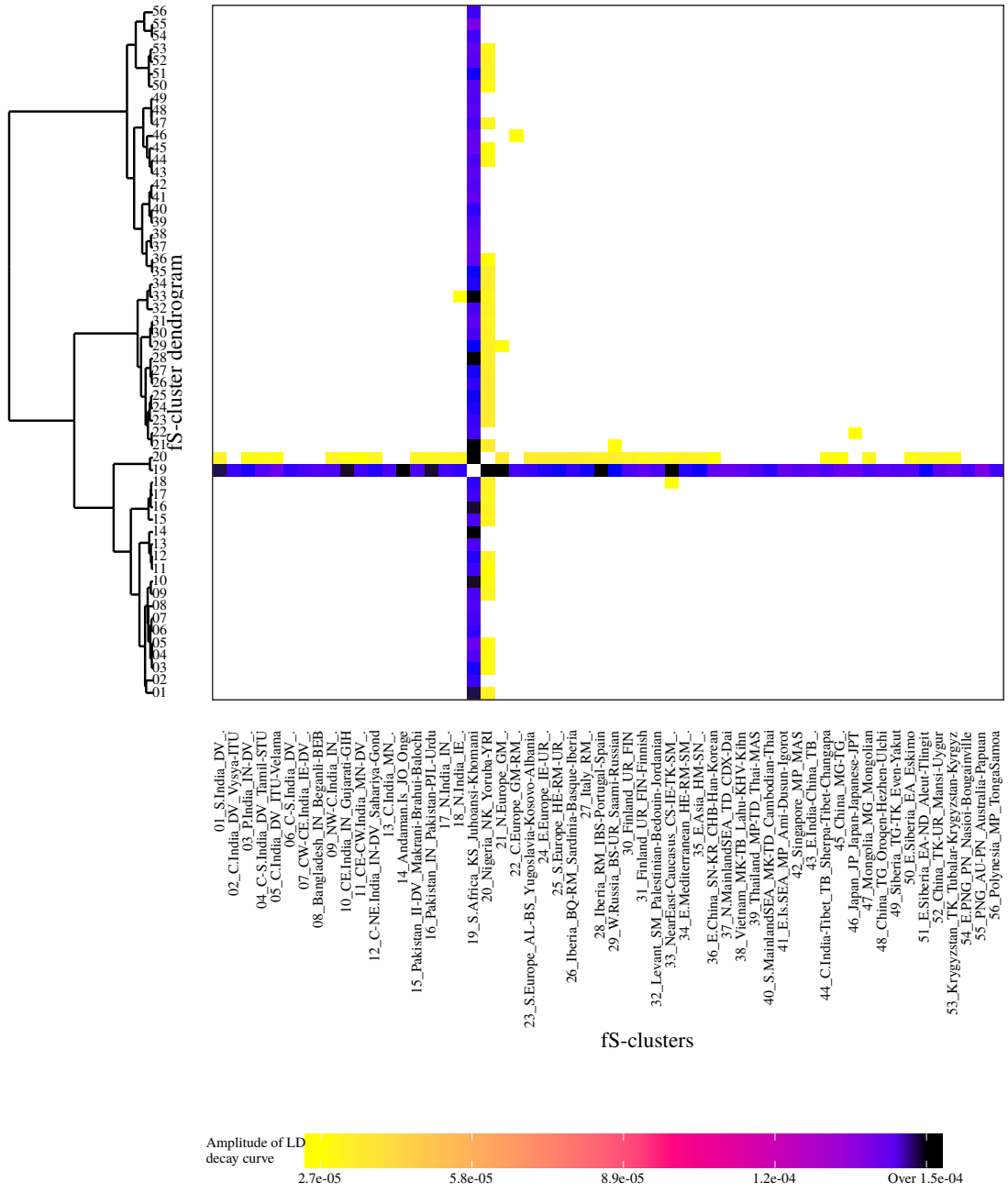


Figure 5.7: Heatmap of LD decay curve amplitudes from curves fit with GR fS clusters to SAC cluster 14_07SAC . Where no curve was fit (no admixture), blocks are left white. GR fS cluster index and fS tree indicated on the left.

5.3.2 f_3 Admixture Test

I estimated the f_3 indices as a formal test for admixture between GR clusters using the SAC fS clusters as target populations [185], [232]. The results contrasts with MALDER, as negative f_3 values ($Z\ score\ <\ -3$) are only detected with $f_3(X, Africa; SAC)$ in all but one SAC cluster, where X is any GR cluster (Figure 5.8 - 5.12). This supports admixture between Eurasia and Africa but not between Eurasian regions. The patterns in $f_3(X, Africa; SAC)$ values were similar across SAC clusters, including in 14_17SAC. Source clusters, 14_Andaman.Is_JO, the Melanesian groups (54 - 55) and clusters with known African/Middle Eastern ancestry (e.g. 32_Levant_SM, 15_Pakistan_II-DV, and 28_Iberia_RM within Europe) consistently produced the largest (but negative) f_3 values. The lowest values across GR clusters are seen with European GR clusters, in particular Finnish and 29_W.Russia_BS-UR. Local minima within sections of the GR dendrogram are observed indicating possible differential contributions. In particular I highlight 10_CE.India_IN, North India/Pakistani groups (16 -18) within Southern Asia and 41_E.Is.SEA_MP and 51_E.Siberia_EA-ND in Eastern Eurasia.

Only for 10_67SAC were negative f_3 values detected for $f_3(Eurasia, Eurasia; SAC)$ (Figure 5.13). Here the lowest f_3 values within local sections of the GR dendrogram differ to those above. The lowest f_3 for South Asian clusters included 11_CE-CW.India_MD-DV and 14_Andaman.Is_JO, within Western Eurasia 26_Iberia_BQ-RM, 32_Levant_SM, 34_E.Mediterranean_HE-RM-SM produced the lowest. Within Eastern Eurasia, 41_E.Is.SEA_MP produced the lowest. Notably, though, no negative f_3 values are fit between clusters within regions (e.g. no intra-Europe).

I detected negative f_3 values for $f_3(Eurasia, Africa; 14\_17SAC)$ supporting admixture in 14_17SAC involving a Eurasian group. However, the lowest values were fit with $f_3(19\_S.Africa\_KS, 20\_Nigeria\_NK; 14\_17SAC)$ (Figure 5.12) and were far lower than values fit with Eurasian groups. The neighbouring SAC cluster, 13_26SAC, produced lower $f_3(Eurasia, 19\_S.Africa\_KS; SAC)$ values than $f_3(20\_Nigeria\_NK, 19\_S.Africa\_KS; SAC)$, indicating better support

which is in contrast to 14_17SAC, despite similar proportions of KhoeSan, non-KhoeSan African and Eurasian ancestry (see Figure 4.24). This may indicate that the signal for Eurasian ancestry in 14_17SAC may be in relation to a longer branch length between 19_S.Africa_KS and 20_Nigeria_NK, relative to Eurasia - 20_Nigeria_NK. In this case non-KhoeSan ancestry in 14_17SAC produces a signal for admixture but f_3 values are less negative due to drift within the Eurasians but not in 14_17SAC.

The absence of signals of admixture between Eurasian groups does not reflect genuine absence as is shown from the ADMIXTURE and CP-fS results. The f_3 test is weaker at detecting admixture than MALDER when mixing proportions are low [78]. Similarly, post-admixture drift in the admixed population may decrease sensitivity. Since several ancestral contributions are present and the predominant representation is African ancestry, many sources may fail to produce significant results [78], [232]. Cluster 10_67SAC has the largest non-African mixing proportions (Figure 4.27) such that the greater number of negative f_3 values is in line with the expectation.

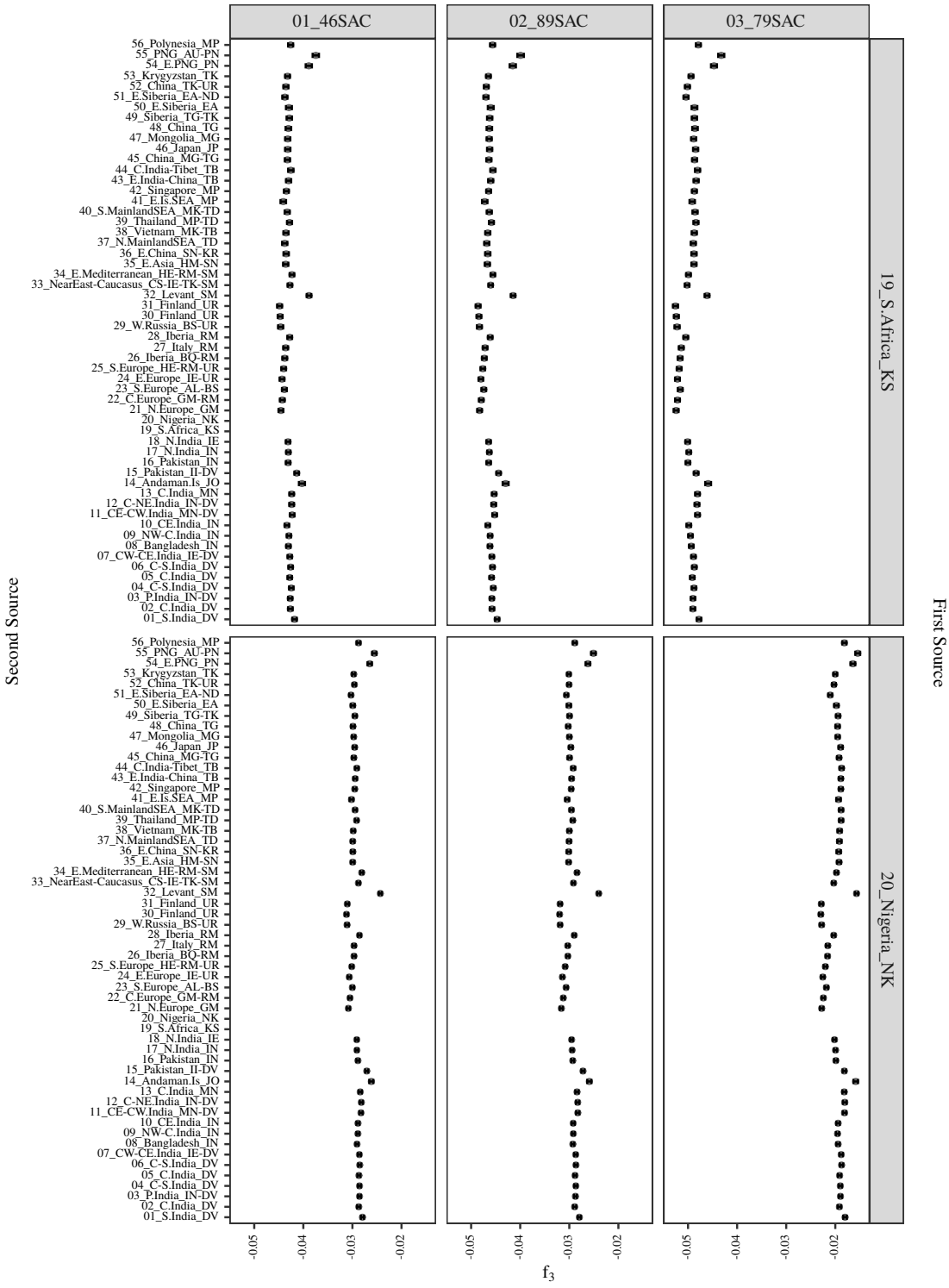


Figure 5.8: Negative f_3 estimates for $f_3(X, Y; SAC)$ for SAC fS clusters 01-03. X and Y are GR fS-inferred clusters, as indicated to the left and right of the plot. SAC clusters indicated above the panels. Black lines indicate standard error. Only values with *Z scores* < -3 are shown. Only estimates including 19_S.Africa_KS and 20_Nigeria_NK are shown as all negative value included these as a source.

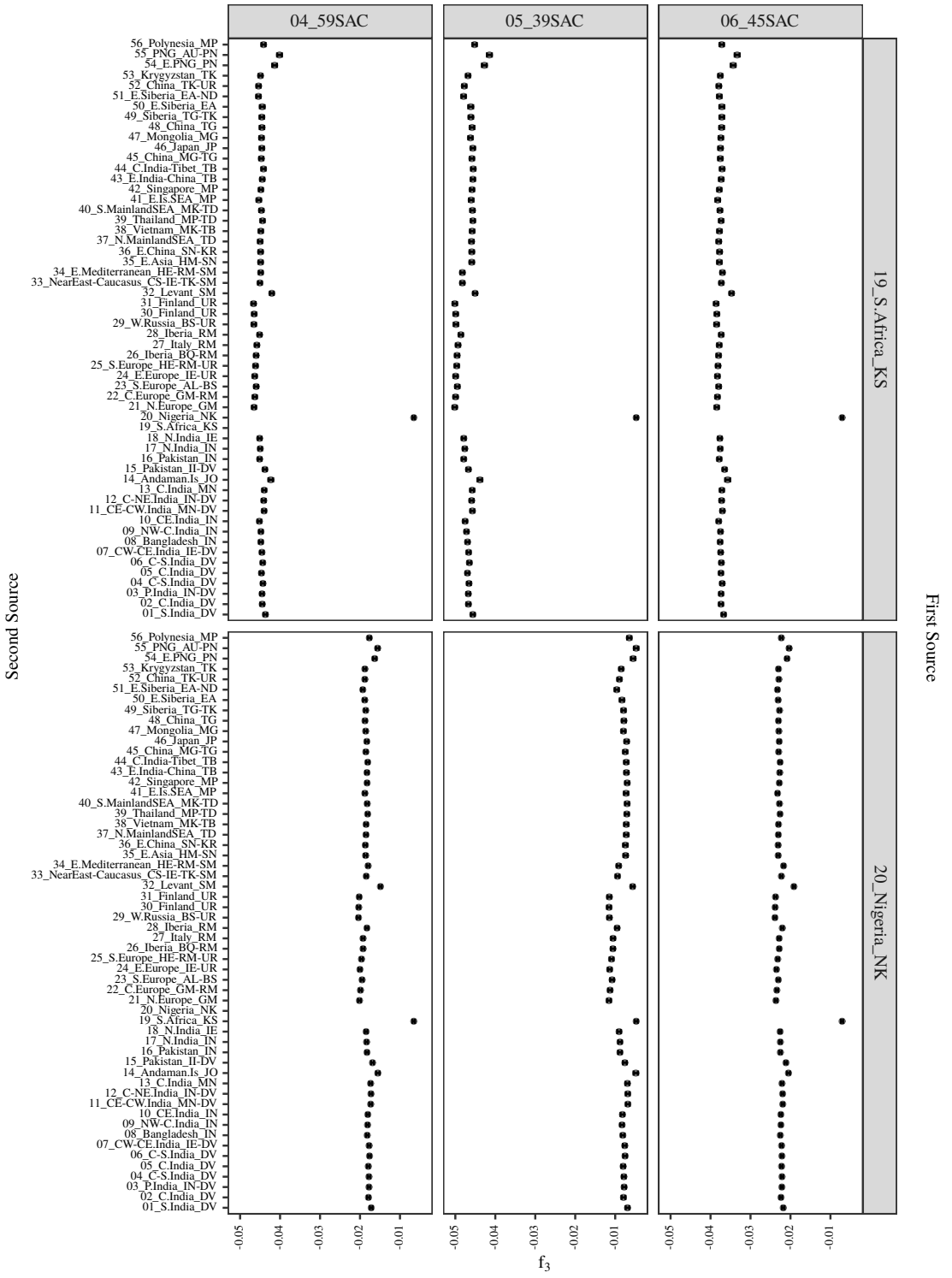


Figure 5.9: Negative f_3 estimates for $f_3(X, Y; SAC)$ for SAC fS clusters 04-06 X and Y are GR fS-inferred clusters, as indicated to the left and right of the plot. SAC clusters indicated above the panels. Black lines indicate standard error. Only values with *Z scores* < -3 are shown. Only estimates including 19_S.Africa_KS and 20_Nigeria_NK are shown as all negative value included these as a source.

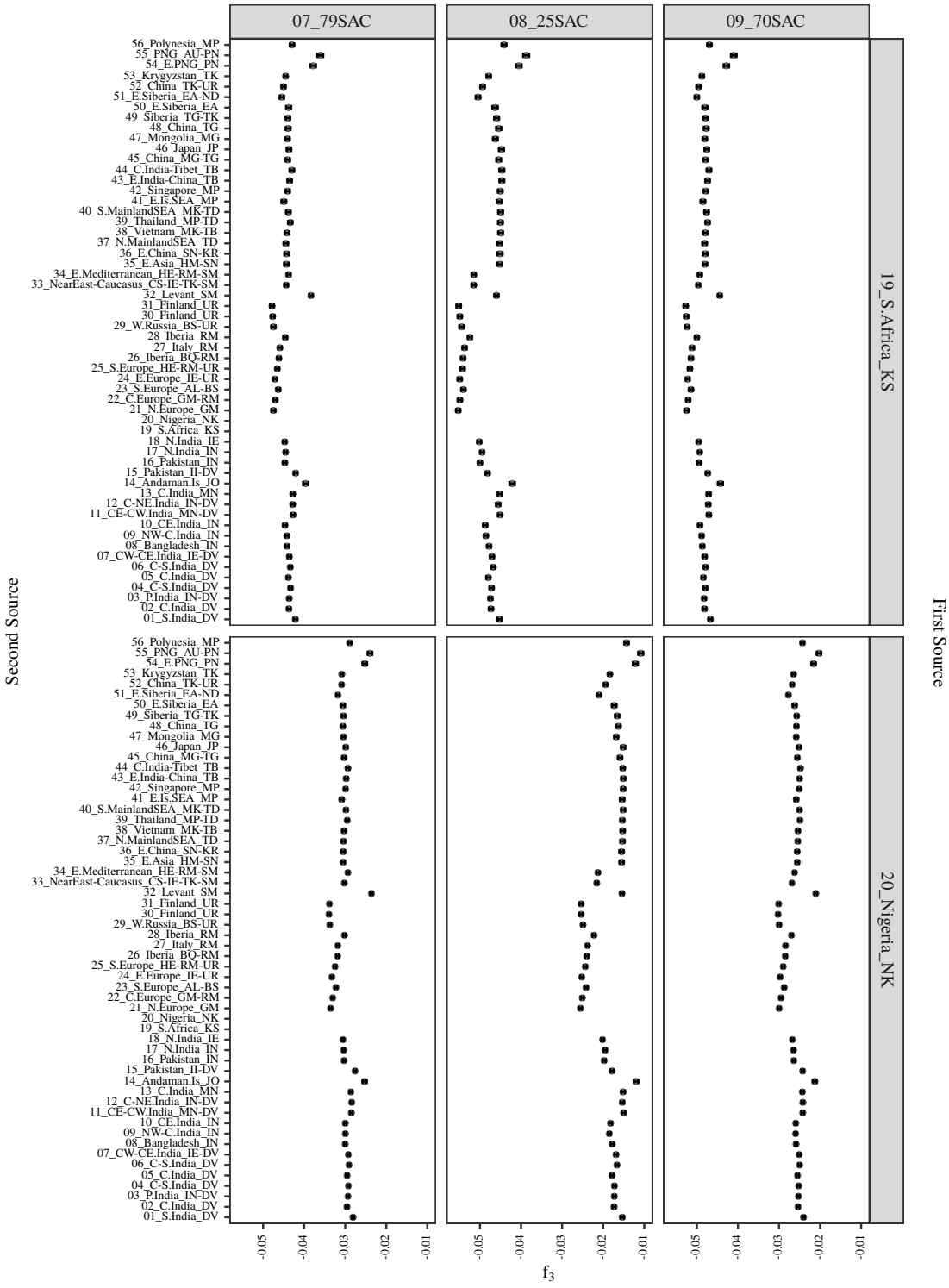


Figure 5.10: Negative f_3 estimates for $f_3(X, Y; SAC)$ for SAC fS clusters 07-09 X and Y are GR fS-inferred clusters, as indicated to the left and right of the plot. SAC clusters indicated above the panels. Black lines indicate standard error. Only values with *Z scores* < -3 are shown. Only estimates including 19_S.Africa_KS and 20_Nigeria_NK are shown as all negative value included these as a source.

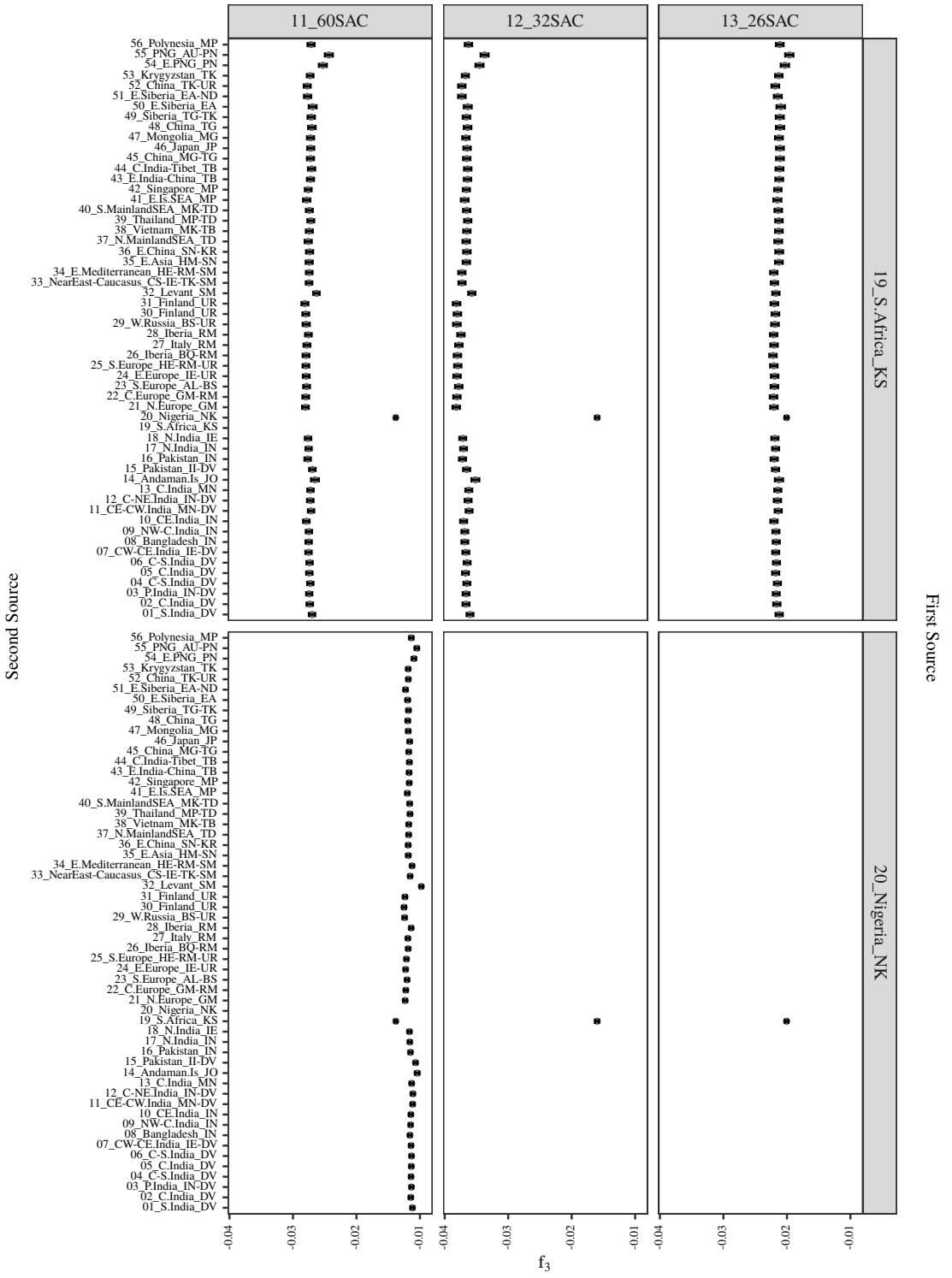


Figure 5.11: Negative f_3 estimates for $f_3(X, Y; SAC)$ for SAC fS clusters 11-13 X and Y are GR fS-inferred clusters, as indicated to the left and right of the plot. SAC clusters indicated above the panels. Black lines indicate standard error. Only values with *Z scores* < -3 are shown. Only estimates including 19_S.Africa_KS and 20_Nigeria_NK are shown as all negative value included these as a source.

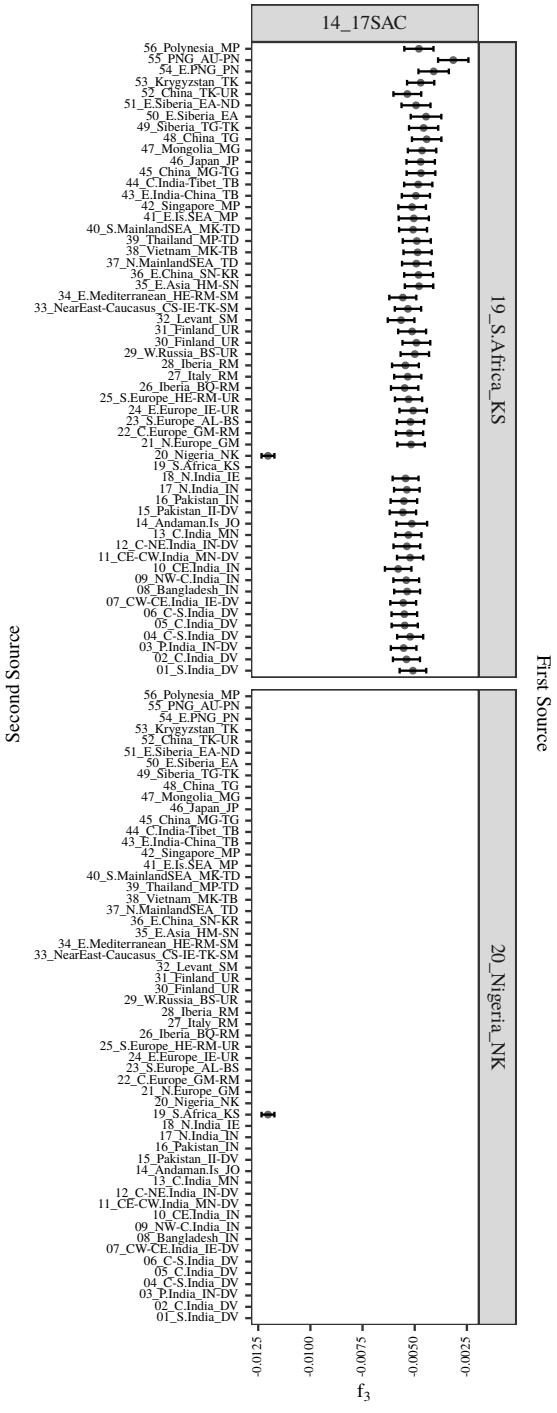


Figure 5.12: Negative f_3 estimates for $f_3(X, Y; SAC)$ for SAC fS cluster 14_17SAC X and Y are GR fS-inferred clusters, as indicated to the left and right of the plot. Black lines indicate standard error. Only values with $Z\ scores\ < -3$ are shown. Only estimates including 19_S.Africa_KS and 20_Nigeria_NK are shown as all negative value included these as a source.

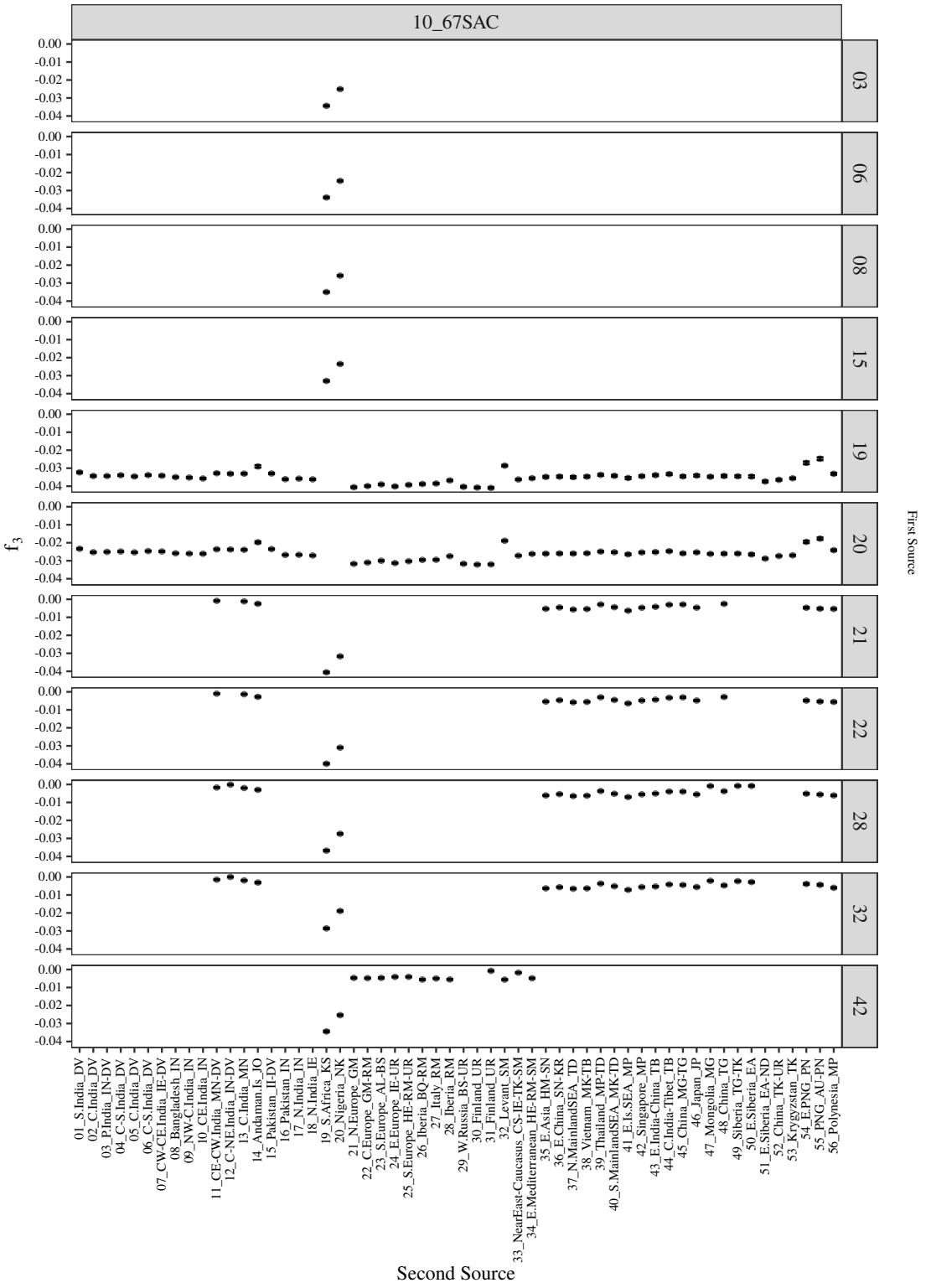


Figure 5.13: Negative f_3 estimates for $f_3(X, Y; SAC)$ for SAC fS cluster 10_67SAC. X and Y are GR fS-inferred clusters, indicated at the bottom and right of the plot. Black lines indicate standard error. Only values with $Z\ scores\ < -3$ are shown. Only estimates including one of the NNLS ancestral contributions (Chapter 4) as a source are shown.

5.4 Discussion

I found that the earliest colonial admixture history for the South African Cape "Coloured" communities started well before the Dutch settlement in 1652, while the strongest signals are from the *VOC* era. Events centred ~15 generations ago (1525 CE) were detected by bootstrapping (Figure 5.3) but as admixture date estimates necessarily post-date the actual arrival, these estimates may reflect an even earlier presence at the Cape. In contrast, I failed to detect clear evidence of a Malagasy contribution as would be indicated by an early Austronesian - West African admixture, nor a clear signal for a Southern Bantu contribution as would be indicated by a early KhoeSan - West African admixture. In both cases the evidence best supports no or limited contributions.

5.4.1 Admixture under the *VOC* Settlement

The most commonly retrieved signals post-date the arrival of the Dutch (5 - 8 generations ago; 1728 - 1815 CE) (Figure 5.3) and the establishment of the Cape colony (~11 generations ago) but is consistent with other reports [108], [177]. Unlike in many other migration events, admixture at the Cape Colony was likely initiated from the date of first settlement. The skew in the sex ratios among the earliest settlers and the overt power dynamic allowed men to take slave women to be their wives from the very start of the Cape history [16], [21]. For slaves, marriage was prohibited until 1823 thus restricting marriages to those between "freed black" and KhoeSan women with white men [21], [152].

The fS-inferred SAC clusters 02-13 produced signals for "single events" which likely reflect the midpoint of multiple admixture events. The random resampling detected two overlapping events (Figure 5.3), lending support to the arguement that the fS-cluster events are midpoints of a wider spread.

In the bootstrapping as well, pre-1652 dates are detected (Figure 5.3). I thus interpret the majority of the iterations (87%, 867/993), where a single event is detected, as the average of the spread of events including events pre-dating the *VOC* settlement.

The bootstrap single event point estimates are centred on the abolishment of slavery in 1807. As two events are infrequently recovered (13%, 126/993) it appears that the strongest signal remains for the recent admixtures, thus the date estimates centred on 1807 likely reflect the midpoint between the average date of birth of participants (~1960) and the earliest settlement (~1652) and not necessarily admixture related to the abolishment itself.

The top ranked amplitudes identified through the LD decay curves strongly support a Northern European or Finnic source in combination with the KhoeSan (Figure 5.3), supporting the observed high proportion that both regional groups have contributed to the ancestry and the relatively high F_{ST} between them (Figure 4.14) [19], [110].

This indicates that the earliest and most common admixtures involving European men may have not necessarily have been the recorded marriages with South Asian slaves. Inter-marriage with the KhoeSan was less often recorded but is known to have happened as indicated by the marriage between Krotoa and Danish Pieter van Meerhoff [16]. Children born out of wedlock would assimilate into the mother's 'racial group' and we may expect that such children were less often acknowledged and baptised [16], as such their genetic contributions may not correspond to historic accounts.

By 1807 the white settler community consisted of people of various European descent however the F_{ST}-based binning of MALDER amplitudes and f_3 estimates did not provide support for multiple European contributions in the SAC (Figure 5.13 and Supp. Figure B.3 - B.9). This contrasts with the four sources identified through the NNLS procedure in the previous chapter (Figure 4.3.3). Allele frequency-based linkage-disequilibrium decay curve methods such as MALDER are known to perform poorly in scenarios of complicated ancestral contributions [87]. The layering of allele frequency influences may have dampened many signals of admixture from sources which contribute small fractions of ancestry [78], [232] and from multiple sources which have low genetic divergence, in particular the low genetic variability within Europe [167], [196], and the relatively low SNP density employed here.

In contrast, there was good support for multiple South Asian sources based on the F_{ST}-based binning of MALDER curve amplitudes (Supp. Figure B.3 - B.9). This included admixture between linguistically/ethnically divergent groups (e.g. Dravidian and Indo-Iranian language groups) as well as between genetically similar groups (e.g. 15_Pakistan_II-DV and 16_Pakistan_IN). The amplitudes were small, suggesting minor contributions which have possibly drifted. Alternatively, some of the signals may be artefactual. For example, 15_Pakistan_II-DV and 16_Pakistan_IN could reflected detecting East African ancestry in the SAC which is shared with 15_Pakistan_II-DV [82] but which is not present in 16_Pakistan_IN. Overall however, the detection of multiple instances of signals with MALDER, in conjunction with the sources identified by CP-NNLS, does lend good support. A diversity of South Asian communities are reported to have arrived [16], [133], [149] as discussed in Chapter 4.

As found in previous work [185], there was poor support for multiple South-East Asian sources nor support for direct East Asian contributions. Only in one instance was an LD decay curve recovered by the F_{ST}-based binning for a fit between a pair of Eastern Eurasians (Supp. Figure B.6 (a)), which is in agreement with the f_3 results (Figure 5.13). Furthermore, the NNLS in Chapter 4 supported only a Malaysian component and no Eastern Asian contributions. The source for the youngest Eastern Eurasian admixture found here would most likely be the recent slaves imported [133], [149]. The origins and ethnic affinities of the arrivals are obscure but the Indian Ocean trade in general involved a great variety of South-East Asian groups, the largest part of which were sourced from South Sulawesi (42%) and Bali (24%) [35] and a number of slaving cities including Batavia, Moluccas and Macassar were sourced for the Cape [133]. Despite a possible diversity of origins there is apparent low diversity in the SAC. This may be a consequence of a bottleneck since arrival at the Cape as any particular ethnic group may not have exceeded 20 individual over a few hundred years [133], making it likely that they had drifted out (see Section 4.2.4). Alternatively, there may be structure within the SAC not well reflected in the present sample set. For example, the Cape Malay are a known cultural

sub-sect to the "Coloured" communities of Cape Town with a history linking them to the earliest Indonesian arrivals [112], [162]. Some reports account for "Chinese" among the slaves and slave traders [e.g. 152], but this may reflect the lax use of ethnic identities by casual observers. For example, Taylor [160] references a 'Malay' along the Southern Coast in the 1700s who was most probably Javanese. Indeed 18th century private slaving records do not report any enslaved Chinese or other mainland East Asian groups [133], [149] though Botha and Pritchard [152] report support for Chinese ancestry based on Diego factor from sero-genetic analysis.

5.4.2 Evidence for Pre-Settlement Admixture

A pre-settlement contribution was detected in 13% of bootstrap iterations with a prevalent signal for Northern Europe or Finnish as a contributor (Figure 5.3). This concurs with the lowest $f_3(Eurasia, Africa; SAC)$ estimates (Figure 5.8 - 5.11).

I suggested in Chapter 4 that the 28_Iberia_RM component identified may have been introduced via pre-settlement admixture, either through admixed South Asian slaves brought from Portuguese colonies or an early admixture between Portuguese and Khoikhoi circa 1488 [142]. Here, however, the SAC cluster in which a pre-Cape Colony admixture was identified (01_46SAC) was not notably enriched in Iberian ancestry (Figure 4.27) instead had lower levels of 28_Iberia_RM compared to other SAC clusters, lending little support for this hypothesis.

Collectively from this chapter, there is no clear support for a specific Portuguese contribution to the early signal.

Many of the early admixture dates identified the 41_E.Is.SEA_MP cluster or the 37_N.MainlandSEA_TD cluster as a source (13/126 curves, 10%) (Figure 5.3) but the dates post-date the Malagasy admixture [86], [217], [230] and pre-date the arrival of the Indonesian slaves at the Cape [16], [133]. The GR 42_Singapore_MP was found at high prevalence in Chapter 4 (at least 1% ancestry in 77% of individuals; mean ~5.7%; Supp. Table A.12), supporting an earlier admixture than many of the other ancestries identified which is in line with MALDER date estimates found here. South-East Asian sailors may have reached Madagascar as early as

700 CE with later arrivals associated with the 12th - 15th century kingdoms of Srivijaya and Majapahit [55], [230]. Genetic estimates of admixture from Malagasy necessarily post-date these dates; 675 BP, 1285 CE [224], [230], which are too early to be the event detected here.

As with my discussion of Iberian ancestry in Chapter 4, the most likely alternative sources for these pre-Cape Colony admixtures are either 1. The arrival of settlers of mixed descent and 2. admixture with the KhoeSan during the period of Portuguese activity.

The Spanish and Portuguese held colonies in South and South-East Asia from the late 15th century onward. When the Dutch and French rose to power in the 17th and 18th century, they took over these territories [35], [36]. Citizens and slaves of these colonial cities would have undoubtedly included people of mixed descent. Some of the earliest accounts of the settlers at the Cape in South Africa include people of mixed descent, such as Simon van der Stel, the founder of Stellenbosch, who was a "European" but had a South or South-East Asian grandmother [21].

An alternative and more direct contribution may have come from shipwreck survivors. There are at least 300 known shipwrecks along the Southern Cape pre-1900s [234]. While often passengers are presumed to have died [see 142], there are sufficient instances where survivors were found subsequently [103], [142], [160] to allude to the possibility of cryptic contributions from shipwreck survivors. For example, in 1647 a Dutch ship, *Nieuwe Haerlem*, sunk at Table bay *en route* from Batavia. The surviving 60 individuals interacted with the local communities until being rescued in 1648 [235]. Several Portuguese shipwrecks are recorded from as early as the 1500s, but these are often further East [142], [236]. The sailing crews, from as early as Vasco da Gama's trip in 1488, were an amalgamation of people from different regions including lascars (Arab/Asian mercenaries) employed by European ships as crewman and to navigate [16], [35], [160], [237]. There are thus multiple possible sources for this ancestry which at least supports the likelihood of a pre-Cape Colony signal being a genuine observation.

5.4.3 Predominant Slave-related African Contributions but No Clear Malagasy Signal

I identified an older admixture event in 14_17SAC with a KhoeSan - Yoruba curve which pre-dates the earliest accounts of any Modern period Eurasian contact (~18 generations ago) and a substantially older KhoeSan - South-East Asian admixture in one bootstrap iteration (17 - 30 generations ago) (Figure 5.3). These events may be related to two possible pre-historic admixture events of interest for the SAC.

Firstly, the detection of an ancient African - Indonesian event would support possible Malagasy descent but the infrequency of the observation in the iterations is difficult to reconcile with the reports that 24% of slaves arriving to the Cape were from Madagascar [35], [133]. The vast majority of the date estimates involving South-East Asian GR clusters indicate recent events (Figure 5.3). To date there has been no clear evidence of the genetic legacy of the Malagasy in the SAC [110], [185], making these results increasingly curious.

Records of the lives of the enslaved after arrival are nearly non-existent such that it is difficult to account for the dynamics of their experience. Botha and Pritchard [152] suggested that the Mozambicans would have had little genetic legacy within the SAC based on their representation in official records. However, there is evidence of continued importation of Mozambicans and Malagasy well after slavery was abolished in 1807 as the British captured slave ships from rivals *en route* to the Americas [16]. At this point in the region's history, there is no reason to think that several hundred individuals would leave no genetic trace. Early mortality without reproduction among slaves is a possible cause although, in contrast to the Atlantic Ocean slave trade, the Indian Ocean slave trade has been argued to have been less brutal and even 'paternalistic' [35]. Further work is still needed to resolve this mismatch in information.

The second pre-historic event of interest would be a Bantu - KhoeSan admixture related to the Southern Bantu. The date estimates identified in SAC cluster 14_17SAC for 19_S.Africa_KS - 20_Nigeria_NK admixture were substantially older than the other West African events from bootstrapping (Figure 5.3. The

f_3 values identified admixture with Eurasians and African groups for this SAC cluster, but the two African sources produced f_3 values far larger (Figure 5.8 - 5.13). Dates were reproducible at low frequency (2% of iterations, Figure 5.3) but the same sources were not recovered, indicating that this may be a genuine pre-Colony admixture signal somewhat unique to this cluster. The dates of 14_17SAC correspond to the expected arrival of the Southern Bantu to South Africa ~1300 CE [64], [102], [177], [238], which by its recency and nature was likely a continuous admixture with further movement into South Africa [64].

The absence of a similar signal in the remaining SAC clusters and bootstrapping, despite the predominance of KhoeSan and West African components in the SAC, indicates that their Bantu-related ancestry has another origin which did not have substantial KhoeSan ancestry. The dates support the proposal from Chapter 4 that 14_17SAC may be descended from recent South African Bantu migrants to the region and would match well the early accounts that imported Africa slaves were predominantly from regions further North where KhoeSan admixture is lower [21], [64], [125]. This included the Mozambican, West African and Malagasy slaves [16], [17], [21] and in disagreement with the suggestion of Botha and Pritchard [152] that the Mozambicans would have had little contribution.

5.4.4 Conclusion

I have added some resolution to the diverse set of ancestries that characterise the South African "Coloured" community and demonstrated that admixture was continuous and at its earliest pre-dated the formal establishment of the European Cape settlement. I found the KhoeSan, Bantu-related, South-East Asian, South Asian and Northern European contributions identifiable by LD-decay curve fitting and f_3 tests. The South-East Asian ancestry is of recent Indonesian origin which I suggest better reflects a contribution from pre-Cape Colony shipwrecks than either a Malagasy contribution or soley a recent slave contribution. The proposed Iberian ancestry does not appear to be associated with a pre-Cape Colony admixture and remains rather unplaced.

6

Geography and Ethno-racial Affinities Influence "Coloured" Genomics

6.1 Introduction

The development of the Cape "Coloured" community in Southern Africa is tied to the arrival of the European settlers and their slaves and servants in the 17th century [16]. Today over 4.5 million people identify with the "Coloured" ethno-racial affinity which is often associated with genetic admixture and cultural creolisation.

The term "Coloured", however, only saw increased usage from the 1840s after the emancipation of the slaves and the beginnings of the British administered population census [15], [109], [111]. The debate regarding the use of the word "Coloured" and its appropriateness and acceptance within the community is ongoing, this is despite the common assumption that it is widely accepted among the so-called "Coloured" people. Some sociologists see it as wholly imposed by governments prior to and during Apartheid, to implement and maintain colonial-style race-based indirect rule [15], [109]. The term is still used in forensic, medical and administrative work and research, again reflecting the assumption that "Coloured" people share either or both socio-economic environments and intrinsic biological characteristics [e.g. 110], [187], [188], [239]–[243]. This is in part due to its continued use in population censuses

because of the reluctance of the present (and past) governments to accommodate other ethnic identities, e.g. Nama and Griqua [111], [244].

Implicitly the notion of a shared 'mixedness' among "Coloured" people assumes some relative homogeneity in ancestry and culture across the identity, however blended that culture and ancestry may be. Considering the extent of the geographic and temporal presence of the "Coloured" people and their predecessors in the region's history, a homogeneous population seems unlikely [16], [21], [142]. Several factors make it clear that cultural and geographic variation should be expected.

Diverse ancestors

The founding groups for the "Coloured" community were a global representation of human diversity [16], [19], [99], [108] which included a large number of Europeans, South East Asians, West and East Africans, and South Asians [16], [112], [133], [140]. The expanding colony also incorporated whatever Khoikhoi and San communities existed across the regions North and East of the Colony. The earliest settlers included the Dutch, Germans and French as recorded in the history of the Afrikaners [21], [61]. A substantial number of slaves were imported by the *VOC* but many more arrived illegally through private trading networks [133], [149]. Subsequently, the British occupied the Cape Colony from 1807 and several thousand British settlers arrived and facilitated two waves of immigration from South Asia affecting mostly the East coast and the inland Transvaal areas [103], [112]. The continuous arrival and dispersal of different groups is expected to have affected the patterns in ancestry composition of the South African "Coloured" both temporally and geographically.

Socio-cultural stratification

Socio-economic, religious and linguistic differences may have created non-random mating among the early settlers and their descendants. Class structure existed among the earliest slaves as a consequence of the means by which they were acquired, their skill set, and the racial stereotypes within the Cape Colony [16], [35], [112]. As an example, African slaves were put to work in the fields or provided with menial tasks, as in the case of Mozambican women who were hired as washerwomen [16].

The majority of these slaves were owned by the VOC, and thus housed together on VOC properties [149]. We may expect that African slaves would most likely have paired up with each other more often than with other slaves prior to emancipation.

In contrast, the Indonesian and Bengali were known for artisanal skills, including fishing and needle work [16], [112]. There is historic evidence that these arrivals from Indonesia are thought to have founded the Cape Malay population [112], [161]. At least in sociological descriptions, the Cape Malay have a distinct history from the other sections of the "Coloured" community in their education, Islamic religion and being favoured for marriage by European settlers [16], [112], [161]. Altogether suggesting a possible case of assortative mating and endogamy.

Social mobility in the early Cape Colony was not racially restricted by law as it became in later centuries but was highly ascriptive in general and strongly determined by one's ethnic origin. I should therefore anticipate that pairing between early settlers was strongly influenced by some aspect of ethnic identity and culture.

Geographic expansion and spatial genetic variation

Different circumstances have influenced migration from the Cape Colony. The establishment, development and history of the communities beyond the Cape Colony would have been variable and more influenced by local events rather than events somehow centralised and specific to 'Colouredness' (e.g. Apartheid legislature in more recent history).

European settlements began expanding eastward as land ownership was allowed from 1657 onward. Many of the Colony's city areas had larger numbers of slaves compared to KhoeSan, as much as a 7:1 ratio, as the Colony initially excluded the KhoeSan [16], [21], [121]. While reluctant to work for the early settlers at first, the KhoeSan were eventually subjugated as the growth of the Colony and the loss of their pastoral land and water sources necessitated integration [21], [141], [148]. They were assimilated into servile and serf positions including household workers and farm staff tied to their ancestral land. Most of the Khoikhoi and San did not live in the cities (e.g. Stellenbosch and Cape Town) but in the surrounding rural farm areas where slaves were less readily available, here they were employed as

shepherds and cowherds [21], [121]. Thus, there could be geographic structure to slaves and KhoeSan distributions during the first few centuries. A form of expansive stock farming also picked up in the Cape Colony, resulting in *trekboeren* (moving farmers) expanding the Colony frontier further inland and eastward [141]. These semi-nomadic farmers would have lived in close quarters with their servants and slaves and were often of mixed descent themselves [21], [103], [141]. In this way, a signal of Cape Colony introgression may be expected beyond the Colony frontier.

Following the emancipation of the slaves in 1838, many slaves would have moved to find a new life elsewhere [103], [142]. There are several mass migration events reported soon after emancipation in which admixed communities fled persecution and encroachment from European settlers [21], [142]. The Griqua (or Griekwa) are possibly the best know example of this [142], [156]. The Griekwa migration over several decades resulted in the establishment of towns in the interior of South Africa. The Baster of Namibia are another refugee community which fled. They departed from the area around the Orange River and established several towns in Namibia in the late 1800s which are still occupied today [99], [158]. More recently "Coloured" people would have migrated in search of labour and as a consequence of urbanisation [103].

Does Identity Reflect Genetic History?

The greatest detail on the development of the "Coloured" community is centred on the former Cape Colony leaving the history beyond this area poorly recorded and patchy. Undoubtedly the founding, identity and progress across Southern Africa has differed and would contribute a layer of genetic variability over and above what has been discussed for the Cape (see Chapter 4 and 5).

I here characterise the variation in genetic ancestry of people who self-identify as "Coloured", Griewka, Cape Malay or Baster from across South Africa and Namibia with particular emphasis on possible substructure related to cultural heritage and geography. I use new samples from self-identified Griekwa and Cape Malay, as

well as samples from self-identified "Coloured" individuals from the eastern half of South Africa where "Coloured" communities are minorities [105].

I ask, to what extent does the genetics of the Cape Town "Coloured" communities [110], [186] reflect the genetics beyond the former Cape Colony? Do communities further out show the same admixing sources and dates? Does the genetic ancestry in the Cape Malay, Griekwa and Namibian Baster reflect the historical descriptions of cultural distinction and endogamy, or do communities share similar genetic patterns to their geographic neighbours? How do patterns across the region relate to the developments in the expansion of the "Coloured" communities?

6.2 Methods and Data Analysis

All supplementary material for this chapter is available in Appendix C.

6.2.1 Sample Collection

Ethics Approval

Ethics approval was obtained from Oxford Tropical Research Ethics Committee (Oxford University, UK) (ref. No. 8-16) and the University of the Free State (South Africa) NatAgri Ethics Committee (ref. No. UFS-HSD2016/1210). Export permits were approved by the South African Department of Health.

Interviews

The selection of towns for sampling was informed by literature review [16], [21], [104], [142] but ultimately determined by the availability and willingness of collaborators within each town. All interviews were performed during March - April 2017 in South Africa. An outline of the stages from interview to DNA extraction is supplied in Figure 6.1.

Communities were contacted via a local community leaders (e.g. church pastors, principal of local schools or chief of the village) up to 15 months ahead of arrival. The local leader informed participants ahead of time of the purpose of the study when possible. Interviews were conducted with willing participants irrespective of

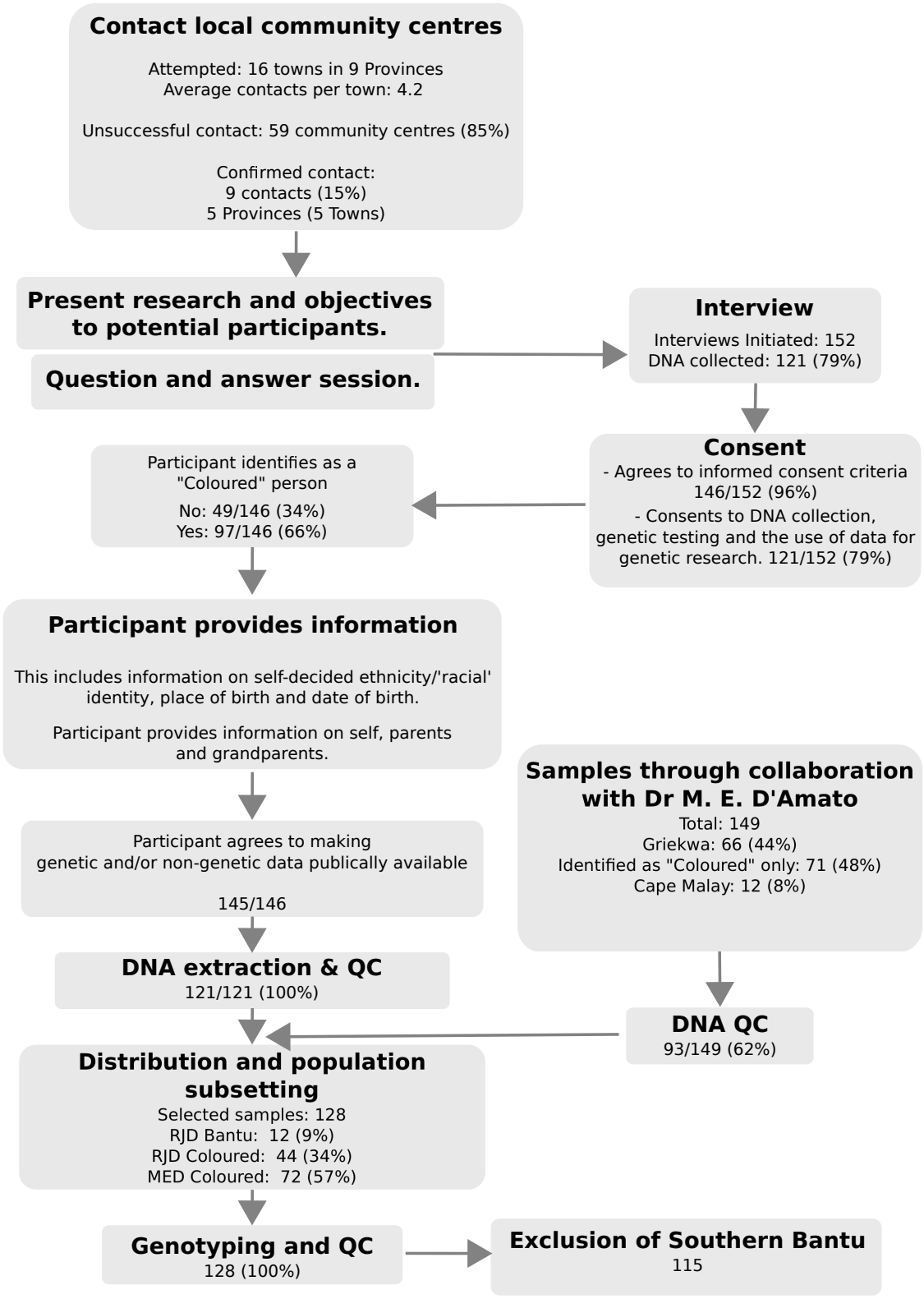


Figure 6.1: Outline of the interview process from initial contact to genotyping. Changes in available samples indicated at each step as a proportion(%). Abbr. MED - Dr Maria E. D'Amato, RJD - Ryan J. Daniels, QC - Quality Control

their identity or descent (151 interviews) following either a presentation on the work or a one-on-one discussion about the work. Interviews were conducted in English or Afrikaans by RJD as these are the two main languages for the "Coloured" community in South Africa. Participants could respond in either language though many chose English and a mix of Afrikaans and English. The questionnaire consisted of open and closed questions but was a guided discussion but not entirely semi-structured. This style of interviewing was preferred due to time constraints and the need for large sample sizes for genetic work on heterogeneous populations. All persons who attended were informed of the limits of what could be said on ancestry as well as the fact that there would be no personalised feedback. Written informed consent was obtained from each participant before discussion began.

Sampling was supplemented through collaboration with Dr Maria Eugenia D'Amato from the University of the Western Cape, South Africa (UWC). This included 149 samples from "Coloured", Cape Malay and Griqua communities. These samples were collected by a similar process by Dr D'Amato's team.

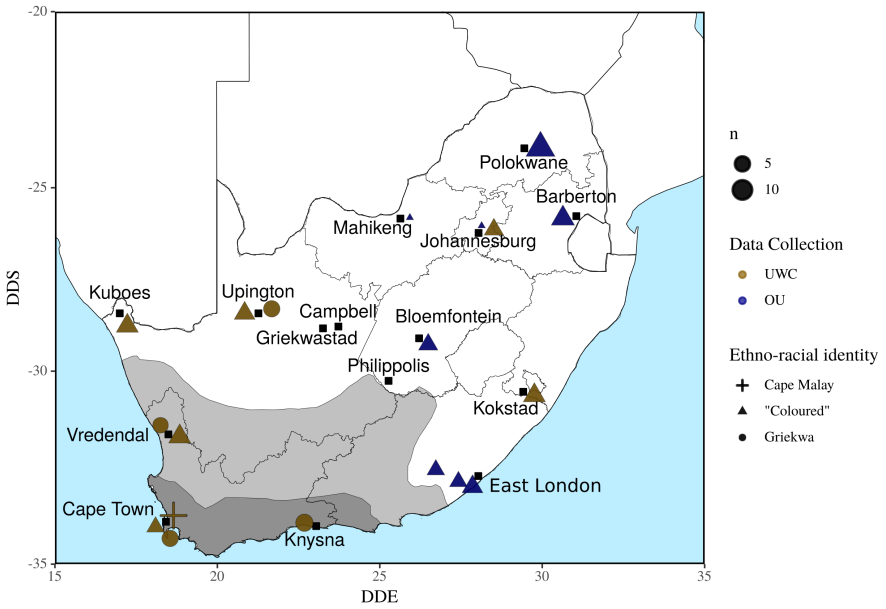
Genotyping

To subset the available samples for genotyping, priority was placed on sample sizes per ethnic/population group followed by the region of sampling and finally the place of birth. This was repeated for information on parents and then grandparents of the participant, resulting in a selection of 115 samples (Figure 6.2 and Table 6.1).

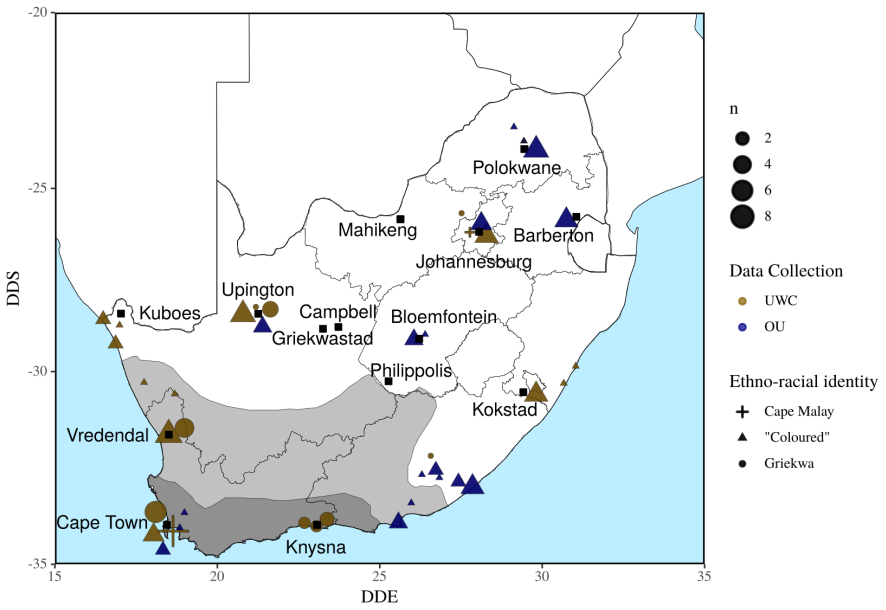
Table 6.1: Summary of the genotyped samples by ethno-racial affinity.

Identifies as	Indicated sub/other identity	Samples genotyped
Cape Malay	Cape Malay	8
Coloured	Baster	4
	Coloured	70
	English	1
	KhoeSan	3
	Mixed ancestry	1
Griekwa	Coloured	1
	Griekwa	27

Samples were clustered geographically based on the place of sampling/residence (POS) and the place of birth (POB) of the participants. Information on place of



(a) Participants by place of sampling/residence



(b) Participants by place of birth

Figure 6.2: Distribution of the places of origin of the new samples collected. Shown are the (a) place of residence/sampling and (b) place of birth. For (b), samples were clustered based on place of birth, with samples within 80 Km of each other pooled to represent a region for their respective subgroups. Sample sizes indicated by plot symbol size. Colours indicate sample collection; Oxford University, (OU) or provided by collaborators at the University of the Western Cape, South Africa (UWC). Cape Colony administrative zones indicated by grey polygons. Dark grey indicates boundary of the Cape Colony administered districts by 1795 and lighter grey shows the extent of the Cape Colony by 1824 [after 156]. Abbr. DDS/E - Decimal Degrees South/East.

birth of the parents and grandparents was unfortunately frequently unavailable, for example missing parent information in 30% of "Coloured" (UWC and OU data). Too few samples clustered by place of birth to form reliable sample sizes. I therefore used place of residence as an indication of the contemporary distribution of the "Coloured" and other communities.

Samples were collected using the Oragene-500 saliva collection pots (DNA Genotek, Canada) for the Oxford University (OU) collection and for the UWC samples, an in-house custom saliva preservation system was used. Approximately 2ml of saliva was collected in field from each person following the manufacturer's instructions. For the OU samples, I followed the recommended procedure for DNA extraction using the prepIT.L2P salt extraction and ethanol precipitation kits (catalog# PT-L2P, DNA genotek, Ottawa Canada). For the UWC samples, an in-house salting out process was used. I quantified DNA as well as impurities using both absorbance and fluorescence methods. Absorbance was measured using a nanodrop spectrophotometer (ThermoFisher Scientific, USA). Samples identified as low quality or DNA quantity were run through the Qiagen PCR purification kits (Qiagen, Germany) to remove further contaminants. I quantified final DNA available using Picogreen fluorescence and standardised samples to ~35 - 100 ng/ul for genotyping. Samples were genotyped on the Illumina Omni2.5-8 Beadchips v1.3 at the Wellcome Trust Centre for Human Genomics, Oxford University. The quality control was performed using GenomeStudio software (Illumina, USA). All samples passed a call rate of 97%. I confirmed that there were no mismatch in labels and data by comparing sex-imputation from PLINK (male F<0.2 and female F>0.8) to that recorded in the interviews (option: `-check-sex`). No samples were discordant.

Below I provide a brief synopsis of the main characteristics of the samples genotyped. The distribution of date of birth (DOB) between sexes within each ethno-racial identity was similar (median DOB "Coloured" ~1964 and Griekwa ~1973) (Figure 6.3), except for the Cape Malay where the Females (n=2) were notably younger (median DOB 1995 vs 1967).

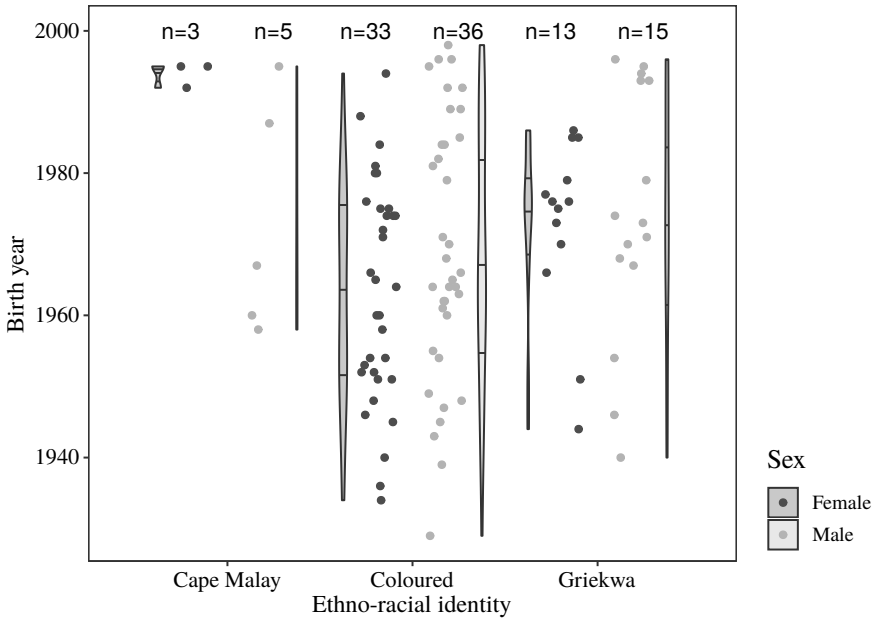


Figure 6.3: Distribution of year of birth by ethno-racial affinity and sex. Sample sizes indicated above the violin plots. Indicated are the quartile values (black horizontal lines) and the scatter of data points is provided alongside (jittered to aid visualisation).

6.2.2 Datasets, Merging and Quality Control

SNP chip Data

Data were merged as described in Chapter 4. In this chapter only data on the Illumina SNP arrays were included to preserve SNP density.

Variant-only Genome Sequence Data

Whole genome sequence data (WGS) from the Simon's Genome Diversity Project [13] (SGDP), the 1000 Genome Project (1KGP) Phase 3 [245] and a set of previously published KhoeSan genomes [114] were merged as described in the Chapter 4. Congruence between WGS and Illumina SNP array data were confirmed using samples genotyped on both platforms (SGDP and 1KGP samples only) and discordant variants were removed (12 453 from 1KGP and 7 849 from SGDP). A PCA was performed to confirm the absence of a SNP array batch effect (Supp. Figures C.1 ,C.2, C.3).

Data Curation

Further to the data curation described in Chapter 4, SNPs out of Hardy-Weinberg equilibrium ($p < 1 \times 10^{-5}$) were removed. This was performed on *a priori* populations with >8 individuals.

With the exception of the focal SAC admixed sample sets, all *a priori* populations were restricted to 25 individuals. Where multiple datasets included the same population (e.g. Yoruba from HGDP and YRI from 1KGP), a random subsample from each available dataset was taken, typically of 10 - 15 individuals.

Linkage Disequilibrium Pruning

When performing any of the Principal Component Analyses (PCA) and ADMIXTURE analyses, I trimmed the datasets for linkage disequilibrium removing SNPs with R^2>0.2 for a 50bp frame with a 5bp sliding window [after 19] with the command `-indep-pairwise 50 5 0.2` using PLINK v1.9 [191].

Kinship Assessment

Kinship was assessed within and between *a priori* populations as previously described. This resulted in 338 individuals from 60 populations being removed.

Population Exclusion from GR Dataset

I excluded population deemed unnecessary for this work, largely groups from the New World and populations of known recent admixture. Secondly, I identified populations with large inter-individual distances based on Principal Component Analysis as described in Chapter 4, estimating distances as a Euclidean distance on the first four eigenvectors resulting in the removal of four populations (Supp. Table C.1). Lastly, I sub-setted the GR populations to conform to the population size restrictions in Eigensoft (100 populations). I retained a global representation of populations.

Outlier Detection in GR dataset

Outliers were detected by three processes. Individuals were identified as outlying based on identity-by-state (IBS) values as employed as by [167] and in Chapter 4.

I further identified outliers iteratively using `smartpca` from the software package Eigensoft [165]. Due to the limit on the number of populations that Eigensoft will accept, the process was performed on each population. Specifically, a population j was merged with three global populations (Han Chinese (CHS), British (GBR) and Yoruba from Nigeria (YRI)) and the four groups defined the PC space. Only outliers detected in population j were removed. I identified 24 individuals from 8 populations as outliers (Supp. Table C.1).

The final outlier detection was performed on the Southern African KhoeSan groups to identify individuals with potential recent admixture. As the KhoeSan and the Southern African "Coloured" populations have a tied history, many KhoeSan populations may have recent admixture as well [see 28], [177]. Using `smartpca` I defined a PC space based on four GR datasets (Oromo from Ethiopia (OROMO_ag14), British (GBR), Mende from Sierra Leone (MSL) and the Ju|hoan (Ju_hoan)). The KhoeSan groups were then projected onto the PC space and a Euclidean distance from the Ju|hoan was estimated for each individual KhoeSan. Individuals who were Tukey outliers with regard to excess distance were excluded. This resulted in the exclusion of three individuals from three sample sets (Supp. Table C.1). In several Southern African KhoeSan groups there was notable variation in ancestry comparable to the "Coloured" groups investigated in this study. I did not remove outliers from these groups. Instead, the data were split into two sets based on proximity to the Ju|hoan. In each sample sets, the closest 25% were removed and the remaining 75% were considered in the focal study set along with the "Coloured", Griekwa, Baster and Cape Malay, which I have referred to collectively as the SAC. By removing the 25% of individuals closest to the Ju|hoan I focus on the subset of the data which most likely reflects recent admixture and is most comparable to the "Coloured" sample sets. This prevents the population averages for the Southern African admixed KhoeSan from being skewed toward high values of KhoeSan ancestry because of the individuals closest to the Ju|hoan who may be unadmixed. I acknowledge that the SAC KhoeSan sample sets discussed in this thesis are thus subsets of KhoeSan data and do not reflect the population history

of the entire dataset. Despite the artificial 'enrichment' of non-KhoeSan ancestry, I found the SAC KhoeSan groups were still distinguished from the "Coloured" by their elevated KhoeSan ancestry overall (see Section 6.3).

I use the term 'sample sets' to refer to the SAC data, as the sample sizes are unlikely to be good representations of the populations. The word 'population' is used for the GR data as these are larger numbers and likely better representations of the populations. The abbreviation SAC is used here for the blanket term 'Southern African "Coloured" and KhoeSan' to discuss the set of self-identified "Coloured", Griekwa, Cape Malay and Baster communities as well as six KhoeSan sample sets which showed mixed descent (SAC KhoeSan), while the word "Coloured" is used here explicitly for individuals who self-identify as "Coloured" and for sample sets composed of predominantly self-identified "Coloured" individuals (i.e. not Griekwa, Cape Malay, Baster or any of the KhoeSan groups).

The final dataset retained 191K SNPs from the 22 autosomal chromosomes, and including 70 *a priori* global reference (GR) populations alongside 29 Southern African "Coloured" and KhoeSan sample sets, for a total 1480 GR individuals (Table 6.3). A map of the distribution of the SAC samples is available in Figure 6.4.

6.2.3 Ancestral Variation in the SAC

I investigated patterns in global reference (GR) population structure to discuss the relationship of the South African "Coloured" (SAC) groups to the GR data.

A principal component analysis was conducted as implemented in Eigensoft 7.2.1 [165]. I estimated a Euclidean distance between individuals based on the first four eigenvectors to quantify inter-individual distances and variation across populations. This provides a measure of variability within populations. Calculations were performed using the `dist` function in R [199].

The relationship among the SAC *a priori* sample sets and among the individuals was visualised using a hierarchical clustering of the PCA distances. Clustering using the `hclust` function in R [199] was performed on individual samples. I use the

Table 6.2: Summary of the grouping of the *a priori* sample sets within the SAC. Indicated are the ethno-racial affinities used to group the sample sets including the "Broad" affinity and the more specific self-identified affinity. The number of samples included after curation indicated (n). All samples are from South Africa unless otherwise indicated. Sampling localities and abbreviations shown in Figure 6.4. Abbreviations not presented in the figure: EC - Eastern Cape, NC - Northern Cape, KHM - ≠Khomani, SA - South Africa, he11 - Henn *et al.* 2011

Broad affinity	specific affinity	Sampling label/locality	Source	n
Blanket term "Coloured"	Basters	Basters (Namibia)	[99]	30
	Cape Malay	CAPEMALAY_CPT	This Study	7
	"Coloured"	COLOURED_BFN	This Study	5
		COLOURED_BRT	This Study	8
		COLOURED_CPT	This Study	4
		COLOURED_EL	This Study	6
		COLOURED_JHB	This Study	6
		COLOURED_KB	This Study	7
		COLOURED_KS	This Study	7
		COLOURED_KWT	This Study	4
		COLOURED_MHK	This Study	1
		COLOURED_PTG	This Study	14
		COLOURED_TBV	This Study	4
		COLOURED_UTN	This Study	6
		COLOURED_VRE	This Study	7
		Coloured-D6	[99]	8
		Coloured-EC	[99]	7
		Coloured-NC	[99]	10
		ColouredColesberg	[115]	19
		ColouredWellington	[115]	19
	Griekwa	GRIEKWA_CPT	This Study	6
		GRIEKWA_KNY	This Study	6
		GRIEKWA_UTN	This Study	6
		GRIEKWA_VRE	This Study	6
KhoeSan	Karretjie	Karretjie	[115]	15
	Nama	Nama (Namibia)	[115]	14
		NAMA_SA	[107]	7
	≠Khomani	KHM_SA	[107]	42
		Khomani	[115]	21
		SAN_he11	[113]	16

clustering to test if genetic similarity is based on shared ethno-racial identities. To compare *a priori* groups and geographic distributions to individual-based clustering results, I measure the sample overlap between the derived clusters and the *a priori* sample sets (i.e. racio-ethnic groups). Secondly, I consider geographic distributions in relation to the Cape Colony administrative region at three time periods associated with changes in the adminstrative practises concerning ethnicity: c.1795, before the onset of British administration, c.1824, before the emancipation of the slaves in

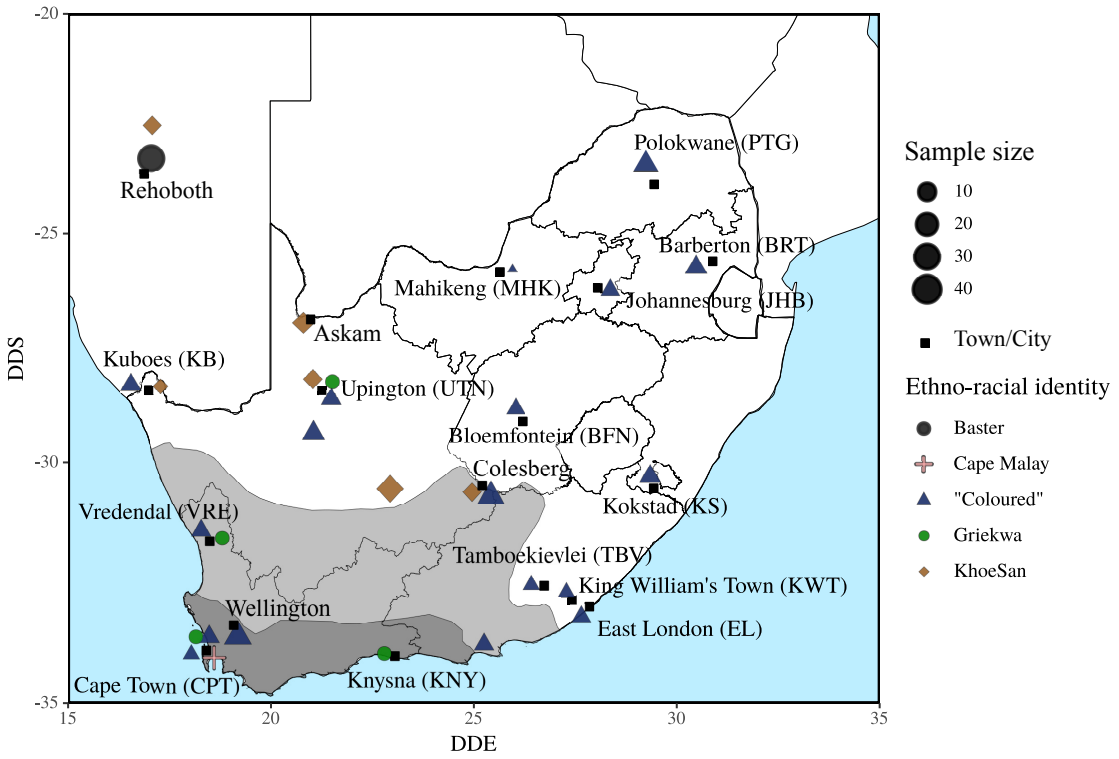


Figure 6.4: Geographic distribution of SAC sample sets included in the analyses. Sample sets as indicated in Table 6.2. Localities of interest indicated and relevant abbreviations shown in brackets. Abbr. DDS/E - Decimal Degrees South/East.

Table 6.3: Changes in the dataset through quality control.

Stage	No. Individuals	No. Populations	No. SNPs
Merged data	8 796	497	192 208
PCA, F_{ST}, ADMIXTURE	1 480 (1 102 GR + 378 SAC)	70 + 29	191 971
MALDER		24 + 29	191 971
LD trimmed $R^2 = 0.2$	-	-	100 561

1838, and c.1900, ahead of the establishment of the Union of South Africa.

To assess similarity among *a priori* clusters I use the mean PCA positions to perform clustering of sample sets following the same procedure above. Statistical significance for the differences of the mean positions of each population in PC space was assessed using the Tracy-Wisdom statistic and an ANOVA as implemented in Eigensoft 7.2.1 [165]. Considering that hierarchical clustering provides a tree-like evolutionary account of the relationship, it cannot take into account the reticulate

nature of relationships among admixed groups [246]. I therefore examine the clustering using a NeighbourNet analysis [247] as implemented in the R package `Phangorn` [248].

I further evaluated population structure using the Bayesian clustering algorithm in ADMIXTURE v1.3.0 [172]. I performed 10 replicates of K between $2 \ldots 20$ with a random seed, 5-fold CV estimation with 100 bootstraps for estimating standard errors (options `-s time -B100 -cv INPUTFILE.bed 2..20`). Results were processed as discussed in Chapter 4. The F-indices of pairwise distances between populations was estimated as implemented in Eigensoft.

I investigated the trend in ancestry with increasing distance from Cape Town using a linear correlation of ADMIXTURE components. I focus on the six components which make up the largest ancestral contributions. The distance from Cape Town was estimated for each individual using the R function `distVincentyEllipsoid` from package `geosphere` v.1.5. [249]. Correlations were implemented in R using the linear regression in the `ggplot2` package and were performed separately for KhoeSan, Griekwa and "Coloured" SAC.

6.2.4 Admixture Dates

I date the admixture events within the SAC dataset by examining the exponential decay of linkage disequilibrium (LD) with the increase in distance between SNP pairs as done in Chapter 5. I perform this using the software MALDER (Multiple Admixture Linkage Disequilibrium for Evolutionary Relationships) [78]. I here use 29 years as the inter-generation time, in line with other researchers [see 108], [177], [223], [225]. Dates are calculated as $1960 - 29\ years.gen.^{-1} \times no.\ of\ generations$, where 1960 is the median date of birth of the sampled SAC.

Bootstrap Evaluation of Signals

As shown in Chapter 5, bootstrapping over the same dataset may produce varying results when the population admixture is multi-faceted and recent. For this reason, I performed a bootstrapping process over the samples within each dataset.

This was done using a subset of the GR data as potential sources (Supp. Table C.2). The minimum distance between SNPs used for curve fitting was set to 1.9 cM as in Chapter 5. I performed 30 bootstrap resamplings from each of the SAC sample sets without replacement. Where $n > 6$, 30 unique combinations are possible, where $n < 6$ the entire sample set was repeatedly utilised.

I investigated the trend in admixture timing with increasing distance from Cape Town using a linear correlation as discussed for ADMIXTURE proportions. Correlations were implemented in R separately for KhoeSan, Griekwa and "Coloured" SAC.

To investigate possible contributions from the Malagasy slaves and the local Southern Bantu groups, I included the Vezo from Madagascar (VezoM) and amaXhosa as recipient populations in the MALDER analysis. The rational here is that the signal for an ancient admixture in the SAC which corresponds to the signal seen in these populations may provide evidence of a genetic contribution. Including either the Vezo or amaXhosa as a recipient population would be uninformative. In a multi-way admixed population such as the SAC multiple populations-pairs can be successfully identified as sources with LD decay curve fitting even if neither contributed directly, as I have shown in Chapter 5. Thus I would likely detect a admixture curve with the amaXhosa and Vezo when fit with several of the GR populations.

6.3 Results

6.3.1 High Ancestry Variability in the KhoeSan and "Coloured" Ethno-racial Affinities

I use a Principal Component Analysis to examine the heterogeneity within the SAC sample sets. Firstly, I examine the genetic structure of the GR samples to validate the merging and quality control processes and to evaluate the relationship between the focal SAC sample sets (including the SAC KhoeSan) and the GR data. The Principal Component Analysis was conducted on the full dataset (GR+SAC) (Figures 6.5 and 6.6). I found the results were in line with those previously

published. The first four principal components (PCs) account for more than 91% of the total genetic variation.

Table 6.4: Summary of the abbreviations used for linguistic groups. Note that groups are not at the same hierarchical level. Note that abbreviations do not necessarily correspond to Table A.9 in Chapter 4

Abbr.	Language Grouping	Abbr.	Language Grouping
AA	Austro-Asiatic	KH	Central/Khoe
BG	Basap-Greater Barito	KK	Khoekhoe
BI	Biaka	LD	Land Dayak
BQ	Basque	ME	Mande
BS	Balto-Slavic	MG	Mongolic
CU	Cushitic	MH	Hmong-Mien
DS1	South Dravidian 1	MP	Malayo-Polynesian
DS2	South Dravidian 2	NB	North Borneo Malayo-Polynesian
ED	Edoid	NE	Nilotic
FN	Finnic	NS	Southern Nilotic
GB	Gbe	PN	Papuan
GM	Germanic	RM	Romance
HE	Hellenic	SB	Southern Bantoid
HZ	Hadza	SD	Sandawe
IG	Igboid	SE	Semitic
II	Indo-Iranian	SN	Sinitic
IN	Indic	TC	Common Turkic
IR	Iranian	TU	Tungusic
JP	Japonic	UT	Southern/Tuu/Taa-!Ui
JU	Northern/!Kung	VM	Viet-Muong
KD	Tai-Kadai	YB	Yoruboid

Classification is based on information from [250]–[252].

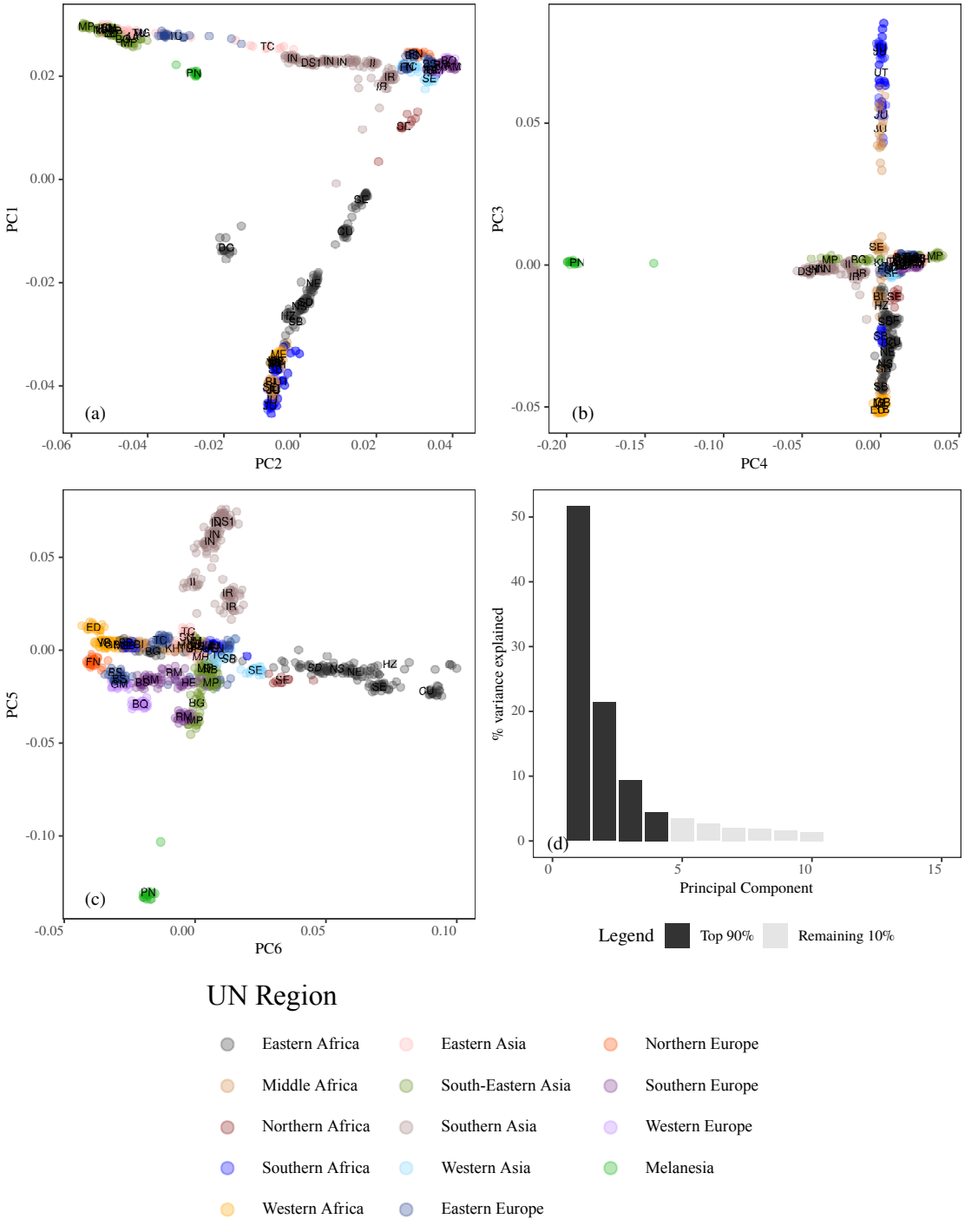


Figure 6.5: Global population structure as evaluated by Principal Component (PC) Analysis. Only the GR data are shown. (a-c) Pairwise comparison of PCs and (d) variance explained by each PC shown in sub-plots. Colours correspond to UN global regions and abbreviations correspond to linguistic groups (see Table 6.4)

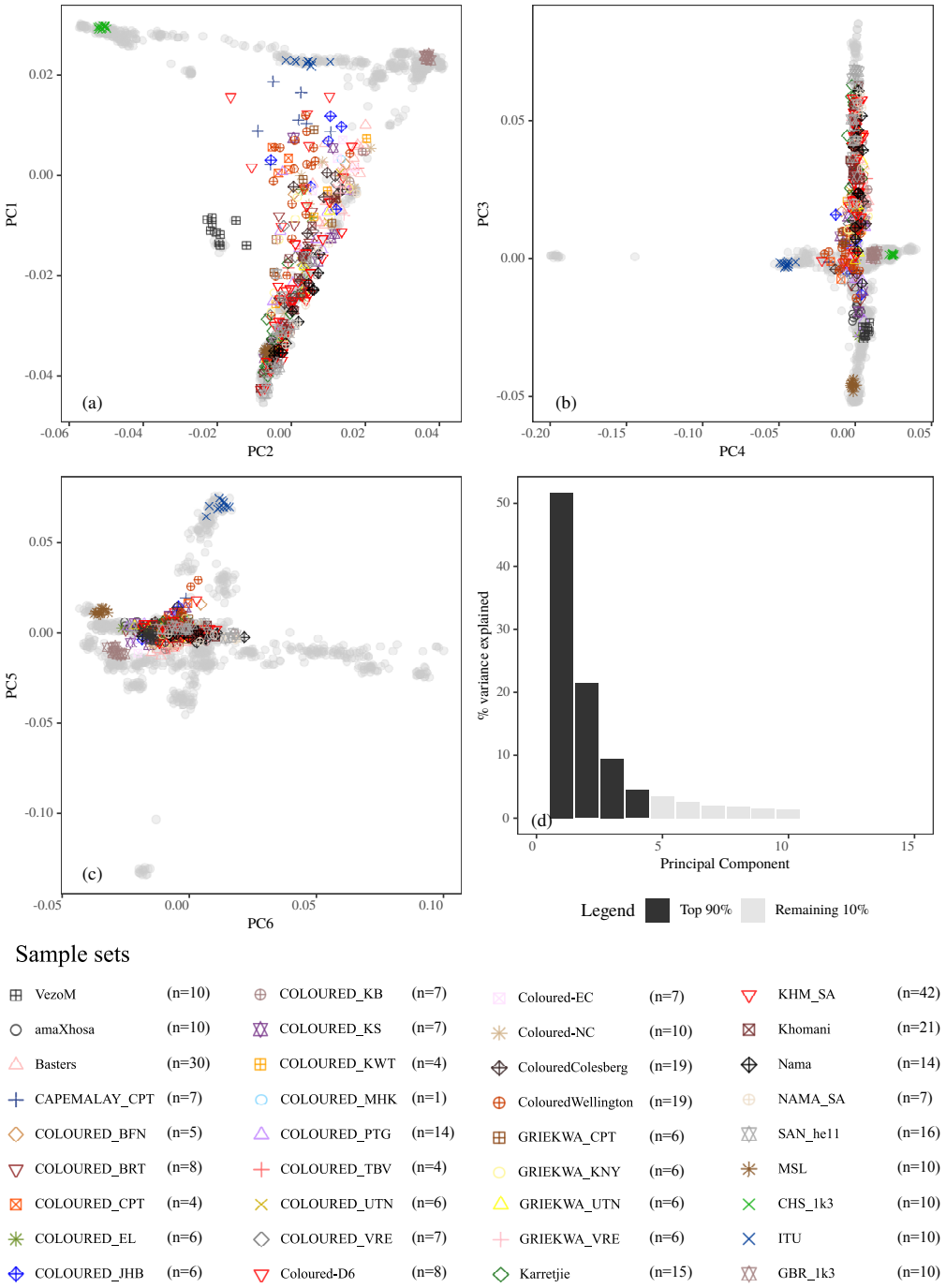


Figure 6.6: Global scale Principal Component (PC) Analysis of the South African "Coloured" (SAC) individuals showing the first 6 PCs. (a-c) Pairwise comparison of PCs with the SAC and representative GR data shown in colour. The remaining GR populations are shown in grey. (d) Variance explained by each PC. All samples (GR + SAC) included in establishing eigenvectors as in Figure 6.5. Abbreviation for reference populations; MSL - Mende from Sierra Leone, ITU - Indian Telugu from the UK, GBR_1K3 - British from the UK, CHS_1K3 - Han Chinese from China. Plots separated by sample set available as Supp. Figures C.4 - C.33

The first eigenvector summarised the divergence between Sub-Saharan Africans and Eurasians, and the second differentiated Eastern and Western Eurasians as reported in previous studies [4], [165], [245]. The third eigenvector described variation between non-KhoeSan Africans from KhoeSan. The fourth eigenvector described variation along the South-East Asian - Melanesian cline, mostly drawing out the Papuan samples. The fifth PC produced a cline roughly along South-East Asia - Southern Asia but also differentiating African and Eurasian groups. The sixth PC produced a gradient which drew out variation which differentiated some East African groups from European and West African groups. The Esan, Yoruba and Finnish shared a common side and Hadza, Oromo and Somali were on the other end.

The South African "Coloured" groups formed a cloud stretching between global populations, in particular between Eurasians and Africans along PC1 and between KhoeSan and Non-KhoeSan Africans along PC3, mirroring the results for the Cape Town SAC in Chapter 4 (Figure 6.6, plots by sample sets available in Supp. Figures C.4 - C.33). On PC2, there is notably little pull toward the Malagasy (VezoM individuals) for any of the SAC groups indicating that no SAC individuals have predominant Malagasy ancestry.

Along PC3, the SAC groups clearly overlap with several non-KhoeSan African and KhoeSan groups. Despite high variation in ancestry among the SAC KhoeSan groups, there was a distinct shift along PC3 toward the Ju|hoansi compared to most of the "Coloured", Cape Malay and Griekwa groups, indicating a strong influence from the KhoeSan component to their ancestry. Among the non-KhoeSan SAC, some were notably further from the Ju|hoan (e.g. COLOURED_CPT) compared to other sample sets (e.g. Griekwa and "Coloured" from Vredendal (VRE) and Upington (UTN)). On PC6, distinguishing Eastern African-related ancestry, some of the Nama and San (SAN_he11) show a slight pull toward the Semitic/Cushitic side of the axis suggestive of Eastern African ancestry. Overall, the SAC sample sets had extensive overlap, indicative of the variation present within each group.

The affinity of the SAC groups with South Asian groups is seen on PC4 and 5 by the skew toward the Telugu samples from the United Kingdom (ITU, [2]). The

Cape Malay in particular clustered closer to the South Asian groups compared to most of the other SAC groups but have discernible overlap with the "Coloured" from Wellington, Cape Town and District six (Cape Town) as well as the Khomani (KHM_SA), Griekwa from Cape Town (CPT) and "Coloured" from Johannesburg (JHB) individuals. Most other clusters lie more linearly on the African - Western Eurasian cline along the first two PCs.

Inter-individual Euclidean distances based on the top four principal component positions showed that all SAC groups were highly variable (Figure 6.7).

Notably, the individuals from the KhoeSan groups (≠Khomani, Karretjie, Nama) which were among the 75% furthest from the Ju|hoansi (see Section 6.2) were as variable as the "Coloured", Griekwa, Baster and Cape Malay. The Euclidean distances between individuals were comparably large (range 0.01 - 0.06). This is also evident from the PCA plots (Supp. Figure C.4 - C.33). The scale of this variability is infrequently seen for the GR data after outliers have been removed (mean PCA position distances ~0.005).

Table 6.5: Analysis of variance results of comparisons of the inter-individual distances between *a priori* sample sets of SAC. Includes the Southern African KhoeSan and 'Coloured' groups. Abbreviations: Df - Degrees of Freedom, Sum Sq. - Sum of Squares, Mean Sq. - Mean of Squares.

	Df	Sum Sq.	Mean Sq.	F value	Pr(>F)
a priori sample set	29	0.01	0.00	5.22	1×10^{-10}
Residuals	330	0.02	0.00		

A single-factorial ANOVA for comparisons across the SAC groups was significant ($p<1 \times 10^{-10}$; Table 6.5) and a post-hoc Tukey honest significant difference pairwise comparison indicated significant differences among a few groups. Notably, between the Baster or COLOURED_CPT (using $\alpha = 0.01$) and to a lesser extent COLOURED_KS (using $\alpha = 0.05$) when compared to several other SAC sample sets (Supp. Table C.3). Both the former groups have lower variability compared to other sample sets, while COLOURED_KS has elevated variability. From this I can say that only two of the "Coloured" sample sets show moderate variability in

distances and thus a closer genetic affinity among members within the sample. Most of the SAC KhoeSan parallel the variability in the non-KhoeSan SAC and this is distinctly different to what is seen in the GR KhoeSan groups (e.g. Juhoansi_sc12).

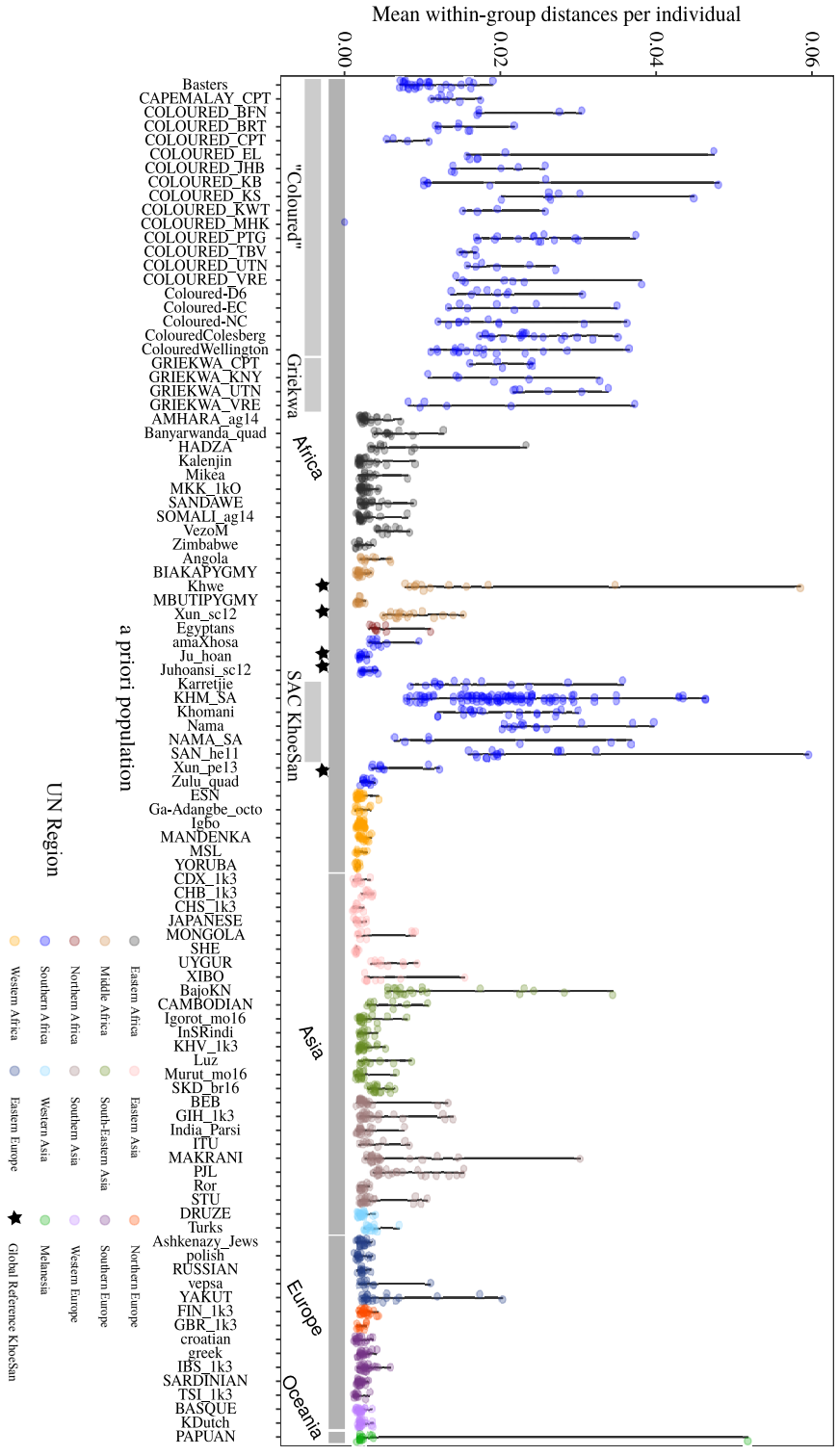


Figure 6.7: Variability of the inter-individual Euclidean distances within each *a priori* population. Points are the mean of distances between individuals for each individual as estimated from the top four eigenvectors. Dots coloured by UN Global Region. Vertical lines indicate the range of values per population. Further information, e.g. sample sizes, available in Supp. Table C.1

6.3.2 Divergences among the SAC Ethno-racial Affinities

I tested for genetic similarity among individuals within *a priori* sample sets and among *a priori* sample sets with shared ethno-racial identities using hierarchical clusters based on Euclidean distances between the PC positions.

By clustering the *a priori* sample sets by mean PCA position (Figure 6.8), I highlight the similarity across the SAC sample sets without disregarding the possibly informative identity affinities. With the exception of the Nama, I find all KhoeSan groups to cluster separately from the non-KhoeSan SAC groups (branch e and f, Figure 6.8)). The clusters are in alignment with the geographic origins of the sample sets as the SAC KhoeSan (Khoi-kwadi and Tuu speakers; branch e) form a separate branch from the Juu-speaking KhoeSan (branch f). The Griekwa from Upington (UTN) cluster with the Nama as a separate branch (branch b) among the non-KhoeSan SAC.

There is evidence of structure based on the clustering of the non-KhoeSan SAC. The groups from the Cape Town region cluster together (along with the "Coloured" from Johannesburg; COLOURED_JHB, branch a). Branches (d) and (c) consist of various regions. Branch (c) included the Griekwa groups from Cape Town and Knysna (CPT, KNY) and "Coloured" sample sets from the Eastern Cape and the eastern interior of South Africa. Branch (d) contained the Namibian Baster as well as groups from the Northern Cape and both "Coloured" and Griekwa from Vredendal (Western Cape).

As admixture is a predominant feature of the SAC, the relationship among the samples sets is likely reticulated through the influence of global contributions. To account for this I examine the clustering using a NeighbourNet [247] analysis which allows for additional edges in the dendrogram.

The distinction of sample sets from the Cape Town region branch (a) and the KhoeSan on branch (e) is again identifiable (Figure 6.9). The identified branches (b-d) on the NeighbourNet network are less clearly distinguished and do not correspond exactly to that of the hierarchical dendrogram. I do find that the branch (c) on

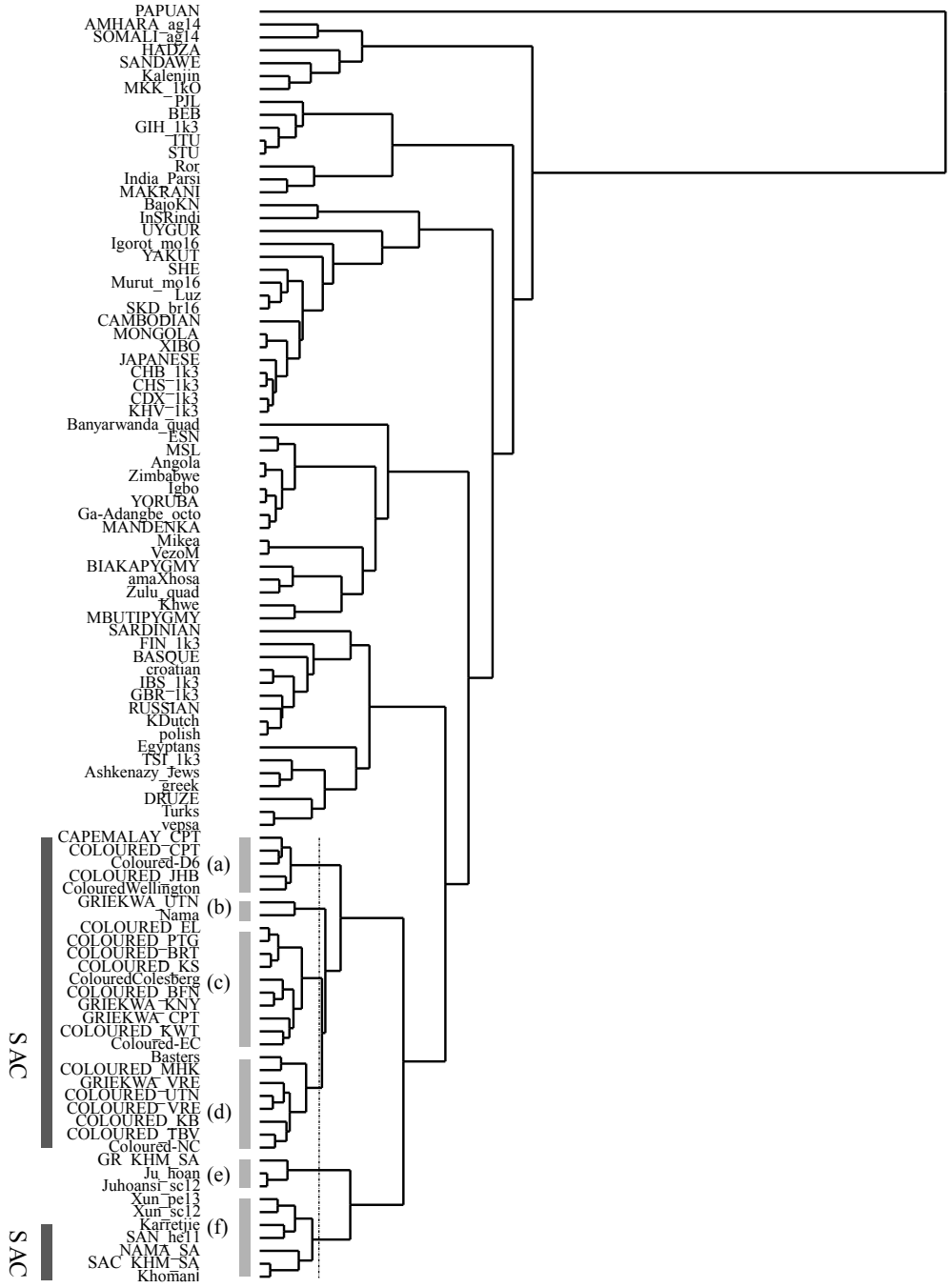


Figure 6.8: Hierarchical clustering of *a priori* populations based on Euclidean distances between pairs informed by mean PCA coordinates. Note, KHM_SA was split into two sets as described in Section 6.2; Shown here are both the 25% of samples closest to the Ju|hoansi (GR) and the 75% furthest (SAC). Dotted line indicates the position used to cut the dendrogram based on visual inspection.

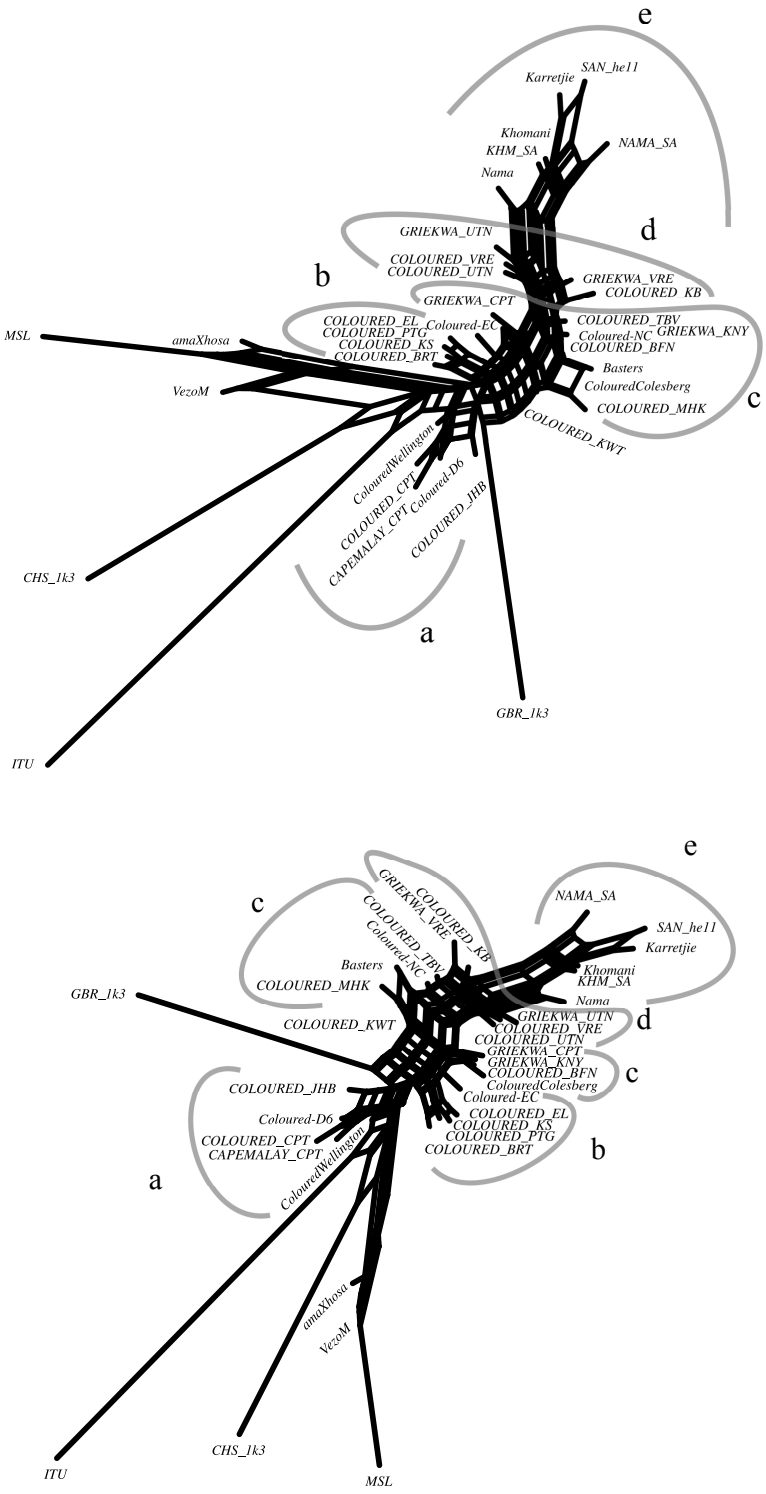


Figure 6.9: NeighbourNet clustering of *a priori* sample sets based on Euclidean distances between pairs informed by mean PCA coordinates. The network shown from two different angles. Proposed clusters are indicated by letters and bars.

the hierarchical dendrogram reflects the clusters (b) and (c) on the network, while (d) corresponds approximately to (d) on the network. We can again see the close relationship of the GRIEKWA_UTN with the Nama, and in general the closer affinities of the Griekwa to the KhoeSan, compared to most of the "Coloured" groups. Considering the close relationship among the SAC as seen in the network, the distinction between clusters (b - d) is likely subjective. For further discussion, I focus on the clusters identified by the hierarchical dendrogram when relevant.

Clustering by individual PCA position produced a dendrogram with minor similarity to the *a priori* sample sets. I retained an arbitrary number of clusters on the individual-based dendrogram (36 to match the number of *a priori* sample sets). The clusters were heterogeneous in their inclusion of samples from *a priori* sample sets and sample overlap with the *a priori* sample sets was most frequently below 25%, indicating that very few *a priori* sample sets could be reproduced (Figure 6.10 and 6.11).

The exceptions to this were among the SAC KhoeSan where individual-based PCA clusters 5 - 11 were predominantly KhoeSan. A second possible exception is seen for the Cape Town region sample sets which were well represented in clusters 22 -24, forming a somewhat coherent block (Figure 6.10). Cluster 26 was predominantly Cape Malay and "Coloured" from District six. Clusters based on geographic position (Figure 6.10) do not show better correspondance to *a priori* groups beyond what is produced by the clustering of the Cape Town region groups.

The Griekwa did not form a coherent block of clusters. The Nama and Griekwa from Upington (UTN) co-cluster (e.g. clusters 12, 13, 19, 30 and 31) but individuals in both groups are split between several clusters across the dendrogram. The Griekwa from Knysna, Cape Town and Vredendal (KNY, CPT, VRE) along with the "Coloured" from Vredendal co-cluster in cluster 27 suggesting some similarity.

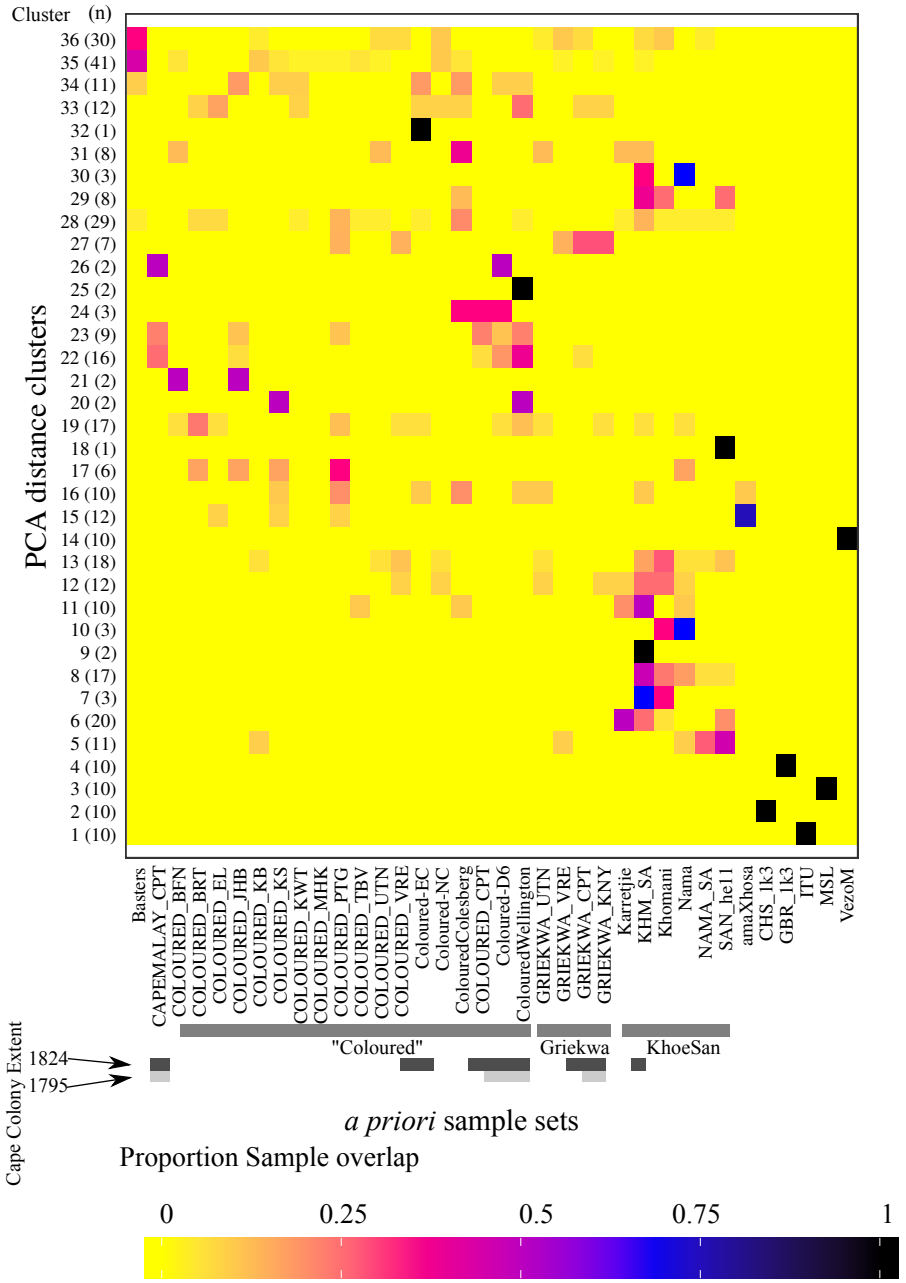


Figure 6.10: Sample overlap between the *a priori* sample sets and ethno-racial affinity clusters. SAC and five GR representative sample sets included. The *a priori* sample sets (x axis) are arranged by ethno-racial affinity. I indicate whether sampling localities are within the geographic extent of the Cape Colony circa 1795 and 1824 or beyond this with the lower grey bars. The ethno-racial affinities are shown by the top grey bars. I retain an arbitrary number of clusters (36; the number of *a priori* groups) on the hierarchical clustering (y axis). Number of individuals per cluster (n) shown. Abbreviation for reference populations; MSL - Mende from Sierra Leone, ITU - Indian Telugu from the UK, GBR_1K3 - British from the UK, CHS_1K3 - Han Chinese from China, VezoM - Vezo Malagasy.

The overall similarity in principal component position was assessed using an Armitage trend χ^2 statistic pairwise across the *a priori* populations (Figure 6.12 and 6.13). Statistically significant differences were common among GR populations whilst comparisons among the SAC groups were most often not significant. The overall pattern was in alignment with the dendrogram clustering, as expected. Significant differences are seen among Griekwa (e.g. GRIEKWA_KNY/UTN vs GRIEKWA_VRE) and among "Coloured" groups. Conversely, non-significant differences are seen between "Coloured" and Griekwa groups demonstrating that genetic divergences are not overtly tied to ethno-racial identity.

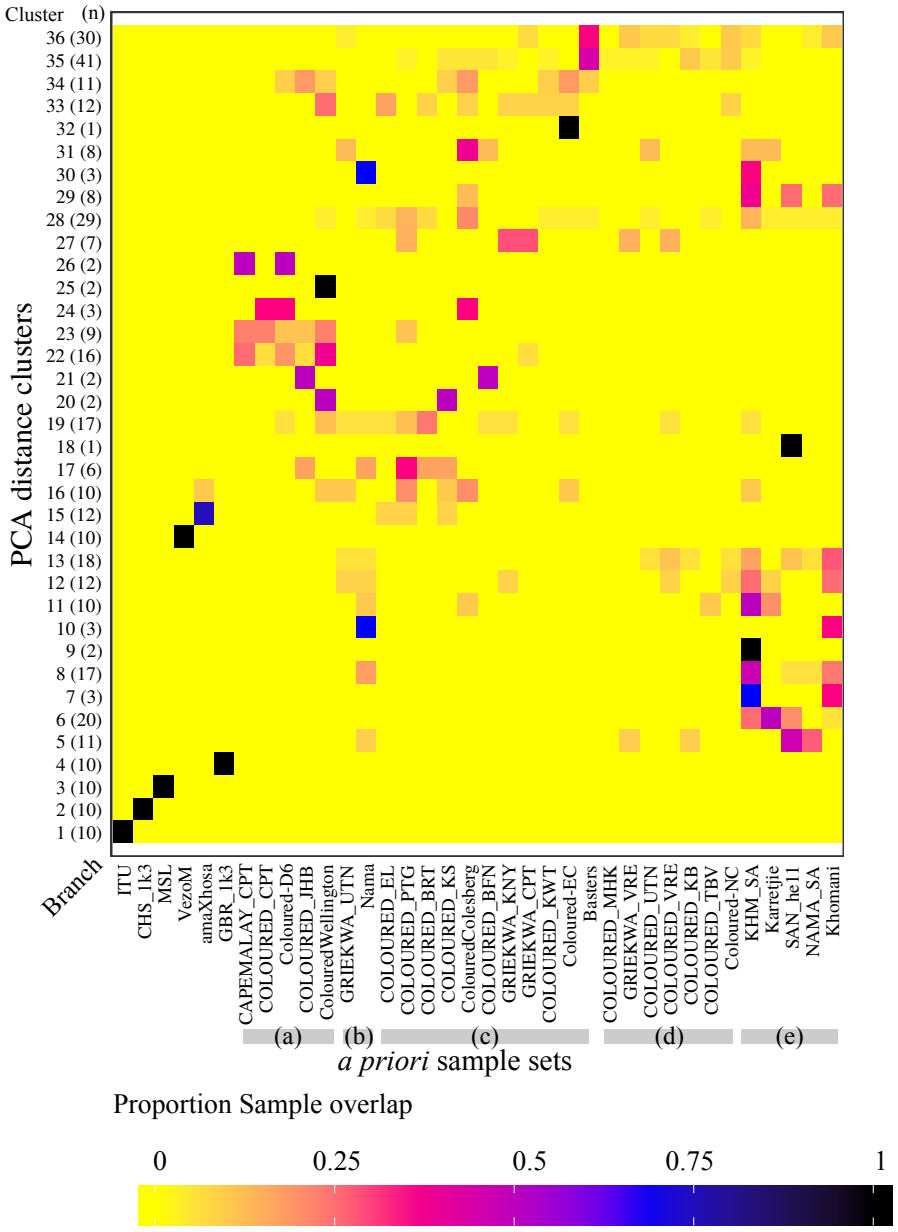


Figure 6.11: Sample overlap between the *a priori* sample sets and clusters of individuals based on Euclidean distances informed by PCA coordinates. SAC and five GR representative sample sets included. The *a priori* sample sets (x axis) are arranged by the hierarchical dendrogram of *a priori* sample sets (Figure 6.8). Branch letters indicated, corresponding to Figure 6.8. I retain an arbitrary number of clusters (36; the number of *a priori* groups) on the hierarchical clustering (y axis). Number of individuals per cluster (n) shown. Abbreviation for reference populations; MSL - Mende from Sierra Leone, ITU - Indian Telugu from the UK, GBR_1K3 - British from the UK, CHS_1K3 - Han Chinese from China, VezoM - Vezo Malagasy.

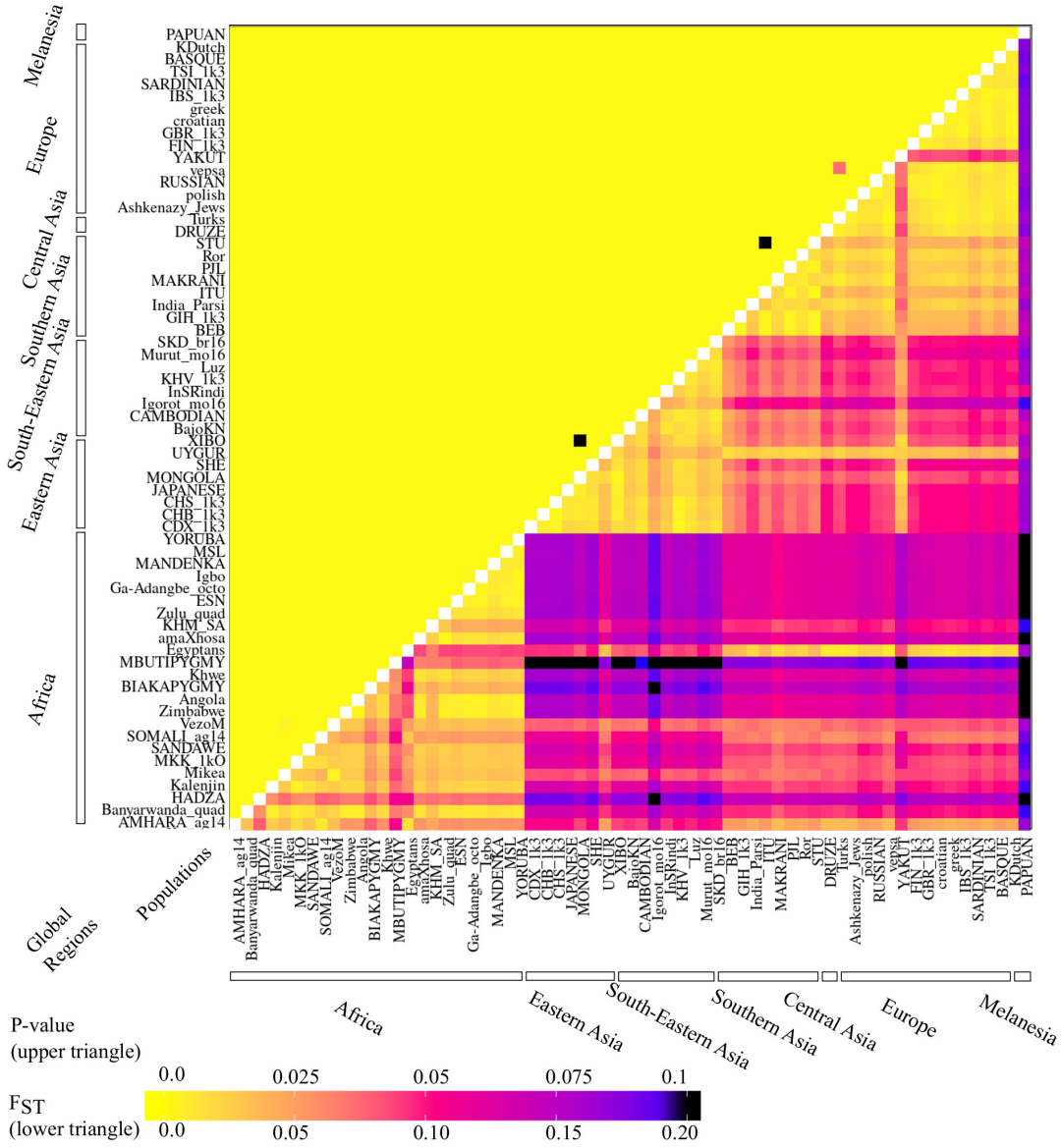


Figure 6.12: Values for pairwise F_{ST} and p-values for differences in PC position for the GR populations. Indicated are p-values after adjusting for multiple comparisons (upper triangle) and pairwise estimates of F_{ST} between the *a priori* populations (lower triangle) (70 GR). Bars below and to the left of population labels indicate global regions for the GR data.

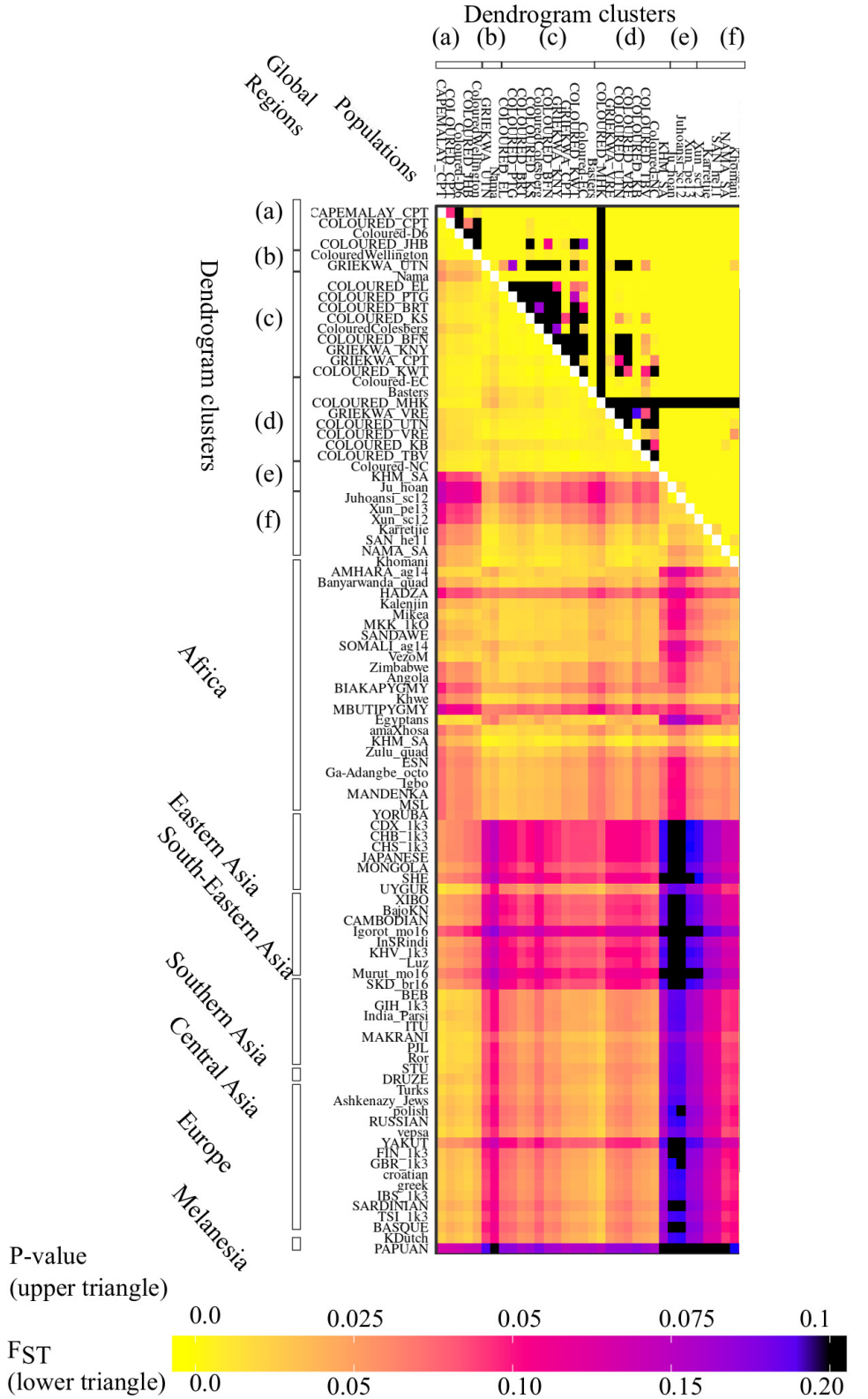


Figure 6.13: Values for pairwise F_{ST} and p-values for differences in PC position for the SAC. Indicated are p-values after adjusting for multiple comparisons (upper triangle) and pairwise estimates of F_{ST} between the *a priori* populations (lower triangle) (70 GR and 29 SAC). Bars above and to the left of population labels indicate PCA distance dendrogram cluster for the SAC and global regions for the GR data.

The patterns in pairwise F_{ST} estimates correspond to the dendrogram structure for some sample sets (Figure 6.13).

The F_{ST} estimates between the "Coloured" and Griekwa were moderate at most, (-0.001 - 0.03) and comparable to estimates between "Coloured" groups (-0.003 - 0.02), demonstrating that distances between "Coloured" and Griekwa are often lower than distances among Griekwa and among "Coloured". Between the Griekwa groups, the largest divergence is seen between the Upington (UTN) and Cape Town (CPT) or Vredendal (VRE) groups. These values again are in agreement with GRIEKWA_UTN being clustered with the Nama rather than other non-KhoeSan SAC in the PCA dendrogram.

Among the "Coloured", the largest distances are found between a variable set of comparisons but values corresponding to divisions seen in the PCA distance dendrogram.

Pairwise distances with the Cape Malay did not correspond to clustering in the dendrogram (Figure 6.13). Divergences within branch (a) (0.01 - 0.05), (b) (~0.046) and (c, d ; 0.012 - 0.031) were similar. Results for the Baster were also discordant from the dendrogram.

The upper values between the non-KhoeSan SAC sample sets (~0.03) are comparable to values between global populations separated by hundreds of kilometers (e.g. Punjabi - Croatian, 0.028; amaXhosa - San, 0.03) suggesting notable differences. However, considering the range of global contributions to ancestral components and the divergences between the source populations, simple variation in the proportion of components may cause large F_{ST} divergences.

Overall, the F_{ST} values indicate that there is structure to the ancestral contributions between sample sets, and in some comparisons, divergences are substantial (likely reflecting proportion of ancestries). These differences may reflect different histories of admixture and drift, and within population heterogeneity.

6.3.3 Ancestral Contributions to the SAC

When estimating F_{ST} distance between the SAC and GR datasets, the values provide some indication of differences in ancestral contributions. Despite the high inter-individual variation in principal component space, the KhoeSan groups show some coherence in having the lowest F_{ST} values for pairwise comparisons within their dendrogram branch and with GR KhoeSan groups. Divergences of the SAC KhoeSan from the Eurasian GR populations were lower compared to the GR KhoeSan - Eurasian divergences (e.g. Ju|hoan, !Xun), supporting additional ancestral contributions to the SAC KhoeSan.

With the non-KhoeSan SAC, the lowest F_{ST} values are consistently with the Southern African KhoeSan groups followed by East Africans and Europeans (Figure 6.13). The Nama, Khomani and Karretjie (0.001 - 0.007) produced the lowest values (Figure 6.13).

Branch (a), the Baster and "Coloured" from the Eastern Cape (Coloured-EC) had the lowest F_{ST} values with GR groups other than the KhoeSan. The pairwise F_{ST} values for these groups were lowest with non-Africa GR groups; Specifically these non-KhoeSan SAC included the Cape Malay (compared to Uygur, 0.013), the "Coloured" from Johannesburg, Coloured-D6 and the Baster (with Egyptians, 0.014, 0.017 and 0.021, respectively), the "Coloured" from Cape Town (with Vezo Malagasy, 0.016) and the Coloured-EC (branch c) and the "Coloured" from Wellington (against Amhara, 0.017).

The Bayesian clustering algorithm in ADMIXTURE was used to identify ancestral components present in the SAC more specifically. The ADMIXTURE cross validation errors continued to drop until K = 10, after which the change in error levelled off (Figure 6.14). I consider any value from 10 - 20 to be suitable, with preferences for lower K where results can be linked to global populations less ambiguously.

ADMIXTURE produces a matrix of group assignment probabilities, each entry is a probability that each individual is assigned to one of the K groups. I refer to

these explicitly as group assignment probabilities but also as 'ancestral components' to reflect that mixed group assignments may reflect shared ancestry with the ancestral K populations.

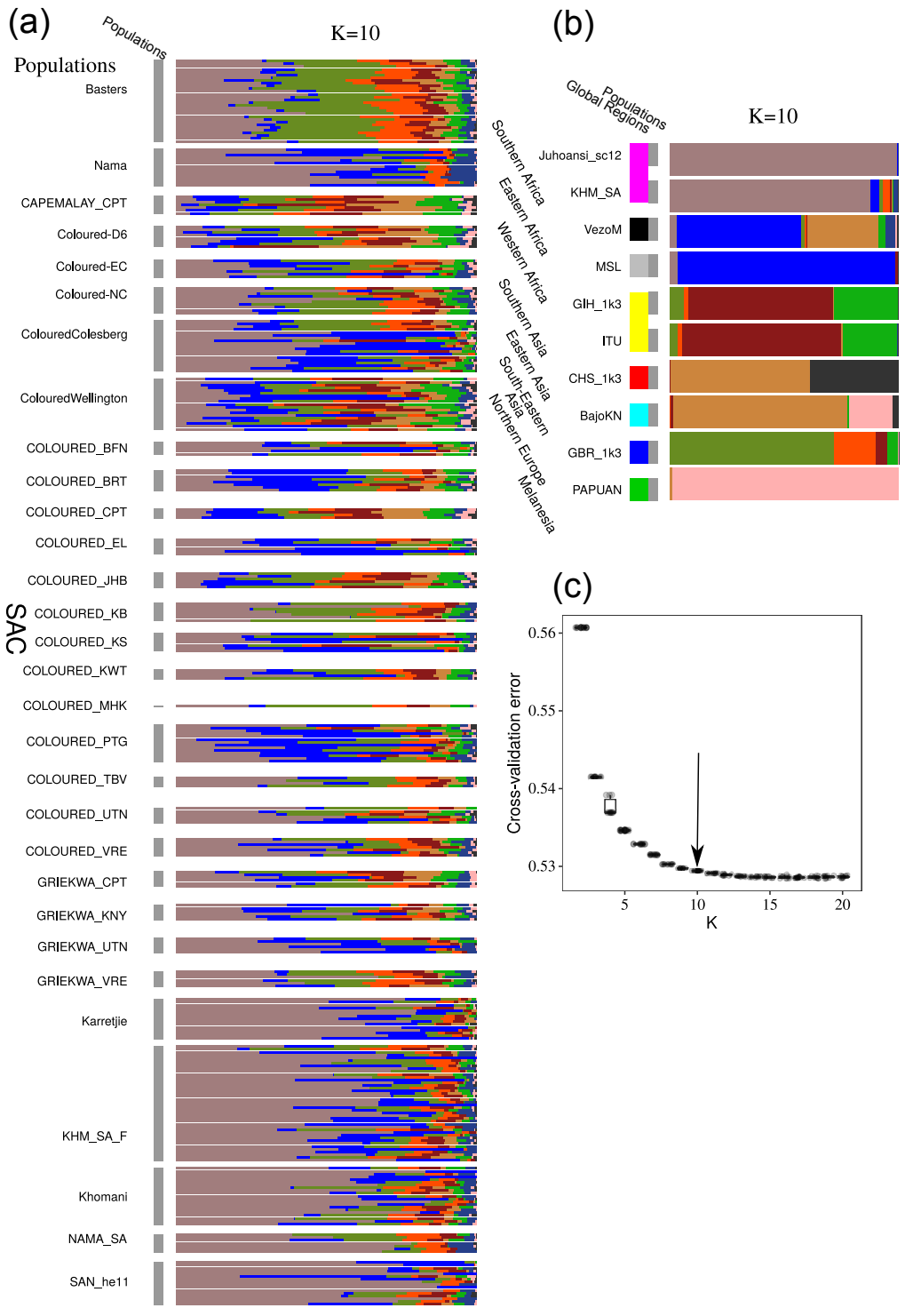


Figure 6.14: ADMIXTURE profiles (K = 10) for SAC and representative GR *a priori* groups. Shown (a) are the profiles for each SAC sample arranged by *a priori* sample sets and, below, (b) the selection of *a priori* populations from GR data averaged over samples. Immediately right of the population labels are bars which indicate groups of samples. Global region labels are to the far left of (b). (c) Boxplot of the cross-validation errors at different K values 2...20 with 10 runs each. K = 10 indicated by an arrow.

Results for the GR data are comparable to that found in Chapter 4 and in line with previous work [e.g. 38], [97], [99], [100], [115]. The earliest split, K = 2, occurs between African and non-African populations (Figure 6.14 and Supp. Figure C.35). The Eurasian component splits into an Eastern - Western component at K = 3 and the KhoeSan related ancestry separates from non-KhoeSan African ancestry at K = 4. At K = 5 a joint Southern, South-East and Central Asian component separates from the Eastern Asian component and further differentiates at K = 6 highlighting Melanesian and South Asian shared ancestry. From K = 7 ... 9, a second component for Western Eurasian, Eastern Asian and African (Eastern) separates out. Finally, at K = 10, the Western Asian samples are distinguished from European samples. The sequence of separation of ADMIXTURE components approximates the order of the principal components recovered.

ADMIXTURE components best represent distal ancestral components and not the proximal ancestries that can be recovered by haplotype-based clustering as in CHROMOPAINTER [174]. Thus, the presence of specific ancestries such as Malagasy, Chinese or Malay is unfortunately rather ambiguous based on group assignment probabilities alone. Various combinations of contributions could produce similar profiles as discussed regarding the specific South Asian contributions in Chapter 4. Contributions to the SAC include diverse populations such that distal ancestries are informative still.

In the SAC, at K = 10 all ten assignments are observed at some probability. This is again suggestive of contributions from multiple global regions (Figure 6.14), including Southern Africa (KhoeSan), Western Africa, South Asia, East Asia, South-East Asia and Europe. For each of the ten components, I use the GR sample set which has the highest assignment probabilities to that cluster as a proxy for the ancestral source.

The largest components in the SAC are shared with Ju/'hoansi (Juhoansi_sc12), Esan from Nigeria (ESN), Sri-Lankan Tamil from the UK (STU), Igorot from the Philippines (Igorot_mo16) and Finnish from Finland (FIN_1k3) (Figure 6.15). Two of the ten components are near zero across all clusters; the Papuan-related

component and the Yakut-related component (Figure 6.15). In both cases these may be indirect contributions from other GR groups as both are present in more likely source populations (e.g. Bajo from South-East Asia and Vezo from Madagascar). A further three components (represented by the Makrani, Hadza and Druze) are minor contributions to group assignment and again are most likely tied to one of the major components.

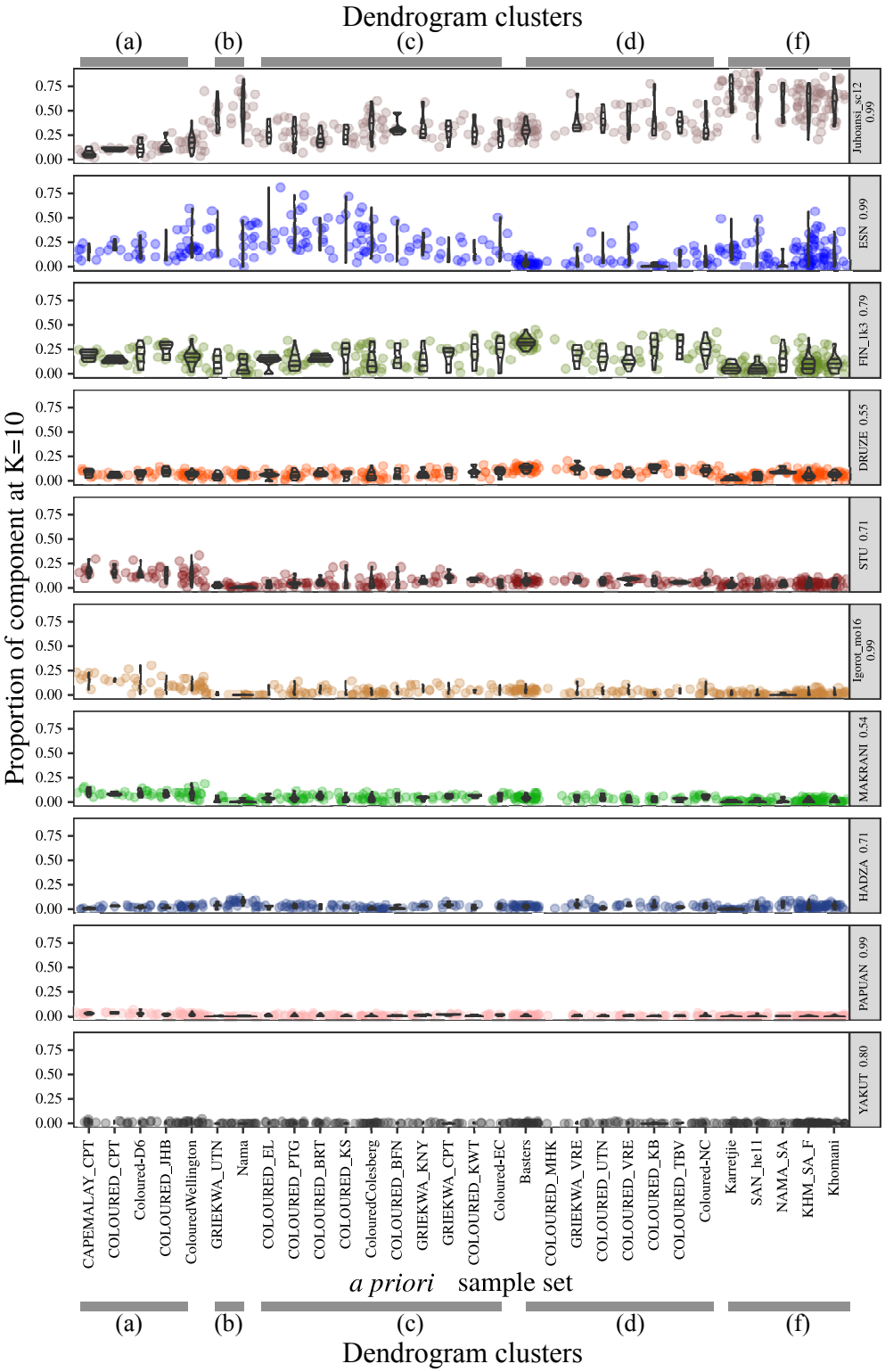


Figure 6.15: Variation in the group assignment probabilities of the SAC from ADMIXTURE (K = 10). Sample sets arranged by the PCA distance dendrogram (Figure 6.8). Individual dots represent individual samples. Violin plots show the range of values with 25, 50 and 75th percentiles shown where possible. Panel labels (right) indicate in which GR *a priori* population the group assignment was the highest, and what the average assignment was within that *a priori* population, e.g. ESN 0.99 means that Esan from Nigeria were assigned to this K cluster with 99% probability. PCA distance dendrogram branches indicated above and below the plot.

The KhoeSan, European and Western African ADMIXTURE components are clearly the most common and abundant, present in most SAC individuals. These result are in line with previous work [12], [19], [99], [110], [185].

Variation in the ADMIXTURE assignments among the SAC show structure corresponding to the branching on the PCA distance dendrogram, suggestive of a distinction between the Cape Town region groups and the KhoeSan from the other SAC groups (Figure 6.15), but do not clearly support differences between branches (c) and (d) of the *a priori* sample sets dendrogram, nor a distinction between "Coloured" and Griekwa.

The Cape Town Region groups have higher levels of South-East and South Asian components and the lowest KhoeSan and West African components out of the SAC. This pattern was common to the "Coloured" and Cape Malay from the region.

I find no clear support for a distinction between branch (c) and (d) of the PCA distance dendrogram. Members of the branches, both "Coloured" and Griekwa, have overlapping variability in European, non-KhoeSan African and KhoeSan ancestry.

With the exception of GRIEKWA_UTN which is most similar to the Nama, Griekwa are not distinct from the "Coloured". Nor is there a clear distinction among the Griekwa. While the Griekwa from the Upington (UTN) and Vredendal (VRE) clustered separately from the Griekwa from the Cape region (CPT,KNY), they do not have divergent ADMIXTURE profiles.

Neither of the Griekwa from within the 1795 Cape administrative boundary (KNY, CPT) are similar to the other Cape Town region groups. Both Griekwa sample sets had notably higher levels of KhoeSan ancestry, reflected in their position among the "Coloured" of branch (c) of the *a priori* clustering dendrogram.

The Baster have the highest proportions of European ancestry and among the lowest proportions of non-KhoeSan African as well as among the lowest variation in all ancestry components, consistent with the low inter-individual PCA distances.

The SAC KhoeSan have consistently lower proportions of the European component than other SAC groups, while the non-KhoeSan African ancestry is variable.

The non-KhoeSan African component differs between the two Nama datasets, it is nearly absent from NAMA_SA [from 113] but variable in the Nama [from 19], this suggesting different histories or different data collection criteria.

I tested each major ADMIXTURE component for a trend with distance from Cape Town which may reflect evolution related to the expansion from the settler colony. The European component showed no correlation with distance for any of the SAC ethno-racial affinities (Figures 6.16 - 6.18). Only in the SAC KhoeSan did the KhoeSan component change significantly, decreasing ($r^2 = 0.85$).

Two components had a constant trend across ethno-racial identities. The South-Asian component correlates (weakly) with distance from Cape Town, declining (Griekwa, "Coloured" and SAC KhoeSan, $r^2 = 0.46$, 0.14, 0.10; coefficient = -0.00004 - -0.0001Km^{-1}) (Figures 6.16 - 6.18). However, both the South Asian and South-East Asian components are higher in the Cape Town region groups and uniformly low in all other groups. In contrast, non-KhoeSan African ancestry increased significantly ($p < 0.05$) but weakly (coefficient = 0.0001 - 0.00026) across sample sets (SAC KhoeSan, "Coloured" and Griekwa; $r^2 = 0.08$ - 0.16).

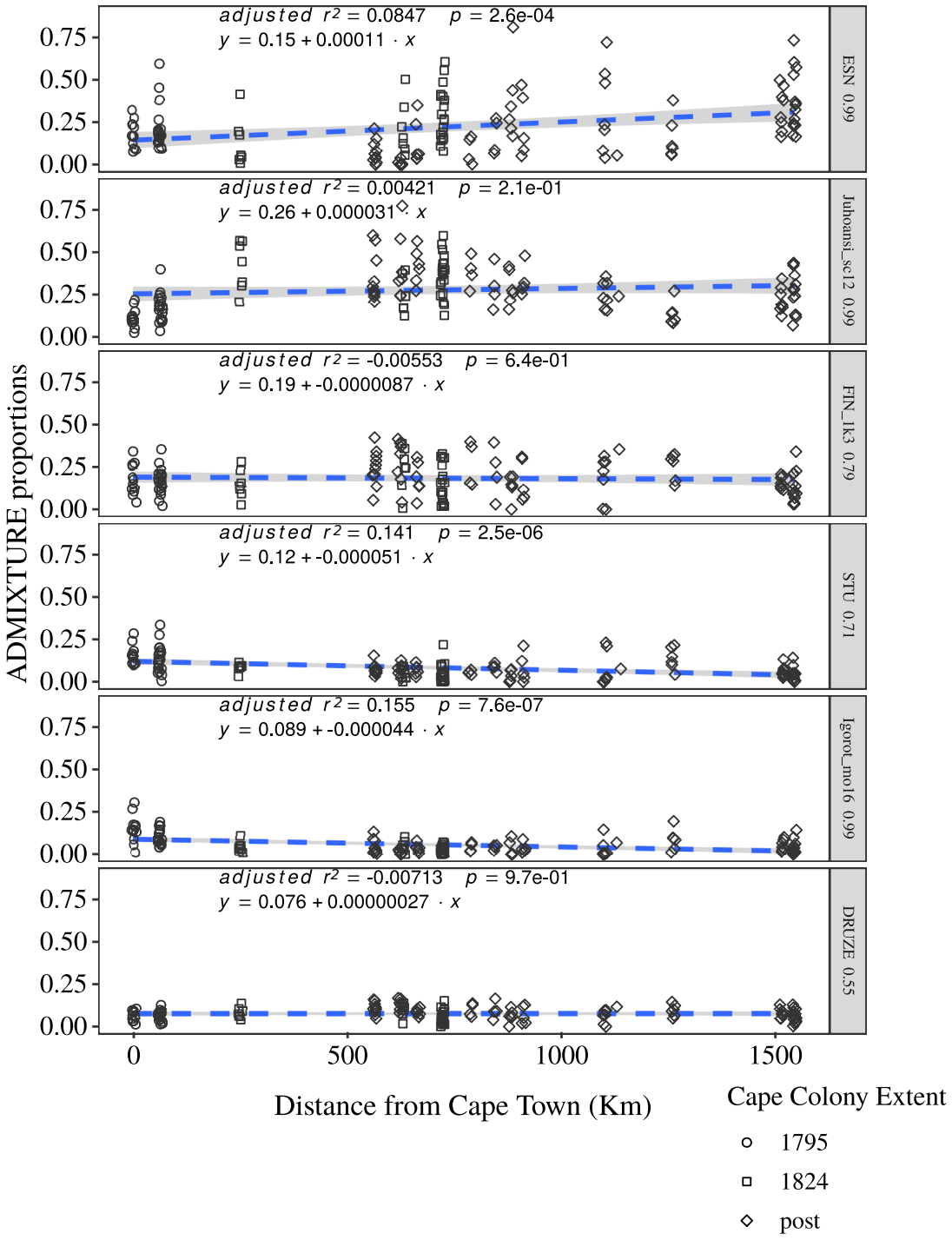


Figure 6.16: Correlation of ADMIXTURE components with distance from Cape Town for 'Coloured' individuals. Only the six largest group assignment probabilities based on ADMIXTURE from K = 10 components are shown. Individual dots represent individual samples, symbol indicates position of sample site relative to Cape Colony extent in 1795, 1824 or post-1824. Panel labels (right) indicate in which GR *a priori* population the group assignment was the highest, and what the average assignment was within that *a priori* population, e.g. ESN 0.99 means that Esan from Nigeria were assigned to this K cluster with 99% probability.

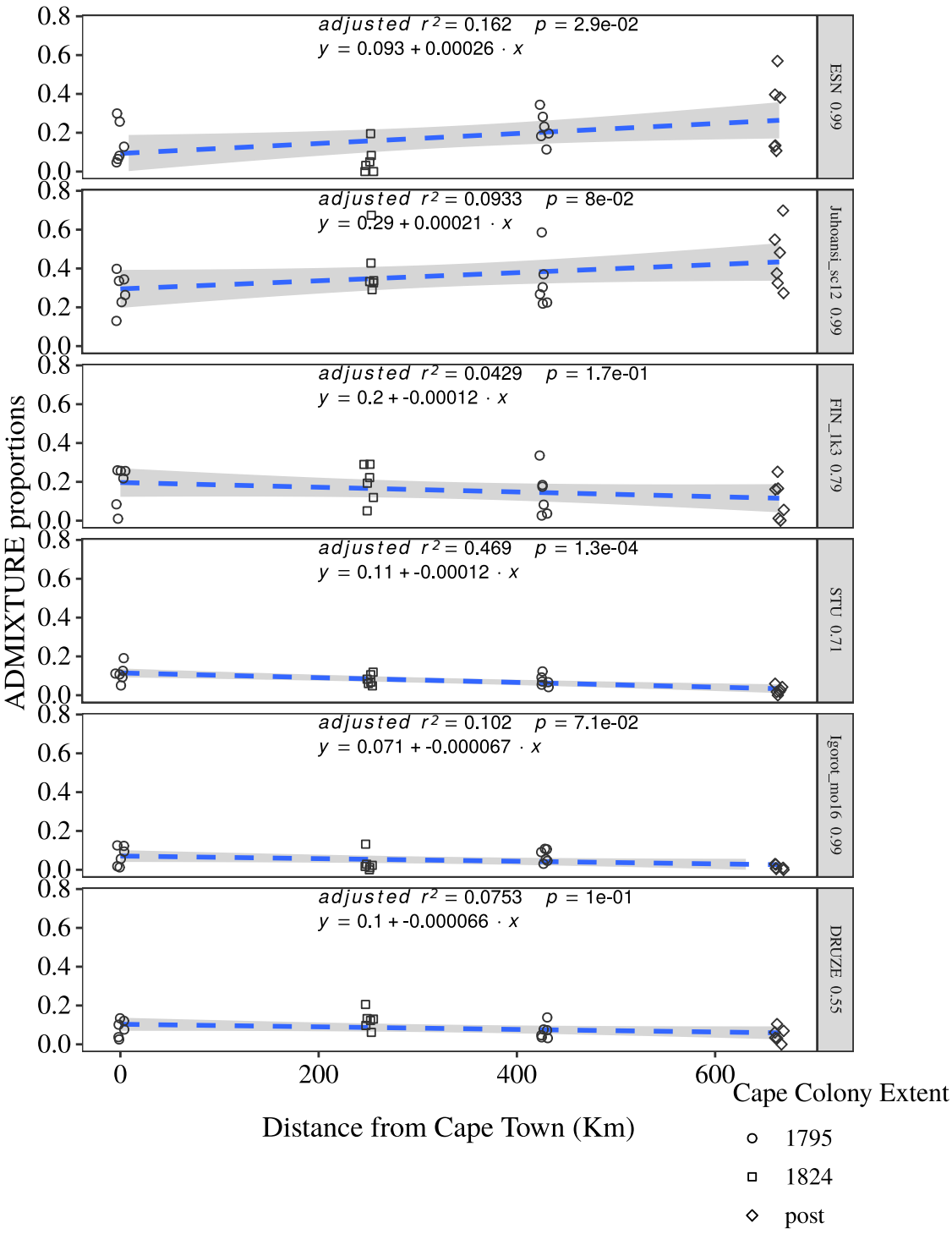


Figure 6.17: Correlation of ADMIXTURE components with distance from Cape Town for Griekwa individuals. Only the six largest group assignment probabilities based on ADMIXTURE from K = 10 components are shown. Individual dots represent individual samples, symbol indicates position of sample site relative to Cape Colony extent in 1795, 1824 or post-1824. Panel labels (right) indicate in which GR *a priori* population the group assignment was the highest, and what the average assignment was within that *a priori* population, e.g. ESN 0.99 means that Esan from Nigeria were assigned to this K cluster with 99% probability.

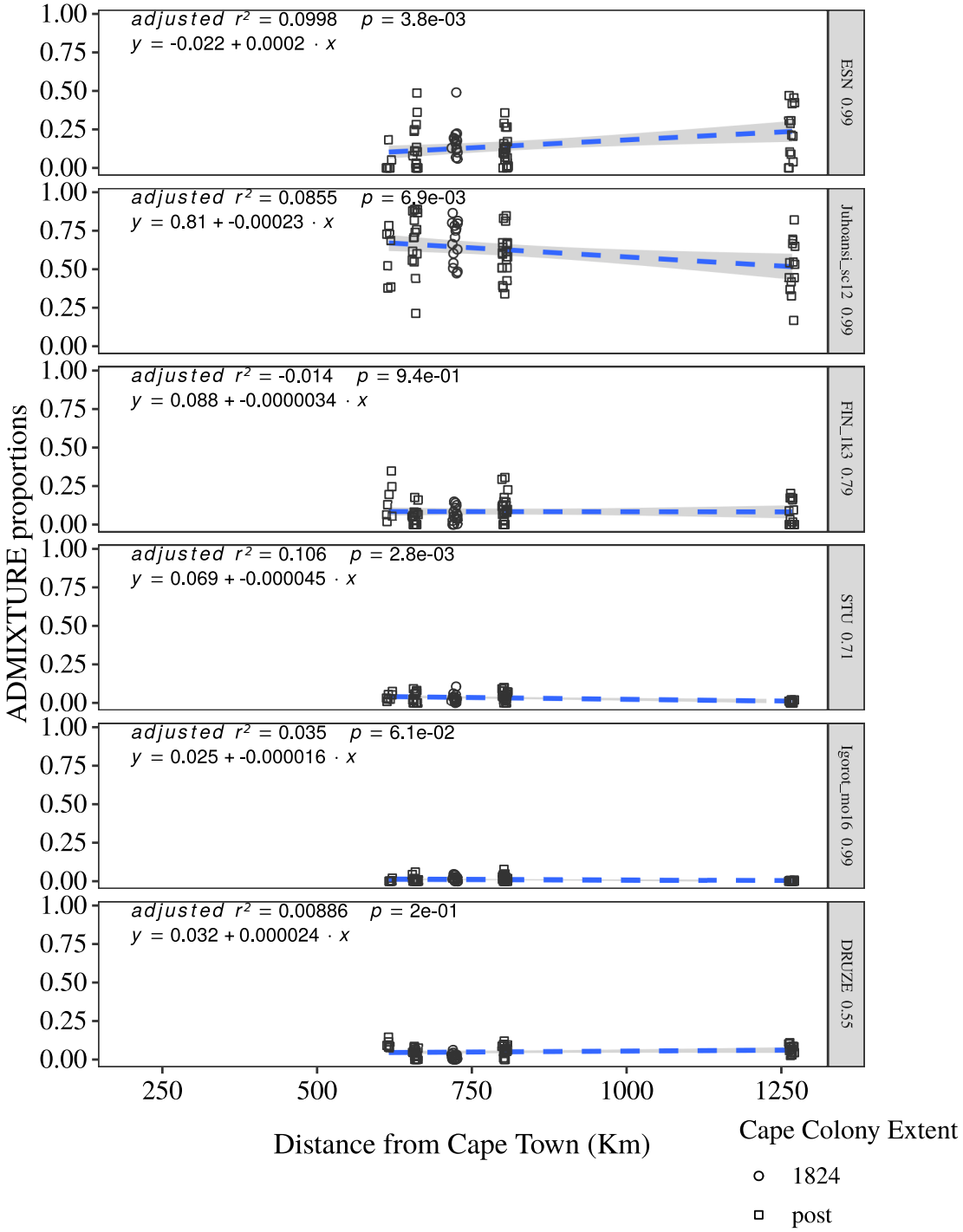


Figure 6.18: Correlation of ADMIXTURE components with distance from Cape Town for KhoeSan individuals. Only the six largest group assignment probabilities based on ADMIXTURE from K = 10 components are shown. Individual dots represent individual samples, symbol indicates position of sample site relative to Cape Colony extent in 1795, 1824 or post-1824. Panel labels (right) indicate in which GR *a priori* population the group assignment was the highest, and what the average assignment was within that *a priori* population, e.g. ESN 0.99 means that Esan from Nigeria were assigned to this K cluster with 99% probability.

6.3.4 Admixture Dating

The results from fitting linkage disequilibrium decay curves produced similar date estimates across the SAC groups, suggesting similar recent events, but some geographic variation was present (Figure 6.19).

Most of the bootstrap iterations across the SAC produced date estimates which predated the emancipaiton of the slaves in ~1838 but in all cases post-dated the 1652 arrival of the Dutch to Southern Africa. There was no support in any of the sample sets for the pre-Cape Colony admixture between KhoeSan and Europeans which was found in Chapter 5.

Sample sets on branch (c) of the PCA distance dendrogram produced date estimates which were younger than estimates from other sample sets (barring the KhoeSan groups on branch (e) and the "Coloured" from Johannesburg). The sample sets on branch (e) and several of the "Coloured" sample sets (e.g. JHB, EC, TBV, EL, PTG) produced date estimates which were younger than those from the Cape Town region "Coloured" groups.

There was a correlation of admixture date with distance from Cape Town for the "Coloured" groups (Figure 6.20), however this was not the case for the Griekwa or KhoeSan. While significant ($p < 0.007$), the decline was very gradual (-0.0013 $generations.Km^{-1}$ or -0.037 $years.Km^{-1}$).

Several groups produced date estimates with low variability across jack-knife iterations , in particular the Baster (n = 30 samples), "Coloured" from Colesburg (n = 19) and all of the sample sets on branch (e) of the dendrogram. This is suggestive of a rather homogenous sample of individuals who share a similar pulse of admixture in their history. The Baster have a distinctly older admixture estimate which is *en par* with the older estimates for the "Coloured" and Griekwa.

Identifying sources from the top MALDER curves will reflect larger mixing proportions for these ancestries and thus allowing more power to MALDER for detections. This does not preclude additional contributions in other SAC sample

sets where these sources are not found but does indicate some quantitative difference between sample sets.

With regards to the identified sources, the Dutch were commonly identified as the best proxy while the GBR population (British) was less frequently identified and, French and IBS (Iberian from Spain) were recovered once each. These results provide support for multiple European contributions to the SAC sample sets. Iterations for the "Coloured" groups from the Cape Town region (branch (a) on the dendrogram) have top decay curves fit with Asian sources. A South-East Asian population features in younger dates for the Cape Malay, in older dates for the "Coloured" from Wellington and variably for the "Coloured" from District six (D6). The best fit GR population for the Cape Malay differs from the other Cape Town region SAC sample sets, Dusun_mo16 and Igorot_mo16 are identified whilst in other sample sets the Dusun_mo16 source is infrequently identified (e.g. COLOURED_PTG) and Igorot_mo16 is predominant. This suggests that there is differences in the admixture history between these groups.

Furthermore, The Griekwa produce curves involving more than the Dutch - Ju|hoan sources which is most commonly observed in other "Coloured" sample sets. Two Griekwa groups produce curves featuring South-East Asian populations at their earlier dates (GRIEKWA_VRE & KNY).

In several SAC groups ("Coloured" from Kokstad (KS), Wellington and D6), the Mozambicans were identified as the best non-KhoeSan African source, indicating Bantu ancestry. The Esan (Nigeria, ESN) were infrequently identified. The "Coloured" from District six (D6) show a distinction from other Cape Town region sample sets as frequently the ESN or Mozambique sources are identified for the youngest dates.

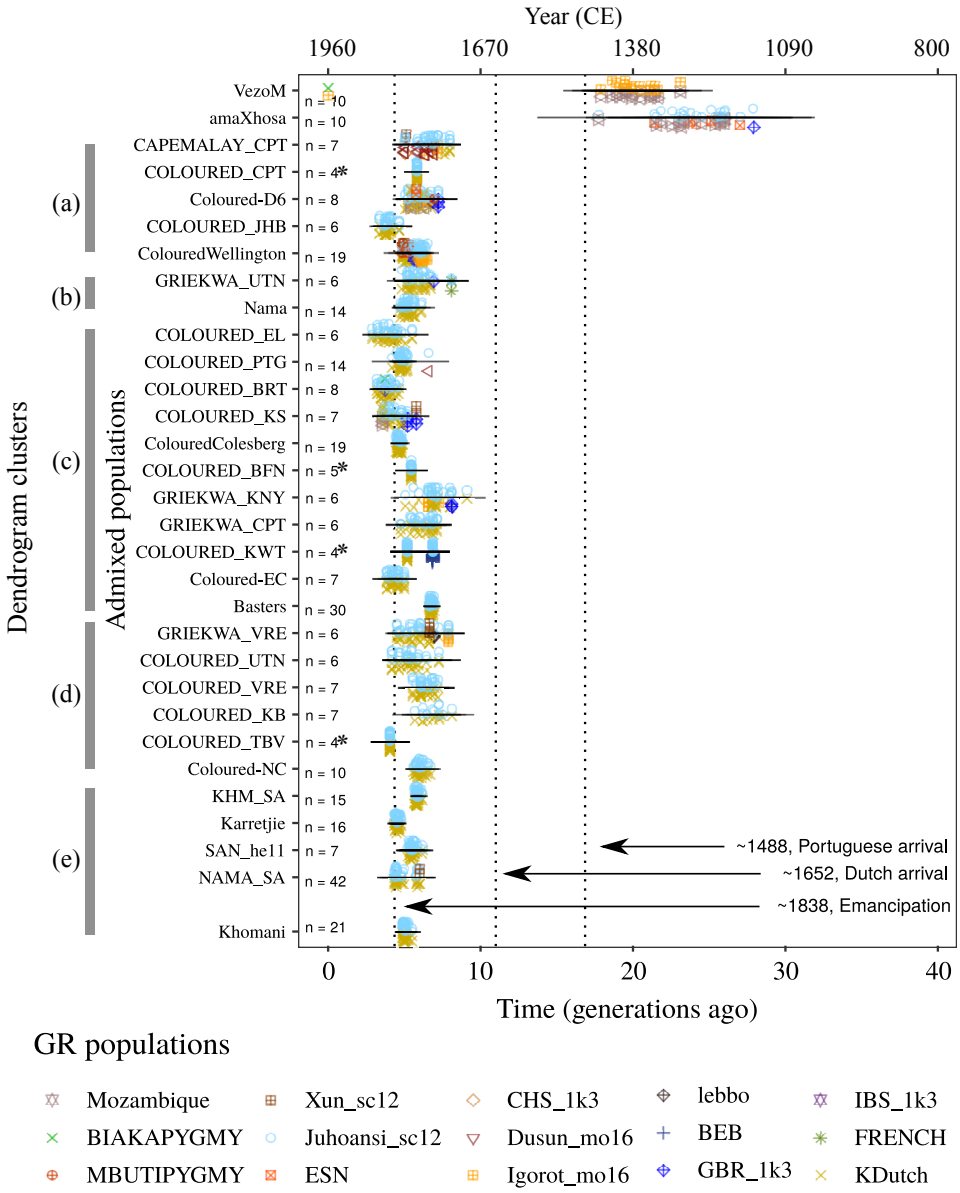


Figure 6.19: Top-ranked MALDER curves from bootstrap iterations for SAC *a priori* sample sets. Shown are iterations (30 each) from each SAC set and two additional populations, Vezo (Malagasy) and amaXhosa (Southern Bantu). Pairs of sources identified are plotted one beneath the other at the estimated date of admixture for that iteration. Sample size (n) per sample set indicated right of the population labels. Data presented includes single and two event admixtures. Where n $<$ 6 (indicated with *), MALDER iterations were of the entire sample set. Where n $>$ 6, at least 30 unique permutations were possible. Data include only events $<$ 50 generations ago. PCA distance dendrogram branches indicated with grey bars. Arrows indicate historic events of interest; 1838 - Emancipation of the slaves, 1652 - Arrival of the Dutch, 1488 - Arrival of the Portuguese [217].

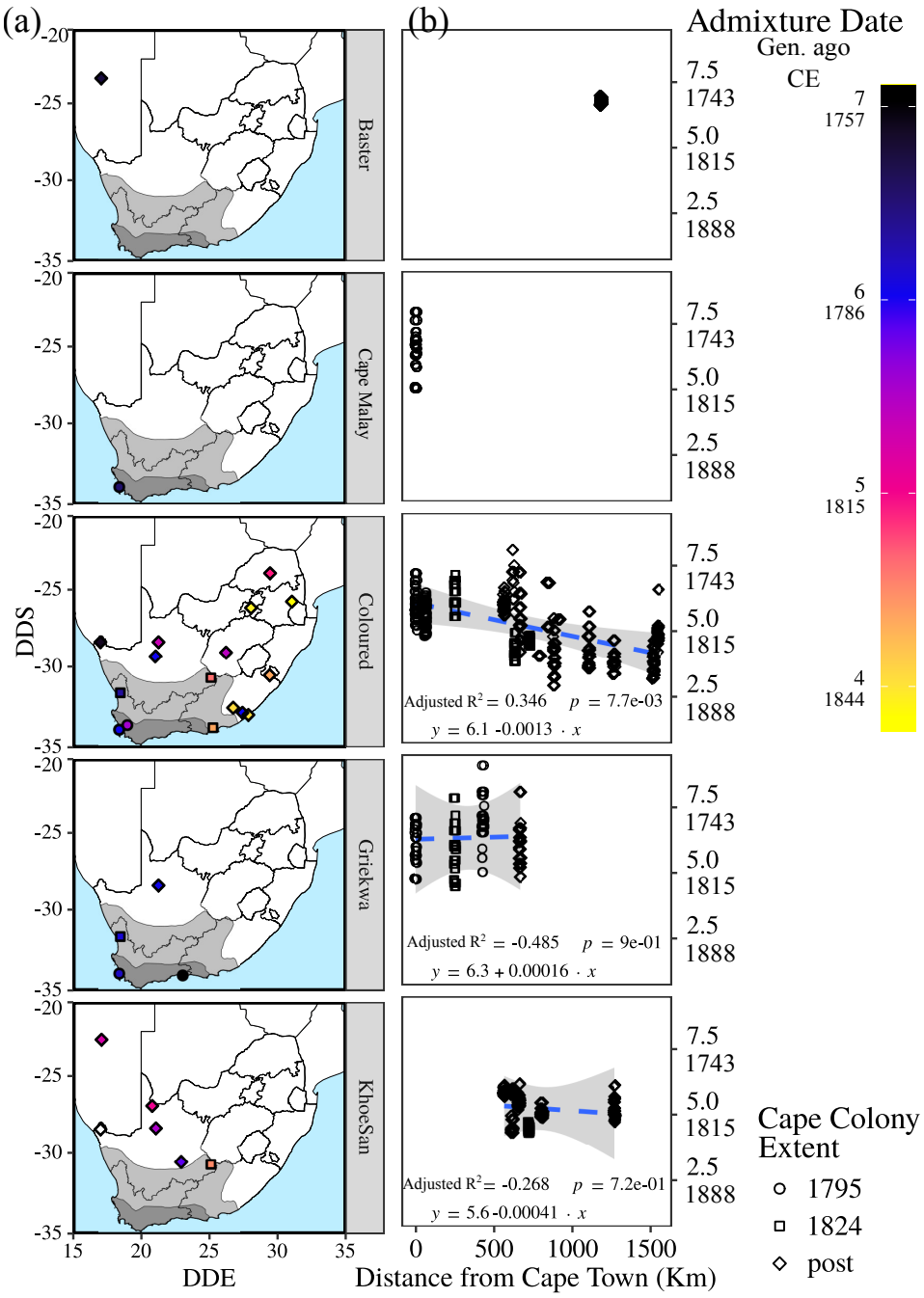


Figure 6.20: Change in MALDER admixture date estimates with distance from Cape Town mapped by ethno-racial affinity. (a) Mapped admixture date estimates averaged across bootstrap iterations. Map polygons reflect the extent of the Cape Colony in 1795 (dark grey) and 1824 (light grey). (b) Correlations of the change in date estimate with distance from Cape Town. These were fit on SAC *a priori* group averages (not individual iterations). Where n<6, MALDER iterations were of the entire sample set. Bootstrap iterations (points plotted) include only events <50 generations ago. Plot symbols in (a) and (b) indicate the position of the sampled site in relation to the extent of the Cape Colony; within the 1795, 1824 and post-1824 ("Post") boundary. Abbr. DDS/E - Decimal Degrees South/East, gen. ago - generations ago, CE - Common Era.

To test for specific contributions from the Malagasy (VezoM) and Southern Bantu (amaXhosa), I tested if the SAC produced a corresponding ancient admixture event. This would provide support for a contribution.

I found that no SAC sample set produced a top-ranking curve fit between a Bantu and KhoeSan group with date estimates similar to what is seen in the Southern Bantu admixture (amaXhosa). This is despite the predominance of KhoeSan and Bantu-like ancestry such that MALDER has sufficient power to detect this admixture [78]. The admixture dates for the amaXhosa were far older than any of the SAC dates (892 - 386 BP, 1068 - 1574 CE vs ~ 300 - 150 BP, 1660 - 1810 CE).

In all SAC sample sets the admixture involving South-East Asians are far younger than that identified in the Malagasy (VezoM; from ~706 BP, 1254 CE), indicating no support for a Malagasy contribution despite a shared South-East Asian best proxy (Igorot_mo16).

Detecting ancient admixtures in the SAC should be possible as seen here in the amaXhosa and Vezo. As a particular example, I identified a GBR - Juhoansi_sc12 admixture in the amaXhosa which predates the Bantu - KhoeSan admixture. Furthermore, I detected an ancient Dutch - Juhoansi_sc12 event ~95±29 generations ago (1914 - 3596 BP; 64 CE - 1636 BCE; 2/30 iterations) in the Nama (NAMA_SA) (results not plotted). Both of these events may reflect the back-to-Africa admixture found in East Africans and some KhoeSan groups [28], [93], [118], alternatively the GBR admixture in the amaXhosa may reflect more recent admixture [see 12]. No other ancient signals were retrieved in any other sample set.

6.4 Discussion

The term "Coloured" has been used to blanket over a diverse group of people characterised by mixed ancestry. This underplays possibly important genetic distinctions between groups with regards to ancestral contribution and post-admixture developments. I show here that the historic expansion of "Coloured" communities beyond the former Cape Colony has resulted in geographic genetic structure. Furthermore, the Cape Malay and Baster sub-identities appear to

have some distinction to the general "Coloured" identity while the Griekwa do not despite their politicised history.

6.4.1 Geographic Structure Reflects the Expanding Frontier

The development of the Colonial frontier was in many ways driven by people of mixed heritage and reciprocally, the expansion had genetic consequences for these communities [21], [142], [157]. To date the discussion of geographic structure however has been restricted to the extent of the former Cape Colony as the history of the "Coloured" communities is better recorded there [19], [99], [110], [152]. I show that the inclusion of post-Cape Colony settlements demonstrates the presence of clear geographic trends in distinctions across "Coloured" communities.

"Coloured" groups from the Cape Town area, where the settler colony was established in 1652, are the most similar to each other by the principal component distances for both *a priori* sample sets and individual clustering (Figure 6.10 - 6.11), by F_{ST} estimates (Figure 6.13), and by the diversity of ancestral contributions identified through MALDER (Figure 6.19). Distance from the former Cape Colony predicts a decline in South Asian ancestry (Figure 6.16) and the Cape Town region sample sets demonstrate elevated South-East Asian ancestry and the lowest KhoeSan ancestry of all groups (Figure 6.15). This pattern likely reflects the geographic structure of access to slaves during the frontier era, resulting in a bias of slaves over KhoeSan as labourers ($\sim 7:1$) on farms in the Cape Town and neighbouring areas (Drakenstein, Wellington, Stellenbosch) [16], [21], [121].

I found a significant trend of increase in non-KhoeSan Africa ancestry away from Cape Town but this does not appear related to Southern Bantu admixture (e.g. amaXhosa, Basotho). MALDER dates found no evidence in the SAC for a pre-Cape Colony KhoeSan - Bantu admixture which would be associated with the arrival of the Bantu to Southern Africa [102], [125], [177].

This result would suggest that the Bantu ancestry is slave-related even in groups as far North as Barberton. During interviews with the Barberton (BRT)

and Pietersburg/Polokwane (PTG) communities, participants frequently indicated that their parents or grandparents had migrated from the coastal regions. Recent migration would explain the similarity and co-clustering with the East London (EL) and Eastern Cape (EC) SAC sample sets with the former and support the likelihood of shared slave-related Bantu-like ancestry.

European contributions seems relatively stable across SAC groups considering the variability in any particular SAC group. Given the history of settlement in Southern Africa, a common European source for the SAC seems likely [21], [157] but a shared timing of admixture seems unlikely as emigration was continuous and several mass migration events are reported [103], [142], [156]. Indeed, the correlation of increasing recency of admixture with distance from Cape Town challenges the notion of a common admixture origin.

The Griekwa, Cape Malay, Baster as well as "Coloured" groups from the Cape Colony and further North share admixture dates which mostly pre-date the emancipation of the slaves c.1838 [16], [21]. This likely reflects shared admixture history associated with the settlement at Cape Town and thus groups beyond the Cape Town area, e.g. the Basters, may have retained their early admixture signal.

The "Coloured" groups from beyond the 1795 Cape Colony boundary (largely dendrogram branch c; Figure 6.8) and many of the KhoeSan communities produced MALDER dates overlapping with the emancipation of the slaves, suggesting a later influence possibly linked to the emigration events from the Cape. For the KhoeSan there appear to be only a few, discrete admixture events which are shared closely among individuals within sample sets. This is indicated by the young dates and the low variability in bootstrap estimates despite the large sample sizes (7 - 42 individuals per group). Thus, while ancestry proportions are variable (Figure 6.15), this may be a consequence of the recency of admixture rather than multiple contributions. Such results emphasise that ancestry proportions are not necessarily a useful characteristic in assigning individuals to populations, particularly for recently admixed groups.

6.4.2 Ethno-racial Affinities Share Genetic Characteristics

I found evidence for common genetic characteristics shared by sample sets with shared *a priori* ethno-racial affinities. The distinctions, however, were not easily made using mixing proportions alone (e.g. PCA positions and ADMIXTURE proportions).

The high proportion of KhoeSan ancestry in the "Coloured" and the state of admixture within the KhoeSan reflects the common history between the two groups. However, the KhoeSan groups cluster together and apart from the "Coloured" and other SAC despite their demonstrated recent admixture. The clustering of the KhoeSan based on *a priori* sample sets (branches e and f) correspond to a geographic Northern (e) and Southern (f) division (Figure 6.8). Individual-based clustering too shows common co-clustering of the related ≠Khomani and Karretjie, and the Nama groups (Figure 6.10). This reflects the influence of long-standing genetic divergences between the predominant KhoeSan ancestry in these groups [19], [94], [108]. The apparent genetic distinction between KhoeSan and "Coloured" suggests that there may be some genetic "watershed" along the line of accepting or inheriting the KhoeSan identity which separates the "Coloured" datasets (from this study and [19], [99]) from the KhoeSan datasets (from [19], [99], [113]). During my interviews for this study, participants were openly asked if they identified as KhoeSan and of 150 interviews only three did, and of these cases it was used in addition to "Coloured", "Coloured" being the first choice indicated based on historic ties to the identity. The discrete admixture dates identified for the KhoeSan sample sets suggests that recent admixture may not be a prominent feature of the identity despite high inter-individual variation and ADMIXTURE profiles which are often difficult to distinguish from the non-KhoeSan SAC.

In contrast to the KhoeSan - non-KhoeSan SAC distinction, there was no clear separation of the Griekwa from the 'Coloured'. The Griekwa have a long and well recorded history which features many aspects typically considered markers of an 'ethnic identity' [18], including notions of genealogical links to specific

ancestors, cultural essentialism and a collection of shared historic events. Clustering did not fully support a genetic similarity among the Griekwa, and geographic proximity seemed to drive co-clustering with neighbouring "Coloured" groups for some Griekwa (Figure 6.9).

The Griekwa communities in the Western Cape (Vredendal, VRE) and Northern Cape province (Upington, UTN), may be the descendants of the earliest Griekwa setters to the region and would have been relatively isolated considering the sparcity of the area north of the Boland. The establishment of these communities predates the emancipation of the slaves [142]. If gene flow were to occur, it would likely happen with the neighbouring communities. This could explain the closer clustering of the GRIEKWA_UTN to the Nama than to other Griekwa or "Coloured" communities, and the clustering of the Vredendal Griekwa and "Coloured" (Figure 6.9). However, the similarity may also reflect a common origin with subsequent contributions from other sources.

Griekwa elsewhere in the country would be descendants of those who migrated eastward [142]. The migration would have resulted in further generations of interaction with non-Griekwa communities as has been recorded in their history [18], [142], [156]. These interactions included collaboration and conflict with other 'baster' communities, e.g. Korannas and 'Bushman', and Southern Bantu, e.g. Batlhapin and Barolong [142], [156]. I however, found no clear evidence for admixture with the Southern Bantu for any Griekwa communities, suggesting little gene flow and possible endogamy within the Griekwa.

The coastal Griekwa communities (CPT and KNY) clustered with the groups from branch (c) of the *a priori* sample set PCA dendrogram (Figure 6.8). These communities are descendants of a recent back-migration from Kokstad, thus it is expected that they should be most similar genetically to the "Coloured" groups from the eastern interior and would have had less time to admix since resettling. In 1918, Griekwa Andries Le Fleur convinced a following to abandon the recessing Kokstad in favour of establishing themselves in the former Cape Colony where

land was to be acquired [18], [142]. The emigrants founded the coastal groups in Plettenberg and Cape Town, and likely Knysna too.

The "Malay", and more specifically the Cape Malay, are referred to as evidently distinct from other slave communities from the earliest mentions [16], [112], [161]. There are indeed a few clear distinctions from the other Cape "Coloured" groups suggesting some genetic distinction but, again, gene flow between neighbouring communities seems likely as ADMIXTURE profiles are similar to other Cape Town "Coloured" sample sets. The dates for admixture events best fit with the Dusun from Borneo and Ju|hoansi were suggestively younger than the events involving Dutch - Ju|hoansi (Figure 6.19). In contrast, admixture in the Wellington and District six sample sets suggests an early Igorot (Philippines) - Ju|hoansi admixture followed by a combination of Mozambican, Dutch, Ju|hoansi. While still tentative, this suggests a qualitative difference between the mixing sources.

Islam is an important feature of the purported longstanding distinction of the Cape Malay from other "Coloured" groups. This is in addition to reports of higher levels of literacy, and the retention of artisanal skills (needlework, fishing etc) [112], [153]. Cultural and religious distinctions in other populations have led to endogamy and the development of genetic 'uniqueness' [218], [253], [254], this is not a clear case for the Cape Malay.

Islam at the Cape was founded by Shaikh Yusuf, brother to the King of Gowa, Macassar [112]. Thus the arrival of South-East Asian ancestry is expected to be associated with Islam. In later decades, Islam was promoted as an alternative to Christianity for 'free blacks'; Christianity being associated with European oppressors [15]. The promotion of Islam drew many to adopt the religion [15], [16], conceivably creating a social conduit for the initial gene flow between communities. The overall similarity in regional contributions (European, Bantu, KhoeSan, South-East Asian and South Asian) seen between the Cape Malay and other Cape Town communities may reflect this.

The Baster community of Rehoboth, Namibia show a much clearer signal for endogamy. The sample set is particularly distinct genetically from other "Coloured"

and KhoeSan groups despite a very similar ADMIXTURE profiles across the SAC. The sample sets has low variability in ancestry proportions including low Bantu ancestry and the highest consistent proportions of European descent (Figure 6.15). Along with the very punctuated admixture date which is notably older than most of the KhoeSan admixtures, the low variation in proportions suggest an early pulse admixture for the Baster which has been maintained since.

Nurse, Jenkins, Africa, and Stellmacher [157] suggested there has been little introgression from neighbouring Herero and Nama, and our results suggest the same. The proportions of ancestry are suggestive of a near equal split between European and KhoeSan ancestry (~30% each) in line with earlier sero-genetic proportions of Nurse, Jenkins, Africa, and Stellmacher [157]. To maintain the near equal KhoeSan - European ancestry ratio, any KhoeSan admixture would require further European immigration, however there is little historic evidence of European immigration into the town of Rehoboth [158].

Researchers have suggested that there was reluctance to assimilate even non-Baster "Coloured" individuals [157]. This apparent endogamy may reflect social taboos on inter-racial marriage present in the Afrikaner Dutch reform church. Such views may have been carried over to the "Coloured" NGK in one form or other, resulting in assortative mating practises which persisted with the migration northward. Following the arrival of the French Huguenots in 1688, European women pressured the Dutch Reformed church (NGK) to be more restrictive on inter-racial marriages to prevent their husbands from taking slave women [21], [22].

Alternatively, there may have been admixture with KhoeSan groups and drift may have lowered non-KhoeSan and European ancestry. Nurse, Jenkins, Africa, and Stellmacher [157] proposed that the large elevations in Duffy allele frequencies and the K allele in the Kell system, well beyond the KhoeSan and European levels, were due to drift and possibly through inbreeding. KhoeSan north of the Cape Colony, such as !Xun and Ju|hoan, have notably reduced Eurasian ancestry [99] and geographic distances between settlements increase further north as a consequence of the region's aridity and population sparsity. This would make gene flow more

infrequent and inbreeding more common. Historic literature supports the possibility of high inbreeding. By 1981, the Baster community had reached over 18,000 individuals from the initial ~800 in 1876 [158]. Each of the founding families was approximately ten members strong [158], suggesting a sizable contribution from each family at the onset. Furthermore, the birth rates in the early settlement were reportedly high despite high war casualties during the conflict between the Baster and surrounding communities (Damara, Oorlam, Afrikaner (KhoeSan creole) and Nama) in the 1800s. In the initial nine years of the settlement war claimed ~20% of all deaths, yet the birth rate still exceeded the death rate [158].

6.4.3 Pre-Cape Colony Admixture

Discussion of the admixture history of the "Coloured" communities have been exclusive to the Cape Colony period, overlooking the possibility of admixture pre-dating this period. In this chapter I have added further resolution to this possibility.

MALDER did not detect a pre-Cape Colony admixture event involving a Bantu - KhoeSan combination in any of the SAC. This suggests predominant African slave ancestry rather than Southern Bantu. Indeed, the amaXhosa were first encountered by frontiersman in the early 1700s near the Kei river and only from the 19th century onward were there migrants to the Colony region as the mineral revolution began [103], [104]. Furthermore, on the expanding frontier Bantu-speakers are seldom recorded as labourers and servants suggesting that assimilation of KhoeSan was more frequent [21], [121], [255]. This may reflect that KhoeSan were more readily disposed of land and autonomy or that the settlers would form alliances along existing amaXhosa - KhoeSan tensions during the fronteir wars [103].

Similarly, MALDER produced no ancient date estimates for an Africa - Asia admixture which would otherwise have suggested a Malagasy contribution. As discussed in Chapter 5, as much as 1/3 of the earliest slaves were Malagasy and should thus be well represented in the genome of the SAC. ADMIXTURE profiles of the "Coloured", Cape Malay and Griekwa support a possible South-East Asian source but do not clearly suggest Malagasy contributions.

It seems increasingly unlikely that a Malagasy contribution is widespread considering that the signal is absent from both a comprehensively sampled area of Cape Town (733 individuals; Chapter 5) and here too from a geographically comprehensive sample. A signal for Mozambican ancestry persists in the "Coloured" from District six and from Kokstad which indicates that African slaves have left a genomic signal in the SAC. Indeed most "Coloured" sample sets have clear evidence of non-KhoeSan African ancestry as indicated in the ADMIXTURE profiles. It raises an interesting question on the fate of the Malagasy slaves at the Cape Colony and perhaps the reliability of MALDER for detecting admixture in complex scenarios [see 233].

Pre-historic admixture between the Ju|hoansi and a European source was detectable in the Nama and amaXhosa and may reflect the known back-to-Africa migration which has influenced the genome of some Southern African KhoeSan groups [19], [28], [93]. However, the KhoeSan SAC in this analysis, overlooking the Nama, do not show signals associated with the same pre-Colonial admixture (Figure 6.19). My results are similar to that found by Busby, Band, Si Le, *et al.* [177] where the Karretjie, ≠Khomani (MALDER and GLOBETROTTER) and Nama (only GLOBETROTTER) produced a single date estimate for a recent event while the !Xun and Ju|hoansi produced a single admixture date for an ancient admixture circa 30 - 40 generations ago. The inconsistent detection of the signal in the Nama with MALDER here and in [108], [177], suggests that MALDER may not reliably detect ancient signals when a more recent event exists, even when the separation between the events exceeds 20 generations. Considering that the extent of the back-to-Africa introgression in Southern Africa is still incompletely characterised, the absence of the signal may reflect a genuine absence from the SAC KhoeSan and by extension, the absence of a Southern Bantu and Malagasy signal may too be genuine. Further resolution will require improved statistical techniques for dealing with multi-way admixture events.

6.5 Conclusion

From this investigation of a geographically representative sample of the South African "Coloured" communities and their parallels among the KhoeSan, I have shown that the region is structured with regards to ancestral contributions. The structure reflects the history of expansion of the frontier. Specifically, the greater diversity of contributions in the former Cape Colony reflects greater slave contributions while beyond the Cape Colony, KhoeSan contributions dominate. The ethno-racial affinities in the SAC have some bearing on genetic similarities and uniqueness as indicated by the Baster endogamy. This is not uniformly true as the Griekwa from geographically separate communities more closely resemble their neighbouring non-Griekwa communities than their identity-kin elsewhere. The Cape Malay, similarly, are almost indistinguishable from other Cape Town "Coloured" groups but with possible qualitative differences in the source populations which are not characterised in this work.

7
Concluding Remarks

"The immediate past is in many ways more difficult to describe and analyse than those times that appear to have receded into history ... writing about the immediate past [one] tend[s] to have a much greater awareness ... of ways in which things could have turned out differently" - Hermann Giliomee [21].

Oral or written history may not be available for many groups in society where a power imbalance and subjugation, presently or historically, has prevented people from passing on their own history and where few records have been kept elsewhere. The "Coloured" communities of South Africa are a perfect example of this scenario, in particular the Cape Town slave descendants. This thesis highlights the possibility of using genetic data to supplement other tools for recovering details of complex historical scenarios.

Overshadowed History

The entire history of the modern South African "Coloured" communities has been established in an atmosphere of subjugation and discrimination; from the slave arrivals and dispossession of the KhoeSan at the Cape, to the disgruntlement with the current dispensation under the African National Congress. As expected, much of their own history was recorded by outsiders and much more is potentially lost forever.

By exploring the DNA of contemporary individuals, however, I was able to shed light on some aspects of this history. Firstly, I show that the expansion of the Coloured groups beyond the former Cape Colony resulted in the establishment of geographically structured diversity. The timing of admixture events reflects historic events related to the abolishment of slavery and emancipation of the slaves as well as the economics of slave and KhoeSan ownership in the Cape Colony.

I also add a genetic aspect to the discussion of the commonly understood polities and sub-identities; the Griekwa, Cape Malay and Baster identities. Genealogical coherence is a significant aspect of the Cape Malay and Griekwa identity. In both cases communities understand a link to an ancestral group in which their identity had established its roots. I find support for such connections in both cases but also highlight the influences of geography and endogamy on genetic signals. The Griekwa groups were not always genetically closest to each other. For some groups their nearest genetic kin were geographically proximate non-Griekwa communities. With the Baster and Cape Malay communities I detect some support for endogamy and discuss how religious, geographic and socio-economic influences have maintained such mating patterns. These dynamics to the identity landscape within the "Coloured" are important when considering the use of ethnicity in medical and anthropological work [187], [256].

Secondly, I found evidence for the presence of historic admixture which is often unaccounted for in even recent genetic work. The results indicated a South-East Asian and European contribution pre-dating the Dutch settlement. The work thus showcases the importance of shipwrecks and pre-settlement contact between the indigenous communities and sailors. Admixture within the Southern African region is by no means restricted to the post-1652 era. The results add another layer of information supporting the ongoing developments in population genomics which demonstrate that admixture has been occurring continuously throughout human history, often pre-dating what are otherwise assumed 'important dates' [32], [93], [227].

It is quite exciting to find a result contrary to historic accounts though interpretation then needs much more thought. Indeed, I found that the Malagasy ancestry indicated by literature was not detectable which is difficult to reconcile and may indicate that some sects of slave society did not contribute to subsequent generations. However, the results in this thesis, overall, do something more remarkable - as remarkable as the work from many other research projects. The results concur with numerous other sources of information. A diversity of ancestries was detected in the "Coloured" people, many of which were overlooked as incidental. Filtering down the results provided support for eleven possible source populations based on SNP haplotypes. These results were remarkably congruent with several historic accounts. Finding congruence supports the reliability of the analyses and strengthens support for information that is, at times, simply assumed true.

Caveats for the Use of Appropriate Proxies

The results obtained from analyses such as CHROMOPAINTER and the NNLS are influenced by the available sources in the dataset and there is thus a risk of mis-interpreting the output. This is especially so when the history of either the source population or the study population is obscure or poorly researched. The burden is then doubly, to have a well-represented set of source populations but to also understand the ancestry composition of these populations. In most cases this will be undesirable as large sample sizes and representation incurs a time cost particularly for such iterative and multi-step analyses as applied in Chapter 4. The time taken to become familiar with the study populations is another restraint. Often, sources are included blindly, and peculiarities identified post-hoc [see 88], [97], [257]. Though this reduces user introduced biases in *a priori* selecting likely sources, it does leave interpretation prone to “creative” conclusions to explain anomalous results. A safe guard would be to perform simulations of the proposed admixture events, testing specific scenarios and alternatives for the best match. This may be amenable to repetition and estimating significance however at a substantial computational time cost.

The use of contemporary samples for inferring historic events has obvious caveats as each population has surely undergone some genetic change since their role in the history of the study population. Indeed, during historic times, the scale of global movement has led high rates of admixture across the world. As such it becomes increasingly important to screen populations used in data analysis to prevent the influence of cryptic, recent ancestry.

Conclusion

The patchiness of records may be otherwise irreparable using historic accounts alone, particularly regarding the ethnic and cultural origins of slaves and their eventual fate in the Cape Colony. Proactive investigations for archaeological clues are important and rewarding for this purpose [e.g. 149], [183]. In the case of Kootker, Mbeki, Morris, Kars, and Davies [149], isotopic analysis from burial remains were compared to known regional geochemistry to help identify possible histories of movements associated with Cape Colony burials [149]. Genetic data can be used in conjunction with such data as it directly tracks the movement and reproductive interactions of people by the process of transmission of DNA to future generations. Thus, we can use the genetic information to trace origins of people and processes that may have influenced the Cape Coloured population after arrival and to the present. The prospects of using historic DNA are even more exciting. Examining historic burials will allow us to track the changes in population structure and admixture dynamics using actual 'time slices' in the development of the "Coloured" and other communities. Many of the signals observed and conclusions arrived at in this thesis can be more decisively understood when such historic DNA data is made available.

Acknowledgements

Big data projects almost always involves large collaborations. I can't name, nor have I met, many of the people who have contributed to this project, but I extend my deepest gratitude for helping keep this project afloat. This extends to research groups which have made their data publicly available. I list some of the collaborators more directly involved in the project in table 7.1 and their contributions.

I thank Drs Miguel Gonzalez-Santos, Francesco Montinaro, Alessandro Raveane and Serena Aneli for the company and academic insight both of which has helped immensely. I thank Dr Cristian Capelli for all of his guidance and patience over the past four years.

I thank Drs Thomas Cousins, Hans Heese and Sam Challis for the invaluable guidance with navigating South Africa's social and political history during field work preparation, interviews and subsequently during the write up. Their enthusiasm for the history of South Africa's people and for the views of the people today has been a motivational force on its own.

Dr Karen Ehlers deserves a huge amount of thanks for facilitating the ethics approval in South Africa despite a plethora of delays including two bouts of nation-wide student protests. This thesis would not exist without Dr Ehlers contributions and persistence.

Mr Puseletso Lecheko and Mr Paballo Chauke assisted with field work across South Africa. Their patience in the field, willingness to work longer-than-usual hours and their welcoming, friendly and involved approach to interviews did wonders for participant interactions.

All of the community leaders and participants deserve to be acknowledged. The willingness to participate in this research amazed me. Despite their history of being exploited and oppressed, often justified on scientific terms, the participants

were welcoming, engaging and open to share their stories. The community leaders who have made sample collection and interviews possible, for which I am deeply thankful, are listed in Table 7.1.

I extend deep gratitude to Dr Maria D'Amato and Dr Eileen Hoal for allowing me access the samples and data of South African 'Coloured' people used here. Their contributions have made up a substantial part of the success of this thesis.

Drs Greger Larson and Rosalind Harding provided much needed perspective on the project at critical points in its progress. There was much valuable insight from them which I can only hope I made full use of.

I thank Dr Kirk Rockett for warmly allowing me access to his group's lab space after the Tinbergen closure.

I personally thank the Canon Collins Education and Legal Assistance Trust for the social buffering that eased the transition into the UK. I thank Dr Craig Peter for his continued moral support and interest in my development as a researcher. My dearest friends have made my time at Oxford and the Zoology department the most pleasant it could have been: Erica Aiazzi, Sneha Menon, Manar Marzouk, Cindy Santander, Andres Ojeda Laguna, Lynn Lewis-Bevan, Nhlakanipho Mkhize and Joanna Bessa.

I thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 203141/Z/16/Z) for the generation of the data. I thank the Oxford University Advanced Research Computing Service for computational resources.

Finally, I acknowledge the Boise Trust Fund, the Commonwealth Commission to the UK, the Oxford University Zoology Department and the Santander Trust for the financial support which has made this project possible.

Table 7.1: List of contributers to the work in this thesis.

Contributers	Contribution	Chapter
Ryan Joseph Daniels Oxford University	Project conception and development Project funding Project budgeting Field work planning Field work sample collection Laboratory work Data preparation Data analysis Write up Administrative work	Entire thesis
Prof Cristian Capelli Oxford University	Project development Project funding Advisory Data acquisition Administrative work	Entire thesis
Dr Francesco Montinaro Oxford University	Advisory: Coding and data analysis	4
Dr Alessandro Raveanne Oxford University	Advisory: Coding and data analysis	4
Dr Maria Eugenia D'Amato University of the Western Cape, ZA	Access to samples	6
Dr Karen Ehlers University of the Free State, ZA	Local ethical clearance	6
Dr Mahaimin Kazo University of the Western Cape, ZA	Cape Malay sample collection	6
Dr Peter Ristow University of the Western Cape, ZA	Griqua sample collection	6
Dr Sam Challis University of the Witwatersrand, ZA	Field work planning Field work sample collection	6
Mr Puseletso Lecheko University of the Witwatersrand, ZA	Field work sample collection and translator (Sotho/Xhosa/Phuthi)	6
Mr Paballo Chauke University of Cape Town, ZA	Field work sample collection and translator (Zulu/Tswana/Sotho/Venda)	6
Dr Eileen Hoal Stellenbosch University, ZA	Access to data	4,5
Dr Hans Heese Stellenbosch University, ZA	Write up Advisory: History of the Cape Coloured community	4,5
Community Leaders/Facilitators		
Mr Michael Hutton NGO Khoisan Nation Barberton	Contacting participants	6

Mrs Janie Grobler Barberton Museum	Contacting participants	6
Father Banarbas Mahikeng	Contacting participants	6
Mrs Stephanie Victor and staff Amatole Museum, King Williams Town	Contacting participants	6
Mrs Jackie Dreyer and Fatima Daniels East London, Tambookiesvlei	Contacting participants	6

References

[1] N Patterson, P Moorjani, Y Luo, *et al.*, "Ancient admixture in human history.", *Genetics*, vol. 192, no. 3, pp. 1065–1093, Nov. 2012, ISSN: 1943-2631. DOI: 10.1534/genetics.112.145037.

[2] 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes", *Nature*, vol. 491, no. 7422, pp. 56–65, Nov. 2012, ISSN: 0028-0836. DOI: 10.1038/nature11632.

[3] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing", *Nature*, vol. 467, no. 7319, pp. 1061–1073, Oct. 2010, ISSN: 0028-0836. DOI: 10.1038/nature09534.

[4] D Gurdasani, T Carstensen, F Tekola-Ayele, *et al.*, "The African Genome Variation Project shapes medical genetics in Africa", *Nature*, vol. 517, no. 7534, pp. 327–332, Dec. 2014, ISSN: 0028-0836. DOI: 10.1038/nature13997.

[5] DM Altshuler, RA Gibbs, L Peltonen, *et al.*, "Integrating common and rare genetic variation in diverse human populations", *Nature*, vol. 467, no. 7311, pp. 52–58, Sep. 2010, ISSN: 0028-0836. DOI: 10.1038/nature09298.

[6] JW Belmont, A Boudreau, SM Leal, *et al.*, "A haplotype map of the human genome", *Nature*, vol. 437, no. 7063, pp. 1299–1320, Oct. 2005, ISSN: 0028-0836. DOI: 10.1038/nature04226.

[7] ES Lander, LM Linton, B Birren, *et al.*, "Initial sequencing and analysis of the human genome", *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001, ISSN: 0028-0836. DOI: 10.1038/35057062.

[8] HM Cann, C de Toma, L Cazes, *et al.*, "A human genome diversity cell line panel.", *Science*, vol. 296, no. 5566, pp. 261–2, Apr. 2002, ISSN: 1095-9203.

[9] MR Nelson, K Bryc, KS King, *et al.*, "The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research", *American Journal of Human Genetics*, vol. 83, no. 3, pp. 347–358, 2008, ISSN: 00029297. DOI: 10.1016/j.ajhg.2008.08.005.

[10] AM Brues, *People and Races*, 1st ed. New York: Macmillan Publishing Co., 1977, p. 336, ISBN: 0-02-315670-8.

[11] PW Hedrick, *Genetics of Populations*, 3rd ed. Tempe, Arizona, USA: Jones and Bartlett Publishers, 2005, ISBN: 9780763757373.

[12] A Choudhury, M Ramsay, S Hazelhurst, *et al.*, "Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans", *Nature Communications*, vol. 8, no. 1, p. 2062, 2017, ISSN: 2041-1723. DOI: 10.1038/s41467-017-00663-9.

[13] S Mallick, H Li, M Lipson, *et al.*, "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations", *Nature*, vol. 538, no. 7624, pp. 201–206, 2016, ISSN: 0028-0836. DOI: 10.1038/nature18964. arXiv: NIHMS150003.

[14] YY Teo, X Sim, RT Ong, *et al.*, "Singapore Genome Variation Project: A haplotype map of three Southeast Asian populations", *Genome Research*, vol. 19, no. 11, pp. 2154–2162, Nov. 2009, ISSN: 1088-9051. DOI: 10.1101/gr.095000.109.

[15] M Adhikari, *Not White Enough, Not Black Enough: Racial Identity in the South African Coloured Community*. Cape Town, South Africa: Double Storey Books and Athens, 2005.

[16] H Heese, *Groep sonder grense: Die rol en status van die gemengde bevolking aan die Kaap, 1652-1795*. Bellville : Wes-Kaaplandse Instituut vir Historiese Navorsing, Universiteit van Wes-Kaapland: Protea Boekhuis, 1984, p. 89, ISBN: 978-0909075989.

[17] M Adhikari, "Contending Approaches to Coloured Identity and the History of the Coloured People of South Africa", *History Compass*, vol. 3, pp. 1–16, 2005, ISSN: 14780542. DOI: 10.1111/j.1478-0542.2005.00177.x.

[18] P Erasmus, "Vote for real people: the making of Griqua and Korana identities in Heidedal", *Anthropology Southern Africa*, vol. 33, no. 1-2, pp. 65–73, 2010, ISSN: 2332-3256. DOI: 10.1080/23323256.2010.11499994.

[19] CM Schlebusch, "Genetic variation in Khoisan-speaking populations from southern Africa", Doctor of Philosophy, University of the Witwatersrand, 2010, p. 379.

[20] W Howells, *Mankind in the Making: The Story of Human Evolution*. UK: Pelican Books, 1967.

[21] H Giliomee, *The Afrikaners: Biography of a People*. Cape Town: Tafelberg Publishers Limited, 2010, ISBN: 0-8139-2237-2.

[22] JM Greeff, FA Greeff, AS Greeff, L Rinken, DJ Welgemoed, and Y Harris, "Low nonpaternity rate in an old Afrikaner family", *Evolution and Human Behavior*, vol. 33, no. 4, pp. 268–273, 2012, ISSN: 10905138. DOI: 10.1016/j.evolhumbehav.2011.10.004.

[23] Statistics South Africa, "General household survey 2012", Statistics South Africa, Pretoria, South Africa, Tech. Rep., 2013.

[24] S Kent, "Interethnic Encounters of the First Kind: An Introduction", in *Ethnicity, Hunter-Gatherers, and the "Other": Association or Assimilation in Africa*, S Kent, Ed., 1st ed., Washington, USA: Smithsonian Institution Press, 2002, ch. 1, ISBN: 1588340600.

[25] T McCarthy and B Rubidge, *The Story of Earth and Life: A southern African perspective on a 4.6-billion-year journey*, 1st ed. Cape Town, South Africa: Struik Publishers, 2005, ISBN: 9781770071483.

[26] M Jobling, E Hollox, M Hurles, T Kivisild, and C Tyler-Smith, *Human Evolutionary Genetics*, 2nd ed. Abingdon, UK: Garland Science, Taylor & Francis Group, 2014, p. 690, ISBN: 9780815341482.

[27] J Lachance, B Vernot, CC Elbers, *et al.*, "Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers.", *Cell*, vol. 150, no. 3, pp. 457–469, Aug. 2012, ISSN: 1097-4172. DOI: 10.1016/j.cell.2012.07.009.

[28] CM Schlebusch, H Malmström, T Günther, *et al.*, "Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago", *Science*, vol. 6266, no. September, eaao6266, 2017, ISSN: 0036-8075. DOI: 10.1126/science.aao6266. arXiv: science.aao6266(2017) [10.1126].

[29] LR Berger, J Hawks, DJ de Ruiter, *et al.*, "Homo naledi, a new species of the genus Homo from the Dinaledi Chamber, South Africa", *eLife*, vol. 4, no. September, 2015, ISSN: 2050084X. DOI: 10.7554/eLife.09560.

[30] C Barras, "Genes of the undead", *New Scientist*, vol. 230, no. 3070, pp. 38–39, 2016, ISSN: 02624079. DOI: 10.1016/S0262-4079(16)30716-3.

[31] V Slon, F Mafessoni, B Vernot, *et al.*, "The genome of the offspring of a Neanderthal mother and a Denisovan father", *Nature*, vol. 561, no. 7721, pp. 113–116, Sep. 2018, ISSN: 0028-0836. DOI: 10.1038/s41586-018-0455-x.

[32] D Reich, RE Green, M Kircher, *et al.*, "Genetic history of an archaic hominin group from Denisova Cave in Siberia", *Nature*, vol. 468, no. 7327, pp. 1053–1060, 2010, ISSN: 0028-0836. DOI: 10.1038/nature09710.

[33] P Frankopan, *The Silk Roads: A New History of the World*, 1st ed. London, UK: Bloomsbury Publishing, 2015, ISBN: 9781408839966.

[34] B Lewis, *The Middle East*, 6th ed. London: Phoenix Press, 2004, ISBN: 1842121391.

[35] M Vink, "The World 's Oldest Trade: Dutch Slavery and Slave Trade in the Indian Ocean in the Seventeenth Century", *Journal of World History*, vol. 14, no. 2, pp. 131–177, 2003. DOI: 10.1353/jwh.2003.0026.

[36] N Ostler, *Empires of the Word: A Language History of the World*. London, UK: Harper Collins Publishers, 2005, ISBN: 0007118708.

[37] RA Goldsby, *Race and Races*, 1st ed. New York, USA: Macmillan Publishing Co., 1971, p. 158, ISBN: 0-02-344310-3.

[38] NA Rosenberg, JK Pritchard, JL Weber, *et al.*, "Genetic structure of human populations.", *Science*, vol. 298, no. 5602, pp. 2381–2385, Dec. 2002, ISSN: 1095-9203. DOI: 10.1126/science.1078311.

[39] RA Brown and GJ Armelagos, "Apportionment of Racial Diversity: A Review", *Evolutionary Anthropology*, vol. 40, pp. 34–40, 2011.

[40] G Barbujani and V Colonna, "Human genome diversity: Frequently asked questions", *Trends in Genetics*, vol. 26, no. 7, pp. 285–295, 2010, ISSN: 01689525. DOI: 10.1016/j.tig.2010.04.002.

[41] R Owen, "Karl Landsteiner and the first human marker locus", *Genetics*, vol. 155, no. 3, pp. 995–998, 2000, ISSN: 00166731. DOI: 10.1590/S0074-02762003000400004.

[42] CS Coon, *The Living Races of Man*. New York: Random House, 1973, ISBN: 0394433726.

[43] LL Cavalli-sforza and AWF Edwards, "Analysis of human evolution", *Genetics Today*, vol. 3, pp. 923–933, 1964.

[44] A Klung, "Rosalind Franklin and the Discovery of the Structure of DNA", *Nature*, vol. 219, pp. 808–844, 1968.

[45] IR Lehman, "Discovery of DNA polymerase", *Journal of Biological Chemistry*, vol. 278, no. 37, pp. 34 733–34 738, 2003, ISSN: 00219258. DOI: 10.1074/jbc.X300002200.

[46] LL Cavalli-sforza, *Genes, Peoples and Languages*. London, UK: Allen Lane The Penguin Press, 2000, p. 227.

[47] JK Pickrell and D Reich, "Toward a new history and geography of human genes informed by ancient DNA", *Trends in Genetics*, vol. 30, no. 9, pp. 377–389, 2014, ISSN: 13624555. DOI: 10.1016/j.tig.2014.07.007.

[48] CJ Venter, MD Adams, EW Myers, *et al.*, "The Sequence of the Human Genome", *Science*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001, ISSN: 0036-8075. DOI: 10.1126/science.1058040.

[49] DL Hartl and EW Jones, *Genetics - Principles and Analysis*, 4th ed. London, UK: Jones and Bartlett Publishers, 1998, ISBN: 0-7637-0489-X.

[50] DJ Depew and BH Weber, "The Fate of Darwinism: Evolution After the Modern Synthesis", *Biological Theory*, vol. 6, no. 1, pp. 89–102, 2011, ISSN: 1555-5542. DOI: 10.1007/s13752-011-0007-1.

[51] Y Liu and SC West, "Timeline: Happy Hollidays: 40th anniversary of the Holliday junction", *Nature Reviews Molecular Cell Biology*, vol. 5, no. 11, pp. 937–944, Nov. 2004, ISSN: 1471-0072. DOI: 10.1038/nrm1502.

[52] CC Gillispie, "Lamarck and Darwin in the history of science", *American Scientist*, vol. 46, no. 4, pp. 388–409, 1958.

[53] MA Jobling and C Tyler-Smith, "The human Y chromosome: an evolutionary marker comes of age", *Nature Reviews Genetics*, vol. 4, no. 8, pp. 598–612, Aug. 2003, ISSN: 1471-0056. DOI: 10.1038/nrg1124.

[54] WS Moore, "Inferring phylogenies from mtDNA variation: Mitochondrial-gene trees versus nuclear-gene trees", *Evolution*, vol. 49, no. 4, pp. 718–726, Apr. 1995.

[55] P Kusuma, MP Cox, D Pierron, *et al.*, "Mitochondrial DNA and the Y chromosome suggest the settlement of Madagascar by Indonesian sea nomad populations", *BMC Genomics*, vol. 16, no. 1, p. 191, Dec. 2015, ISSN: 1471-2164. DOI: 10.1186/s12864-015-1394-7.

[56] C de Filippo, C Barbieri, M Whitten, *et al.*, "Y-Chromosomal Variation in Sub-Saharan Africa: Insights Into the History of Niger-Congo Groups", *Molecular Biology and Evolution*, vol. 28, no. 3, pp. 1255–1269, Mar. 2011, ISSN: 0737-4038. DOI: 10.1093/molbev/msq312.

[57] IV Ovchinnikov, A Götherström, GP Romanova, VM Kharitonov, K Lidén, and W Goodwin, "Molecular analysis of Neanderthal DNA from the northern Caucasus", *Nature*, vol. 404, no. 6777, pp. 490–493, Mar. 2000, ISSN: 00280836. DOI: 10.1038/35006625.

[58] LL Cavalli-sforza, P Menozzi, and A Piazza, *The history and geography of human genes*. New Jersey, USA: Princeton University Press, 1994, ISBN: 9780691029054.

[59] L Pagani, S Schiffels, D Gurdasani, *et al.*, "Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians", *The American Journal of Human Genetics*, vol. 96, no. 6, pp. 986–991, 2015, ISSN: 00029297. DOI: 10.1016/j.ajhg.2015.04.019.

[60] M Ingman, H Kaessmann, S Pääbo, and U Gyllensten, "Mitochondrial genome variation and the origin of modern humans.", *Nature*, vol. 408, no. 6813, pp. 708–13, Dec. 2000, ISSN: 0028-0836. DOI: 10.1038/35047064.

[61] JM Greeff, "Deconstructing Jaco: genetic heritage of an Afrikaner.", *Annals of Human Genetics*, vol. 71, no. 5, pp. 674–688, Sep. 2007, ISSN: 0003-4800. DOI: 10.1111/j.1469-1809.2007.00363.x.

[62] GBJ Busby, F Brisighelli, P Sanchez-Diz, *et al.*, "The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269", *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1730, pp. 884–892, Mar. 2012, ISSN: 0962-8452. DOI: 10.1098/rspb.2011.1044.

[63] L Quintana-Murci, C Harmant, H Quach, *et al.*, "Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture.", *American Journal of Human Genetics*, vol. 86, no. 4, pp. 611–20, 2010, ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2010.02.014.

[64] SJ Marks, F Montinaro, H Levy, *et al.*, "Static and Moving Frontiers: The Genetic Landscape of Southern African Bantu-Speaking Populations", *Molecular Biology and Evolution*, vol. 32, no. 1, pp. 29–43, Jan. 2015, ISSN: 0737-4038. DOI: 10.1093/molbev/msu263.

[65] E Heyer, R Chaix, S Pavard, and F Austerlitz, *Sex-specific demographic behaviours that shape human genomic variation*, Feb. 2012. DOI: 10.1111/j.1365-294X.2011.05406.x.

[66] P Francalacci and D Sanna, "History and geography of human Y-chromosome in Europe: a SNP perspective.", *Journal of anthropological sciences*, vol. 86, pp. 59–89, Jan. 2008, ISSN: 1827-4765.

[67] LS Emery, KM Magnaye, AW Bigham, JM Akey, and MJ Bamshad, "Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup", *American Journal of Human Genetics*, vol. 96, no. 2, pp. 183–193, 2015, ISSN: 15376605. DOI: 10.1016/j.ajhg.2014.12.015.

[68] A Auton, "The Estimation of Recombination Rates from Population Genetic Data", Doctor of Philosophy, University of Oxford, Oxford, 2007, p. 202.

[69] M Modesti and R Kanaar, "Homologous recombination: from model organisms to human disease.", *Genome biology*, vol. 2, no. 5, pp. 1–5, Apr. 2001, ISSN: 1474-760X (Electronic). DOI: 10.1186/gb-2001-2-5-reviews1014.

[70] CA Buerkle and C Lexer, "Admixture as the basis for genetic mapping", *Trends in Ecology and Evolution*, vol. 23, no. 12, pp. 686–694, 2008, ISSN: 01695347. DOI: 10.1016/j.tree.2008.07.008.

[71] A Kong, DF Gudbjartsson, J Sainz, GM Jonsdottir, and SA Gudjonsson, "A high-resolution recombination map of the human genome", *Nature Genetics*, vol. 31, pp. 241–247, 2002.

[72] AJ Jeffreys and R Neumann, "Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot", *Nature Genetics*, vol. 31, no. 3, pp. 267–271, Jul. 2002, ISSN: 1061-4036. DOI: 10.1038/ng910.

[73] KP Donnelly, "The probability that related individuals share some section of genome identical by descent", *Theoretical Population Biology*, vol. 23, no. 1, pp. 34–63, Feb. 1983, ISSN: 00405809.

[74] CD Huff, DJ Witherspoon, TS Simonson, *et al.*, "Maximum-likelihood estimation of recent shared ancestry (ERSA)", *Genome Research*, vol. 21, no. 5, pp. 768–774, 2011, ISSN: 10889051. DOI: 10.1101/gr.115972.110.

[75] K Prüfer, F Racimo, N Patterson, *et al.*, "The complete genome sequence of a Neanderthal from the Altai Mountains", *Nature*, vol. 505, no. 7481, pp. 43–49, Jan. 2014, ISSN: 0028-0836. DOI: 10.1038/nature12886.

[76] AM Bowcock, A Ruiz-Linares, J Tomfohrde, E Minch, JR Kidd, and LL Cavalli-Sforza, "High resolution of human evolutionary trees with polymorphic microsatellites.", *Nature*, vol. 368, no. 6470, pp. 455–457, 1994, ISSN: 0028-0836. DOI: 10.1038/368455a0.

[77] SA Tishkoff and SM Williams, "Genetic analysis of African populations: human evolution and complex disease.", *Nature reviews. Genetics*, vol. 3, no. 8, pp. 611–621, Aug. 2002, ISSN: 1471-0056. DOI: 10.1038/nrg865.

[78] PR Loh, M Lipson, N Patterson, *et al.*, "Inferring admixture histories of human populations using linkage disequilibrium", *Genetics*, vol. 193, no. 4, pp. 1233–1254, 2013, ISSN: 19432631. DOI: 10.1534/genetics.112.147330. arXiv: arXiv:1211.0251v2.

[79] R Chakraborty and KM Weiss, "Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci.", *Proceedings of the National Academy of Sciences*, vol. 85, no. 23, pp. 9119–9123, Dec. 1988, ISSN: 0027-8424. DOI: 10.1073/pnas.85.23.9119.

[80] WPC Stemmer, "DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 22, pp. 10 747–10 751, 1994, ISSN: 00278424. DOI: 10.1073/pnas.91.22.10747.

[81] G Coop and M Przeworski, "An evolutionary view of human recombination", *Nature Reviews Genetics*, vol. 8, no. 1, pp. 23–34, 2007, ISSN: 14710056. DOI: 10.1038/nrg1947.

[82] G Hellenthal, GBJ Busby, G Band, *et al.*, "A Genetic Atlas of Human", *Science*, vol. 343, no. February, pp. 747–751, 2014, ISSN: 0036-8075. DOI: 10.1126/science.1243518.

[83] DJ Lawson, G Hellenthal, S Myers, and D Falush, "Inference of population structure using dense haplotype data", *PLoS Genetics*, vol. 8, no. 1, e1002453, 2012, ISSN: 15537390. DOI: 10.1371/journal.pgen.1002453.

[84] N Li and M Stephens, "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data", *Genetics*, vol. 165, pp. 2213–2233, 2003.

[85] D Reich, K Thangaraj, N Patterson, AL Price, and L Singh, "Reconstructing Indian population history", *Nature*, vol. 461, no. 7263, pp. 489–494, Sep. 2009, ISSN: 0028-0836. DOI: 10.1038/nature08365.

[86] P Kusuma, N Brucato, MP Cox, *et al.*, "Contrasting Linguistic and Genetic Origins of the Asian Source Populations of Malagasy", *Scientific Reports*, vol. 6, p. 26 066, 2016, ISSN: 2045-2322. DOI: 10.1038/srep26066.

[87] ER Chimusa, J Defo, PK Thami, *et al.*, "Dating admixture events is unsolved problem in multi-way admixed populations", *Briefings In Bioinformatics*, vol. 45, no. 2, pp. 846–860, Jan. 2018, ISSN: 1467-5463. DOI: 10.1093/bib/bby112. arXiv: 1611.06654.

[88] S Leslie, B Winney, G Hellenthal, *et al.*, "The fine-scale genetic structure of the British population", *Nature*, vol. 519, no. 7543, pp. 309–314, 2015, ISSN: 0028-0836. DOI: 10.1038/nature14230.

[89] LP Wong, RTH Ong, Wt Poh, *et al.*, "Deep Whole-Genome Sequencing of 100 Southeast Asian Malays", *The American Journal of Human Genetics*, vol. 92, no. 1, pp. 52–66, Jan. 2013, ISSN: 00029297. DOI: 10.1016/j.ajhg.2012.12.005.

[90] JZ Li, DM Absher, H Tang, *et al.*, "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation", *Science*, vol. 319, no. 5866, pp. 1100–1104, Feb. 2008, ISSN: 0036-8075. DOI: 10.1126/science.1153717.

[91] JD Wall and MF Hammer, "Archaic admixture in the human genome.", *Current opinion in genetics & development*, vol. 16, no. 6, pp. 606–10, 2006, ISSN: 0959-437X. DOI: 10.1016/j.gde.2006.09.006.

[92] LC Francioli, A Menelaou, SL Pulit, *et al.*, "Whole-genome sequence variation, population structure and demographic history of the Dutch population", *Nature genetics*, vol. 46, no. 8, pp. 818–25, Jun. 2014, ISSN: 1061-4036. DOI: 10.1038/ng.3021.

[93] JK Pickrell, N Patterson, C Barbieri, *et al.*, "The genetic prehistory of southern Africa", *Nature Communications*, vol. 3, no. 1143, p. 1143, Oct. 2012, ISSN: 2041-1723. DOI: 10.1038/ncomms2140.

[94] P Skoglund, JC Thompson, ME Prendergast, *et al.*, "Reconstructing Prehistoric African Population Structure", *Cell*, vol. 171, no. 39, 59–71.e21, 2017, ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.08.049.

[95] ES Lander, DE Reich, S Gnerre, NJ Patterson, and DJ Richter, "Genetic evidence for complex speciation of humans and chimpanzees", *Nature*, vol. 441, no. 7097, pp. 1103–1108, 2006, ISSN: 0028-0836. DOI: 10.1038/nature04789.

[96] P Moorjani, K Thangaraj, N Patterson, *et al.*, "Genetic evidence for recent population mixture in India", *American Journal of Human Genetics*, vol. 93, no. 3, pp. 422–438, 2013, ISSN: 00029297. DOI: 10.1016/j.ajhg.2013.07.006.

[97] F Montinaro, GB Busby, VL Pascali, S Myers, G Hellenthal, and C Capelli, "Unravelling the hidden ancestry of American admixed populations", *Nature Communications*, vol. 6, no. 1, p. 6596, Dec. 2015, ISSN: 2041-1723. DOI: 10.1038/ncomms7596.

[98] Juliana Alves-Silva, MdS Santo, PEM Guimaraes, *et al.*, "The Ancestry of Brazilian mtDNA Lineages Juliana", *American journal of human genetics*, vol. 67, pp. 444–461, 2000.

[99] DC Petersen, O Libiger, Ea Tindall, *et al.*, "Complex patterns of genomic admixture within southern Africa.", *PLoS genetics*, vol. 9, no. 3, e1003309, Mar. 2013, ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003309.

[100] K Bryc, A Auton, MR Nelson, *et al.*, "Genome-wide patterns of population structure and admixture in West Africans and African Americans.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 2, pp. 786–791, 2010, ISSN: 0027-8424. DOI: 10.1073/pnas.0909559107.

[101] AM Shah, R Tamang, P Moorjani, *et al.*, "Indian siddis: African descendants with Indian admixture", *American Journal of Human Genetics*, vol. 89, no. 1, pp. 154–161, 2011, ISSN: 00029297. DOI: 10.1016/j.ajhg.2011.05.030.

[102] L Barham and P Mitchell, *The First Africans: African Archaeology from the earliest toolmakers to most recent foragers*. Cambridge: Cambridge University Press, 2008, p. 623, ISBN: 9780521612654. DOI: 10.1017/CBO9780511817830.

[103] I Berger, *South Africa in World History*, ser. The new Oxford world history. Oxford, UK.: Oxford University Press, 2009, p. 208, ISBN: 978-0-19-515754-3.

[104] L Thompson, *The History of South Africa*, 3rd ed. Yale Nota Bene Books, 2000, ISBN: 0300087764.

[105] Statistics South Africa, "Household Service Delivery Statistics", Statistics South Africa, Pretoria, South Africa, Tech. Rep., 2012.

[106] M Bolaane, "San Cross-border Cultural heritage and identity in Botswana, Namibia and South Africa", *African Study Monographs*, vol. 35, no. 1, pp. 41–64, 2014.

[107] C Uren, M Kim, AR Martin, *et al.*, "Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries.", *Genetics*, vol. 204, no. 1, genetics.116.187369, Sep. 2016, ISSN: 0016-6731. DOI: 10.1534/genetics.116.187369.

[108] F Montinaro, GBJ Busby, M Gonzalez-santos, *et al.*, "Complex ancient genetic structure and cultural transitions in southern African populations.", *Genetics*, vol. 205, no. 1, pp. 303–316, Mar. 2017, ISSN: 19432631. DOI: 10.1534/genetics.116.189209.

[109] V Bicword-smith, "Slavery, emancipation and the question of coloured identity, with particular attention to Cape town, 1875-1910", in *Societies of Southern Africa in the 19th and 20th Centuries: postgraduate seminar papers*, ser. Societies of southern Africa in the 19th and 20th centuries: collected seminar papers, Institute of Commonwealth Studies, 1993, pp. 17–40, ISBN: 00760773.

[110] E de Wit, W Delport, CE Rugamika, *et al.*, "Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape.", *Human Genetics*, vol. 128, no. 2, pp. 145–153, Aug. 2010, ISSN: 1432-1203. DOI: 10.1007/s00439-010-0836-1.

[111] AJ Christopher, "Delineating the nation: South African censuses 1865-2007", *Political Geography*, vol. 28, no. 2, pp. 101–109, 2009, ISSN: 09626298. DOI: 10.1016/j.polgeo.2008.12.003.

[112] EC Mandivenga, "The Cape Muslims and the Indian Muslims of South Africa: A Comparative Analysis", *Journal of Muslim Minority Affairs*, vol. 20, no. 2, pp. 347–352, 2000, ISSN: 1360-2004. DOI: 10.1080/713680371.

[113] BM Henn, CR Gignoux, M Jobin, *et al.*, "Hunter-gatherer genomic diversity suggests a southern African origin for modern humans", *Proceedings of the National Academy of Sciences*, vol. 108, no. 13, pp. 5154–5162, Mar. 2011, ISSN: 0027-8424. DOI: 10.1073/pnas.1017511108.

[114] HL Kim, A Ratan, GH Perry, A Montenegro, W Miller, and SC Schuster, "Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history.", *Nature communications*, vol. 5, p. 5692, 2014, ISSN: 2041-1723. DOI: 10.1038/ncomms6692.

[115] CM Schlebusch, P Skoglund, P Sjodin, *et al.*, "Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History", *Science*, vol. 338, no. 6105, pp. 374–379, Oct. 2012, ISSN: 0036-8075. DOI: 10.1126/science.1227721.

[116] A Timmermann and T Friedrich, "Late Pleistocene climate drivers of early human migration", *Nature*, vol. 538, no. 7623, pp. 92–95, 2016, ISSN: 0028-0836. DOI: 10.1038/nature19365.

[117] T Guldemann and AM Fehn, Eds., *Beyond 'Khoisan': Historical relations in the Kalahari Basin*. Amsterdam, Netherlands: John Benjamins Publishing Company, 2014, ISBN: 978-90- 272-4849-7.

[118] JK Pickrell, N Patterson, PR Loh, *et al.*, "Ancient west Eurasian ancestry in southern and eastern Africa", *Proceedings of the National Academy of Sciences*, vol. 111, no. 7, pp. 2632–2637, 2014, ISSN: 0027-8424. DOI: 10.1073/pnas.1313787111. arXiv: arXiv:1307.8014v1.

[119] AB Smith, *Pastoralism in Africa: origins and development ecology*. Johannesburg, South Africa: Hurst & Co., Ohio University Press, Witwatersrand University Press, 1992.

[120] G Breton, CM Schlebusch, M Lombard, P Sjödin, H Soodyall, and M Jakobsson, "Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists", *Current Biology*, vol. 24, no. 8, pp. 852–858, 2014, ISSN: 09609822. DOI: 10.1016/j.cub.2014.02.041.

[121] J Fourie and E Green, "The Missing People: Accounting for the Productivity of Indigenous Populations in Cape Colonial History", *Journal of African History*, vol. 56, no. 2, pp. 195–215, 2015, ISSN: 14695138. DOI: 10.1017/S002185371500002X.

[122] M Gallego Llorente, ER Jones, A Eriksson, *et al.*, "Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent", *Science*, vol. 350, no. 6262, pp. 820–2, Nov. 2015, ISSN: 0036-8075. DOI: 10.1126/science.aad2879. arXiv: 1011.1669.

[123] G Berniell-Lee, F Calafell, E Bosch, *et al.*, "Genetic and demographic implications of the bantu expansion: Insights from human paternal lineages", *Molecular biology and evolution*, vol. 26, no. 7, pp. 1581–1589, Jul. 2009, ISSN: 1537-1719. DOI: 10.1093/molbev/msp069.

[124] C Barbieri, M Vicente, S Oliveira, *et al.*, "Migration and Interaction in a Contact Zone: mtDNA Variation among Bantu-Speakers in Southern Africa", *PLoS ONE*, vol. 9, no. 6, A Achilli, Ed., e99117, Jun. 2014, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0099117.

[125] M González-Santos, F Montinaro, O Oosthuizen, *et al.*, "Genome-Wide SNP Analysis of Southern African Populations Provides New Insights into the Dispersal of Bantu-Speaking Groups", *Genome Biology and Evolution*, vol. 7, no. 9, pp. 2560–2568, 2015, ISSN: 1759-6653. DOI: 10.1093/gbe/evv164.

[126] J Rocha, "Bantu-Khoisan interactions at the edge of the Bantu expansions: insights from southern Angola", *Journal of Anthropological Sciences*, vol. 88, pp. 5–8, 2010, ISSN: 1827-4765.

[127] L van Dorp, D Balding, S Myers, *et al.*, "Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference", *PLoS Genetics*, vol. 11, no. 8, pp. 1–49, 2015, ISSN: 15537404. DOI: 10.1371/journal.pgen.1005397.

[128] O Patterson, "Slavery", *Annual Review of Sociology*, vol. 3, pp. 407–449, 1977.

[129] DR James and S Heiliger, "Slavery and Involuntary Servitude", in *Encyclopedia of Sociology*, EF Borgatta and RJV Montgmery, Eds., 1st ed., Macmillan Reference USA, 2000, pp. 2596–2610.

[130] O Patterson, *Slavery and Social Death: A Comparative Study*. Cambridge, USA: Harvard University Press, 1982, ISBN: 0674810821.

[131] ED Domar, "The Causes of Slavery or Serfdom: A Hypothesis", *The Journal of Economic History*, vol. 30, no. 1, pp. 18–32, 1970.

[132] R Ross, *Cape of Torments: Slavery and Resistance in South Africa*, 3. London, UK: Routledge & Kegan Paul Press, 1984, vol. 54, p. 110, ISBN: 0710094078. DOI: 10.2307/1160743.

[133] L Mbeki and M van Rossum, "Private slave trade in the Dutch Indian Ocean world: a study into the networks and backgrounds of the slavers and the enslaved in South Asia and South Africa", *Slavery & Abolition*, vol. 9523, no. April, pp. 1–22, 2016, ISSN: 0144-039X. DOI: 10.1080/0144039X.2016.1159004.

[134] N Worden, *Slavery in Dutch South Africa.* Cambridge, UK: Cambridge University Press, 1985.

[135] PE Lovejoy, *Transformations in Slavery: A History of Slavery in Africa.* Cambridge: Cambridge University Press, 1983.

[136] MPM Vink, "Encounters on the Opposite Coast: Cross-Cultural Contacts between the Dutch East India Company and the Nayaka State of Madurai in the Seventeenth Century", PhD, University of Minnesota, 1998.

[137] G Groenewald, "Slaves and Free Blacks in VOC Cape Town, 1652-1795", *History Compass*, vol. 8, no. 9, pp. 964–983, Sep. 2010, ISSN: 14780542. DOI: 10.1111/j.1478-0542.2010.00724.x.

[138] A Reid, "Introduction; Slavery and Bondage in Southeast Asian History", in *Slavery, Bondage and Dependency*, A Reid, Ed., New York: Palgrave Macmillan, 1983, pp. 30–32, ISBN: 0312728123.

[139] G Campbell, "Madagascar and the slave trade.", *Journal of African History*, vol. 22, no. 1981, pp. 203–227, 1981.

[140] RCH Shell, *Children of Bondage: A Social History of the Slave Society at the Cape of Good Hope, 1652-1838*. Hanover, Netherlands: Wesleyan University Press, 1994, p. 501.

[141] L Guelke, "Frontier settlement in early dutch South Africa", *Annals of the Association of American Geographers*, vol. 66, no. 1, pp. 25–42, Mar. 1976, ISSN: 0004-5608. DOI: 10.1111/j.1467-8306.1976.tb01070.x.

[142] SJ Halford, *The Griquas of Griqualand*. Cape Town, South Africa: Juta and Company, Ltd., 1949.

[143] NSJ Van Rensburg, "Coloured Afrikaans-speakers in Potchefstroom before and after 1950: identity or political ideology?", *African Studies*, vol. 51, no. 2, pp. 261–275, Jan. 1992, ISSN: 0002-0184. DOI: 10.1080/00020189208707760.

[144] A Whyte, P van Duin, and R Ross, "The Economy of the Cape Colony in the Eighteenth Century", *African Economic History*, no. 18, p. 189, 1989, ISSN: 01452258. DOI: 10.2307/3601800.

[145] L Guelke, "The early European settlement of South Africa", PhD thesis, University of Toronto, 1974.

[146] R Elphick, *Kraal and Castle: Khoikhoi and the Founding of White South Africa*. New Haven: Yale University Press, 1977.

[147] R Elphick and VC Malherbe, "The Khoisan to 1828", in *The Shaping of South African Society, 1652-1840*, R Elphick and H Giliomee, Eds., Cape Town, 1989, pp. 3–65.

[148] L Guelke and RCH Shell, "An early colonial landed gentry: land and wealth in the Cape Colony 1682-1731", *Journal of Historical Geography*, vol. 9, no. 3, pp. 265–286, 1983. DOI: 10.1016/0305-7488(83)90183-4.

[149] LM Kootker, L Mbeki, AG Morris, H Kars, and GR Davies, "Dynamics of indian ocean slavery revealed through isotopic data from the colonial era cobern street burial site, cape town, South Africa (1750-1827)", *PLoS ONE*, vol. 11, no. 6, pp. 1–20, 2016, ISSN: 19326203. DOI: 10.1371/journal.pone.0157750.

[150] AK Khalfani and T Zuberi, "Racial classification and the modern census in South Africa, 1911-1996", *Race and Society*, vol. 4, no. 2, pp. 161–176, 2001, ISSN: 10909524. DOI: 10.1016/S1090-9524(03)00007-X.

[151] A Whiten, RA Hinde, KN Laland, and CB Stringer, "Culture evolves", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, no. 1567, pp. 938–948, Apr. 2011, ISSN: 0962-8436. DOI: 10.1098/rstb.2010.0372. arXiv: arXiv:1011.1669v3.

[152] MC Botha and J Pritchard, "Blood group gene frequencies", *Munger Africana Library Notes*, no. 16, pp. 1–27, 1972.

[153] RL Morris, "Spatial Patterns of Social Differentiation among the Coloureds in Greater Cape Town", Masters of Arts, University of Cape Town, 1980.

[154] BJ Liebenberg, "Die vrystelling van die slawe in die Kaapkolonie en die implikasies daarvan", Masters of Art, University of the Orange Free State, 1953.

[155] TRH Davenport, *South Africa: A Modern History*, 2nd ed. Johannesburg, South Africa: Macmillan, 1977, p. 432.

[156] GT Nurse and T Jenkins, "The Griqua of Campbell, Cape Provice, South Africa", *Am J Phys Anthropol*, vol. 43, no. 1, pp. 71–78, 1975. DOI: 10.1002/ajpa.1330430111.

[157] GT Nurse, T Jenkins, BJ Africa, and FF Stellmacher, "Sero-genetic studies on the basters of Rehoboth, South West Africa/Namibia", *Annals of Human Biology*, vol. 9, no. 2, pp. 157–166, 1982, ISSN: 03014460. DOI: 10.1080/03014468200005621.

[158] H Lang, "The Population Development of the Rehoboth Basters", *Anthropos*, vol. 93, no. 4-6, pp. 381–391, 1998, ISSN: 02579774.

[159] Y Da Costa, "Assimilatory processes amongst the cape muslims in South Africa during the 19th century", *South African Journal of Sociology*, vol. 23, no. 1, pp. 5–11, 1992, ISSN: 02580144. DOI: 10.1080/02580144.1992.10520103.

[160] S Taylor, *The Caliban Shore: The Fate of the Grosvenor Castaways*. London: Faber and Faber Ltd., 2005, ISBN: 0571210724.

[161] SA Rochlin, "Aspects of Islam in Nineteenth-Century South Africa", *Bulletin of the School of Oriental and African Studies*, vol. 10, no. 1, pp. 213–221, 1940, ISSN: 14740699. DOI: 10.1017/S0041977X00068312.

[162] P Scott, "Cape Town: A Multi-Racial City", *The Geographical Journal*, vol. 2, no. 120, pp. 149–157, 1955.

[163] PW Laider, *The Growth and Government of Cape Town*. Unie Volkspers Bpk, 1939, p. 525.

[164] KL Sainani, "Introduction to principal components analysis", *American Academy of Physical medicine and Rehabilitation*, vol. 6, no. 3, pp. 275–278, 2014, ISSN: 19341482. DOI: 10.1016/j.pmrj.2014.02.001. arXiv: 9905079 [astro-ph].

[165] N Patterson, AL Price, and D Reich, "Population structure and eigenanalysis", *PLoS Genetics*, vol. 2, no. 12, pp. 2074–2093, 2006, ISSN: 15537390. DOI: 10.1371/journal.pgen.0020190.

[166] D Reich, AL Price, and N Patterson, *Principal component analysis of genetic data*, 2008. DOI: 10.1038/ng0508-491. arXiv: Techniques [Thesis].

[167] J Novembre, T Johnson, K Bryc, *et al.*, "Genes mirror geography within Europe.", *Nature*, vol. 456, no. 7218, pp. 98–101, Nov. 2008, ISSN: 1476-4687. DOI: 10.1038/nature07331.

[168] Z Hofmanová, S Kreutzer, G Hellenthal, *et al.*, "Early farmers from across Europe directly descended from Neolithic Aegeans", *Proceedings of the National Academy of Sciences*, vol. 113, no. 25, pp. 6886–6891, 2016, ISSN: 0027-8424. DOI: 10.1073/pnas.1523951113. arXiv: arXiv:1011.1669v3.

[169] J Novembre and M Stephens, "Interpreting principal component analyses of spatial population genetic variation", *Nature Genetics*, vol. 40, no. 5, pp. 646–649, 2008, ISSN: 10614036. DOI: 10.1038/ng.139. arXiv: NIHMS150003.

[170] JK Pritchard, M Stephens, and P Donnelly, "Inference of population structure using multilocus genotype data.", *Genetics*, vol. 155, no. 2, pp. 945–959, Jun. 2000, ISSN: 0016-6731.

[171] GBJ Busby, "The Peopling of Europe: A Genetic Perspective", DPhil, University of Oxford, 2012, p. 132.

[172] DH Alexander, J Novembre, and K Lange, "Fast Model-Based Estimation of Ancestry in Unrelated Individuals", *Genome Research*, vol. 19, pp. 1655–1664, 2009. DOI: 10.1101/gr.094052.109.vidual.

[173] H Tang, J Peng, P Wang, and NJ Risch, "Estimation of individual admixture: analytical and study design considerations.", *Genetic epidemiology*, vol. 28, no. 4, pp. 289–301, May 2005, ISSN: 07410395. DOI: 10.1002/gepi.20064.

[174] D Falush, L van Dorp, and DJ Lawson, "A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots", *bioRxiv*, p. 066 431, 2016. DOI: 10.1101/066431.

[175] G Evanno, S Regnaut, and J Goudet, "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.", *Molecular ecology*, vol. 14, no. 8, pp. 2611–2620, Jul. 2005, ISSN: 0962-1083. DOI: 10.1111/j.1365-294X.2005.02553.x.

[176] D Hall, EM Wijsman, J Louw Roos, JA Gogos, and M Karayiorgou, "Extended intermarker linkage disequilibrium in the Afrikaners", *Genome Research*, vol. 12, no. 6, pp. 956–961, 2002, ISSN: 10889051. DOI: 10.1101/gr.136202.

[177] GBJ Busby, G Band, Q Si Le, *et al.*, "Admixture into and within sub-Saharan Africa", *eLife*, pp. 1–44, Feb. 2016, ISSN: 2050-084X. DOI: 10.1101/038406.

[178] JP Huelsenbeck, P Andolfatto, and ET Huelsenbeck, "Structurama: Bayesian Inference of Population Structure", *Evolutionary Bioinformatics*, vol. 7, EBO.S6761, Jan. 2011, ISSN: 1176-9343. DOI: 10.4137/EBO.S6761.

[179] FM Busing, E Meijer, and R Van Der Leeden, "Delete-m Jackknife for Unequal m", *Statistics and Computing*, vol. 9, no. 1, pp. 3–8, 1999, ISSN: 09603174. DOI: 10.1023/A:1008800423698.

[180] P Moorjani, N Patterson, JN Hirschhorn, *et al.*, "The history of african gene flow into Southern Europeans, Levantines, and Jews", *PLoS Genetics*, vol. 7, no. 4, 2011, ISSN: 15537390. DOI: 10.1371/journal.pgen.1001373.

[181] S Baharian, M Barakatt, CR Gignoux, *et al.*, "The Great Migration and African-American Genomic Diversity", *PLOS Genetics*, vol. 12, no. 5, G Gibson, Ed., e1006059, May 2016, ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1006059.

[182] AS Burrell and TR Disotell, "Panmixia postponed: ancestry-related assortative mating in contemporary human populations.", *Genome biology*, vol. 10, no. 11, p. 245, 2009, ISSN: 1465-6906. DOI: 10.1186/gb-2009-10-11-245.

[183] S Challis, "Creolisation on the nineteenth-century frontiers of Southern Africa: A case study of the AmaTola 1 'Bushmen' in the Maloti-Drakensberg", *Journal of Southern African Studies*, vol. 38, no. 2, pp. 265–280, 2012. DOI: 10.1080/03057070.2012.666905.

[184] JM Greeff and JC Erasmus, "Three hundred years of low non-paternity in a human population", *Heredity*, vol. 115, no. 5, pp. 396–404, 2015, ISSN: 13652540 0018067X. DOI: 10.1038/hdy.2015.36.

[185] N Patterson, DC Petersen, RE van der Ross, *et al.*, "Genetic structure of a unique admixed population: implications for medical research.", *Human Molecular Genetics*, vol. 19, no. 3, pp. 411–419, Feb. 2010, ISSN: 1460-2083. DOI: 10.1093/hmg/ddp505.

[186] ER Chimusa, M Daya, M Möller, *et al.*, "Determining Ancestry Proportions in Complex Admixture Scenarios in South Africa Using a Novel Proxy Ancestry Selection Method", *PLoS ONE*, vol. 8, no. 9, D O'Rourke, Ed., e73971, Sep. 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0073971.

[187] ER Chimusa, N Zaitlen, M Daya, *et al.*, "Genome-wide association study of ancestry-specific TB risk in the South African Coloured population", *Human Molecular Genetics*, vol. 23, no. 3, pp. 796–809, 2014, ISSN: 0964-6906. DOI: 10.1093/hmg/ddt462.

[188] M Daya, L van der Merwe, U Galal, *et al.*, "A panel of ancestry informative markers for the complex five-way admixed South African coloured population.", *PloS one*, vol. 8, no. 12, e82224, 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0082224.

[189] FE Kritzinger, S den Boon, S Verver, *et al.*, "No decrease in annual risk of tuberculosis infection in endemic area in Cape Town, South Africa", *Tropical Medicine & International Health*, vol. 14, no. 2, pp. 136–142, 2009, ISSN: 13602276. DOI: 10.1111/j.1365-3156.2008.02213.x.

[190] S den Boon, SWP van Lill, MW Borgdorff, *et al.*, "High prevalence of tuberculosis in previously treated patients, Cape Town, South Africa.", *Emerging infectious diseases*, vol. 13, no. 8, pp. 1189–94, 2007, ISSN: 1080-6040. DOI: 10.3201/eid1308.051327.

[191] S Purcell, B Neale, K Todd-Brown, *et al.*, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses", *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007, ISSN: 00029297. DOI: 10.1086/519795.

[192] RM Kuhn, D Haussler, and W James Kent, "The UCSC genome browser and associated tools", *Briefings in Bioinformatics*, 2013, ISSN: 14675463. DOI: 10.1093/bib/bbs038.

[193] A Manichaikul, JC Mychaleckyj, SS Rich, K Daly, M Sale, and WM Chen, "Robust relationship inference in genome-wide association studies", *Bioinformatics*, vol. 26, no. 22, pp. 2867–2873, 2010, ISSN: 13674803. DOI: 10.1093/bioinformatics/btq559.

[194] M Hubert and S Van der Veeken, "Outlier detection for skewed data", *Journal of Chemometrics*, vol. 22, no. 3-4, pp. 235–246, Mar. 2008, ISSN: 08869383. DOI: 10.1002/cem.1123.

[195] M Maechler, P Rousseeuw, C Croux, *et al.*, *robustbase: Basic Robust Statistics. R package version 0.93-3.* 2009.

[196] O Lao, TT Lu, M Nothnagel, *et al.*, "Correlation between Genetic and Geographic Structure in Europe", *Current Biology*, vol. 18, no. 16, pp. 1241–1248, 2008, ISSN: 09609822. DOI: 10.1016/j.cub.2008.07.049.

[197] NM Kopelman, J Mayzel, M Jakobsson, NA Rosenberg, and I Mayrose, "Clumpak: a program for identifying clustering modes and packaging population structure inferences across K", *Molecular Ecology Resources*, vol. 15, no. 5, pp. 1179–1191, Sep. 2015, ISSN: 1755098X. DOI: 10.1111/1755-0998.12387. arXiv: 15334406.

[198] H Wickham and G Grolemund, *R for Data Science*, 1st ed., M Beaugureau and M Loukides, Eds. O'Reilly, 2017.

[199] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2018.

[200] M Nothnagel, D Ellinghaus, S Schreiber, M Krawczak, and A Franke, "A comprehensive evaluation of SNP genotype imputation", *Human Genetics*, vol. 125, no. 2, pp. 163–171, 2009, ISSN: 03406717. DOI: 10.1007/s00439-008-0606-5.

[201] O Delaneau, J Marchini, and JF Zagury, "A linear complexity phasing method for thousands of genomes", *Nature Methods*, vol. 9, no. 2, pp. 179–181, 2011, ISSN: 1548-7091. DOI: 10.1038/nmeth.1785.

[202] The Coop Lab, *How many genetic ancestors do I have?*, 2013.

[203] J Bacon-Shone, "A short history of compositional data analysis", in *Compositional Data Analysis: Theory and Applications*, V Powlowsky-Glahn and A Buccianti, Eds., 2011, ch. 1, pp. 3–11.

[204] TP Quinn, MF Richardson, D Lovell, and TM Crowley, "Propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis", *Scientific Reports*, vol. 7, no. 1, pp. 1–9, 2017, ISSN: 20452322. DOI: 10.1038/s41598-017-16520-0.

[205] V Pawlowsky-Glahn and JJ Egozcue, "Compositional data and their analysis: an introduction", *Geological Society, London, Special Publications*, vol. 264, no. 1, pp. 1–10, 2006, ISSN: 0305-8719. DOI: 10.1144/GSL.SP.2006.264.01.01.

[206] M Templ, K Hron, and P Filzmoser, "robCompositions: An R-package for Robust Statistical Analysis of Compositional Data", in *Compositional Data Analysis*, Chichester, UK: John Wiley & Sons, Ltd, Jul. 2011, pp. 341–355, ISBN: 9780470711354. DOI: 10.1002/9781119976462.ch25.

[207] N Gilbert, S Boyle, H Fiegler, K Woodfine, NP Carter, and WA Bickmore, "Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers", *Cell*, vol. 118, no. 5, pp. 555–566, Sep. 2004, ISSN: 00928674. DOI: 10.1016/j.cell.2004.08.011.

[208] LG Wilming, JG Gilbert, K Howe, S Trevanion, T Hubbard, and JL Harrow, "The vertebrate genome annotation (Vega) database", *Nucleic Acids Research*, vol. 36, no. SUPPL. 1, pp. 753–760, 2008, ISSN: 03051048. DOI: 10.1093/nar/gkm987.

[209] P Skoglund, C Posth, K Sirak, *et al.*, "Genomic insights into the peopling of the Southwest Pacific.", *Nature*, vol. 538, no. 7626, pp. 510–513, Oct. 2016, ISSN: 1476-4687. DOI: 10.1038/nature19844.

[210] G Hudjashov, TM Karafet, DJ Lawson, *et al.*, "Complex patterns of admixture across the Indonesian archipelago", *Molecular Biology and Evolution*, 2017, ISSN: 0737-4038. DOI: 10.1093/molbev/msx196.

[211] A Mörseburg, L Pagani, FX Ricaut, *et al.*, "Multi-layered population structure in Island Southeast Asians", *European Journal of Human Genetics*, vol. 24, no. 11, pp. 1–7, 2016, ISSN: 14765438. DOI: 10.1038/ejhg.2016.60.

[212] J Kere, "Human Population Genetics: Lessons from Finland", *Annual Review of Genomics and Human Genetics*, vol. 2, no. 1, pp. 103–128, 2001, ISSN: 1527-8204. DOI: 10.1146/annurev.genom.2.1.103.

[213] J Yarwood, "With Mixed Feelings : Negotiating Coloured Identity in Post-Apartheid South Africa", Doctor of Philosophy, City University of New York, 2011.

[214] Statistics South Africa, "South Africa Census 2011 and Community Survey 2007", Statistics South Africa, Pretoria, South Africa, Tech. Rep., 2011.

[215] CM Schlebusch, F Prins, M Lombard, M Jakobsson, and H Soodyall, "The disappearing San of southeastern Africa and their genetic affinities", *Human Genetics*, vol. 135, no. 12, pp. 1365–1373, Dec. 2016, ISSN: 0340-6717. DOI: 10.1007/s00439-016-1729-8.

[216] TA Jinam, ME Phipps, and N Saitou, "Admixture patterns and genetic differentiation in negrito groups from West Malaysia estimated from genome-wide SNP data.", *Human Biology*, vol. 85, no. 1-3, pp. 173–88, 2013, ISSN: 1534-6617.

[217] D Pierron, M Heiske, H Razafindrazaka, *et al.*, "Genomic landscape of human diversity across Madagascar", *Proceedings of the National Academy of Sciences*, vol. 114, no. 32, E6498–E6506, Aug. 2017, ISSN: 0027-8424. DOI: 10.1073/pnas.1704906114.

[218] M Haber, D Gauguier, S Youhanna, *et al.*, "Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture", *PLoS Genetics*, vol. 9, no. 2, 2013, ISSN: 15537390. DOI: 10.1371/journal.pgen.1003316.

[219] L Pagani, DJ Lawson, E Jagoda, *et al.*, "Genomic analyses inform on migration events during the peopling of Eurasia", *Nature*, vol. 538, no. 7624, pp. 238–242, 2016, ISSN: 0028-0836. DOI: 10.1038/nature19792. arXiv: NIHMS150003.

[220] J Fourie and J Cilliers, "Die huwelikspatrone van Europese setlaars aan die Kaap, 1652-1910", *New Contree*, vol. 69, no. 2010, pp. 45–70, 2014.

[221] S Swiss and JE Giller, "Rape as a Crime of War", *JAMA*, vol. 270, no. 5, p. 612, Aug. 1993, ISSN: 0098-7484. DOI: 10.1001/jama.1993.03510050078031.

[222] P Scully, "Rape, Race, and Colonial Culture: The Sexual Politics of Identity in the Nineteenth-Century Cape Colony, South Africa", *The American Historical Review*, vol. 100, no. 2, pp. 335–359, 1995.

[223] JN Fenner, "Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies", *American Journal of Physical Anthropology*, vol. 128, no. 2, pp. 415–423, 2005, ISSN: 00029483. DOI: 10.1002/ajpa.20188.

[224] N Brucato, P Kusuma, P Beaujard, H Sudoyo, MP Cox, and FX Ricaut, "Genomic admixture tracks pulses of economic activity over 2,000 years in the Indian Ocean trading network", *Scientific Reports*, vol. 7, no. 1, pp. 5–10, 2017, ISSN: 20452322. DOI: 10.1038/s41598-017-03204-y.

[225] N Brucato, V Fernandes, S Mazières, *et al.*, "The Comoros Show the Earliest Austronesian Gene Flow into the Swahili Corridor", *American Journal of Human Genetics*, vol. 102, no. 1, pp. 58–68, 2018, ISSN: 15376605. DOI: 10.1016/j.ajhg.2017.11.011.

[226] J Prado-Martinez, PH Sudmant, JM Kidd, *et al.*, "Great ape genetic diversity and population history", *Nature*, vol. 499, no. 7459, pp. 471–475, 2014, ISSN: 0028-0836. DOI: 10.1038/nature12228.

[227] JC Chacón-Duque, K Adhikari, M Fuentes-guajardo, *et al.*, "Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance", *Nature Communications*, pp. 1–22, 2018, ISSN: 20411723. DOI: 10.1101/252155.

[228] AS Malaspinas, MC Westaway, C Muller, *et al.*, "A genomic history of Aboriginal Australia", *Nature*, vol. 538, no. 7624, pp. 207–214, 2016, ISSN: 0028-0836. DOI: 10.1038/nature18299. arXiv: NIHMS150003.

[229] T Murray, "The childhood of William Lanne: Contact archaeology and Aboriginality in Tasmania", *Antiquity*, vol. 67, no. 256, pp. 504–519, 1993, ISSN: 17451523. DOI: 10.1017/S0003598X00045725.

[230] N Brucato, P Kusuma, MP Cox, *et al.*, "Malagasy Genetic Ancestry Comes from an Historical Malay Trading Post in Southeast Borneo", *Molecular Biology and Evolution*, vol. 33, no. 9, pp. 2396–2400, 2016, ISSN: 15371719. DOI: 10.1093/molbev/msw117.

[231] Z Gu, *circlize: Circular Visualization. R package v*, 2019.

[232] BM Peter, "Admixture, population structure, and f-statistics", *Genetics*, vol. 202, no. 4, pp. 1485–1501, 2016, ISSN: 19432631. DOI: 10.1534/genetics.115.183913.

[233] J Defo, "Genetic Dating and Pattern of Admixture in Modern Human Evolution", Masters, University of the Western Cape, 2017.

[234] BEJS Werz, "Diving up the human past . Perspectives of maritime archaeology , with specific reference to developments in South Africa until 1996", in *BAR International Series*, Oxford, UK: Archaeopress, 1999, ISBN: 0-86054-98366.

[235] M Hall, "The archaeology of colonial settlement in southern Africa", *Annual Review of Anthropology*, no. 22, pp. 177–200, 1993.

[236] T Maggs, "The Great Galleon Sao Joao: remains from a mid-sixteenth century wreck on the Natal South Coast", *Annals of the Natal Museum*, vol. 26, no. 1, pp. 173–186, 1984, ISSN: 0304-0798.

[237] SDS Jayasuriya, *The Portuguese in the East: A Cultural History of a Maritime Trading Empire*. London, UK: Tauris Academic Studies, 2008, ISBN: 978 1 84511 585 2.

[238] T Huffman, *Handbook to the iron age: The archaeology of pre-colonial farming societies in Southern Africa*. Scottsville: University Kwazulu Natal Press, 2007, ISBN: 1869141083.

[239] A Gaedigk and C Coetsee, "The CYP2D6 gene locus in South African Coloureds: Unique allele distributions, novel alleles and gene arrangements", *European Journal of Clinical Pharmacology*, vol. 64, no. 5, pp. 465–475, 2008, ISSN: 00316970. DOI: 10.1007/s00228-007-0445-7.

[240] A Lucassen, K Ehlers, PJ Grobler, and AL Shezi, "Allele frequency data of 15 autosomal STR loci in four major population groups of South Africa", *International Journal of Legal Medicine*, vol. 128, no. 2, pp. 275–276, 2014, ISSN: 14371596. DOI: 10.1007/s00414-013-0898-4.

[241] G Hefke, S Davison, and ME D'Amato, "Forensic performance of Investigator DIPplex indels genotyping kit in native, immigrant, and admixed populations in South Africa", *ELECTROPHORESIS*, vol. 36, no. 24, pp. 3018–3025, Dec. 2015, ISSN: 01730835. DOI: 10.1002/elps.201500243.

[242] PG Ristow, KW Cloete, and ME D'Amato, "GlobalFiler ® Express DNA amplification kit in South Africa: Extracting the past from the present", *Forensic Science International: Genetics*, vol. 24, pp. 194–201, Sep. 2016, ISSN: 18724973. DOI: 10.1016/j.fsigen.2016.07.007.

[243] K Cloete, L Ehrenreich, ME D'Amato, N Leat, S Davison, and M Benjeddou, "Analysis of seventeen Y-chromosome STR loci in the Cape Muslim population of South Africa", *Legal Medicine*, vol. 12, no. 1, pp. 42–45, Jan. 2010, ISSN: 13446223. DOI: 10.1016/j.legalmed.2009.10.001.

[244] MP Besten, "Transformation and reconstruction of Khoe-San identities: AAS Le Fleur I, Griqua Identities and Post-Apartheid Khoe-San Revivalism (1894-2004)", Doctorate, Universiteit Leiden, 2006.

[245] A Auton, GR Abecasis, DM Altshuler, *et al.*, "A global reference for human genetic variation", *Nature*, vol. 526, no. 7571, pp. 68–74, Sep. 2015, ISSN: 0028-0836. DOI: 10.1038/nature15393. arXiv: 15334406.

[246] DH Huson and D Bryant, "Application of phylogenetic networks in evolutionary studies.", *Molecular biology and evolution*, vol. 23, no. 2, pp. 254–67, Feb. 2006, ISSN: 0737-4038. DOI: 10.1093/molbev/msj030.

[247] D Bryant, "Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks", *Molecular Biology and Evolution*, vol. 21, no. 2, pp. 255–265, Aug. 2003, ISSN: 0737-4038. DOI: 10.1093/molbev/msh018.

[248] KP Schliep, "phangorn: phylogenetic analysis in R", *Bioinformatics*, vol. 27, no. 4, pp. 592–593, Feb. 2011, ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btq706.

[249] RJ Hijmans, E Williams, and C Vennes, *geosphere: Spherical Trigonometry. R package version 1.5-10*, 2019.

[250] H Hammarström, R Forkel, and M Haspelmath, *Glottolog 4.0*, Max Planck Institute for the Science of Human History, Jena, 2019.

[251] K Brown and S Ogilvie, Eds., *Concise Encyclopedia of Languages of the World*, 1st ed. Oxford, UK.: Elsevier Ltd, 2009, ISBN: 978-0-08-087774-7.

[252] AV Lyovin, B Kessler, and WR Leben, *An Introduction to the Languages of the World*. Oxford, UK.: Oxford University Press, 2017, ISBN: 9780195149883.

[253] L Kalaydjieva, D Gresham, and F Calafell, "Genetic studies of the Roma (Gypsies): a review.", *BMC medical genetics*, vol. 2, p. 5, 2001, ISSN: 1471-2350. DOI: 10.1186/1471-2350-2-5.

[254] DM Behar, B Yunusbayev, M Metspalu, *et al.*, "The genome-wide structure of the Jewish people", *Nature*, vol. 466, no. 7303, pp. 238–242, Jul. 2010, ISSN: 0028-0836. DOI: 10.1038/nature09103.

[255] L Guelke and R Shell, "Landscape of conquest: Frontier water alienation and khoikhoi strategies of survival, 1652-1780", *Journal of Southern African Studies*, vol. 18, no. 4, pp. 803–824, 1992, ISSN: 14653893. DOI: 10.1080/03057079208708339.

[256] JB Torres and RA Kittles, "The relationship between "race"and genetics and biomedical research", *Current Hypertension Reports*, vol. 9, no. 3, pp. 196–201, 2007, ISSN: 15226417. DOI: 10.1007/s11906-007-0035-1.

[257] G Hellenthal, GBJ Busby, G Band, *et al.*, "A genetic atlas of human admixture history.", *Science*, vol. 343, no. 6172, pp. 747–751, 2014, ISSN: 1095-9203. DOI: 10.1126/science.1243518.

[258] J Xing, WS Watkins, DJ Witherspoon, *et al.*, "Fine-scaled human genetic structure revealed by SNP microarrays", *Genome Research*, vol. 19, no. 5, pp. 815–825, 2009, ISSN: 1088-9051. DOI: 10.1101/gr.085589.108.

[259] S Holm, "A Simple Sequentially Rejective Multiple Test Procedure", *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979, ISSN: 0218-1959. DOI: 10.1142/S0218195905001683.

[260] D Pierron, H Razafindrazaka, L Pagani, *et al.*, "Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar", *Proceedings of the National Academy of Sciences*, vol. 111, no. 3, pp. 936–941, Jan. 2014, ISSN: 0027-8424. DOI: 10.1073/pnas.1321860111.

[261] TM Karafet, JS Lansing, A Sim, *et al.*, "Small Traditional Human Communities Sustain Genomic Diversity over Microgeographic Scales despite Linguistic Isolation", *Molecular Biology and Evolution*, vol. 33, no. 9, pp. 2273–2284, 2016, ISSN: 0737-4038. DOI: 10.1093/molbev/msw099.

[262] AK Pathak, A Kadian, A Kushniarevich, *et al.*, "The Genetic Ancestry of Modern Indus Valley Populations from Northwest India", *American Journal of Human Genetics*, vol. 103, no. 6, pp. 918–929, 2018, ISSN: 15376605. DOI: 10.1016/j.ajhg.2018.10.022.

[263] G Chaubey, Q Ayub, N Rai, *et al.*, "Like sugar in milk: reconstructing the genetic history of the Parsi population", *Genome Biology*, vol. 18, no. 1, p. 110, Dec. 2017, ISSN: 1474-760X. DOI: 10.1186/s13059-017-1244-9.

[264] B Yunusbayev, M Metspalu, E Metspalu, *et al.*, "The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia.", *PLoS genetics*, vol. 11, no. 4, e1005068, 2015, ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005068.

[265] BP Hoh, L Deng, MJ Julia-Ashazila, *et al.*, "Fine-scale population structure of Malays in Peninsular Malaysia and Singapore and implications for association studies", *Human Genomics*, vol. 9, no. 1, p. 16, 2015, ISSN: 1479-7364. DOI: 10.1186/s40246-015-0039-x.

Appendices

A

Chapter 4 Supplementary

A.1 Supplementary Material

A.1.1 SAC Cluster-Averaged NNLS

Whether the copying vectors are averaged across individuals before or after the NNLS is performed may influence the outcome of the analysis. One possible source of error is that haplotype assignment within individuals may be incorrect when assignments are ambiguous. It may be worthwhile to average the copying vector across a group of individuals to negate the influence prior to performing NNLS regression. The counter argument for not averaging ahead of NNLS is that in admixed populations (and particularly recent and hyper-diverse populations such as the SAC), there will be inherently variable source populations. One would then inadvertently produce incorrect copying vectors and sway the NNLS regression to identify spurious ancestries.

To investigate the differences between procedures we repeated the NNLS procedure on the SAC-GR datasets after averaging the copying vectors within the fourteen fS-inferred SAC clusters. I contrasted these results with that of the individual based NNLS analyses presented in the main text. Except for the copying-vector aggregating step, all other steps were identical.

The results were by-en-large the same, with some notable differences. The relationship of ancestry prevalence in the populations to the average proportion ancestry remained the same (threshold >1% in at least one individual). Far more source populations were possible candidates when using the individual-based NNLS (33 vs 16; A.39), but after applying the criteria of at least 1% ancestry in at least 10% of the individuals (equivalently >1 cluster), the cluster-based NNLS retrieved 15 sources (vs 11 from individual-based) (Supp. Figure A.39).

Four of the newly identified source populations represented a very low proportion ancestry on average (~1-2%) and the other results remain qualitatively the same A.40. There is a decline in the prevalence of 06_C-S.India_DV from eleven to six clusters and for 32_Levant_SM from eleven to five clusters (Supp. Figure A.41). Among the newly identified ancestry, 34_E.Mediterranean_HE-RM-SM is present in nine of 14 clusters, 17_N.India_IN and 41_E.Is.SEA_MP occur in 3 clusters only.

I conclude that the two directions for the analyses produce qualitatively similar results. The novel sources identified through the cluster-based NNLS are mi-

nor contributors with regards to both prevalence and abundance. The source 34_E.Mediterranean_HE-RM-SM is an exception in its prevalence though it appears to have replaced 32_Levant_SM in several clusters though the two are not mutually exclusive.

A.2 Supplementary Figures

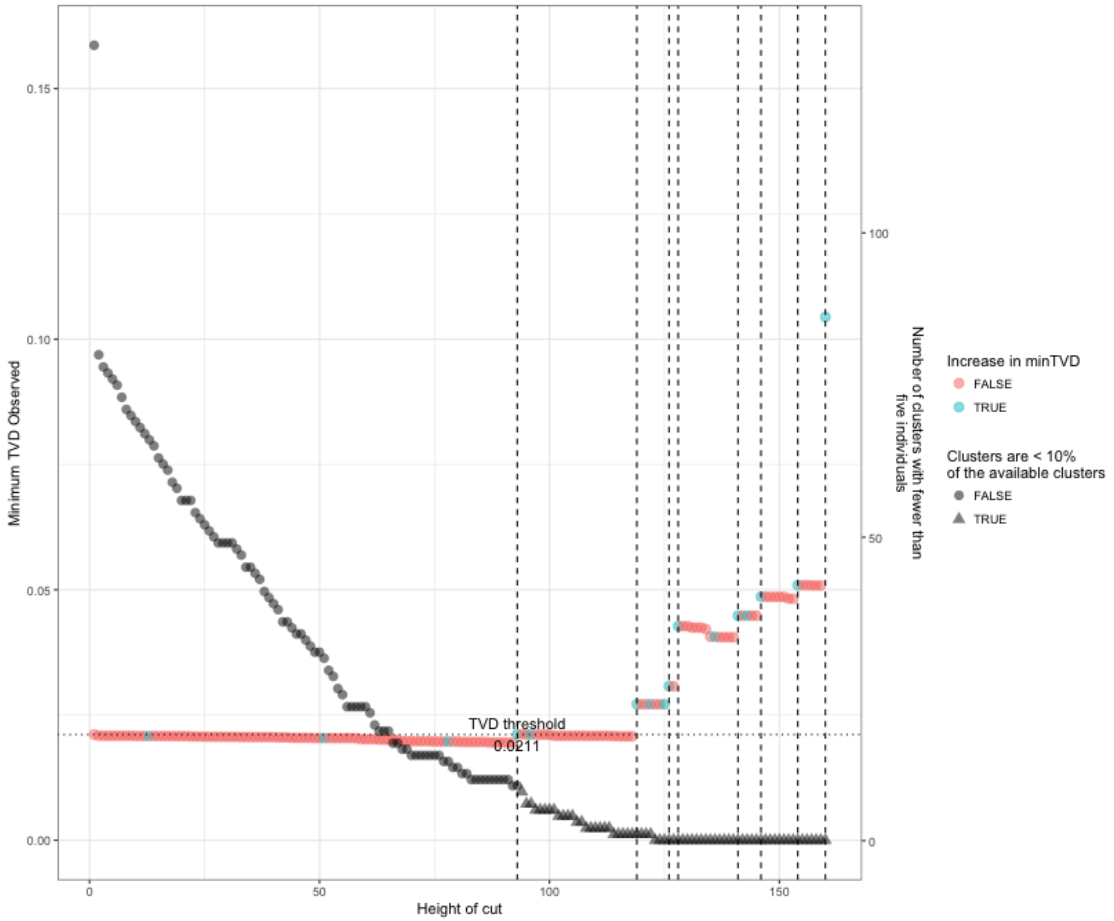


Figure A.1: Changes in minimum Total Variance Distance (TVD) observed between fS-inferred clusters for the GR - GR CP run at varying heights in the dendrogram. The TVD was estimated on total copy length. Minimum TVD values indicated for each height on the tree and colour indicates if that TVD values was an increase or decrease from the previous TVD. Dashed vertical lines indicate the position of increases in minimum TVD which exceeds $3 \times \frac{1}{H} \sum_{1}^{H} |d_h|$ where $|d_h|$ is the change in minimum TVD from heights h to $h-1$ and H is the final height. Indicated in black is the number of fS-inferred clusters with fewer than 20 individuals and if that number of clusters represents <10% of all clusters. Final decision for threshold indicated by the horizontal dashed line.

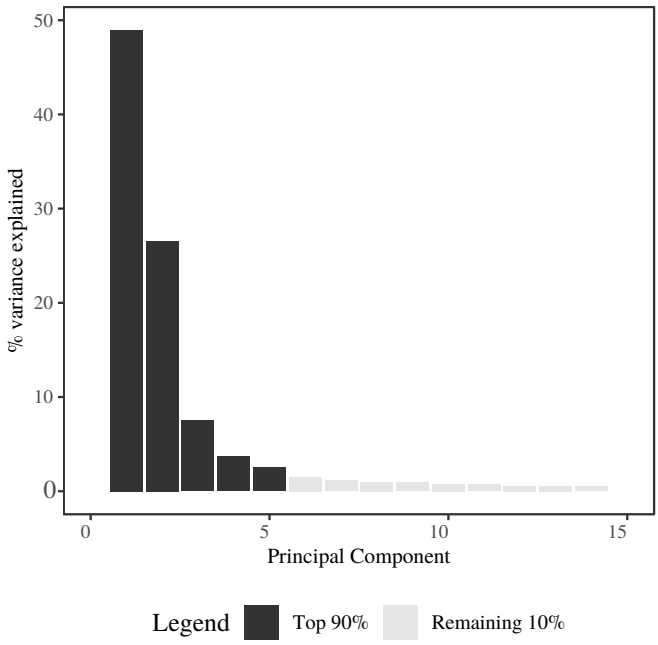


Figure A.2: Variance explained by each principal component (PC) from the GR + SAC analysis. The largest contributing PCs which sum to 90% of the variance are indicated in black. The remaining components contribute to the final 10% variance.

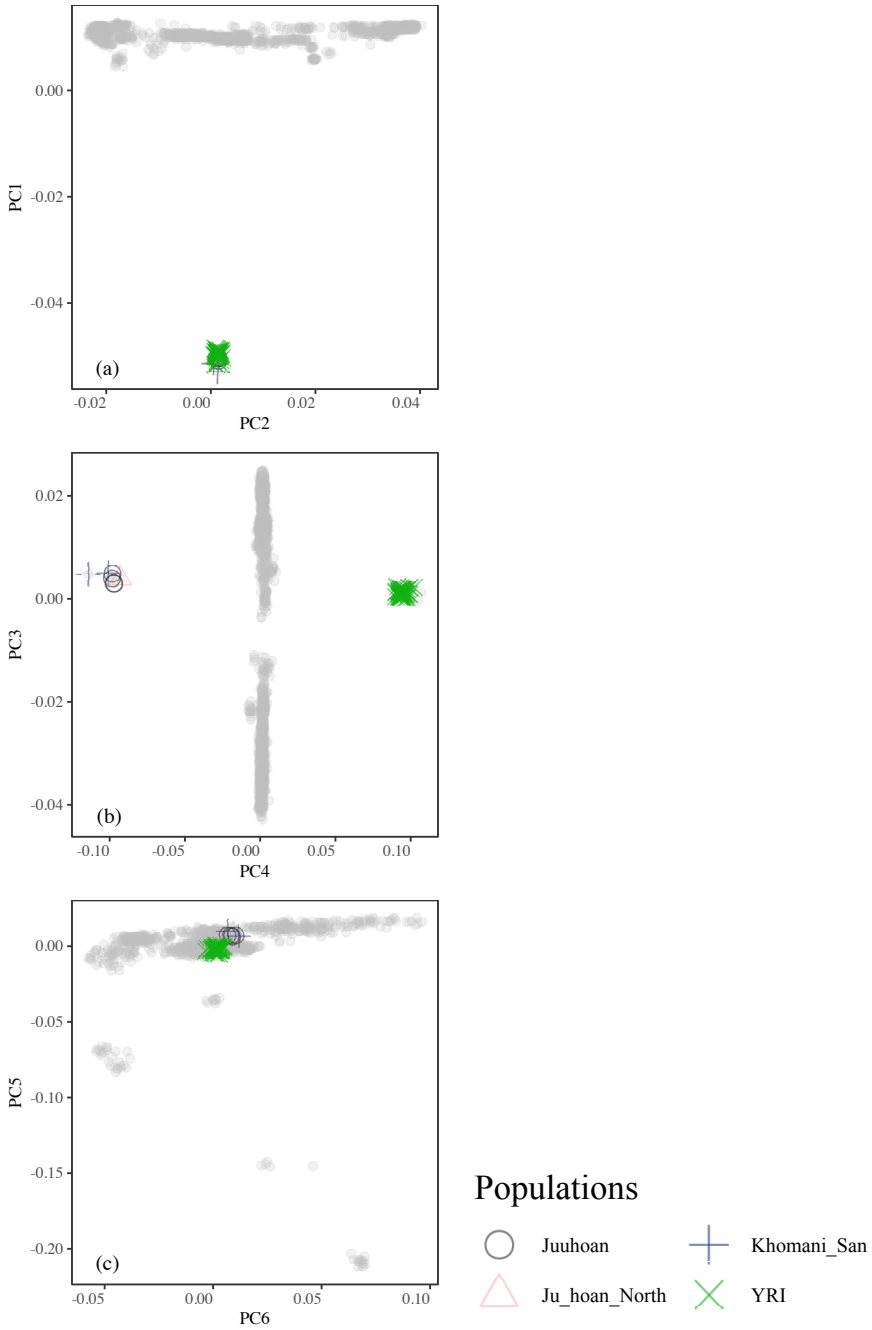


Figure A.3: Global scale PCA focused on Sub-Saharan Africa. Sub-plots (a-c) show pairwise plots of PCs.Samples from the region are presented in colour according to *a priori* population labels while the remaining GR populations are represented by grey circles. . Variance explained by each PC shown in Figure A.2. The SAC individuals not shown. Explanations of abbreviations in Supp. Table A.1

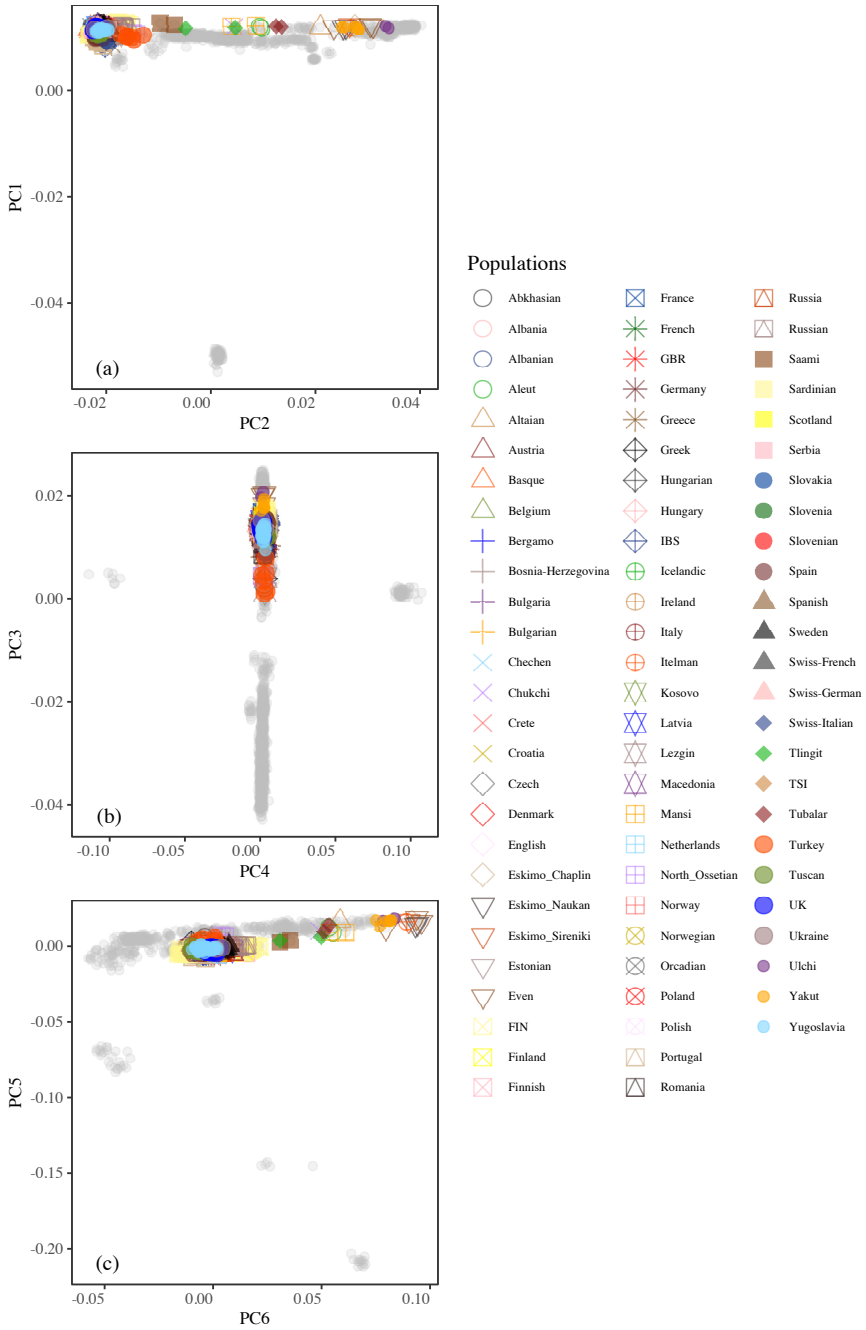


Figure A.4: Global scale PCA focused on Siberia and Europe. Sub-plots (a-c) show pairwise plots of PCs.Samples from the region are presented in colour according to *a priori* population labels while the remaining GR populations are represented by grey circles. . Variance explained by each PC shown in Figure A.2. The SAC individuals not shown. Explanations of abbreviations in Supp. Table A.1

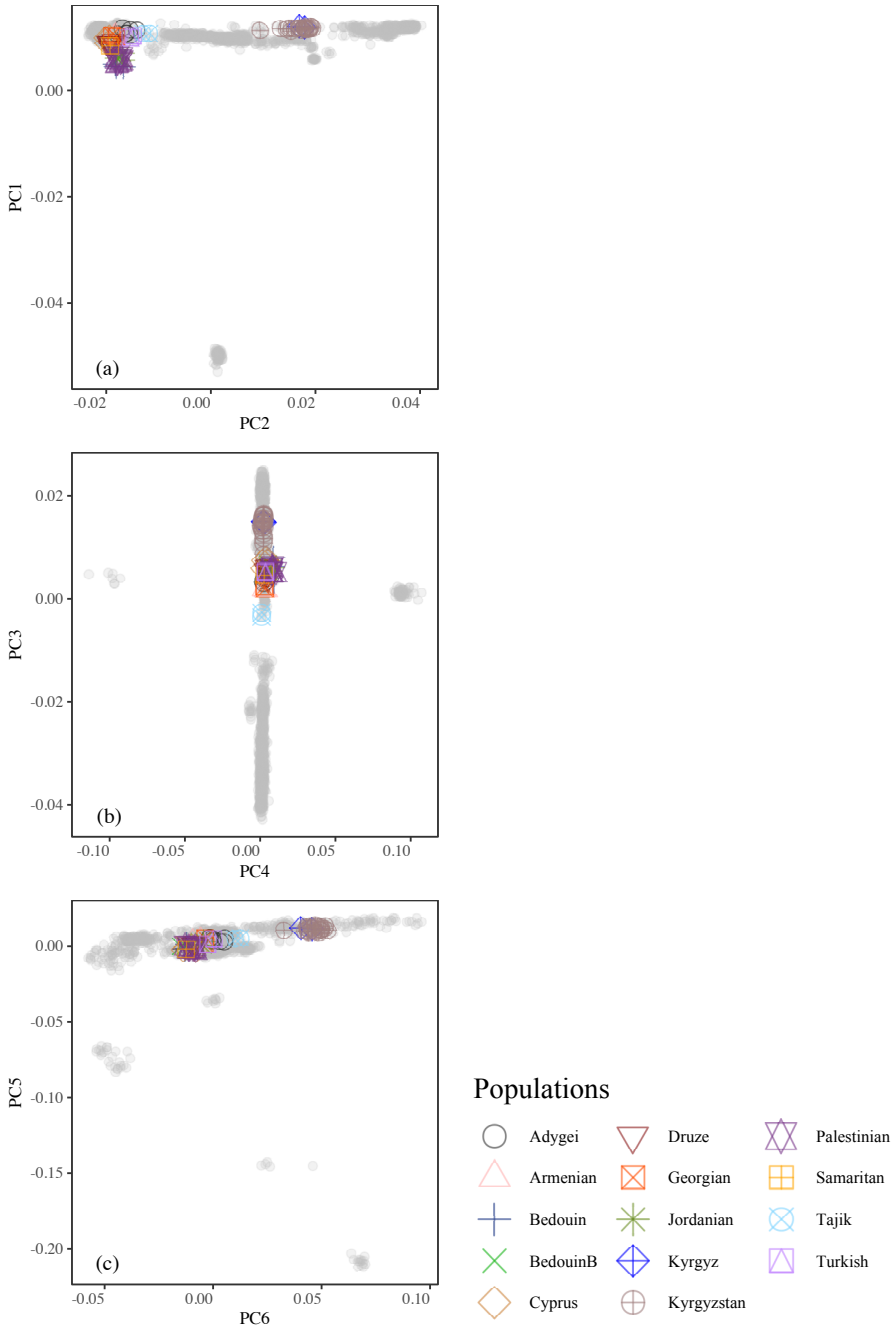


Figure A.5: Global scale PCA focused on Western and Central Asia. Sub-plots (a-c) show pairwise plots of PCs.Samples from the region are presented in colour according to *a priori* population labels while the remaining GR populations are represented by grey circles. . Variance explained by each PC shown in Figure A.2. The SAC individuals not shown. Explanations of abbreviations in Supp. Table A.1

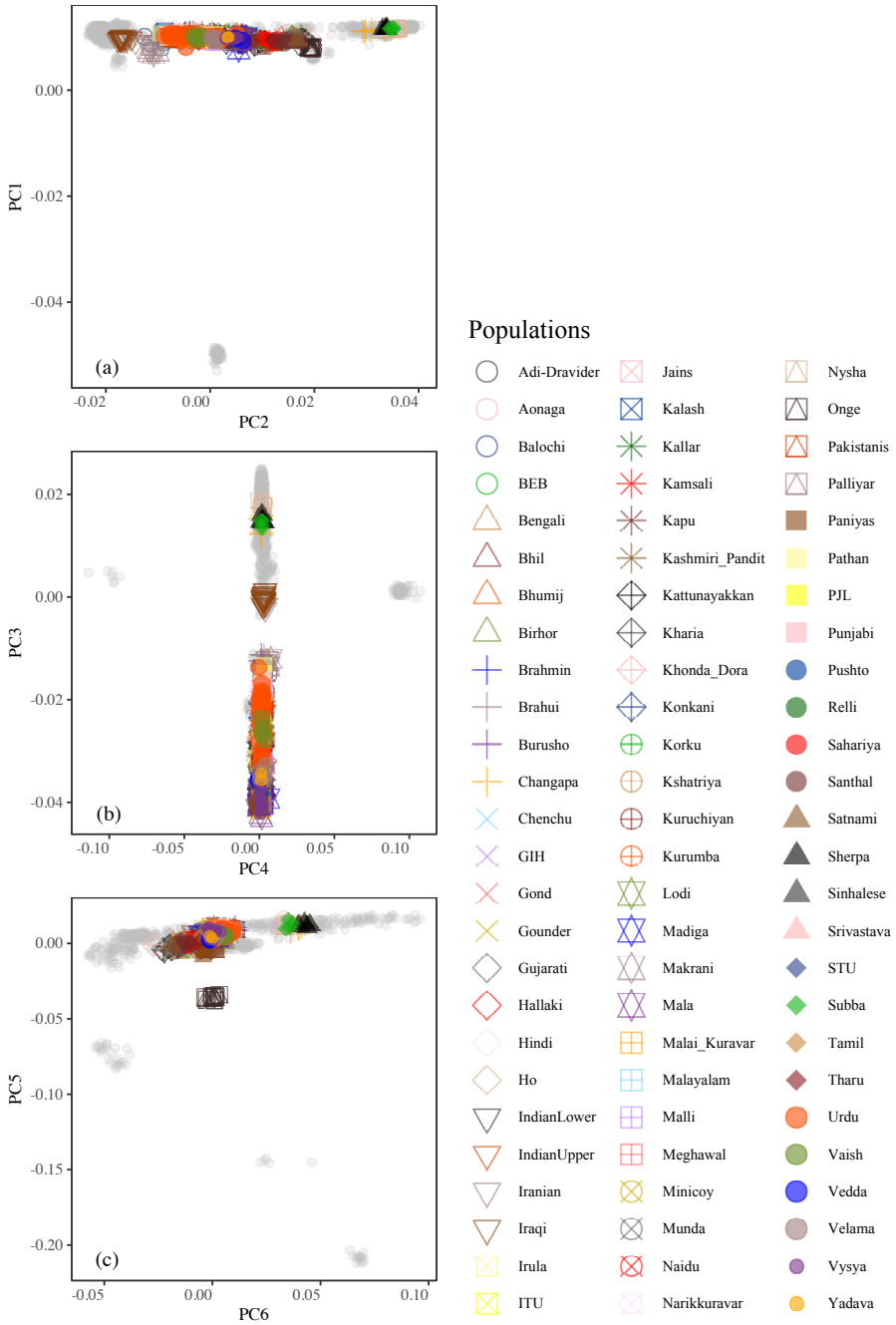


Figure A.6: Global scale PCA focused on Southern Asia. Sub-plots (a-c) show pairwise plots of PCs.Samples from the region are presented in colour according to *a priori* population labels while the remaining GR populations are represented by grey circles. . Variance explained by each PC shown in Figure A.2. The SAC individuals not shown. Explanations of abbreviations in Supp. Table A.1

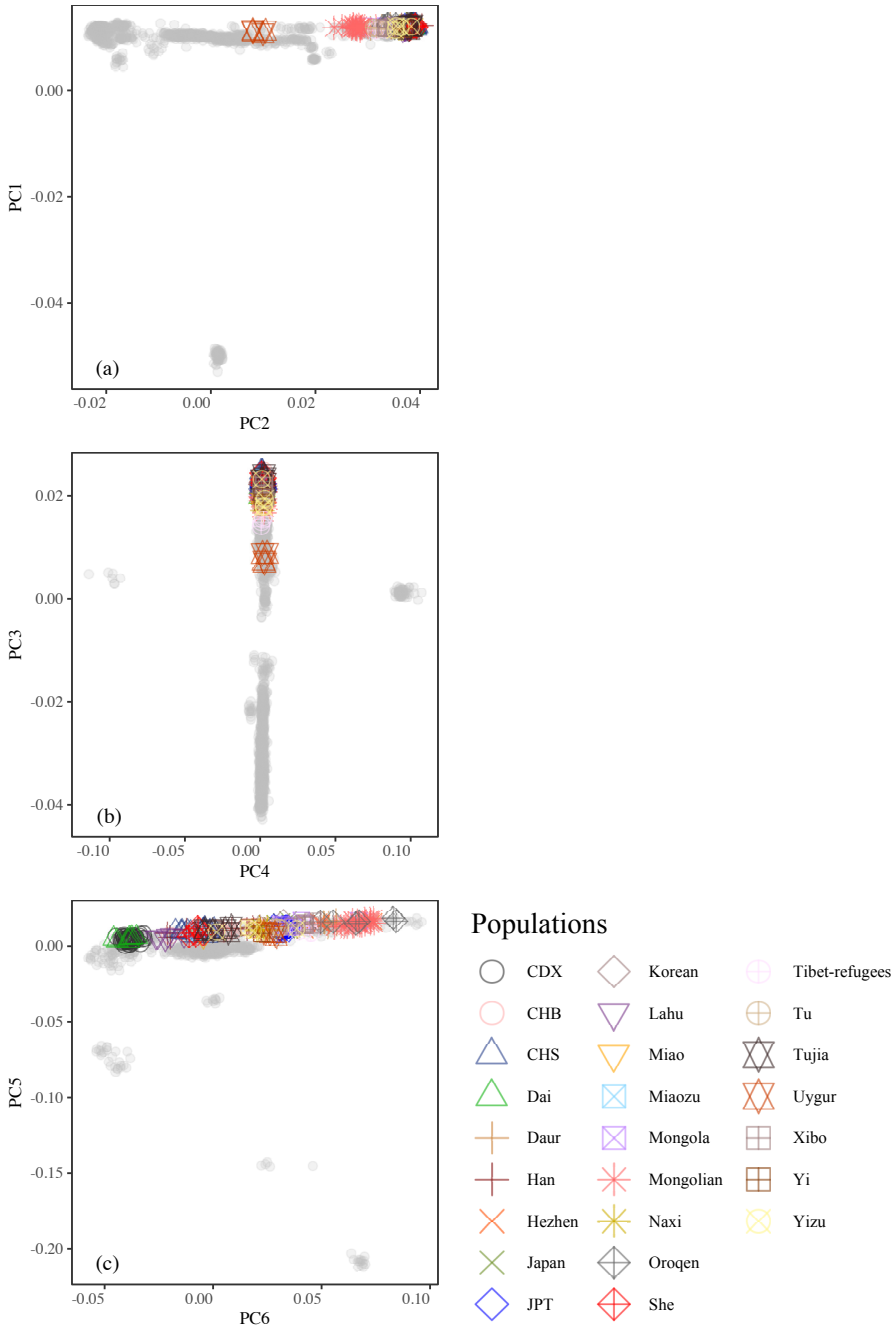


Figure A.7: Global scale PCA focused on Eastern Asia. Sub-plots (a-c) show pairwise plots of PCs.Samples from the region are presented in colour according to *a priori* population labels while the remaining GR populations are represented by grey circles. . Variance explained by each PC shown in Figure A.2. The SAC individuals not shown. Explanations of abbreviations in Supp. Table A.1

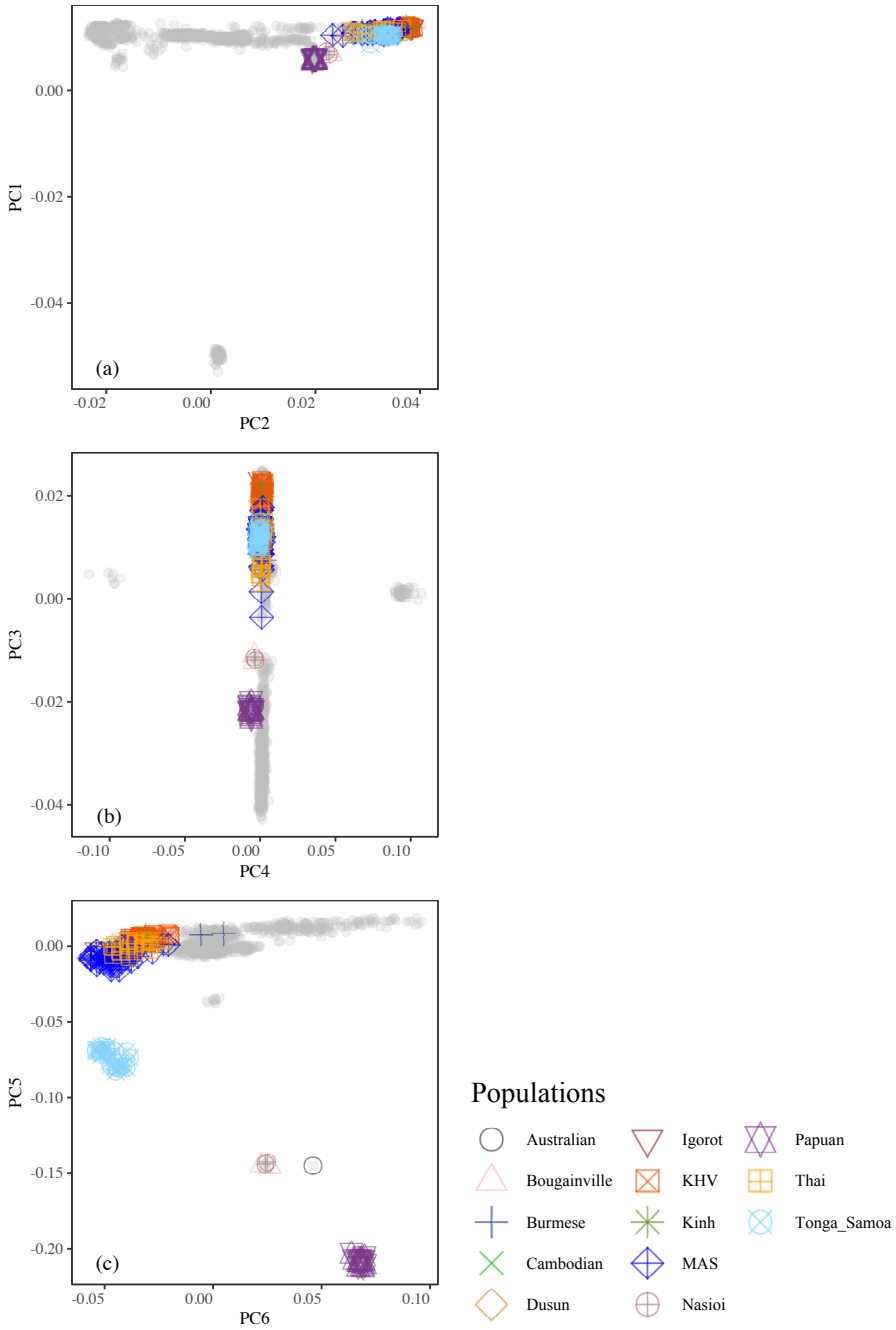


Figure A.8: Global scale PCA focused on South-East Asia and Oceania. Sub-plots (a-c) show pairwise plots of PCs.Samples from the region are presented in colour according to *a priori* population labels while the remaining GR populations are represented by grey circles. . Variance explained by each PC shown in Figure A.2. The SAC individuals not shown. Explanations of abbreviations in Supp. Table A.1

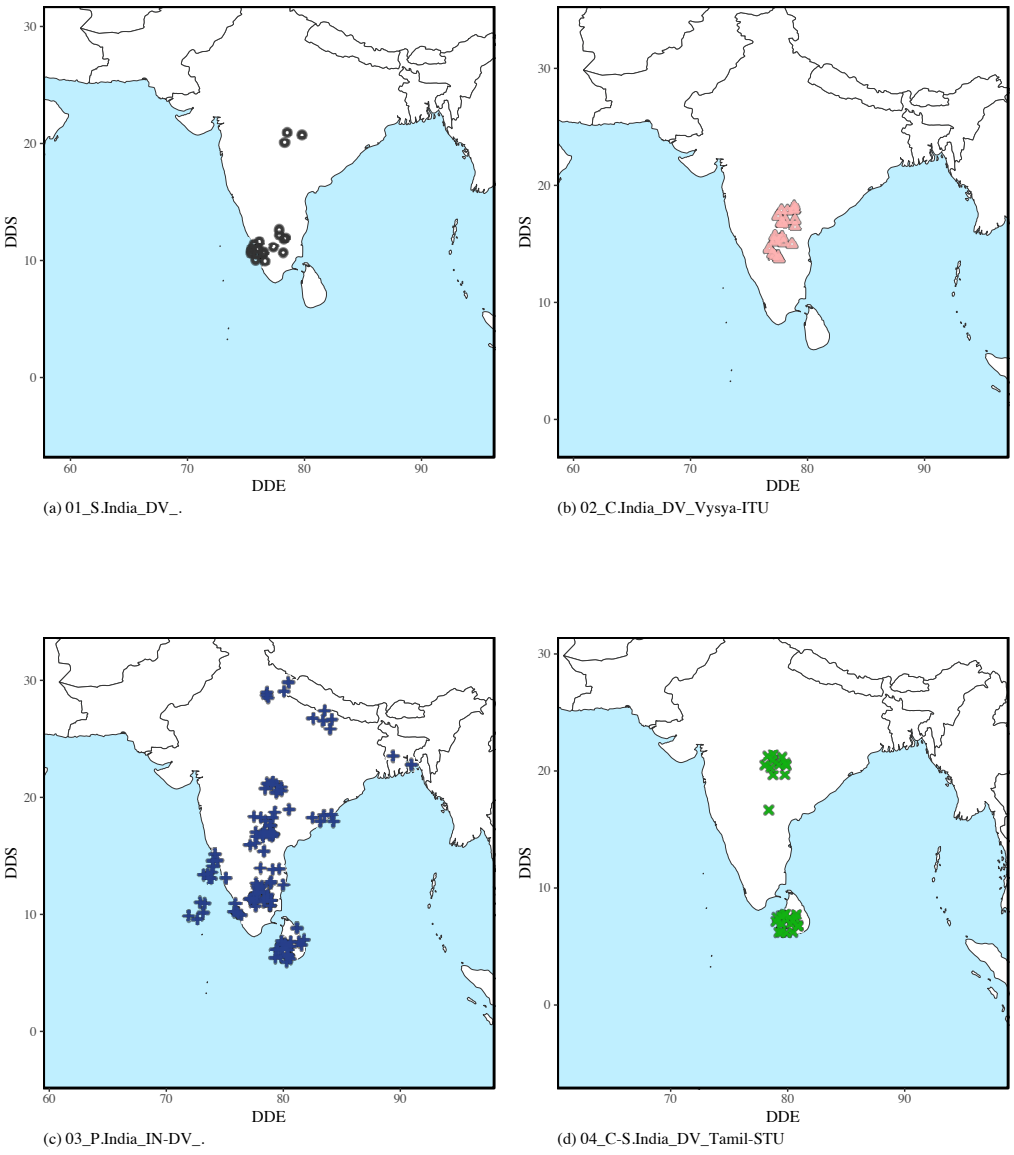


Figure A.9: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

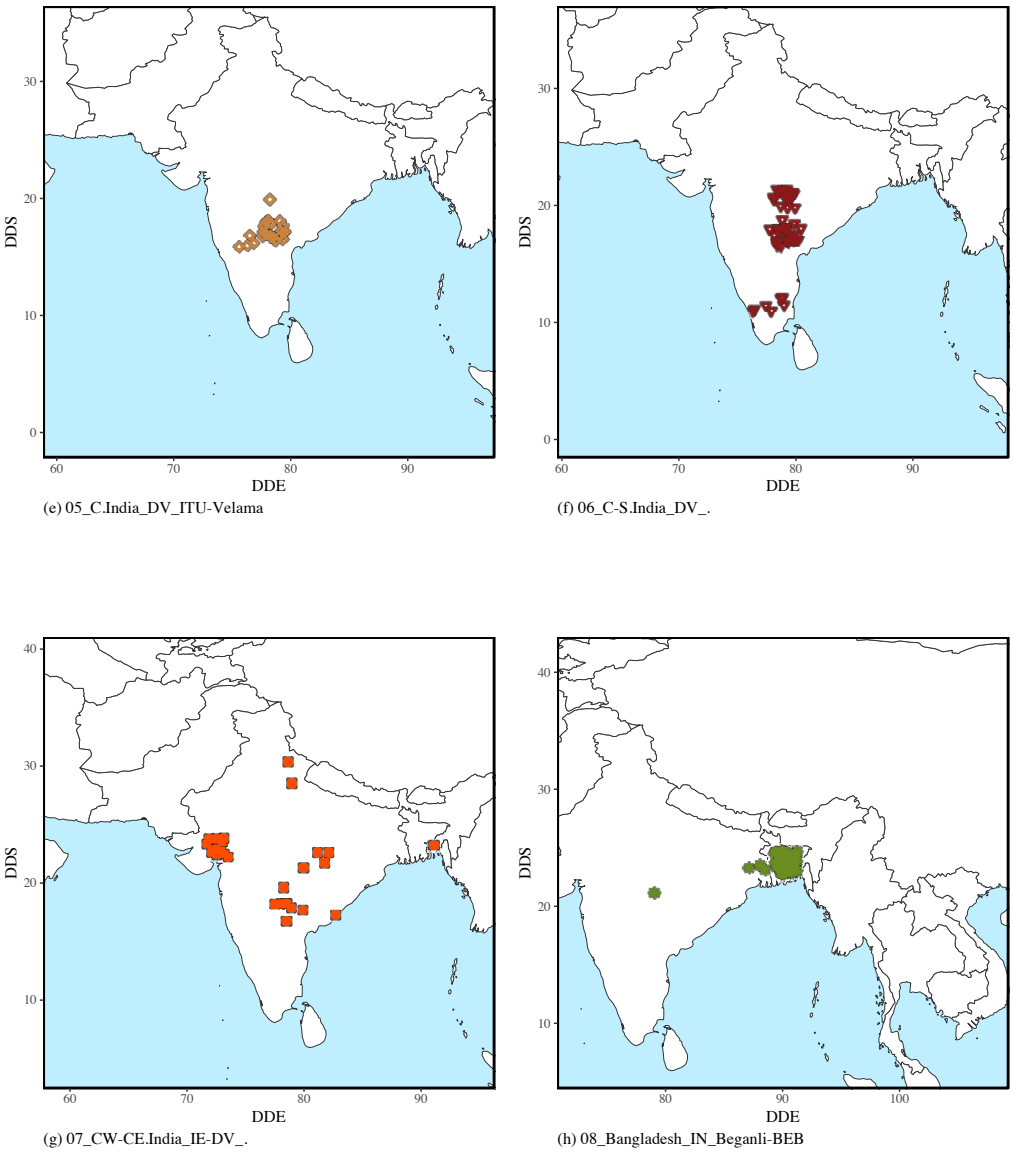


Figure A.10: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

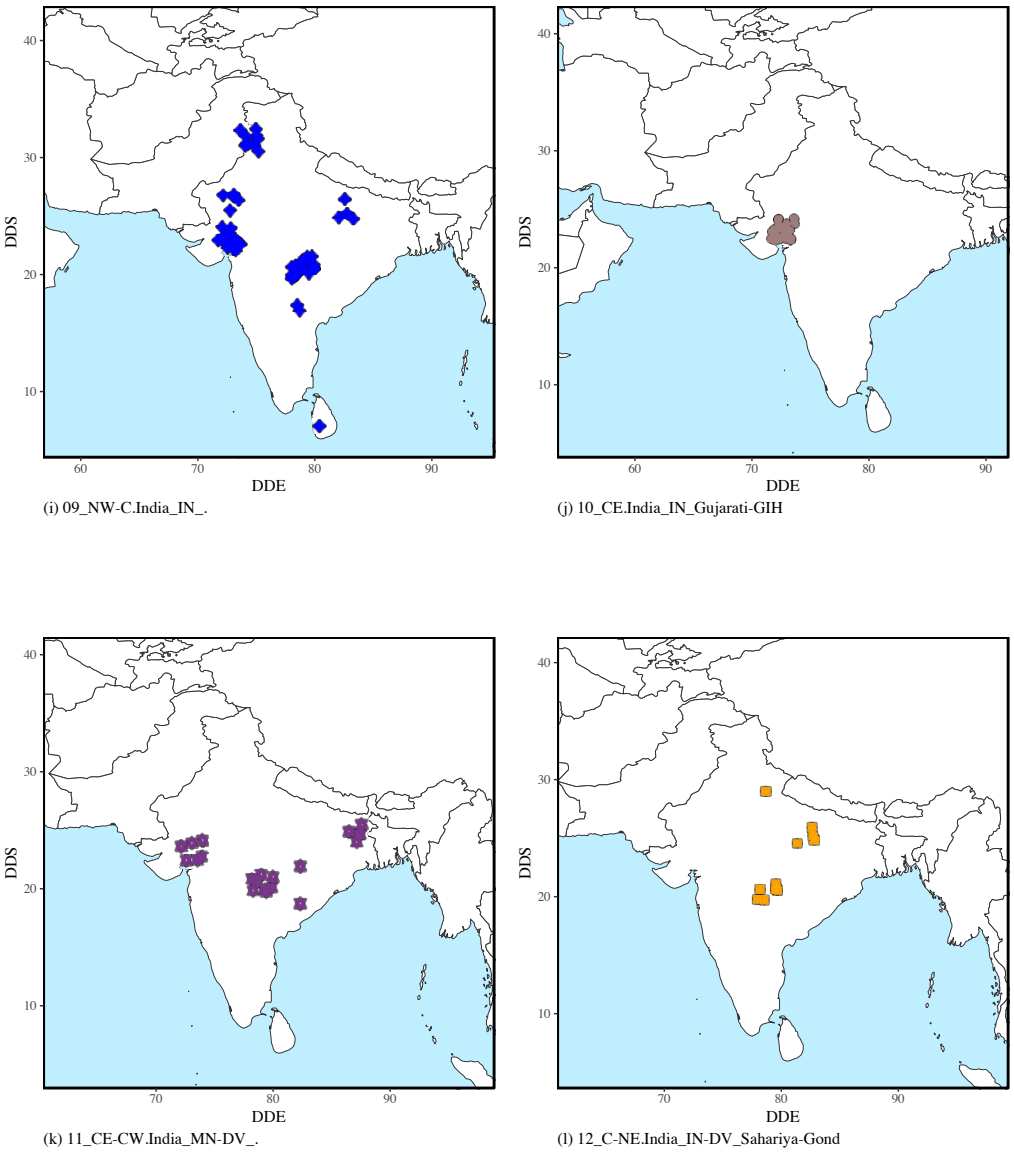


Figure A.11: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

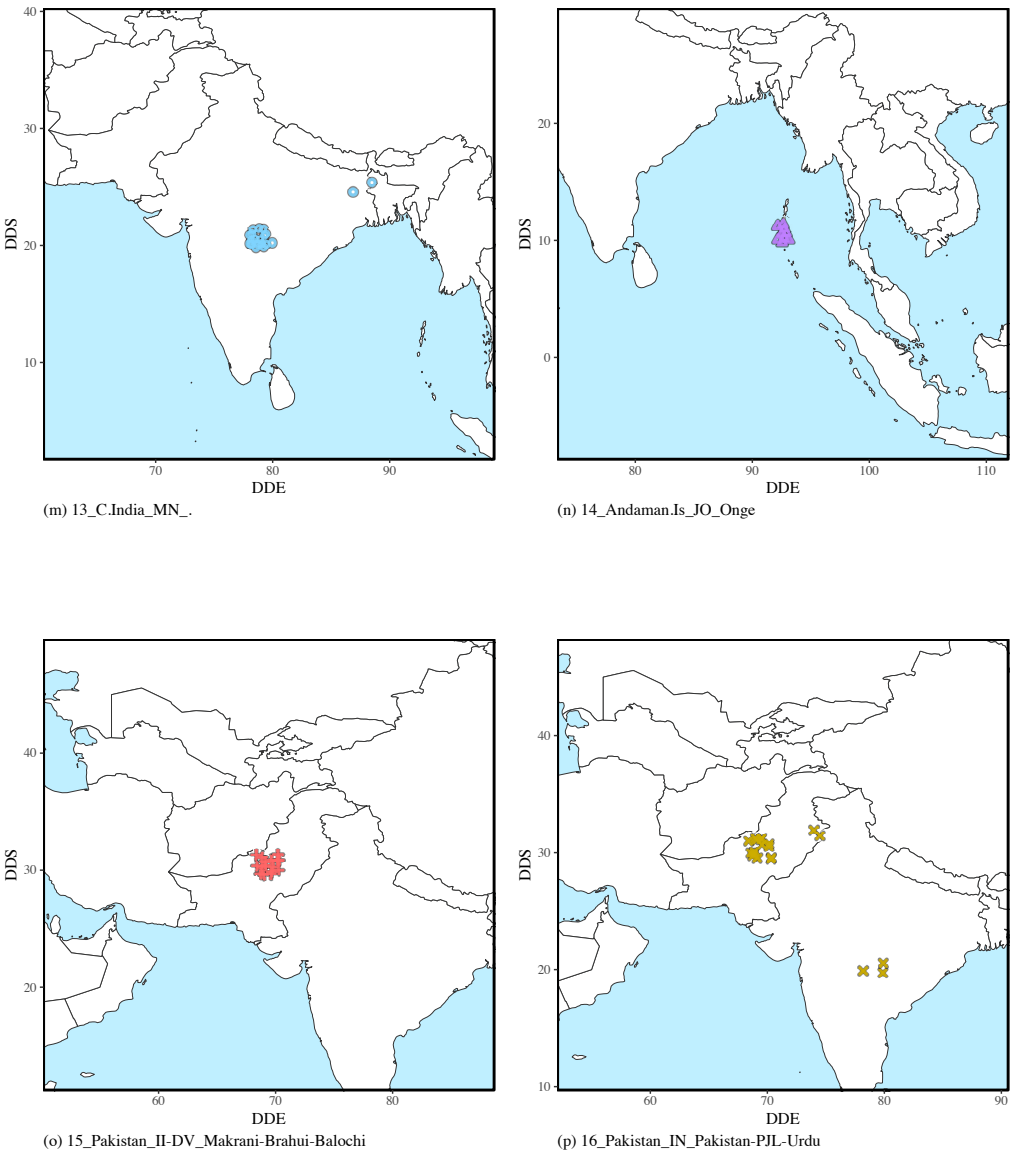


Figure A.12: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

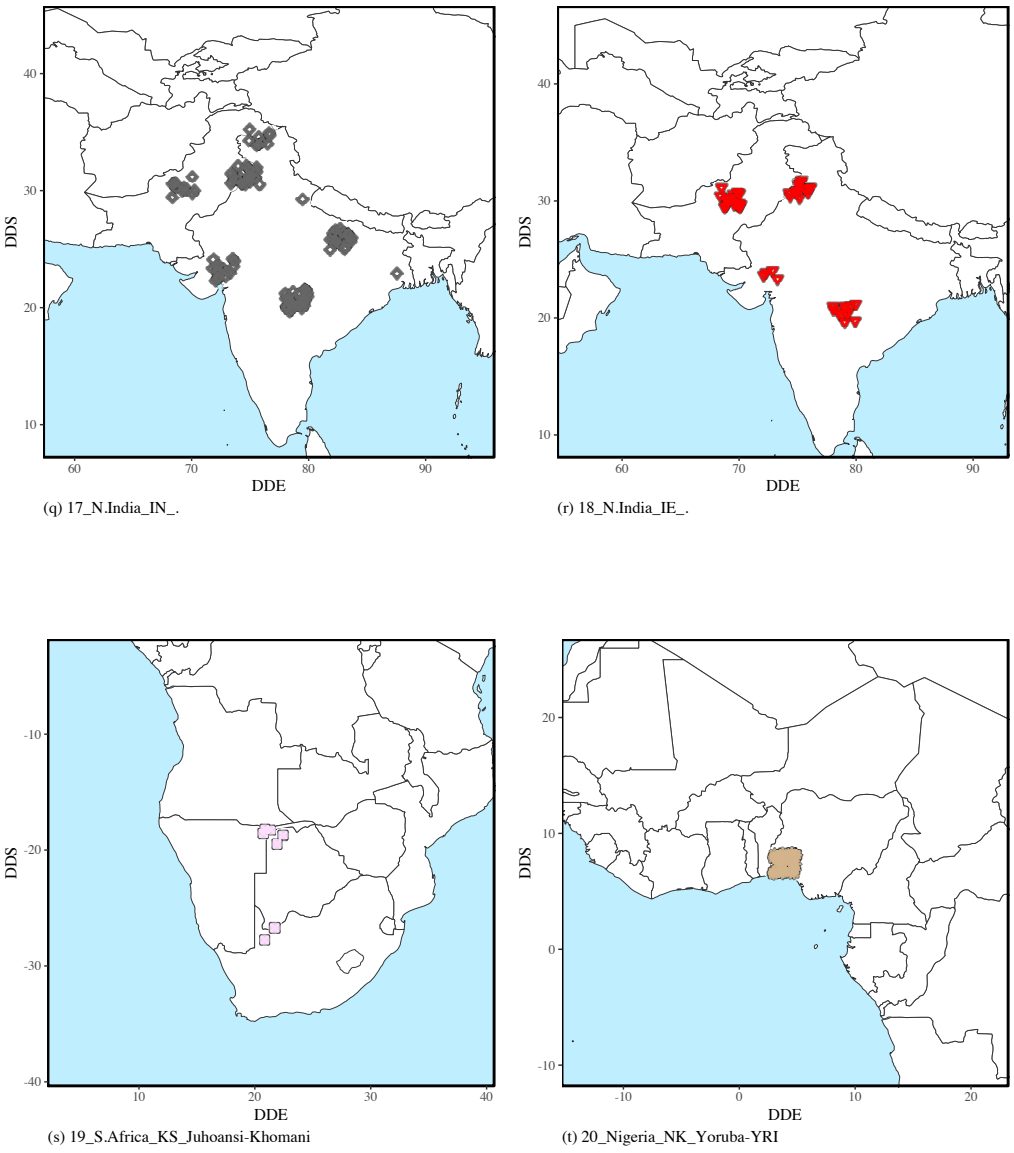


Figure A.13: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

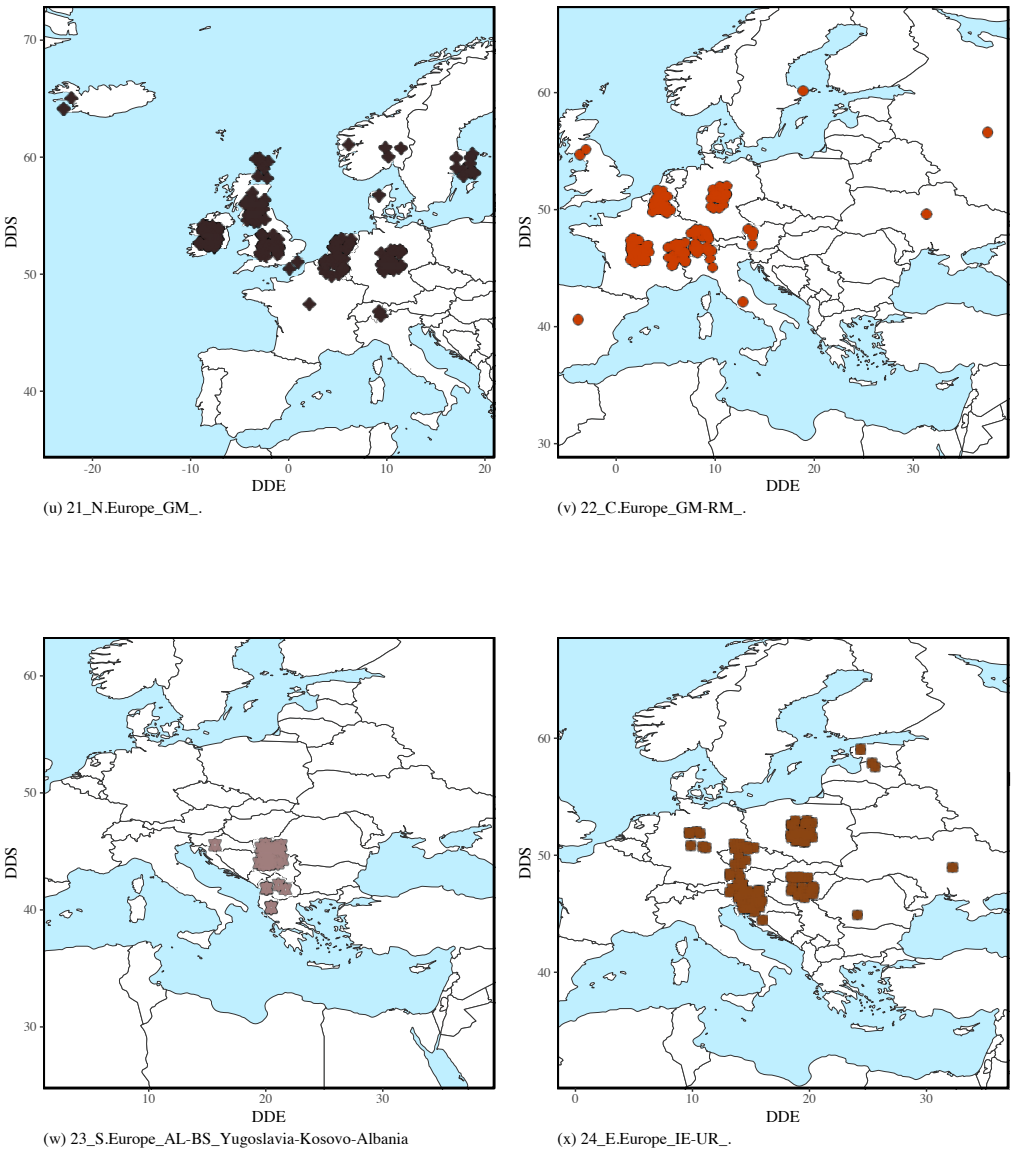


Figure A.14: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

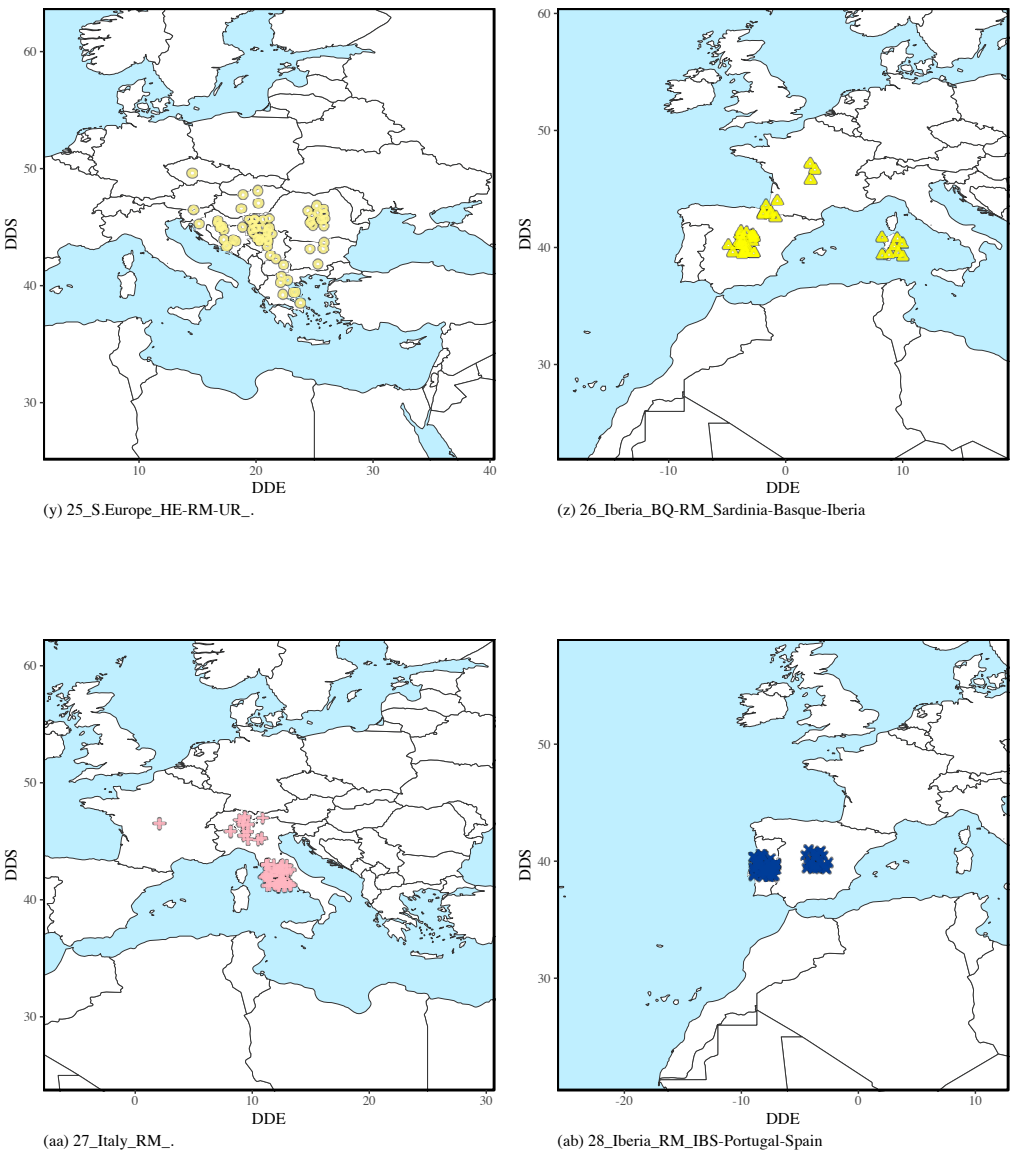


Figure A.15: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

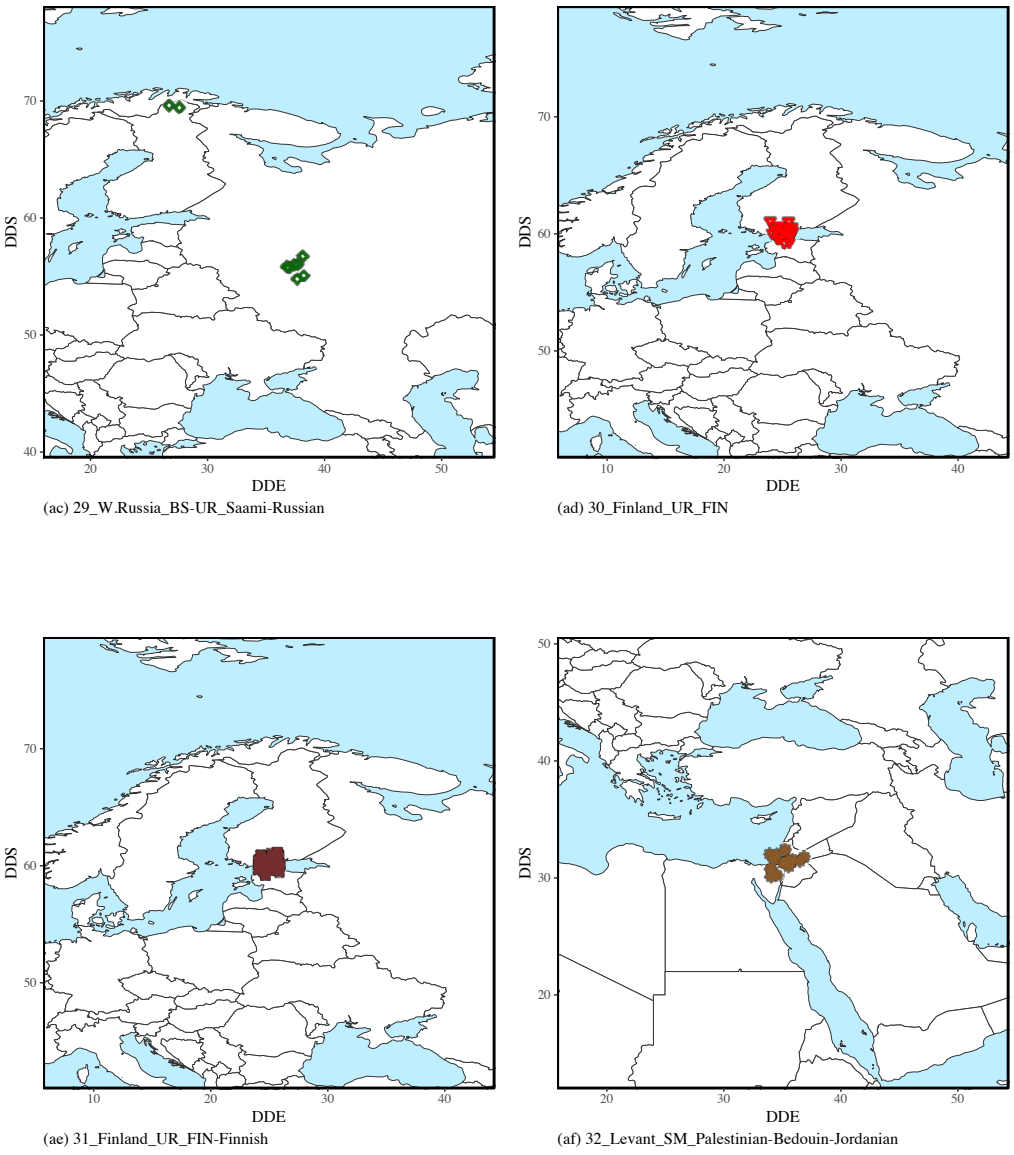


Figure A.16: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

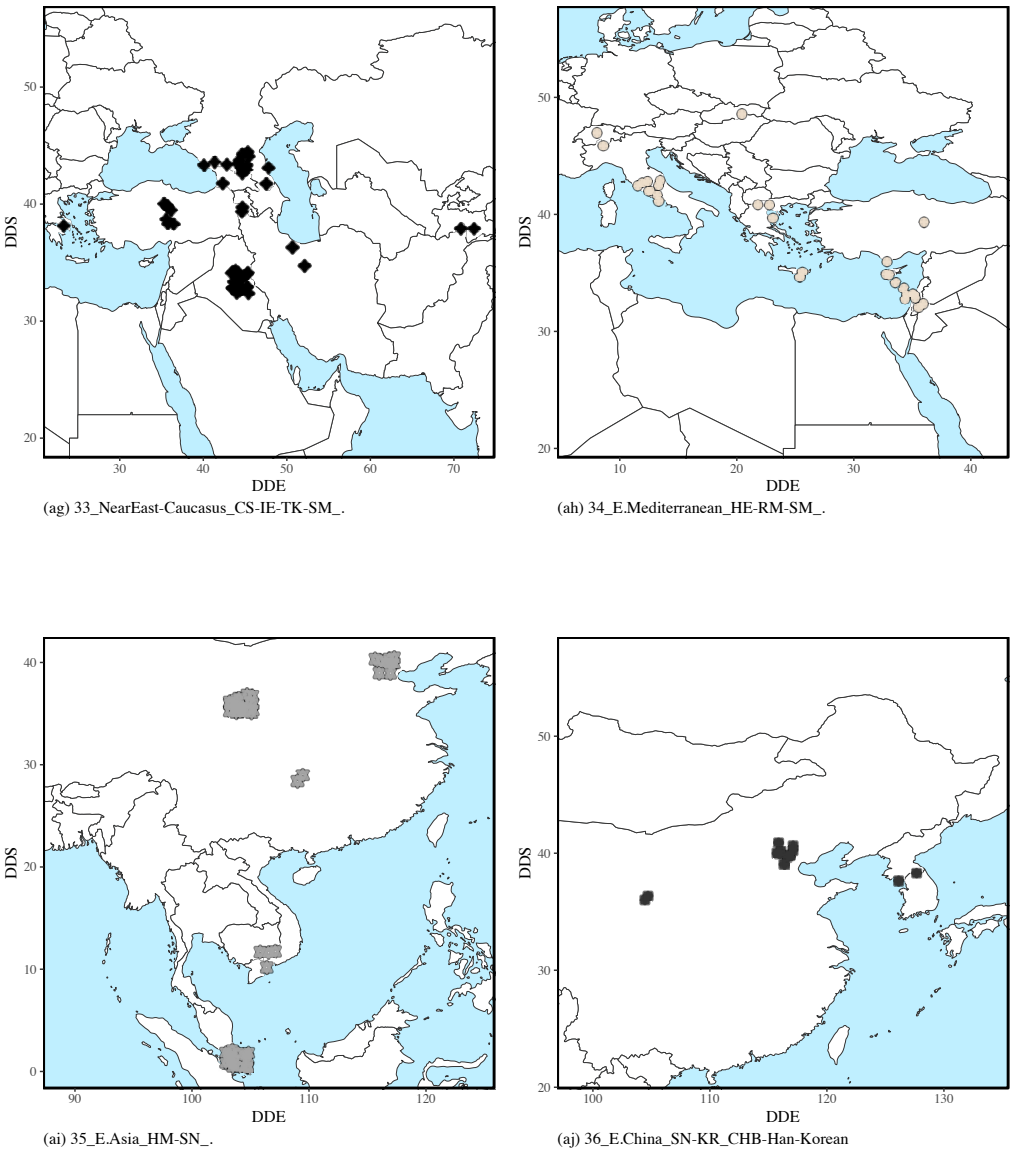


Figure A.17: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

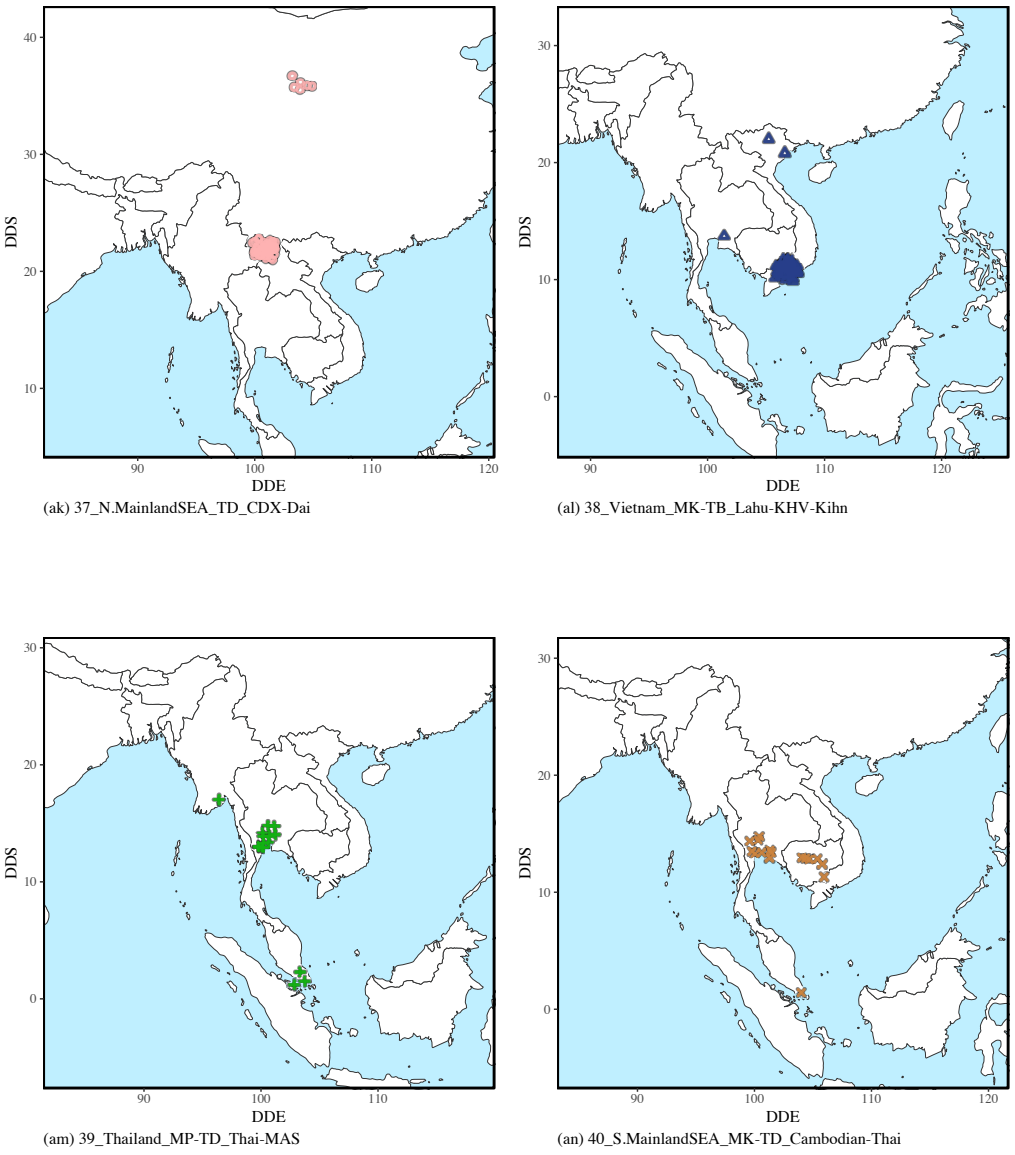


Figure A.18: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

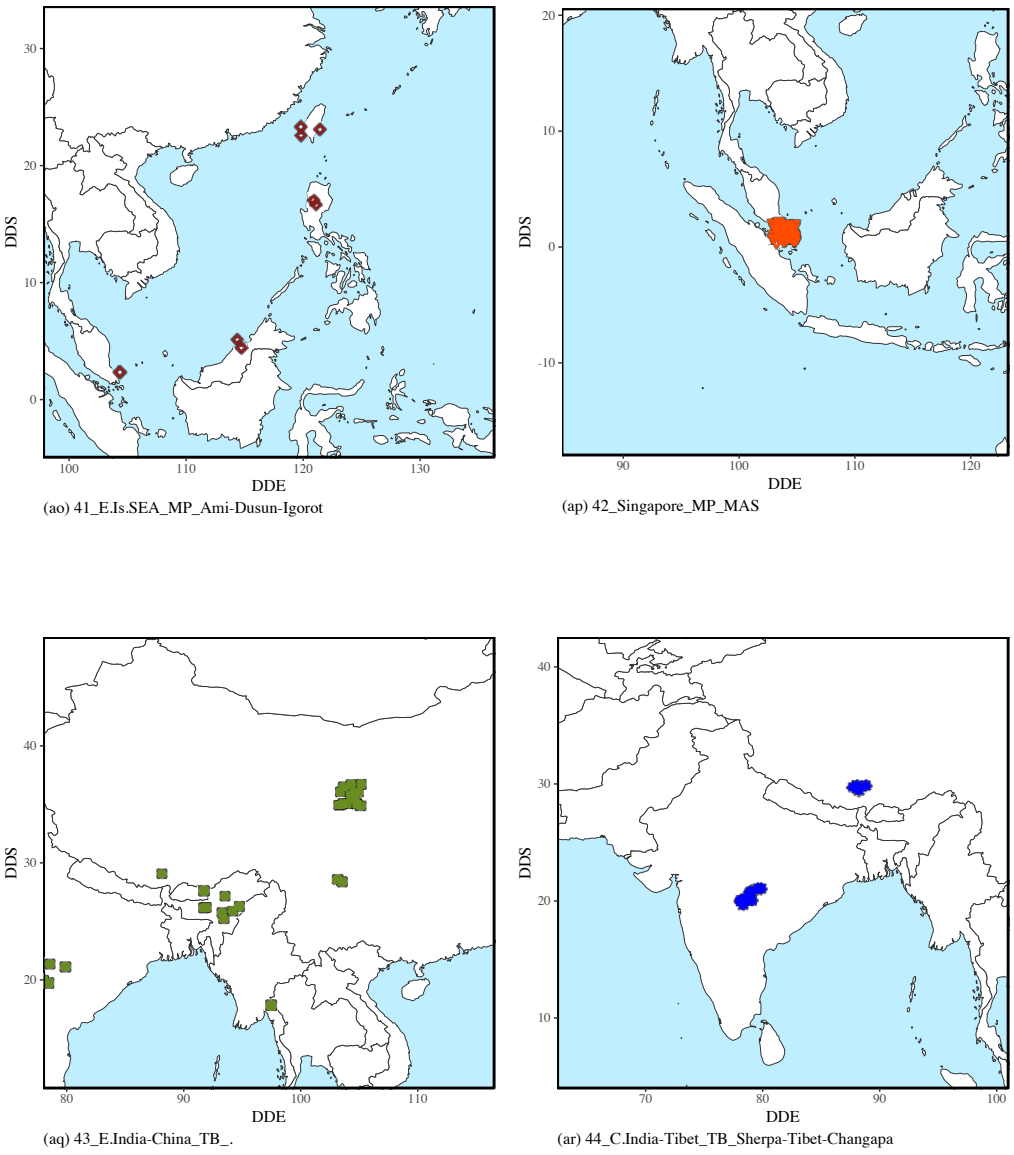


Figure A.19: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

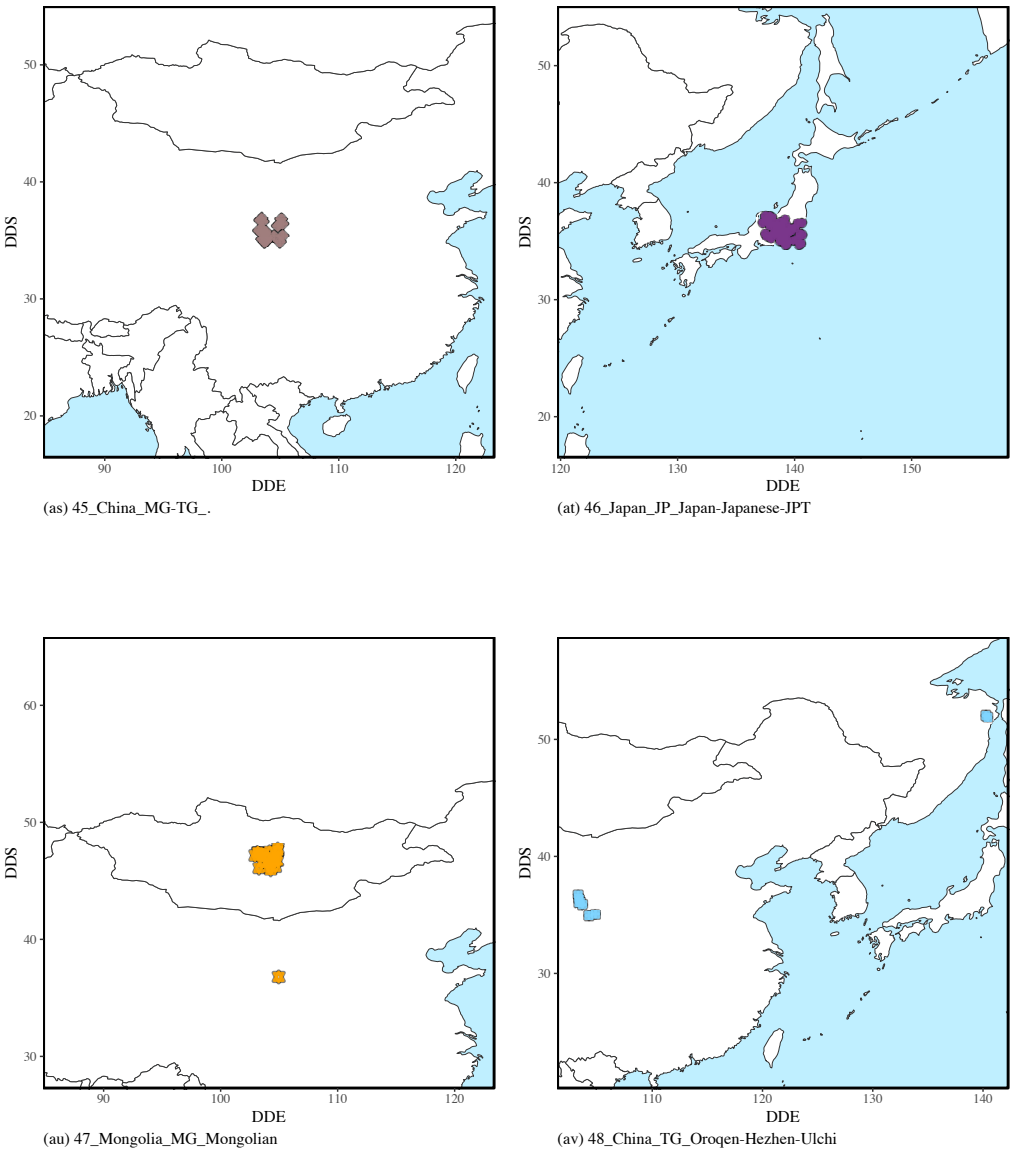


Figure A.20: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

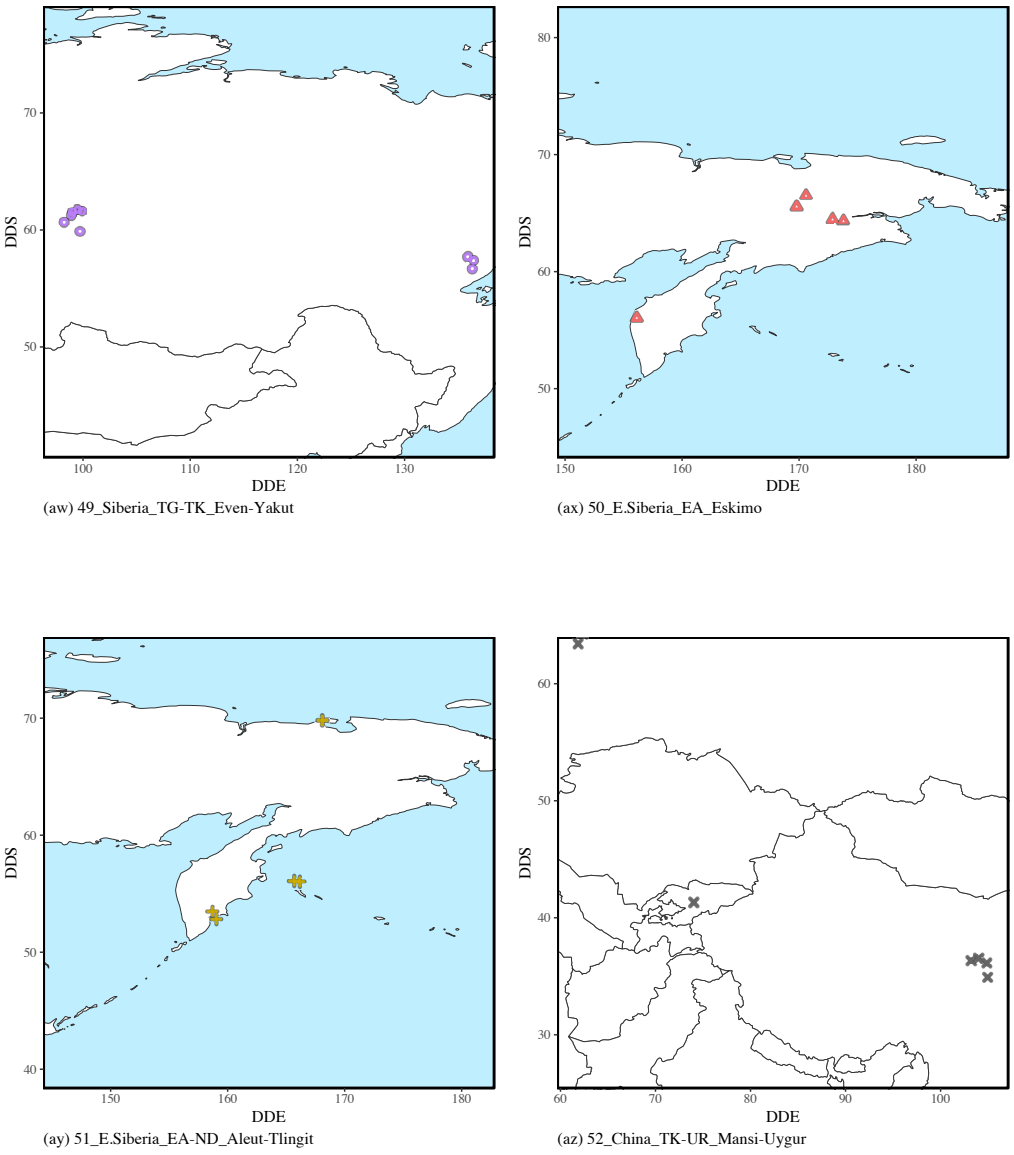


Figure A.21: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

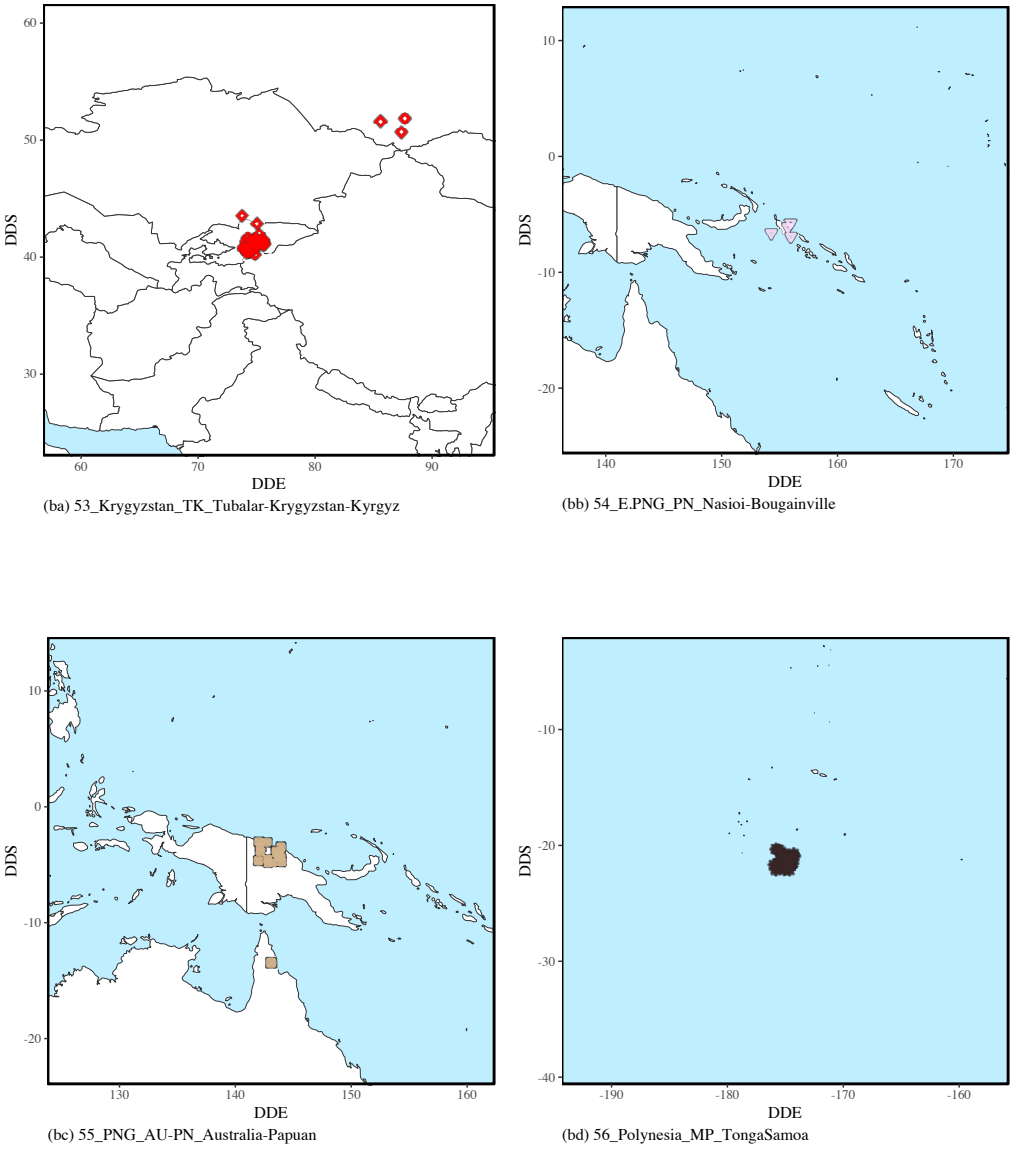


Figure A.22: Global distribution of the FineSTRUCTURE inferred clusters for the Global Reference (GR) data. Subplots are of individual clusters as labelled. Plotted points have been jittered to aid visualisation and symbols match Figure 4.14. Abbr. DDS/E - Decimal Degrees South/East

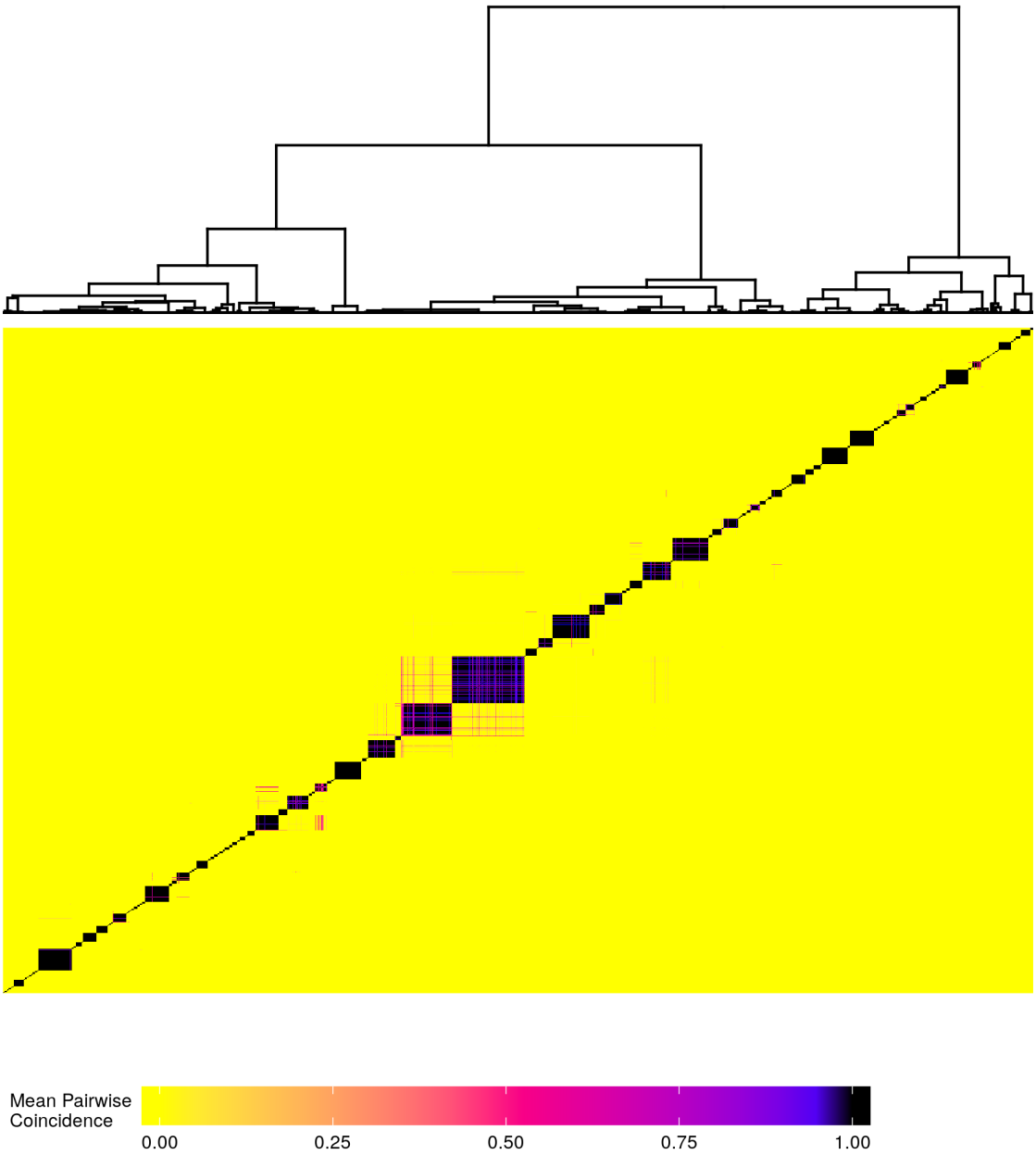


Figure A.23: FineSTRUCTURE inferred maximum concordance trees for CP-fS of Global Reference populations (chain 0). Shown are the associated heatmaps of pairwise coincidence values of individuals as a proportion of the observations in the MCMC runs.

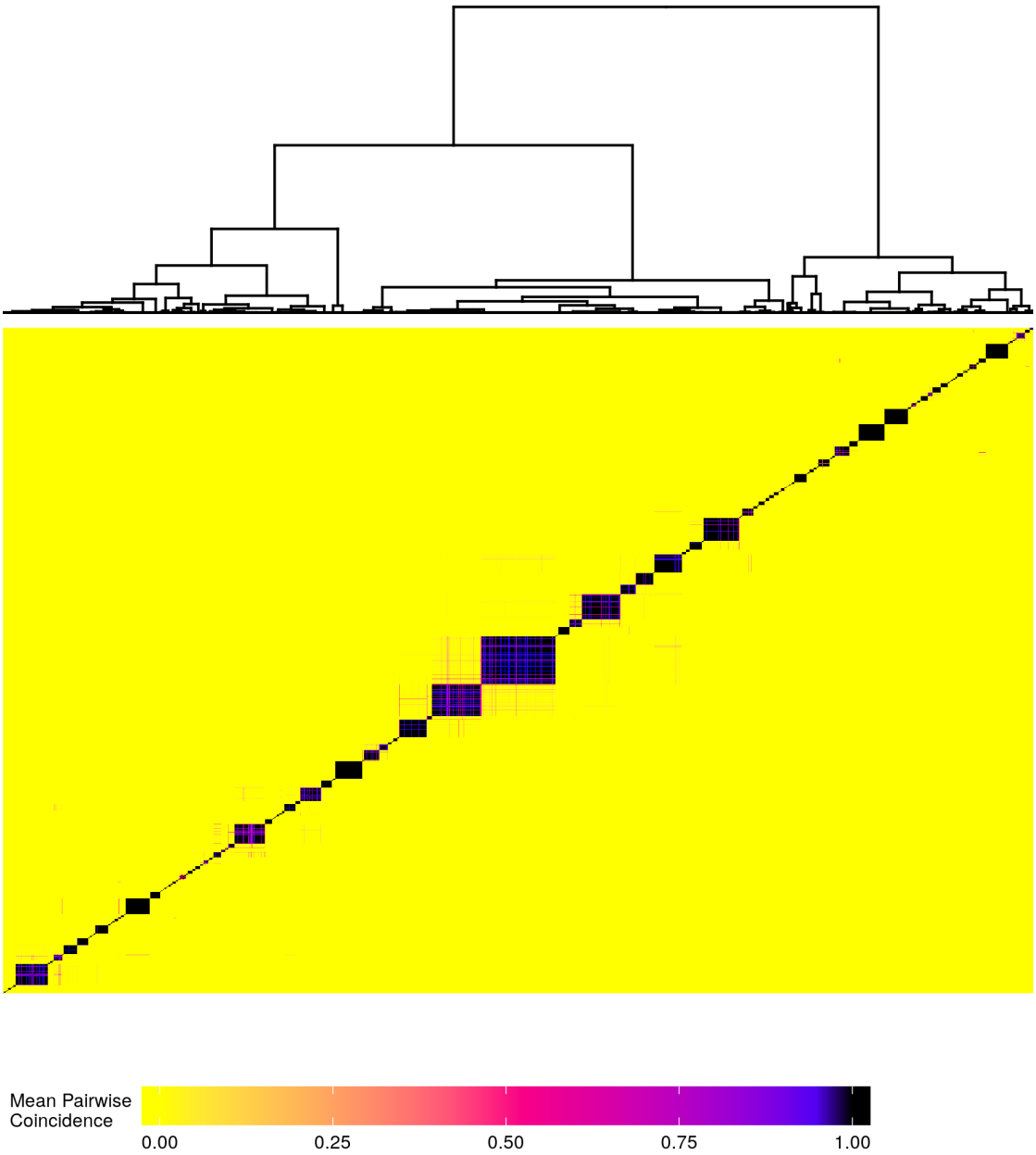


Figure A.24: FineSTRUCTURE inferred maximum concordance trees for CP-fS of Global Reference populations (chain 1). Shown are the associated heatmaps of pairwise coincidence values of individuals as a proportion of the observations in the MCMC runs.

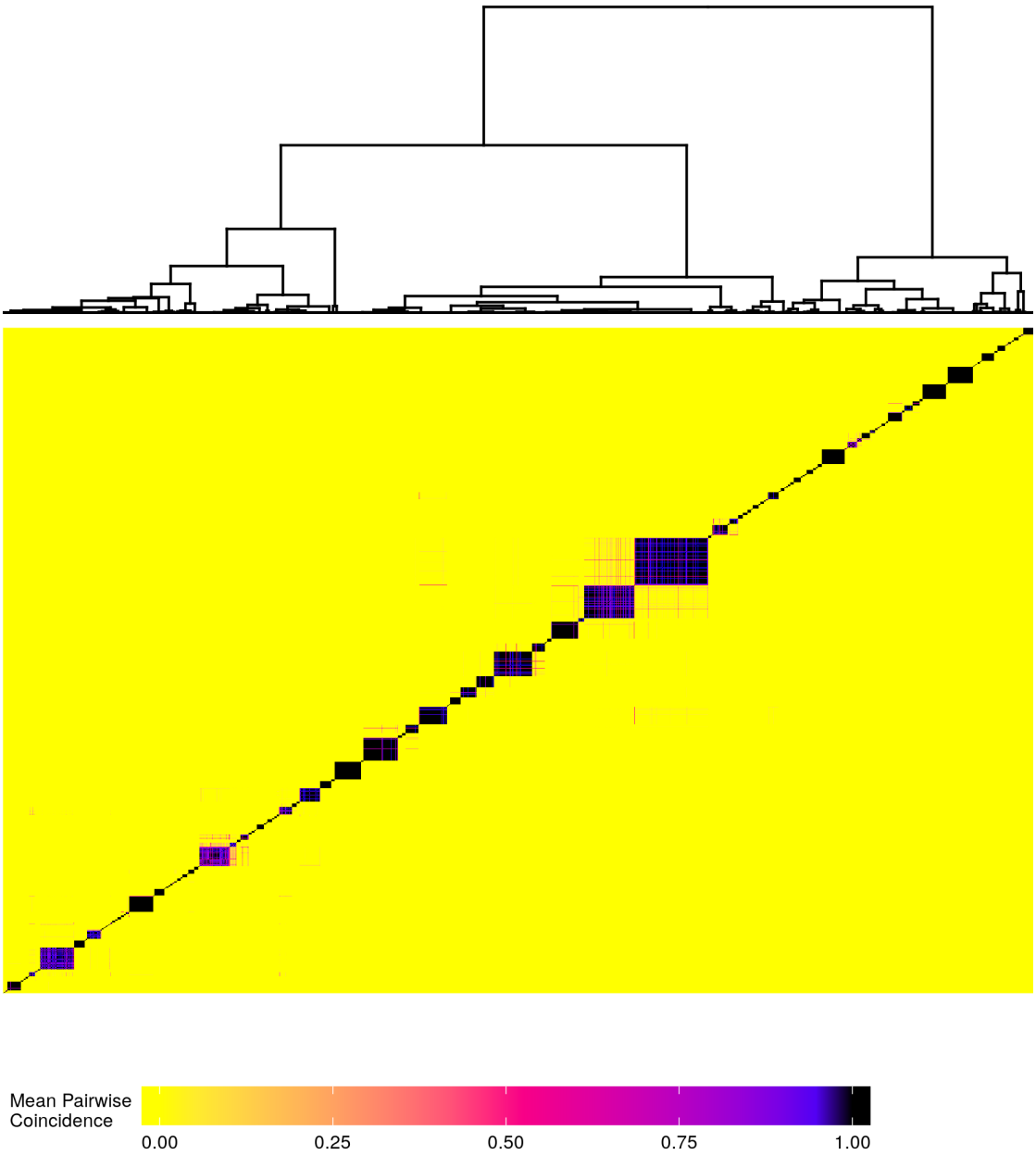


Figure A.25: FineSTRUCTURE inferred maximum concordance trees for CP-fS of Global Reference populations (chain 2). Shown are the associated heatmaps of pairwise coincidence values of individuals as a proportion of the observations in the MCMC runs.

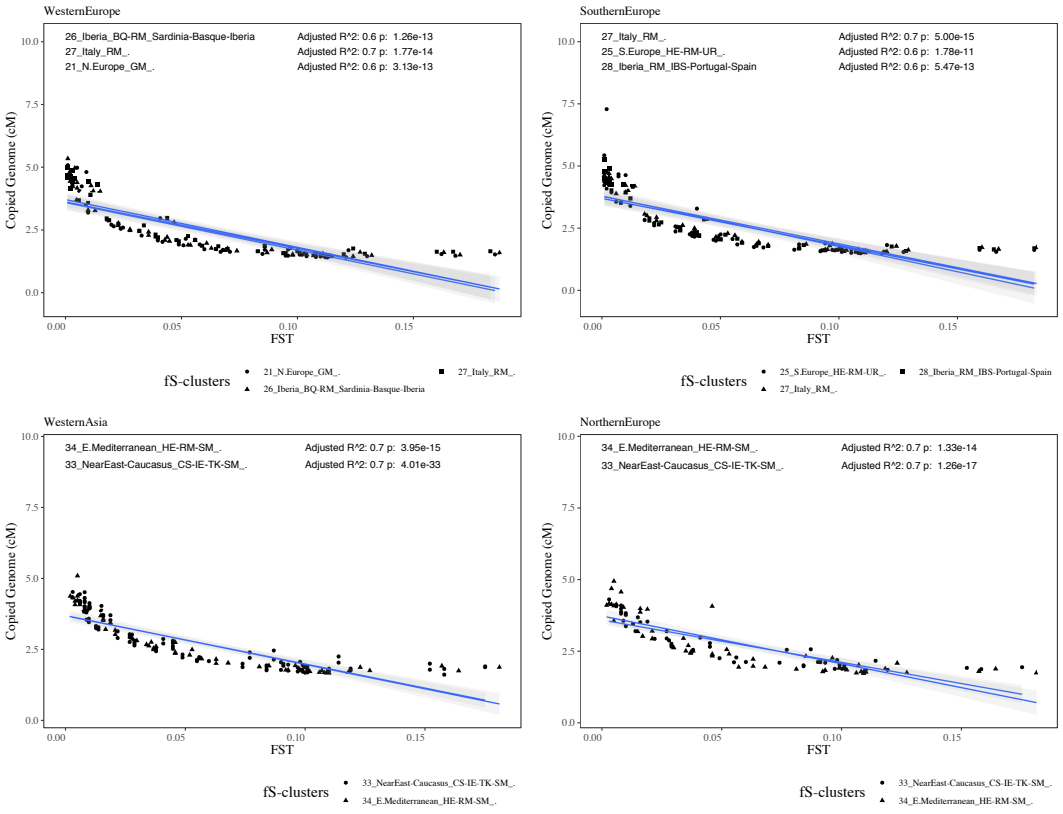


Figure A.26: Correlation between total genome length copied and the F_{ST} distances. Plots highlight informativeness of the "genome copied" values as estimated from CP for recipient GR individuals copying from fS-inferred clusters. In each plot a randomly chosen set of recipient clusters are shown (max 3) for each global region based on the distribution of the samples. All copying from African groups were excluded from the correlation.

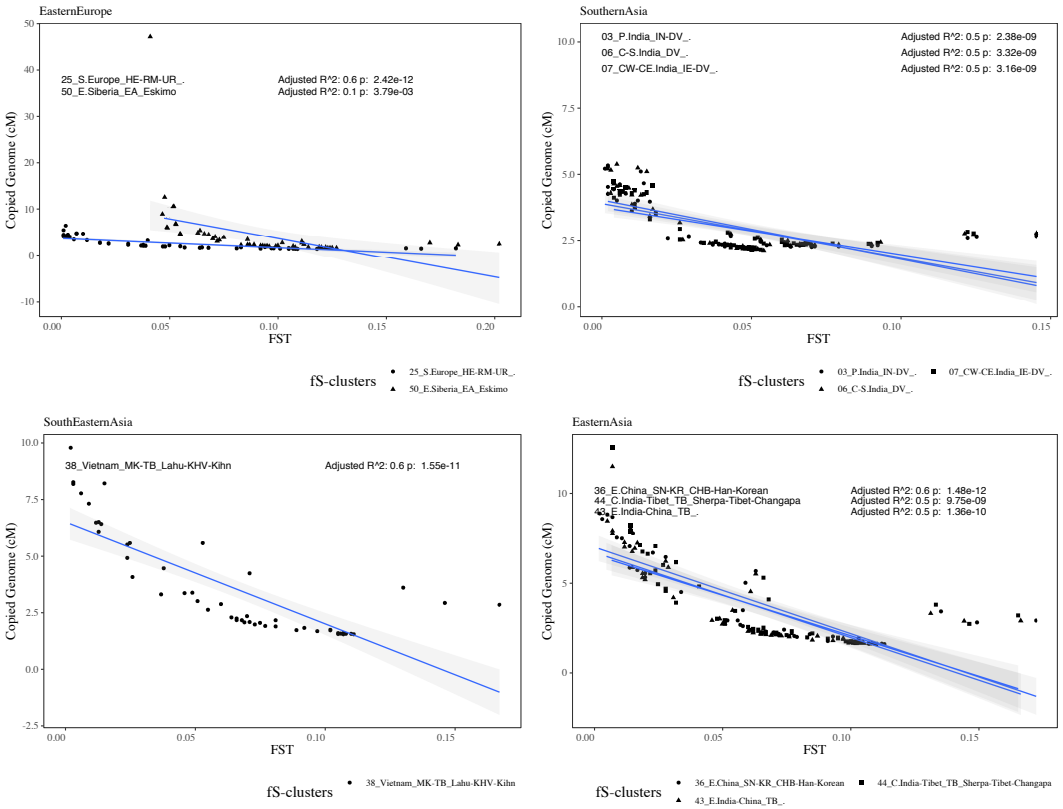


Figure A.27: Correlation between total genome length copied and the F_{ST} distances (continued from previous plot ...) Plots highlight informativeness of the "genome copied" values as estimated from CP for recipient GR individuals copying from fS-inferred clusters. In each plot a randomly chosen set of recipient clusters are shown (max 3) for each global region based on the distribution of the samples. All copying from African groups were excluded from the correlation.

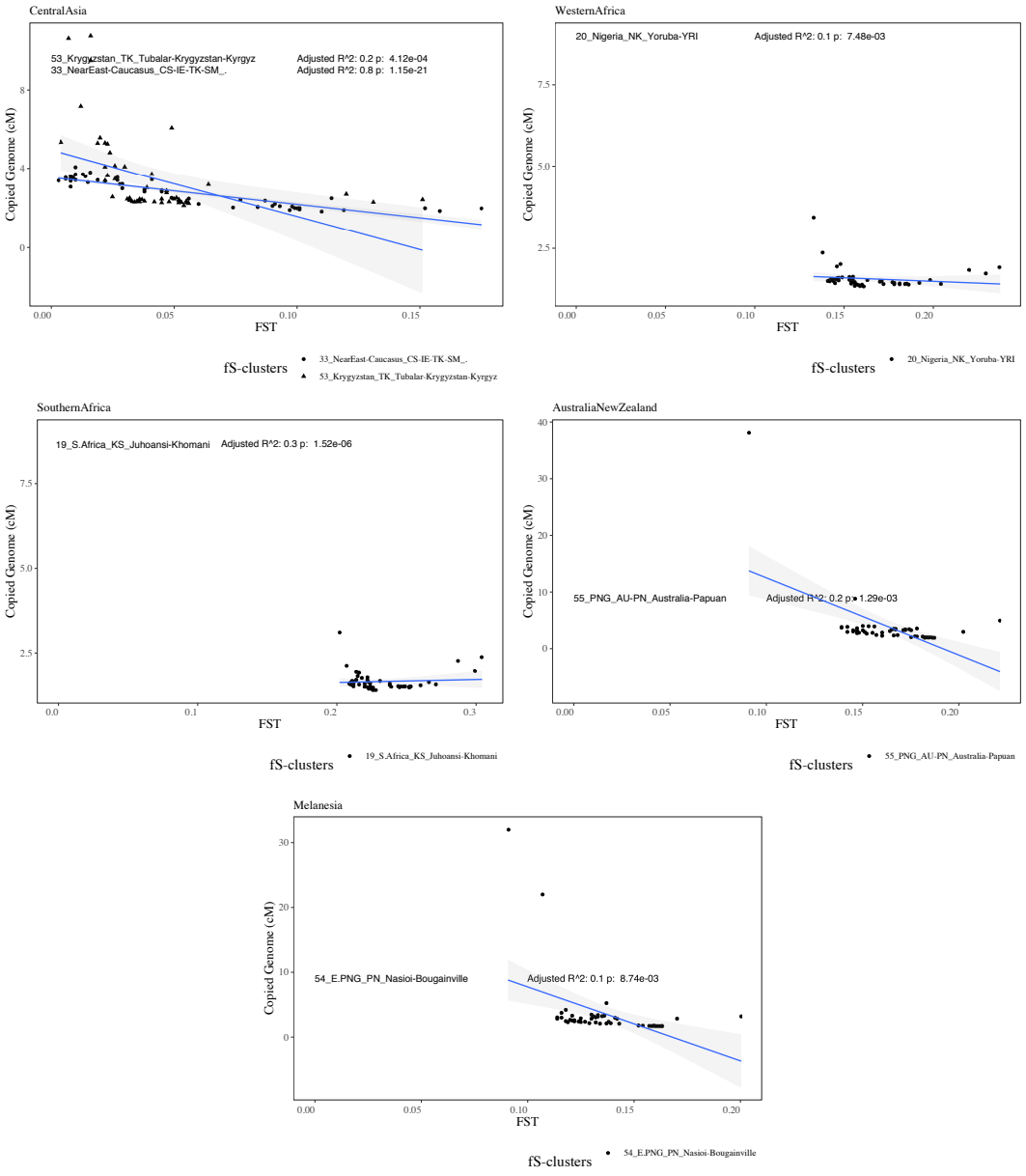


Figure A.28: Correlation between total genome length copied and the F_{ST} distances (continued from previous plot ...) Plots highlight informativeness of the "genome copied" values as estimated from CP for recipient GR individuals copying from fS-inferred clusters. In each plot a randomly chosen set of recipient clusters are shown (max 3) for each global region based on the distribution of the samples. All copying from African groups were excluded from the correlation.

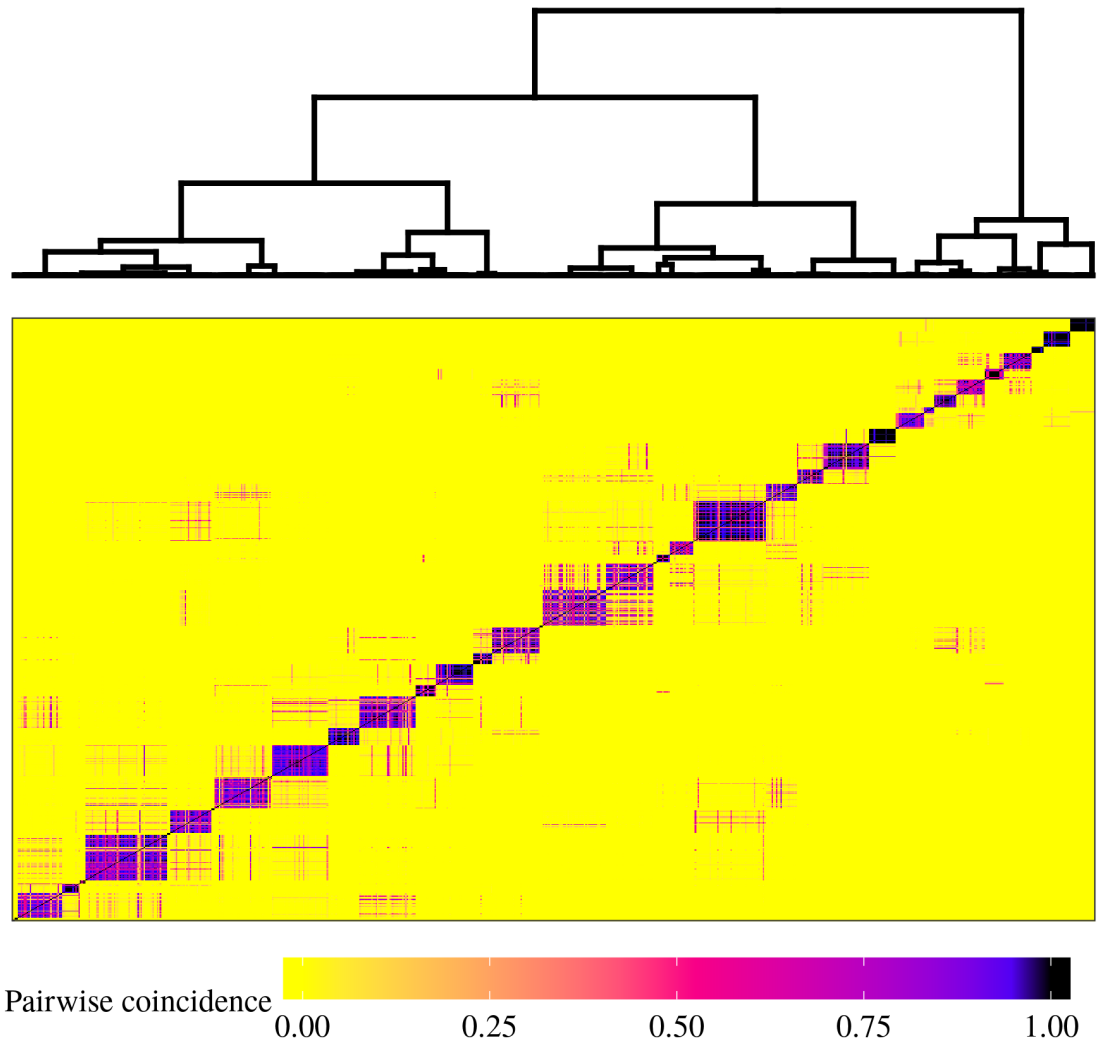


Figure A.29: FineSTRUCTURE inferred Maximum concordance trees for the CP output (chain 0) of the SAC dataset. Shown are the associated heatmaps of pairwise coincidence values of individuals as a proportion of the observations in the MCMC runs.

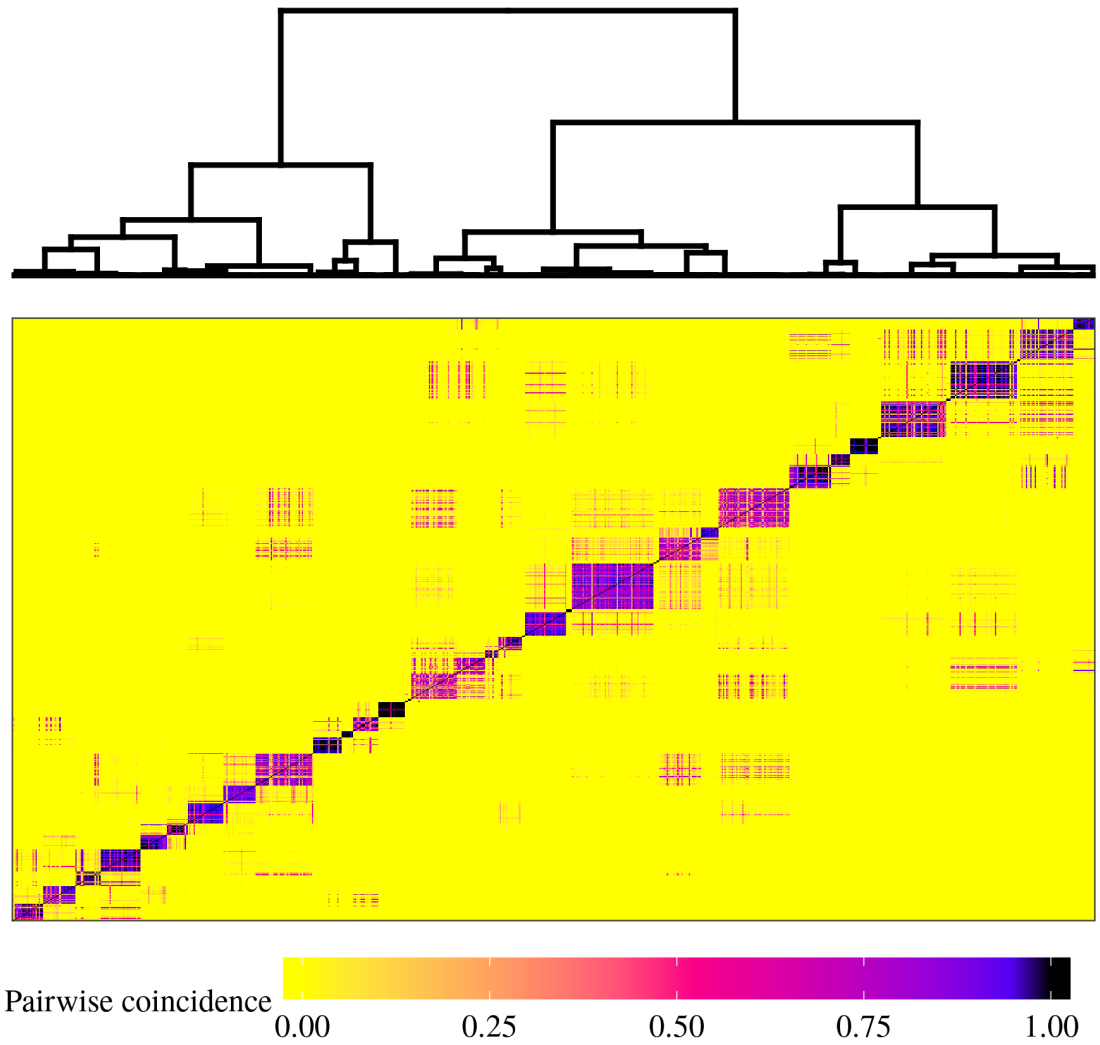


Figure A.30: FineSTRUCTURE inferred Maximum concordance trees for the CP output (chain 1) of the SAC dataset. Shown are the associated heatmaps of pairwise coincidence values of individuals as a proportion of the observations in the MCMC runs.

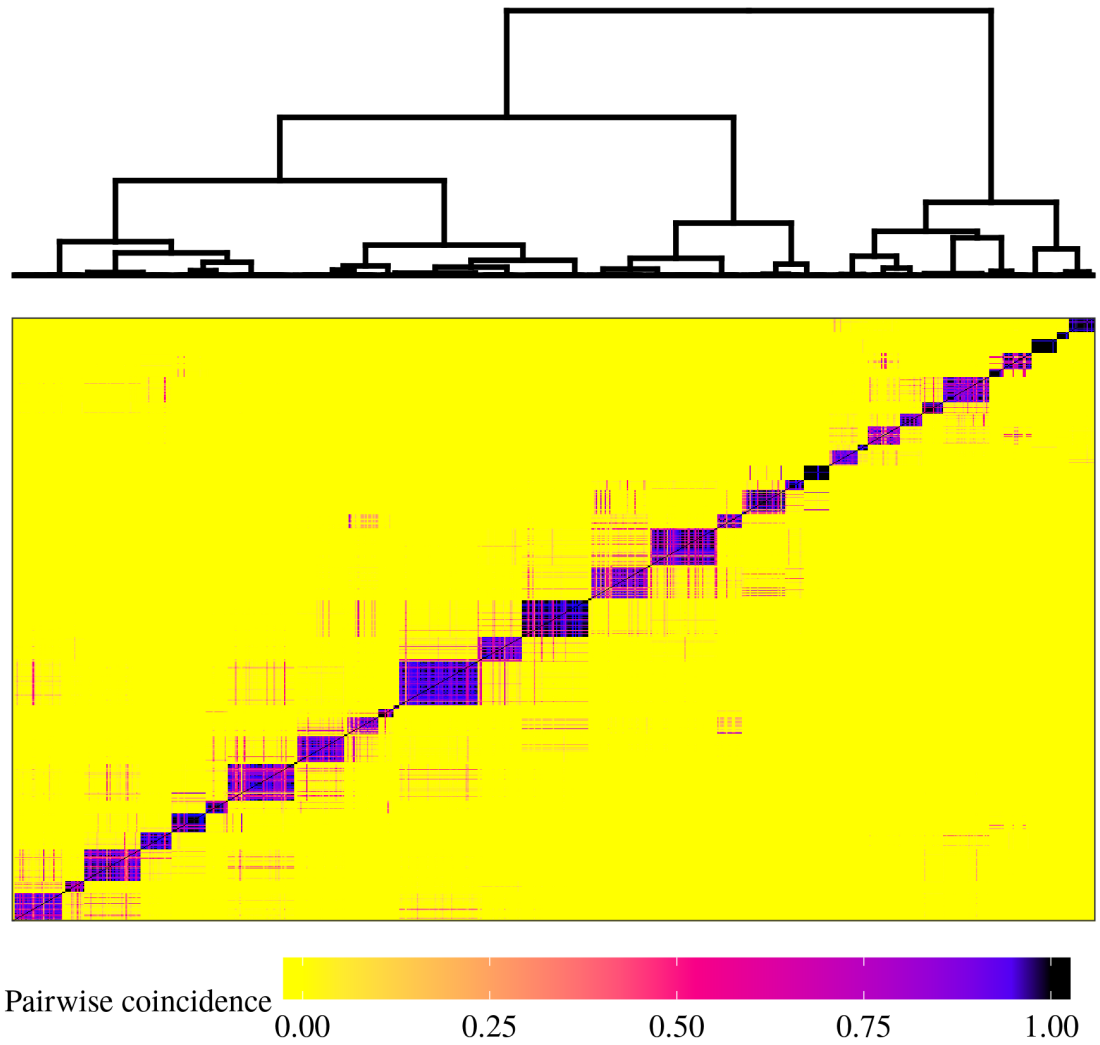


Figure A.31: FineSTRUCTURE inferred Maximum concordance trees for the CP output (chain 2) of the SAC dataset. Shown are the associated heatmaps of pairwise coincidence values of individuals as a proportion of the observations in the MCMC runs.

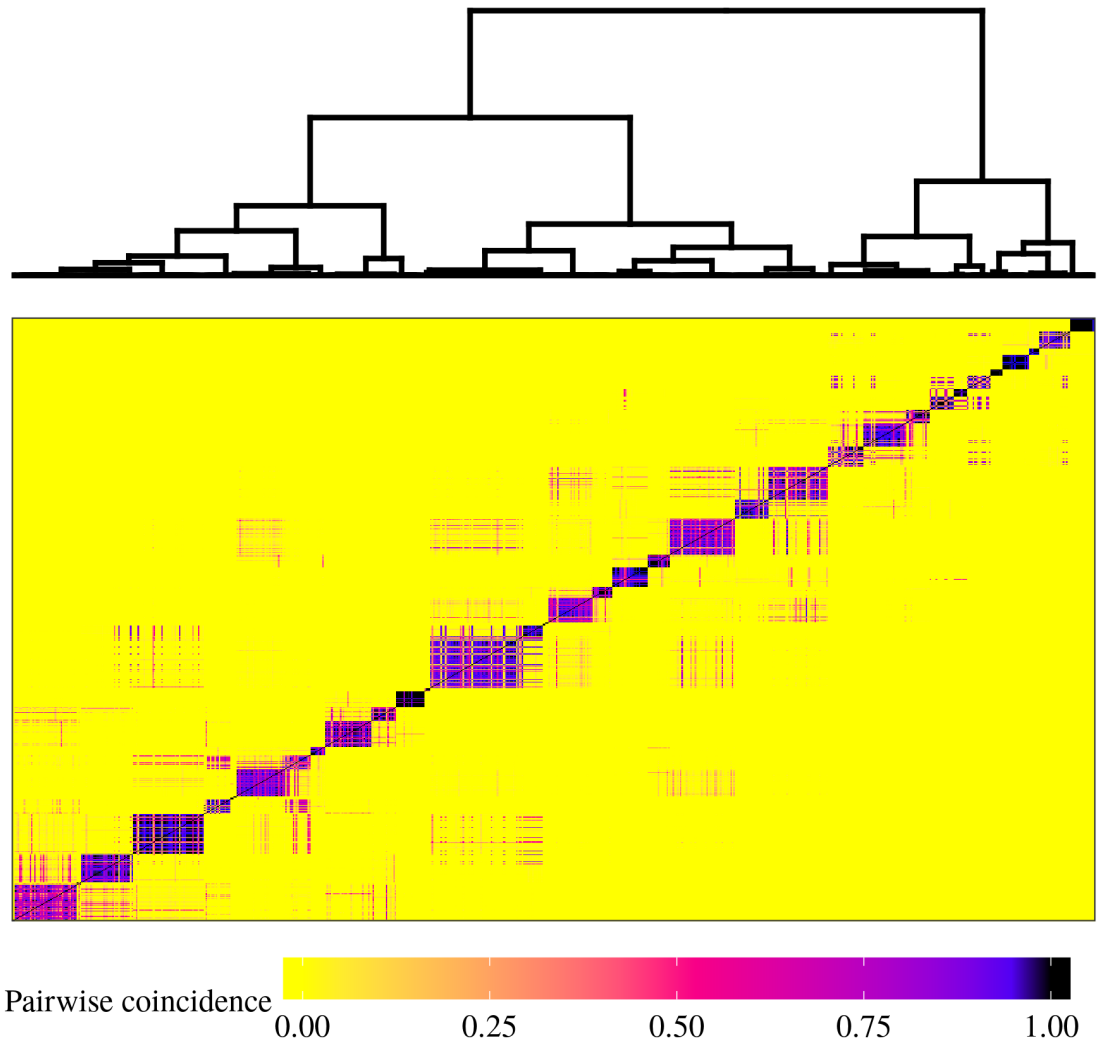


Figure A.32: FineSTRUCTURE inferred Maximum concordance trees for the CP output (chain 3) of the SAC dataset. Shown are the associated heatmaps of pairwise coincidence values of individuals as a proportion of the observations in the MCMC runs.

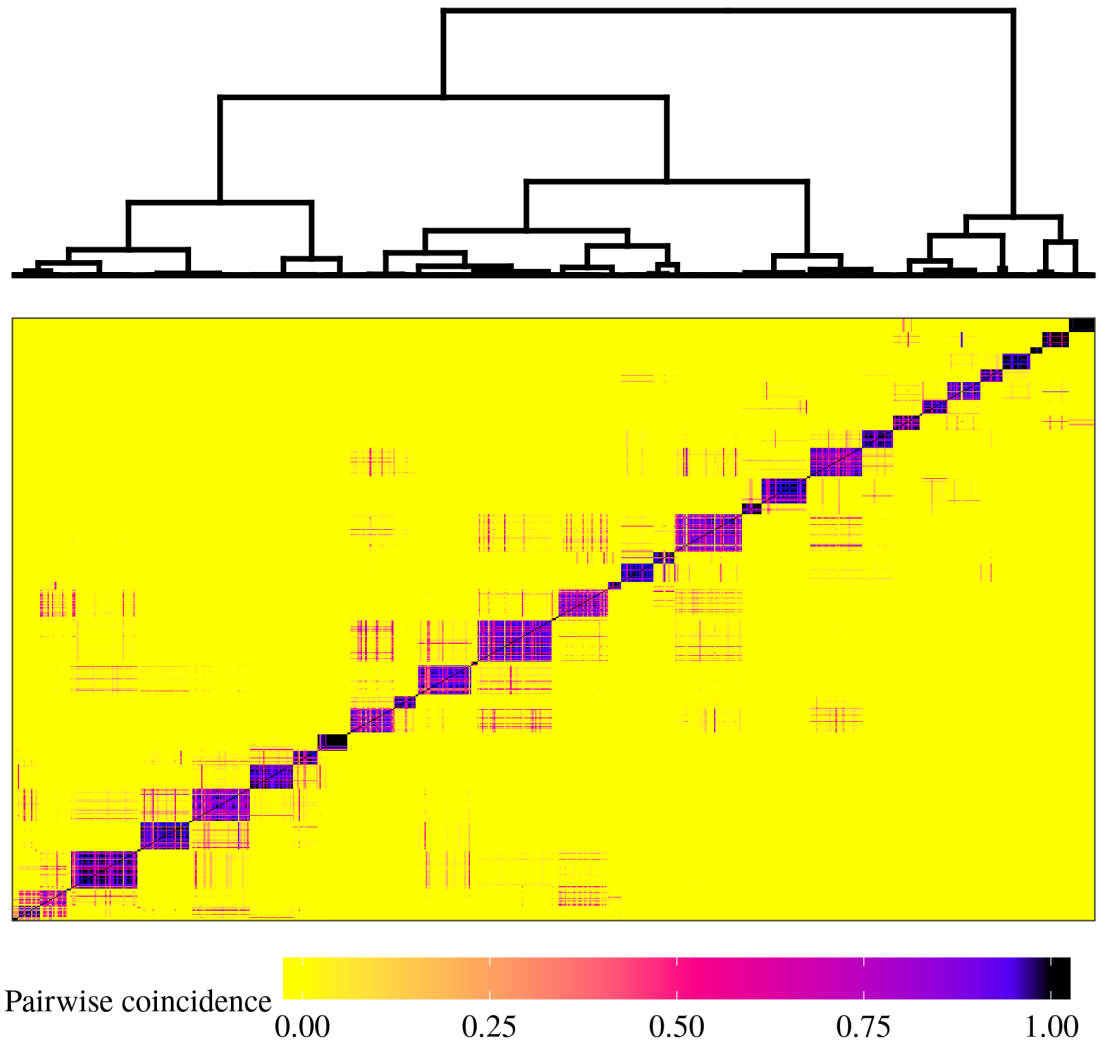


Figure A.33: FineSTRUCTURE inferred Maximum concordance trees for the CP output (chain 4) of the SAC dataset. Shown are the associated heatmaps of pairwise coincidence values of individuals as a proportion of the observations in the MCMC runs.

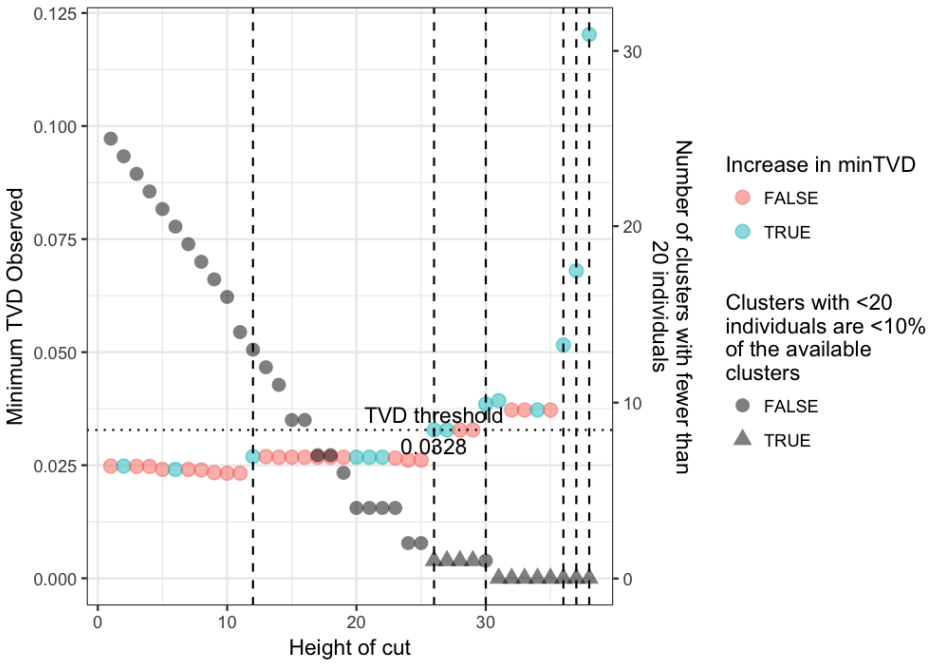


Figure A.34: Changes in minimum Total Variance Distance (TVD) observed across the fS-inferred clusters for the SAC-SAC CP run. The TVD was estimated on total copy length. Minimum TVD values indicated for each height on the tree and colour indicates if that TVD values was an increase or decrease from the previous TVD. Dashed vertical lines indicate the position of increases in minimum TVD which exceeds $3 \times \frac{1}{H} \sum_1^H |d_h|$ where $|d_h|$ is the change in minimum TVD from heights h to $h-1$ and H is the final height. Indicated in black is the number of fS-inferred clusters with fewer than 20 individuals and if that number of clusters represents <10% of all clusters. Final decision for threshold indicated by the horizontal dashed line.

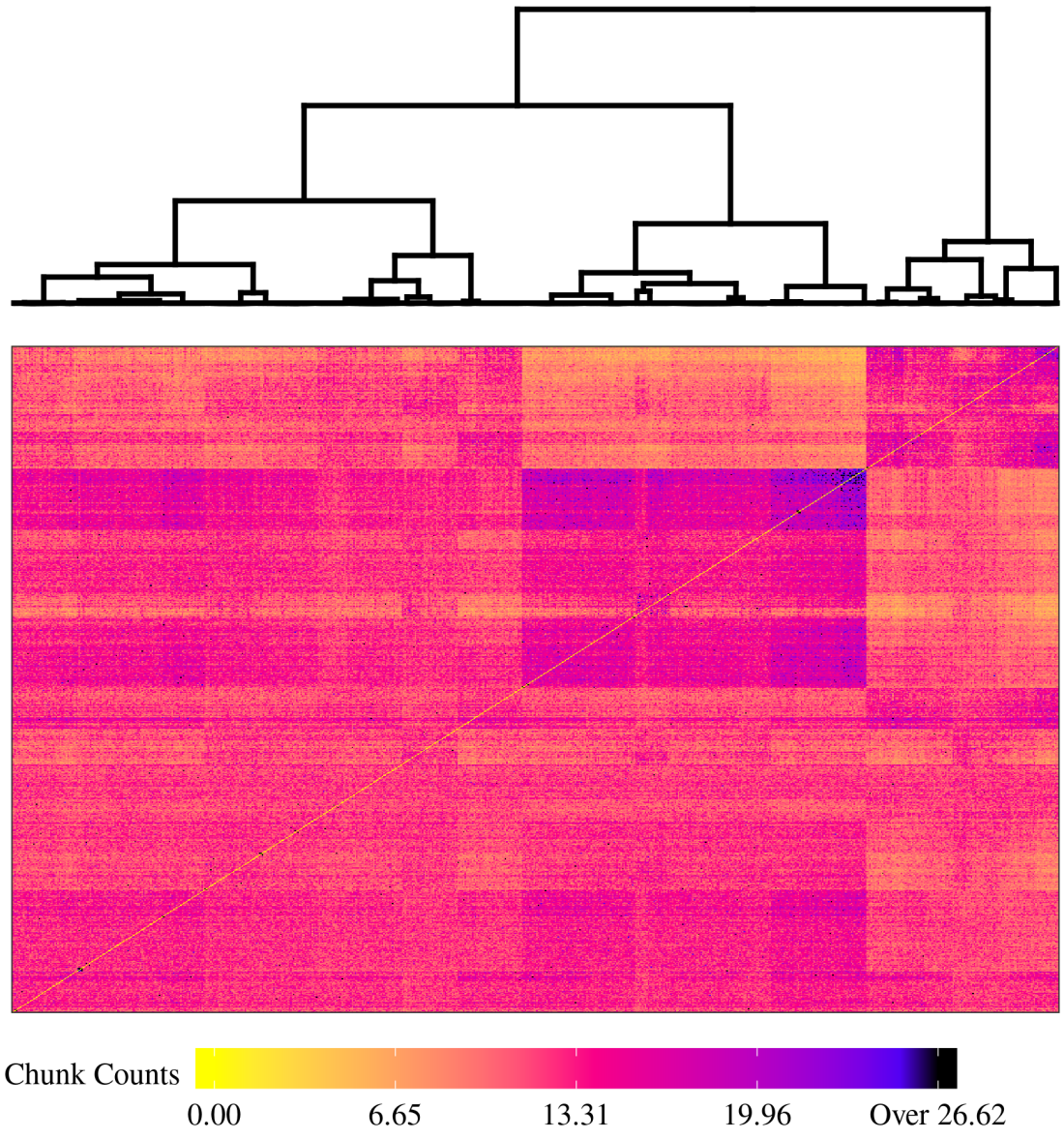


Figure A.35: Pairwise CP chunk counts values for SAC - SAC run. The upper limit on the heatmap was set 1.3 × 99$^{\text{th}}$-percentile value. Recipients in rows and donors in columns. fS dendrogram from chain 0 indicated above.

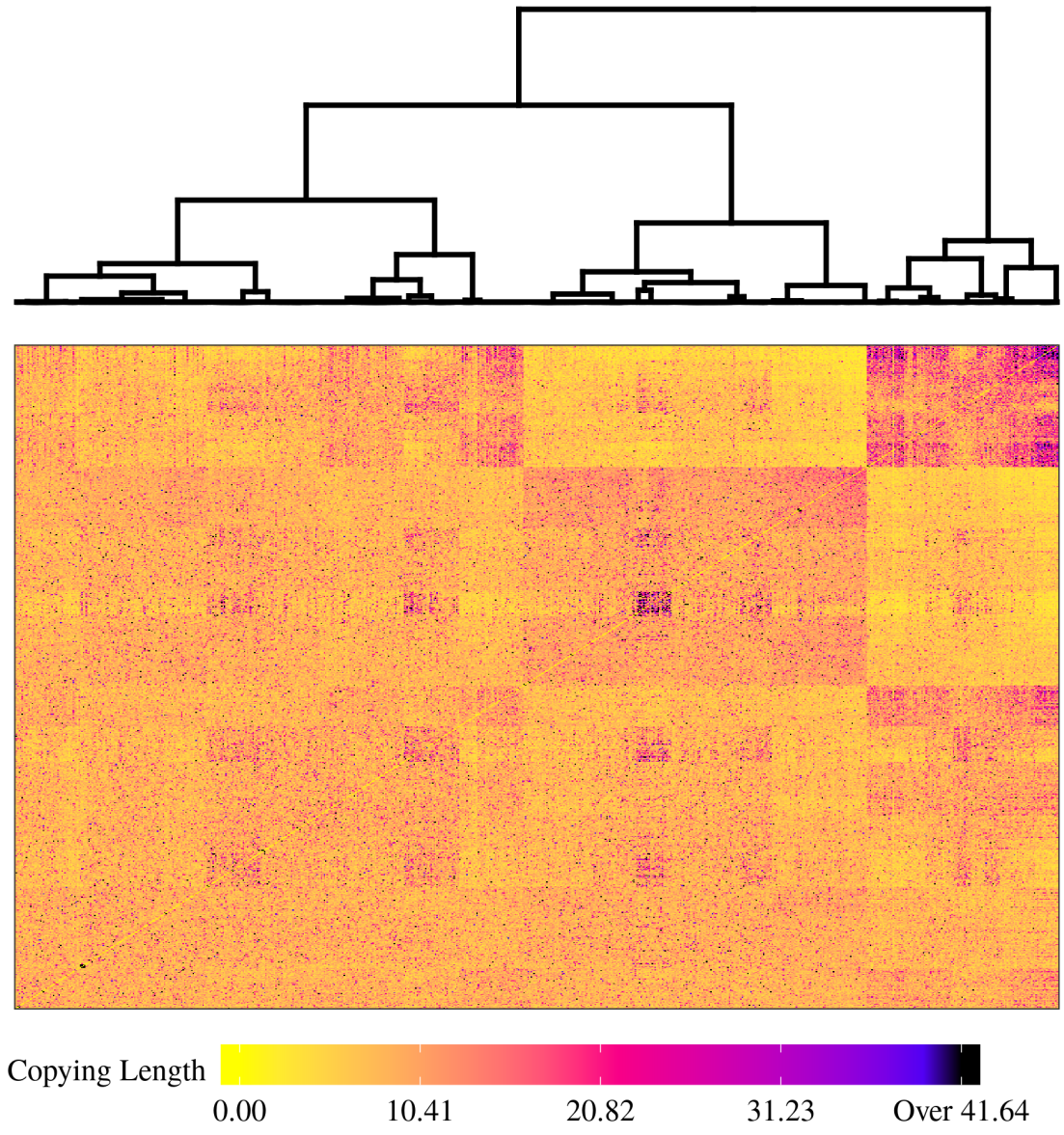


Figure A.36: Pairwise CP chunk lengths values for SAC - SAC run. The upper limit on the heatmap was set 1.3 × 99$^{\text{th}}$-percentile value. Recipients in rows and donors in columns. fS dendrogram from chain 0 indicated above.

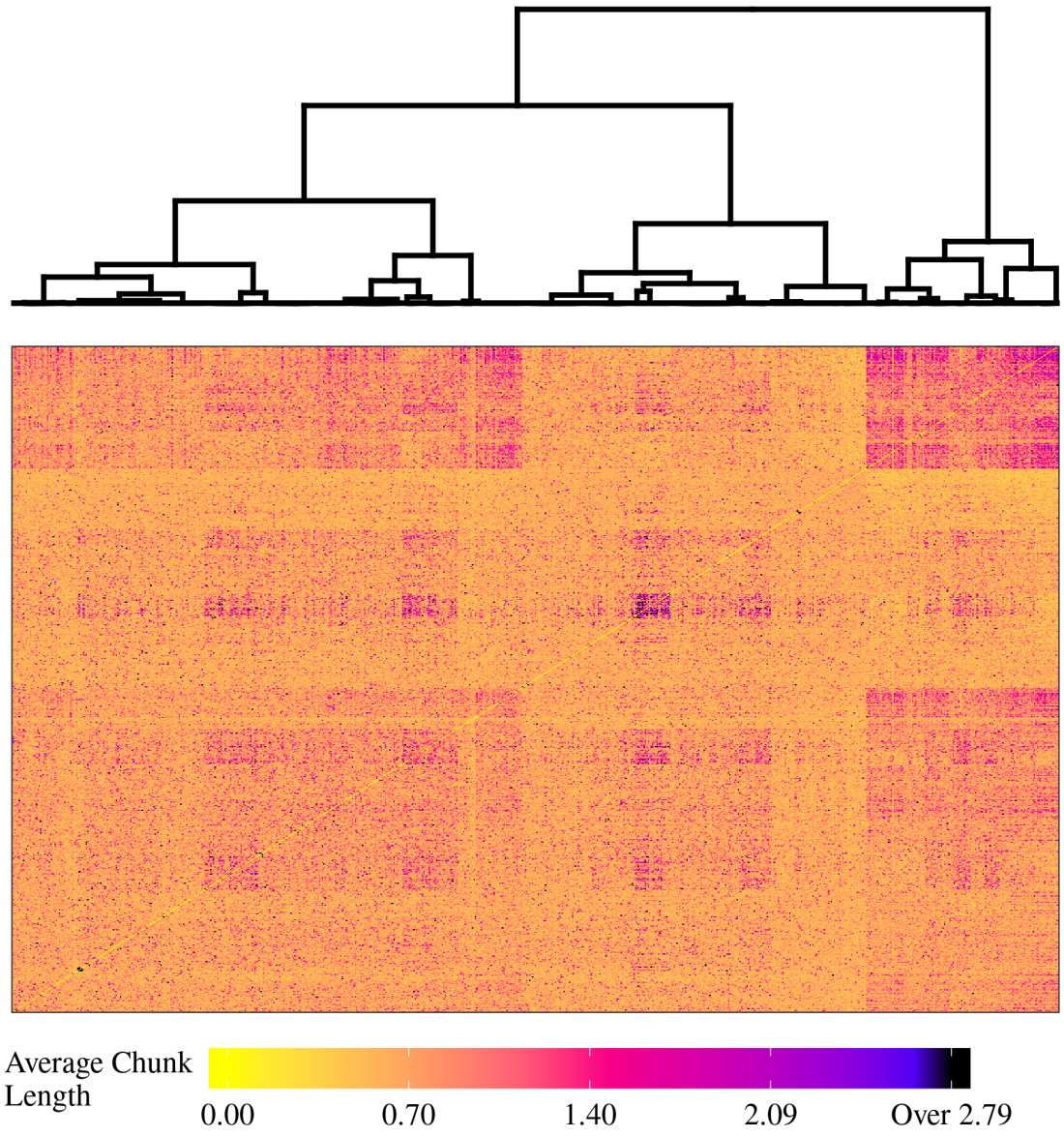


Figure A.37: Pairwise CP average chunk length values for SAC - SAC run. The upper limit on the heatmap was set 1.3 × 99^{th}-percentile value. Recipients in rows and donors in columns. fS dendrogram from chain 0 indicated above.

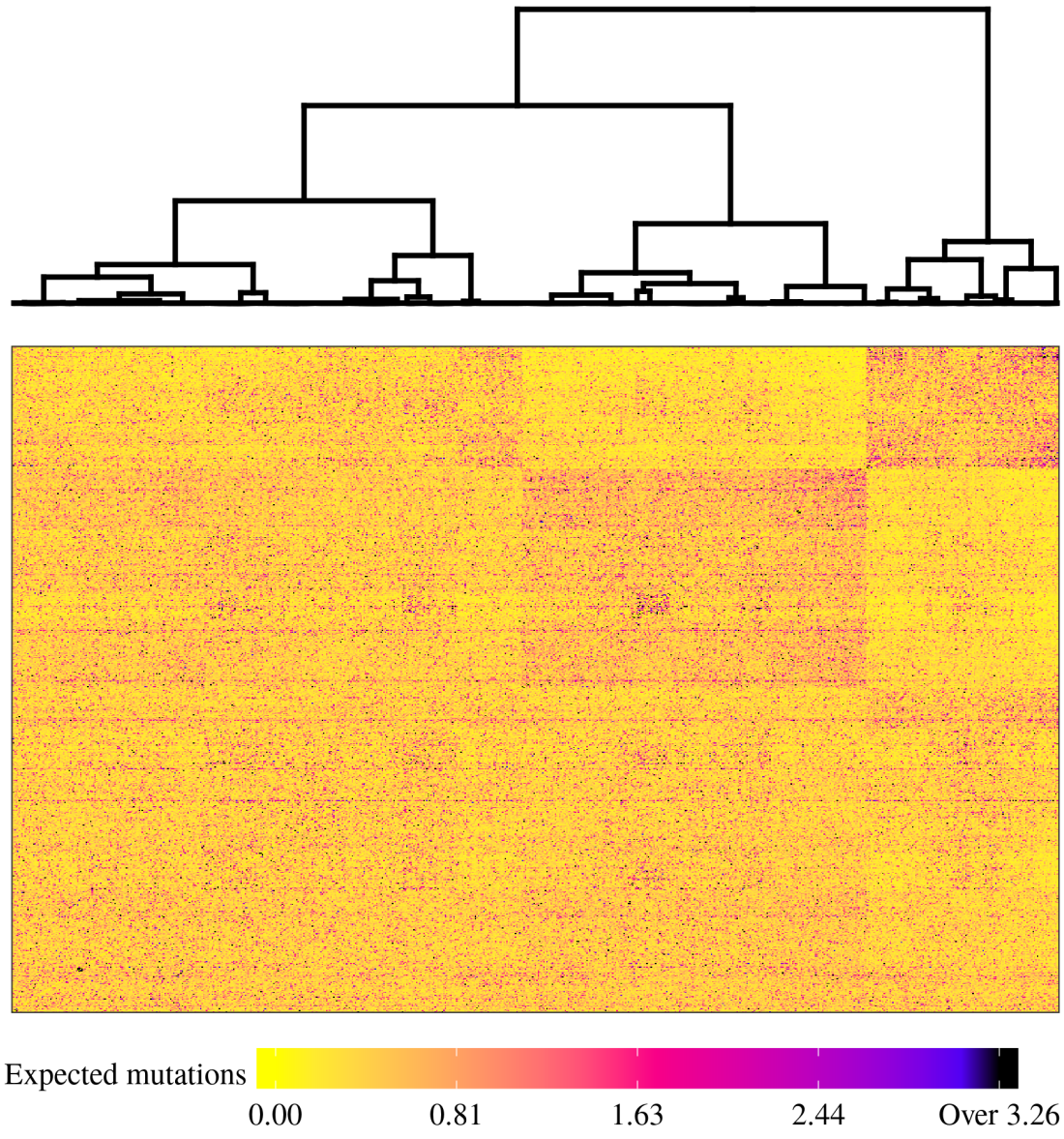


Figure A.38: Pairwise CP expected mutation values for SAC - SAC run. The upper limit on the heatmap was set 1.3 × 99$^{\text{th}}$-percentile value. Recipients in rows and donors in columns. fS dendrogram from chain 0 indicated above.

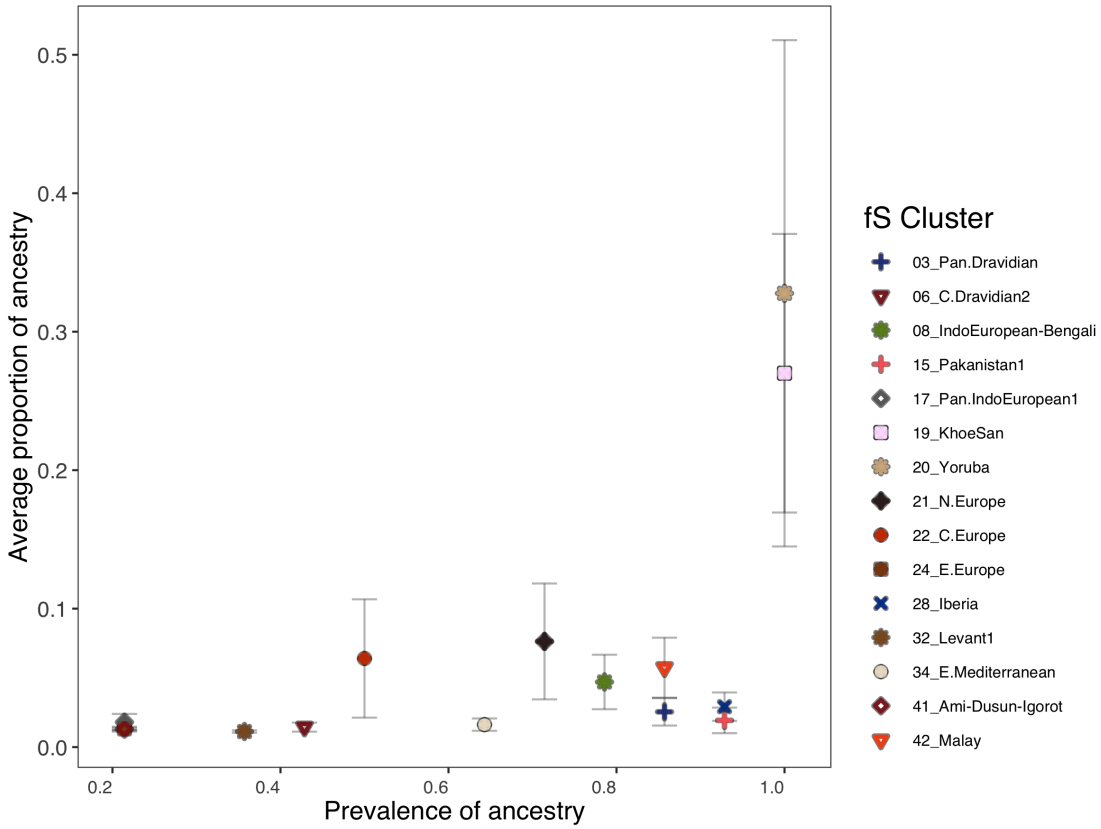


Figure A.39: Prevalence and average abundance of identified NNLS sources from across individuals for averaged SAC fS clusters. Data includes only source populations with at least 1% ancestry in at least 1 cluster.

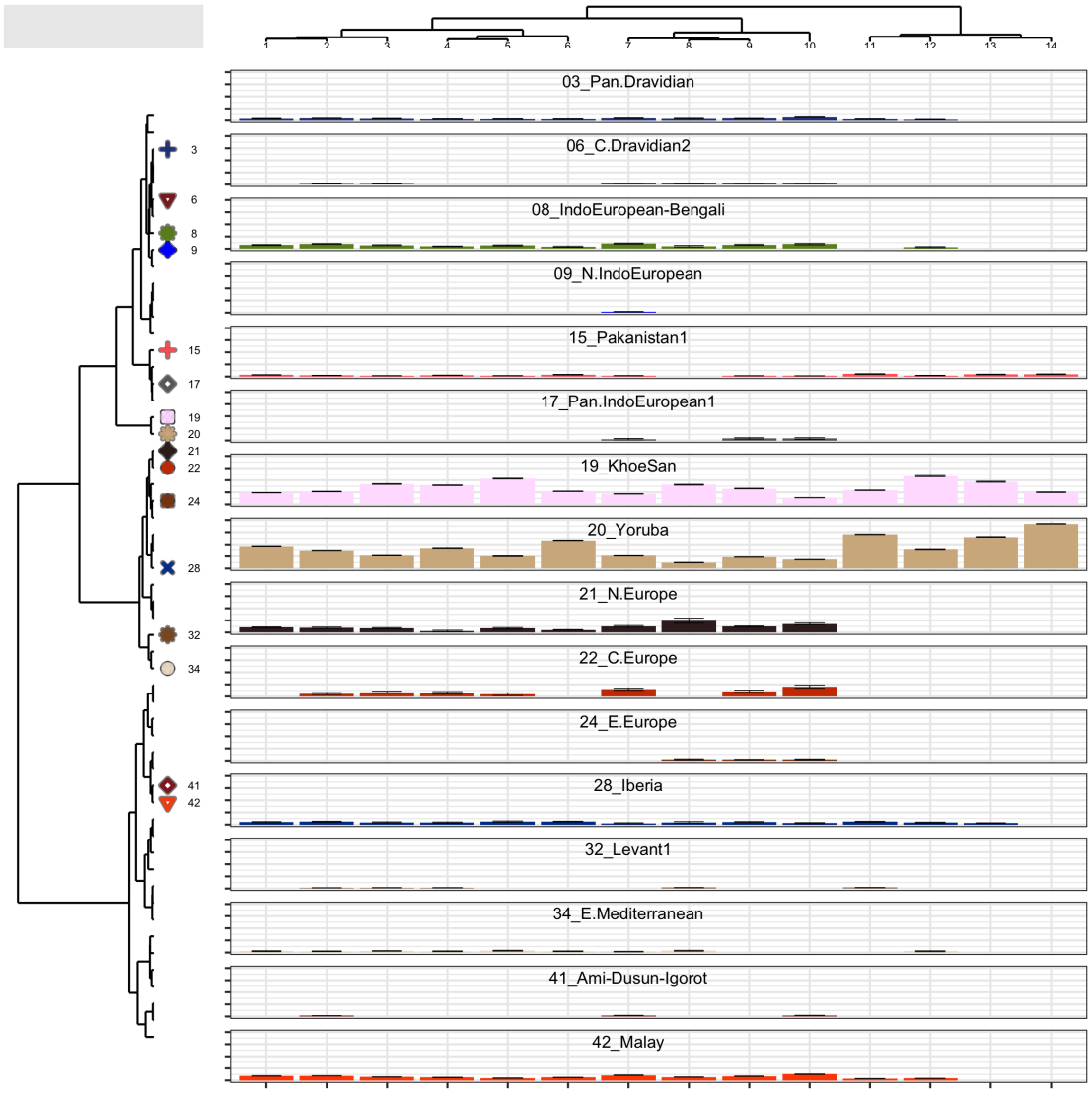


Figure A.40: Ancestral contributions from GR fS-clusters for averaged SAC fS clusters. Identified clusters contributed at least 1% ancestry to at least 1 fS-inferred SAC cluster. Jack-knife s.e. values indicated averaged within each cluster. GR fS-tree on the left side and SAC fS-tree above. Symbols, cluster numbers and colours correspond to Figures 4.14.

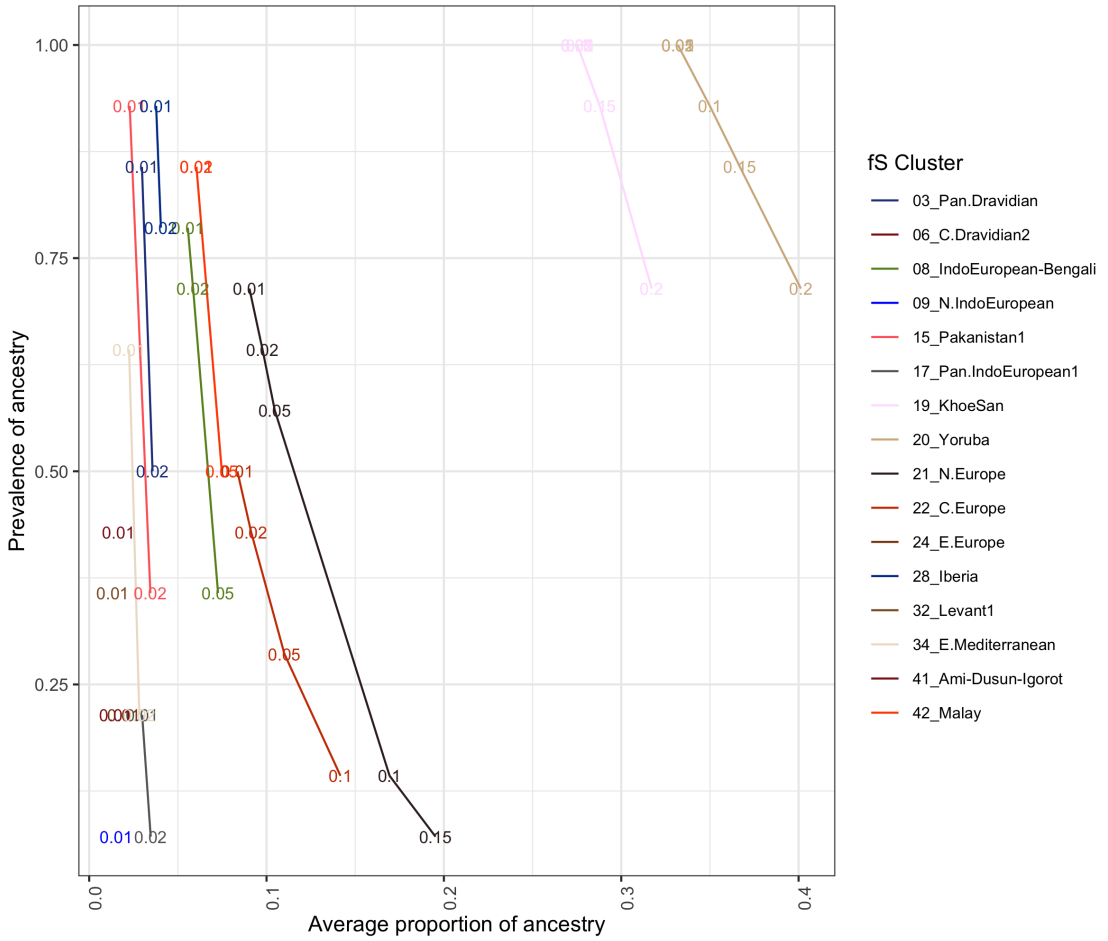


Figure A.41: Changes in prevalence and average proportion ancestry with cut-off criteria for averaged SAC fS clusters. Shown are changes in prevalence of ancestry and average proportion ancestry with proposed ancestry cut-off values for the cluster-based NNLS. Cut-off varied from 0.01 - 0.2. Values based on the mean - jack-knife s.e. of the remaining samples after the cut-off was applied.

A.3 Supplementary Tables

Table A.1: Global reference data used in Chapter 4 Populations arranged by UNESCO global regions. Indicated are the Decimal degrees South (DDS) and East (DDE) and the sample sizes included (n)

no.	*a priori* label	UNESCO Region	DDS	DDE	n
1	Ju_hoan_North	Southern Africa	-19.15	22.31	4
2	Juuhoan	Southern Africa	-19.15	21.02	1
3	Khomani_San	Southern Africa	-26.16	20.65	2
4	YRI	Western Africa	8.03	4.75	50
5	Yoruba	Western Africa	7.21	4.01	6
6	Croatia	Southern Europe	44.82	16.86	8
7	Albania	Southern Europe	40.44	20.23	3
8	Kosovo	Southern Europe	42.42	21.41	2
9	Albanian	Southern Europe	41.21	19.55	1
10	Slovenian	Southern Europe	46.86	14.06	24
11	Slovenia	Southern Europe	45.81	14.54	2
12	Greece	Southern Europe	39.79	22.11	8
13	Bosnia-Herzegovina	Southern Europe	43.86	17.22	9
14	Greek	Southern Europe	37.26	23.26	2
15	Serbia	Southern Europe	42.95	20.88	3
16	Macedonia	Southern Europe	40.97	21.67	4
17	IBS	Southern Europe	39.89	-4.02	25
18	Spanish	Southern Europe	40.19	-4.56	2
19	Sardinian	Southern Europe	39.38	9.62	8
20	TSI	Southern Europe	42.62	12.62	25
21	Bergamo	Southern Europe	46.99	10.52	7
22	Tuscan	Southern Europe	42.73	12.60	6
23	Portugal	Southern Europe	40.07	-7.23	50
24	Crete	Southern Europe	35.37	25.24	2
25	Spain	Southern Europe	41.17	-2.99	25
26	Italy	Southern Europe	41.49	12.72	25
27	Estonian	Northern Europe	58.68	24.45	2
28	Latvia	Northern Europe	56.90	24.81	1
29	Saami	Northern Europe	70.78	26.49	2
30	FIN	Northern Europe	60.80	24.19	50
31	Finland	Northern Europe	61.07	24.56	1
32	Finnish	Northern Europe	60.77	24.94	3
33	Turkey	Northern Europe	39.75	34.40	6
34	Sweden	Northern Europe	59.74	17.30	11
35	Norway	Northern Europe	59.72	11.07	3
36	Denmark	Northern Europe	55.29	8.30	1
37	Norwegian	Northern Europe	59.88	5.44	1
38	Icelandic	Northern Europe	63.48	-22.19	2
39	Russia	Eastern Europe	55.67	36.67	6
40	Yugoslavia	Eastern Europe	44.73	20.78	42
41	Poland	Eastern Europe	52.07	19.23	22
42	Hungary	Eastern Europe	48.18	20.14	19
43	Czech	Eastern Europe	50.36	14.72	12
44	Romania	Eastern Europe	45.62	25.91	14
45	Polish	Eastern Europe	51.58	21.27	1
46	Hungarian	Eastern Europe	48.23	19.14	2
47	Bulgarian	Eastern Europe	42.58	24.41	2
48	Bulgaria	Eastern Europe	42.35	25.87	2
49	Russian	Eastern Europe	56.13	37.14	7

50	Abkhasian	Eastern Europe	43.43	41.48	2
51	Chechen	Eastern Europe	44.06	46.17	1
52	North_Ossetian	Eastern Europe	43.38	45.08	2
53	Lezgin	Eastern Europe	42.42	47.81	2
54	Slovakia	Eastern Europe	48.68	19.22	1
55	Ulchi	Eastern Europe	51.84	141.15	2
56	Yakut	Eastern Europe	61.47	98.83	6
57	Even	Eastern Europe	56.55	135.53	3
58	Eskimo_Naukan	Eastern Europe	65.56	170.13	2
59	Itelman	Eastern Europe	57.21	157.77	1
60	Eskimo_Chaplin	Eastern Europe	65.12	173.17	1
61	Eskimo_Sireniki	Eastern Europe	64.38	173.80	1
62	Chukchi	Eastern Europe	68.97	169.90	1
63	Aleut	Eastern Europe	55.34	166.45	2
64	Tlingit	Eastern Europe	53.07	157.78	2
65	Mansi	Eastern Europe	64.27	60.45	2
66	Tubalar	Eastern Europe	51.54	87.48	2
67	Altaian	Eastern Europe	50.02	85.52	1
68	Ukraine	Eastern Europe	49.62	30.94	2
69	Basque	Western Europe	43.41	-1.78	7
70	Netherlands	Western Europe	51.86	5.98	17
71	Ireland	Western Europe	52.37	-8.92	50
72	GBR	Western Europe	52.37	-2.00	25
73	Belgium	Western Europe	51.35	5.15	43
74	Germany	Western Europe	51.12	11.37	50
75	UK	Western Europe	54.51	-2.84	25
76	Orcadian	Western Europe	58.74	-3.51	7
77	English	Western Europe	50.28	0.47	2
78	Swiss-German	Western Europe	47.64	8.50	25
79	France	Western Europe	46.57	2.65	40
80	Scotland	Western Europe	55.09	-2.43	5
81	Swiss-French	Western Europe	46.24	5.16	25
82	French	Western Europe	45.69	3.05	7
83	Swiss-Italian	Western Europe	45.17	9.38	13
84	Austria	Western Europe	47.82	14.01	14
85	Palestinian	Western Asia	31.93	35.29	8
86	Jordanian	Western Asia	32.53	35.40	3
87	Bedouin	Western Asia	31.47	34.28	3
88	BedouinB	Western Asia	30.62	34.16	2
89	Georgian	Western Asia	43.15	41.59	2
90	Armenian	Western Asia	39.28	44.49	2
91	Adygei	Western Asia	44.49	45.88	7
92	Turkish	Western Asia	39.44	36.27	2
93	Samaritan	Western Asia	31.47	34.72	1
94	Druze	Western Asia	32.36	35.08	7
95	Cyprus	Western Asia	34.99	34.04	4
96	Irula	Southern Asia	21.12	78.70	7
97	Kallar	Southern Asia	12.38	78.24	5
98	Adi-Dravider	Southern Asia	10.94	78.62	5
99	Hallaki	Southern Asia	13.46	73.85	7
100	Gond	Southern Asia	21.41	79.29	14
101	Relli	Southern Asia	17.50	82.81	2
102	Iranian	Southern Asia	35.65	50.89	2
103	Iraqi	Southern Asia	32.62	44.42	24

104	Malli	Southern Asia	10.30	72.33	5
105	Tharu	Southern Asia	28.62	79.76	9
106	Tamil	Southern Asia	21.12	78.12	12
107	STU	Southern Asia	6.81	79.91	38
108	Aonaga	Southern Asia	26.20	93.49	4
109	Subba	Southern Asia	20.42	79.28	4
110	Nysha	Southern Asia	26.93	92.43	4
111	Sherpa	Southern Asia	21.17	78.60	4
112	Changapa	Southern Asia	20.55	78.99	5
113	Kapu	Southern Asia	17.95	83.56	2
114	Paniyas	Southern Asia	11.44	77.08	5
115	Malayalam	Southern Asia	20.36	79.95	2
116	Naidu	Southern Asia	13.53	79.13	4
117	Madiga	Southern Asia	18.08	79.24	14
118	BEB	Southern Asia	23.93	89.48	50
119	Lodi	Southern Asia	26.17	83.38	5
120	Mala	Southern Asia	16.59	78.17	15
121	Vedda	Southern Asia	8.72	80.91	2
122	Minicoy	Southern Asia	20.22	79.84	4
123	Gounder	Southern Asia	12.30	77.89	5
124	Yadava	Southern Asia	18.14	83.71	2
125	Kattunayakkan	Southern Asia	10.81	76.66	5
126	Malai_Kuravar	Southern Asia	12.20	78.94	5
127	Sinhalese	Southern Asia	6.98	81.72	3
128	Narikkuravar	Southern Asia	10.85	78.01	5
129	Konkani	Southern Asia	15.29	73.59	3
130	Urdu	Southern Asia	21.34	79.92	37
131	Hindi	Southern Asia	20.27	78.93	15
132	Velama	Southern Asia	15.83	76.76	4
133	IndianLower	Southern Asia	19.89	78.36	12
134	Satnami	Southern Asia	22.85	81.31	4
135	Chenchu	Southern Asia	17.87	78.60	6
136	Palliyar	Southern Asia	11.57	78.07	5
137	Bhil	Southern Asia	22.23	73.40	17
138	Bengali	Southern Asia	22.63	88.34	4
139	PJL	Southern Asia	32.45	74.03	25
140	Brahmin	Southern Asia	26.56	82.94	7
141	Jains	Southern Asia	23.69	72.00	5
142	Punjabi	Southern Asia	31.54	74.52	25
143	Gujarati	Southern Asia	23.91	71.82	20
144	Meghawal	Southern Asia	25.76	72.77	5
145	IndianUpper	Southern Asia	19.92	78.77	13
146	GIH	Southern Asia	23.30	73.36	30
147	Kurumba	Southern Asia	11.19	75.59	9
148	Srivastava	Southern Asia	25.26	81.66	2
149	Kharia	Southern Asia	23.68	72.84	6
150	Santhal	Southern Asia	24.86	88.50	7
151	Korku	Southern Asia	21.42	78.29	3
152	Khonda_Dora	Southern Asia	18.21	82.00	1
153	Sahariya	Southern Asia	26.20	81.08	4
154	Birhor	Southern Asia	20.23	79.91	4
155	Ho	Southern Asia	19.63	78.63	5
156	Munda	Southern Asia	20.27	79.73	4
157	Bhumij	Southern Asia	21.52	79.38	5

158	ITU	Southern Asia	16.40	79.37	50
159	Onge	Southern Asia	10.48	92.56	9
160	Brahui	Southern Asia	30.26	69.54	6
161	Makrani	Southern Asia	29.80	68.93	5
162	Balochi	Southern Asia	29.93	68.88	6
163	Pakistanis	Southern Asia	30.25	69.15	25
164	Kshatriya	Southern Asia	26.54	82.91	10
165	Vaish	Southern Asia	26.44	83.73	4
166	Kashmiri_Pandit	Southern Asia	33.76	75.08	10
167	Pathan	Southern Asia	29.48	70.08	7
168	Burusho	Southern Asia	29.71	68.76	6
169	Vysya	Southern Asia	14.99	77.39	12
170	Kalash	Southern Asia	30.49	69.87	6
171	Pushto	Southern Asia	19.97	78.42	1
172	Kamsali	Southern Asia	14.90	78.35	4
173	Kuruchiyan	Southern Asia	10.16	76.28	5
174	CHS	Eastern Asia	1.34	103.12	20
175	She	Eastern Asia	35.16	104.01	6
176	Tujia	Eastern Asia	36.02	103.52	6
177	CHB	Eastern Asia	39.65	117.20	20
178	Han	Eastern Asia	35.46	104.07	8
179	Miaozu	Eastern Asia	36.44	104.21	3
180	Yizu	Eastern Asia	34.92	105.06	4
181	Miao	Eastern Asia	27.88	108.05	2
182	Korean	Eastern Asia	37.60	127.06	2
183	CDX	Eastern Asia	22.89	101.75	50
184	Dai	Eastern Asia	36.12	104.60	6
185	Lahu	Eastern Asia	35.58	103.47	6
186	Tu	Eastern Asia	35.41	104.36	5
187	Naxi	Eastern Asia	36.47	104.19	7
188	Yi	Eastern Asia	27.79	102.97	2
189	Tibet-refugees	Eastern Asia	28.97	87.92	5
190	Hezhen	Eastern Asia	36.11	103.96	5
191	Xibo	Eastern Asia	35.69	103.92	6
192	Mongola	Eastern Asia	36.21	104.72	7
193	Daur	Eastern Asia	35.33	104.60	6
194	Oroqen	Eastern Asia	36.79	104.41	6
195	Japanese	Eastern Asia	36.58	139.39	7
196	JPT	Eastern Asia	35.88	140.40	20
197	Japan	Eastern Asia	35.25	137.57	20
198	Mongolian	Eastern Asia	46.98	103.41	25
199	Uygur	Eastern Asia	36.63	103.22	4
200	Tajik	Central Asia	37.88	72.23	2
201	Kyrgyzstan	Central Asia	41.06	73.77	24
202	Kyrgyz	Central Asia	43.16	75.36	2
203	KHV	South Eastern Asia	11.23	106.98	50
204	Kinh	South Eastern Asia	21.96	106.54	2
205	Thai	South Eastern Asia	13.36	101.13	22
206	MAS	South Eastern Asia	0.98	103.52	50
207	Burmese	South Eastern Asia	17.75	97.58	2
208	Cambodian	South Eastern Asia	11.04	104.59	6
209	Igorot	South Eastern Asia	16.87	120.86	2
210	Atayal	South Eastern Asia	23.84	120.88	1
211	Ami	South Eastern Asia	22.93	120.17	2

212	Dusun	South Eastern Asia	5.46	113.85	2
213	Tonga_Samoa	South Eastern Asia	-22.11	-175.24	26
214	Australian	Australia New Zealand	-13.66	143.64	1
215	Nasioi	Melanesia	-6.15	154.87	2
216	Bougainville	Melanesia	-6.08	155.09	2
217	Papuan	Melanesia	-3.65	143.14	15

Table A.2: Samples removed from the Affymetrix dataset based on within-group (*a priori*) kinship.

Population	Source dataset	Initial sample size	Samples removed	% removed
BEB	[2]	141	53	37
Bolivian	[258]	25	3	12
Brahmin	[13], [85], [96]	17	10	59
CDX	[2]	105	6	6
CEU	[2], [6]	184	105	57
CHD	[2], [6]	85	2	2
CHS	[2]	171	103	60
ESN	[2]	170	78	46
GBR	[2]	103	4	4
Germany	[167]	75	2	3
GIH	[2], [6]	112	9	8
Great_Andamanese	[85], [96]	7	3	43
GWD	[2]	178	93	52
IBS	[2]	160	89	56
IndianLower	[85], [96]	13	1	8
Iraqi	[258]	25	1	4
Italy	[167]	225	3	1
ITU	[2]	118	8	7
Kashmiri_Pandit	[85], [96]	20	10	50
KHV	[2]	119	33	28
Korku	[85], [96]	4	1	25
Kshatriya	[85], [96]	20	10	50
Lahu	[6], [13]	7	1	14
LWK	[2], [6]	114	14	12
Madiga	[13], [85], [96]	21	7	33
Mala	[13], [85], [96]	20	5	25
MKK	[6]	170	86	51
MSL	[2]	120	35	29
PJL	[2]	151	73	48
Portugal	[167]	135	3	2
Punjabi	[13], [167]	229	5	2
Slovenian	[167]	26	1	4
STU	[2]	128	22	17
Swiss-French	[167]	760	8	1
Thai	[258]	27	3	11
TSI	[2], [6]	111	1	1
UK	[167]	390	4	1
Vedda	[85], [96]	4	2	50
Vysya	[85], [96]	20	8	40
YRI	[2], [6]	192	104	54

Table A.3: Individuals identified (IID) as outliers in their *a priori* populations (APP) by identity-by-descent (IBD) estimates and Principle Component Analysis (PCA) distances. This included visual inspection (VI) on Principal Component plot and Tukey outlier detection on IBD (IBD) and PCA values (PCD).

APP	IID	PCD	VI	IBD	APP	IID	PCD	VI	IBD
Balochi	HGDP00076	N	Y	N	Swiss-French	ppr27160	Y	N	N
CDX	HG01798	Y	Y	N	Swiss-French	ppr34448	Y	N	N
CHS	HG00673	Y	N	N	Swiss-French	ppr36340	Y	N	N
CHS	HG00717	Y	N	N	Swiss-French	ppr38472	Y	N	N
CHS	HG00592	N	Y	N	Swiss-French	ppr48783	Y	N	N
ESN	HG02971	Y	N	N	Swiss-French	ppr8488	Y	N	N
FIN	HG00181	Y	N	N	Swiss-French	ppr21695	Y	Y	N
FIN	HG00182	Y	N	N	Swiss-French	ppr28339	Y	Y	N
FIN	HG00304	Y	N	N	Switzer-land	ppr20019	Y	N	N
Germany	ppr34088	Y	Y	N	Switzer-land	ppr26297	Y	N	N
GWD	HG02643	N	Y	N	Switzer-land	ppr34058	Y	N	N
Hausa	NGHA017	N	Y	N	Thai	F089339	Y	Y	Y
IndianUp-per	TBR14	N	Y	N	Thai	F066599	N	Y	N
Italy	ppr33242	Y	N	N	UK	ppr11555	Y	N	N
Italy	ppr34049	Y	N	N	UK	ppr14511	Y	N	N
Italy	ppr44221	Y	N	N	UK	ppr14576	Y	N	N
Italy	ppr49500	Y	N	N	UK	ppr20322	Y	N	N
Italy	ppr7623	Y	N	N	UK	ppr24146	Y	N	N
ITU	HG03714	Y	N	N	UK	ppr24170	Y	N	N
ITU	HG03875	Y	N	N	UK	ppr25486	Y	N	N
ITU	HG04214	Y	N	N	UK	ppr26916	Y	N	N
JPT	NA18954	Y	Y	N	UK	ppr27956	Y	N	N
JPT	NA18951	N	Y	N	UK	ppr31475	Y	N	N
Ju-uhoan_South	BOT6.090	N	N	Y	UK	ppr35433	Y	N	N
Kyrgyzstan	F063402	N	Y	N	UK	ppr41981	Y	N	N
MKK	NA21436	Y	N	N	UK	ppr43032	Y	N	N
MKK	NA21717	Y	N	N	UK	ppr6939	Y	Y	N
Mozabite	HGDP01271	N	Y	N	UK	ppr38714	Y	Y	Y
Mozabite	HGDP01270	N	Y	Y	UK	ppr43570	Y	Y	Y
Naro	BOT6.058	N	Y	Y	Urdu	ppr15145	Y	Y	N
Portugal	ppr13600	Y	N	N	Urdu	ppr5361	Y	Y	N
Portugal	ppr43011	Y	N	N	YRI	NA18916	Y	Y	N

Portugal	ppr14374	Y	Y	N	YRI	NA19118	Y	Y	N
Portugal	ppr47614	Y	Y	N	Yugoslavia	ppr31376	Y	N	N
Punjabi	ppr16705	N	N		Yugoslavia	ppr12588	Y	Y	N
Punjabi	ppr4604	Y	N	N	Bamoun	CABM022	N	Y	N
Punjabi	ppr9237	Y	N	N	Bulala	CHBU015A	N	Y	N
Sindhi	HGDP00175	N	Y	N	KHV	HG02131	N	Y	N
Spain	ppr39220	N	N	Y	KHV	HG02142	N	Y	N
Spain	ppr5613	Y	N	N	Sherpa	SHER_WB01	N	Y	N
STU	HG03694	Y	N	N	Slovenian	F038390	N	Y	N
STU	HG03743	Y	N	N	MAS	SGVP00507	Y	N	N
STU	HG03746	Y	N	N	MAS	SGVP00537	Y	N	N
STU	HG03757	Y	N	N	Makrani	HGDP00130	N	Y	N
STU	HG03856	Y	N	N	Minicoy	MINI_LK04	N	Y	N
STU	HG03888	Y	N	N	Brahui	HGDP00029	N	Y	N
STU	HG04003	Y	N	N	Munda	MUND_MP03	N	Y	N
Swiss-French	ppr12747	Y	N	N	Siddi	SIDD_KA03	N	Y	N
Swiss-French	ppr21802	Y	N	N	Subba	SUBB_WB04	N	Y	N
Swiss-French	ppr24318	Y	N	N	Uygur	HGDP01302	N	Y	N

Table A.4: Populations excluded from the dataset due to confounding admixture, non-homogenous PCA results or identified as not relevant.

Removed	Reasons	Removed	Reasons
CEU	Better proxy available	Kongo	Not Relevant
CHD	Better proxy available	Kusunda	Not Relevant
Switzerland	Better proxy available	Luhya	Not Relevant
Bantu	Confounding Admixture	Luo	Not Relevant
BantuHerero	Confounding Admixture	LWK	Not Relevant
BantuKenya	Confounding Admixture	Mada	Not Relevant
BantuSouthAfrica	Confounding Admixture	Mandenka	Not Relevant
BantuTswana	Confounding Admixture	Maori	Not Relevant
Great_Andamanese	Confounding Admixture	Masai	Not Relevant
Himba	Confounding Admixture	Mayan	Not Relevant
Jews	Confounding Admixture	MbororoFulani	Not Relevant
Kgalagadi	Confounding Admixture	Mbukushu	Not Relevant
Shua	Confounding Admixture	Mbuti	Not Relevant
Siddi	Confounding Admixture	MbutiPygmy	Not Relevant
Tshwa	Confounding Admixture	Mende	Not Relevant
Wambo	Confounding Admixture	Mixe	Not Relevant
Xhosa	Confounding Admixture	Mixtec	Not Relevant
Nepalese	Non-homogenous PCA	MKK	Not Relevant
Bambaran	Not Relevant	Mozabite	Not Relevant
Bamoun	Not Relevant	MSL	Not Relevant
Biaka	Not Relevant	Naro	Not Relevant
BiakaPygmy	Not Relevant	Piapoco	Not Relevant
Bolivian	Not Relevant	Pima	Not Relevant
Brong	Not Relevant	Quechua	Not Relevant
Bulala	Not Relevant	Saharawi	Not Relevant
Chane	Not Relevant	Sindhi	Not Relevant
Dinka	Not Relevant	Somali	Not Relevant
Dogon	Not Relevant	Surui	Not Relevant
Esan	Not Relevant	Yemenite_Jew	Not Relevant
ESN	Not Relevant	Zapotec	Not Relevant

Table A.5: Summary statistics of the CHROMOPAINTER output values for GR-GR painting.

Output	Minimum	Mean ± sd	Maximum	25-75 percentile
Total Chunk Count (number of haplotypes)	0	7.48 ± 4.47	783.87	5.73 - 8.66
Total Copy Length (cM)	0	3.24 ± 5.59	1365.04	1.83 - 3.82
Average Chunk Length (cM)	0	0.39 ± 0.12	6.68	0.32 - 0.44
Total Mutations (No. of SNPs)	0	0.11 ± 0.61	198.45	0.049 - 0.10

Table A.6: Summary statistics of the CHROMOPAINTER output values for the SAC-GR painting.

Output	Minimum	Mean ± sd	Maximum	25-75 percentile
Total Chunk Count (number of haplotypes)	0	9.36 ± 13.799	535.89	6.95 - 7.96
Total Copy Length (cM)	0	3.24 ± 6.730	299.13	1.89 - 2.68
Average Chunk Length (cM)	0	0.31 ± 0.071	4.44	0.27 - 0.34
Total Mutations (No. of SNPs)	0	0.61 ± 2.876	137.96	0.23 - 0.36

Table A.7: Summary statistics of the CHROMOPAINTER output values for the SAC-SAC painting.

Output	Minimum	Mean ± sd	Maximum	25-75 percentile
Total Chunk Count (number of haplotypes)	0	12.23 ± 2.93	76.89	10.31 - 13.86
Total Copy Length (cM)	0	9.51 ± 7.56	542.87	6.12 - 10.89
Average Chunk Length (cM)	0	0.75 ± 0.35	9.39	0.55 - 0.82
Total Mutations (No. of SNPs)	0	0.51 ± 0.56	32.28	0.21 - 0.60

Table A.8: fS-inferred GR clusters and their constituent samples and labels used. Numeric prefix used throughout the chapter. Numbers before populations names under 'Component populations' indicates the number of individuals per *a priori* population e.g. 3Kurumba = 3 individuals from Kurubma. The names consist of three parts as follows 00_[Area]_[linguistic group]_[*a priori population*]. Where "00" is a sequential index and "Area" is a description of region in which the samples are distributed, defaulting to the country of sampling when samples are not widespread. "Linguistic group" is a description of the linguistic families present in the cluster. "*A priori population*" is used when the cluster has three or fewer *a priori* subgroups. Abbreviations found in Table A.9 and A.10

Index	Component populations	fS-cluster name
01	01_3Kurumba;5Kattunayakkan;5Paniyas;3Irula;5Palliyar	01_S.India_DV_.
02	02_1Kamsali;11Vysya;11ITU	02_C.India_DV_Vysya-ITU
03	03_19STU;5MalaiKuravar;5Malli;4Minicoy;7Hallaki;5Gounder; 5Kallar;16ITU;2BEB;1AdiDravider;5Kuruchiyan;2Yadava;1Mala;1Tamil; 2Malayalam;3Kurumba;2Sinhalese;4Naidu;3Kamsali;2Madiga;1Relli; 3Konkani;2Vedda;1Urdu;2Kapu;1Vysya;5Narikkuravar;1Gond;5Tharu; 5Lodi	03_P.India_IN-DV_.
04	04_19STU;11Tamil;1ITU	04_C-S.India_DV_Tamil-STU
05	05_18ITU;4Velama;1Hindi	05_C.India_DV_ITU-Velama
06	06_4Irula;4AdiDravider;14Mala;12IndianLower;11Madiga ; 2ITU;3Kurumba	06_C-S.India_DV_.
07	07_12Bhil;6Chenchu;1Madiga;3Satnami;2Tharu;2Gond;1BEB; 1Relli	07_CW-CE.India_IE-DV_.
08	08_3Bengali;47BEB;1Urdu	08_Bangladesh_IN_Bengali-BEB
09	09_5Jains;1Gujarati;12IndianUpper;2Brahmin;9PJL;6GIH;2ITU; 5Urdu;2Srivastava;1Hindi;1Punjabi;1Sinhalese;5Meghawal;5Bhil	09_NW-C.India_IN_.
10	10_17GIH;6Gujarati	10_CE.India_IN_Gujarati-GIH
11	11_6Kharia;5Gond;1Satnami;3Korku;1KhondaDora;5Santhal	11_CE-CW.India_MN-DV_.
12	12_4Sahariya;6Gond;1Tharu	12_C-NE.India_IN-DV_Sahariya-Gond
13	13_4Birhor;5Bhumij;4Munda;2Santhal;5Ho	13_C.India_MN_.
14	14_9Onge	14_Andaman.Is_JO_Onge
15	15_5Makrani;6Brahui;6Balochi	15_Pakistan_II-DV_Makrani-Brahui-Balochi
16	16_15Pakistanis;2PJL;3Urdu	16_Pakistan_IN_Pakistan-PJL-Urdu
17	17_13PJL;10Pakistanis;5Brahmin;21Urdu;8Hindi; 10KashmiriPandit;12Punjabi;6GIH;10Kshatriya;10Gujarati; 1Pathan;1Bengali;1IndianUpper;4Vaish;1Tharu	17_N.India_IN_.
18	18_6Urdu;6Burusho;6Kalash;12Punjabi;6Pathan;1PJL;5Hindi; 1Pushto;3Gujarati;1GIH	18_N.India_IE_.
19	19_4JuhoanNorth;1Juuhoan;2KhomaniSan	19_S.Africa_KS_Juhoansi-Khomani

20	20_6Yoruba;50YRI	20_Nigeria_NK_Yoruba-YRI
21	21_25GBR;7Orcadian;50Ireland;23UK;5Scotland;10Sweden; 3Norway;1Norwegian;1France;2SwissGerman;16Belgium;24Germany; 17Netherlands;1Denmark;2Icelandic;2English	21_N.Europe_GM_.
22	22_6French;36France;27Belgium;4Austria;25SwissFrench; 22SwissGerman;2UK;19Germany;1Sweden;6SwissItalian;1Ukraine; 1Spain;1Russia;1Italy	22_C.Europe_GM-RM_.
23	23_22Yugoslavia;2Kosovo;3Albania;1Croatia;1Albanian	23_S.Europe_AL-BS_Yugoslavia-Kosovo-Albania
24	24_2Estonian;22Poland;23Slovenian;1Ukraine;11Czech; 15Hungary;5Russia;1Latvia;10Austria;2Slovenia;7Germany;2Hungarian; 3Croatia;1Polish;1Romania	24_E.Europe_IE-UR_.
25	25_13Romania;2Bulgaria;4Hungary;2Bulgarian;5Greece; 4Macedonia;1Czech;20Yugoslavia;1Greek;9BosniaHerzegovina;4Croatia; 3Serbia;1Slovenian	25_S.Europe_HE-RM-UR_.
26	26_17IBS;7Basque;8Sardinian;2Spanish;4Spain;2France; 1French	26_Iberia_BQ-RM_Sardinia-Basque-Iberia
27	27_5SwissItalian;7Bergamo;6Tuscan;25TSI;15Italy;1SwissGerman;1France	27_Italy_RM_.
28	28_8IBS;50Portugal;20Spain	28_Iberia_RM_IBS-Portugal-Spain
29	29_2Saami;7Russian	29_W.Russia_BS-UR_Saami-Russian
30	30_20FIN	30_Finland_UR_FIN
31	31_30FIN;3Finnish;1Finland	31_Finland_UR_FIN-Finnish
32	32_8Palestinian;2BedouinB;3Bedouin;3Jordanian	32_Levant_SM_Palestinian-Bedouin-Jordanian
33	33_2Tajik;2Iranian;2Turkish;24Iraqi;5Turkey;1Chechen; 2Armenian;2Abkhasian;7Adygei;2Georgian;2NorthOssetian;2Lezgin; 1Greek	33_NearEast-Caucasus_CS-IE-TK-SM_.
34	34_2Crete;2SwissItalian;9Italy;3Greece;1Slovakia;4Cyprus; 1Turkey;7Druze;1Samaritan	34_E.Mediterranean_HE-RM-SM_.
35	35_6She;2Miao;3Miaozu;20CHS;6Han;6Tujia;10CHB;1Yizu;4KHV	35_E.Asia_HM-SN_.
36	36_10CHB;2Han;2Korean	36_E.China_SN-KR_CHB-Han-Korean
37	37_50CDX;6Dai	37_N.MainlandSEA_TD_CDX-Dai
38	38_6Lahu;46KHV;1Thai;2Kinh	38_Vietnam_MK-TB_Lahu-KHV-Kihn
39	39_1Burmese;12Thai;3MAS	39_Thailand_MP-TD_Thai-MAS
40	40_6Cambodian;9Thai;1MAS	40_S.MainlandSEA_MK-TD_Cambodian-Thai
41	41_1Atayal;2Ami;1MAS;2Dusun;2Igorot	41_E.Is.SEA_MP_Ami-Dusun-Igorot
42	42_45MAS	42_Singapore_MP_MAS
43	43_4Aonaga;4Subba;4Nysha;3Yizu;2Yi;7Naxi;1Tibetrefugees;5Tu;1Burmese	43_E.India-China_TB_.

44	44_4Sherpa;4Tibetrefugees;5Changapa	44_C.India-Tibet_TB_Sherpa-Tibet-Changapa
45	45_1Oroqen;6Daur;6Mongola;3Hezhen;6Xibo	45_China_MG-TG_.
46	46_20Japan;7Japanese;20JPT	46_Japan_JP_Japan-Japanese-JPT
47	47_25Mongolian;1Mongola	47_Mongolia_MG_Mongolian
48	48_5Oroqen;2Hezhen;2Ulchi	48_China_TG_Oroqen-Hezhen-Ulchi
49	49_3Even;6Yakut	49_Siberia_TG-TK_Even-Yakut
50	50_1EskimoChaplin;2EskimoNaukan;1EskimoSireniki;1Itelman	50_E.Siberia_EA_Eskimo
51	51_2Tlingit;1Chukchi;2Aleut	51_E.Siberia_EA-ND_Aleut-Tlingit
52	52_2Mansi;1Kyrgyzstan;4Uygur	52_China_TK-UR_Mansi-Uygur
53	53_2Tubalar;23Kyrgyzstan;2Kyrgyz;1Altaian	53_Krygyzstan_TK_Tubalar-Krygyzstan-Kyrgyz
54	54_2Nasioi;2Bougainville	54_E.PNG_PN_Nasioi-Bougainville
55	55_15Papuan;1Australian	55_PNG_AU-PN_Australia-Papuan
56	56_26TongaSamoa	56_Polynesia_MP_TongaSamoa

Table A.9: Abbreviations used for linguistic groups in naming the fS-inferred clusters

Abbr.	Language Grouping
DV	Dravidian
JO	Jarawa-Onge
NK	Niger-Kordofanian
KS	KhoeSan
IE	Indo-European
IN	Indic (Indo-Aryan)
II	Indo-Iranian
GM	Germanic
AL	Albanian
BS	Balto-Slavic
BQ	Basque
HE	Hellenic
RM	Romance
SM	Semitic (Arabic)
UR	Uralic
CS	Caucasus Area
TK	Turkic
MG	Mongolic
JP	Japonic
TG	Tugunsic
KR	Korean
HM	Hmong-Mein
TD	Tai-Kadai
TB	Tibeto-Burmese
SN	Sinitic
AA	AustroAsiatic
MN	Munda
MK	Mon-Khmer
MP	Malayo-Polynesian
AU	Australian
PN	Papuan
EA	Eskimo-Aluet
ND	Na-Dene

Table A.10: Abbreviations used for geography in naming the fS-inferred clusters.

Abbr.	Areas
N	North
E	East
S	South
W	West
C	Central
P	Pan
PNG	Papuan New Guinea
SEA	South East Asia
Is.	Island(s)

Table A.11: fS-inferred global reference (GR) clusters identified as sources and their contributions to the SAC community . Shown are the mean proportion ancestry in the SAC (mean) and average jack-knife error (ave. j.e.) when including all individuals and when including only SAC individuals that meet the cut-off criteria (i.e. $mean \ \ jack-knife \ s.e. > 1\%$ ancestry). Also shown are the minimum (min) and maximum (max) values among those who meet the criteria, the number of individuals with no ancestry (No. w/ no ancestry) and the prevalence ($> 1\%$ *ancestry*) in the overall SAC dataset (prevalence).

	Only samples that meet the criteria		All samples					
fS GR cluster	mean	ave. j.e.	mean	ave. j.e.	max	min	No. w/ no ancestry	prevalence
20_Nigeria_NK	0.276	0.154	0.276	0.154	0.800	0.027	0	1.000
19_S.Africa_KS	0.235	0.103	0.235	0.103	0.639	0.009	0	1.000
21_N.Europe_GM	0.085	0.069	0.027	0.067	0.422	-0.084	385	0.408
22_C.Europe_GM-RM	0.081	0.061	-0.006	0.056	0.339	-0.167	598	0.171
42_Singapore_MP	0.049	0.030	0.038	0.033	0.188	-0.024	85	0.772
08_Bangladesh_IN	0.042	0.031	0.003	0.027	0.198	-0.042	540	0.218
28_Iberia_RM	0.037	0.026	0.001	0.028	0.146	-0.053	493	0.266
03_P.India_IN-DV	0.031	0.029	0.010	0.025	0.425	-0.024	337	0.387
06_C-S.India_DV	0.020	0.009	0.003	0.010	0.066	-0.011	426	0.172
15_Pakistan_II-DV	0.020	0.009	0.007	0.011	0.056	-0.008	289	0.330
32_Levant_SM	0.015	0.004	0.003	0.006	0.030	-0.009	347	0.142

Table A.12: Summary of NNLS proportions within the SAC without any exclusion criteria . Shown are the fS GR clusters and if they passed the inclusion criteria (IC; Y - Yes, N - No) and the mean, variance (Var.) and standard deviation (Std. dev.) along with percentiles.

					Percentiles				
IC	fS GR cluster	Mean	Var.	Std. dev.	5%	25%	75%	95%	99%
Y	20_Nigeria_NK_Yoruba-YRI	0.300	0.024	0.155	0.106	0.188	0.376	0.605	0.750
	19_S.Africa_KS_Juhoansi-Khomani	0.258	0.011	0.107	0.097	0.179	0.326	0.447	0.522
	21_N.Europe_GM_.	0.075	0.007	0.083	0.000	0.000	0.116	0.244	0.347
	42_Singapore_MP_MAS	0.057	0.002	0.040	0.001	0.027	0.080	0.131	0.168
	22_C.Europe_GM-RM_.	0.048	0.006	0.076	0.000	0.000	0.080	0.214	0.295
	28_Iberia_RM_IBS-Portugal-Spain	0.033	0.001	0.037	0.000	0.000	0.053	0.108	0.153
	08_Bangladesh_IN_Bengali-BEB	0.031	0.002	0.040	0.000	0.000	0.050	0.114	0.164
	03_P.India_IN-DV_.	0.030	0.001	0.032	0.000	0.007	0.045	0.081	0.116
	15_Pakistan_II-DV_Makrani-Brahui-Balochi	0.018	0.000	0.015	0.000	0.005	0.027	0.046	0.056
	17_N.India_IN_.	0.015	0.001	0.034	0.000	0.000	0.016	0.074	0.165
	06_C-S.India_DV_.	0.013	0.000	0.014	0.000	0.000	0.020	0.040	0.054
	32_Levant_SM_Palestinian-Bedouin-Jordanian	0.010	0.000	0.008	0.000	0.003	0.016	0.026	0.032
	24_E.Europe_IE-UR_.	0.012	0.001	0.024	0.000	0.000	0.016	0.065	0.097
	09_NW-C.India_IN_.	0.012	0.000	0.018	0.000	0.000	0.019	0.051	0.069
	34_E.Mediterranean_HE-RM-SM_.	0.010	0.000	0.019	0.000	0.000	0.013	0.046	0.085
	41_E.Is.SEA_MP_Ami-Dusun-Igorot	0.008	0.000	0.011	0.000	0.000	0.012	0.032	0.051
	27_Italy_RM_.	0.007	0.000	0.020	0.000	0.000	0.000	0.051	0.096
	04_C-S.India_DV_Tamil-STU	0.006	0.000	0.008	0.000	0.000	0.010	0.022	0.032
	18_N.India_IE_.	0.005	0.000	0.009	0.000	0.000	0.008	0.023	0.036
	33_NearEast-Caucasus_CS-IE-TK-SM_.	0.005	0.000	0.008	0.000	0.000	0.007	0.023	0.033
	11_CE-CW.India_MN-DV_.	0.004	0.000	0.008	0.000	0.000	0.006	0.022	0.034
	05_C.India_DV_ITU-Velama	0.004	0.000	0.008	0.000	0.000	0.004	0.022	0.039
	26_Iberia_BQ-RM_Sardinia-Basque-Iberia	0.004	0.000	0.010	0.000	0.000	0.000	0.025	0.047
	12_C-NE.India_IN-DV_Sahariya-Gond	0.003	0.000	0.008	0.000	0.000	0.000	0.021	0.039

55_PNG_AU-PN_Australia-Papuan	0.003	0.000	0.003	0.000	0.001	0.004	0.008	0.012
13_C.India_MN_.	0.002	0.000	0.004	0.000	0.000	0.004	0.009	0.015
56_Polynesia_MP_TongaSamoa	0.002	0.000	0.002	0.000	0.001	0.004	0.007	0.010
01_S.India_DV_.	0.002	0.000	0.003	0.000	0.000	0.004	0.008	0.011
25_S.Europe_HE-RM-UR_.	0.002	0.000	0.009	0.000	0.000	0.000	0.014	0.056
07_CW-CE.India_IE-DV_.	0.002	0.000	0.003	0.000	0.000	0.003	0.008	0.012
46_Japan_JP_Japan-Japanese-JPT	0.002	0.000	0.003	0.000	0.000	0.002	0.007	0.011
39_Thailand_MP-TD_Thai-MAS	0.001	0.000	0.002	0.000	0.000	0.002	0.006	0.009
38_Vietnam_MK-TB_Lahu-KHV-Kihn	0.001	0.000	0.004	0.000	0.000	0.000	0.010	0.020
10_CE.India_IN_Gujarati-GIH	0.001	0.000	0.002	0.000	0.000	0.002	0.007	0.011
02_C.India_DV_Vysya-ITU	0.001	0.000	0.006	0.000	0.000	0.001	0.004	0.006
35_E.Asia_HM-SN_.	0.001	0.000	0.003	0.000	0.000	0.000	0.008	0.019
37_N.MainlandSEA_TD_CDX-Dai	0.001	0.000	0.003	0.000	0.000	0.000	0.007	0.012
54_E.PNG_PN_Nasioi-Bougainville	0.001	0.000	0.001	0.000	0.000	0.002	0.003	0.006
30_Finland_UR_FIN	0.001	0.000	0.002	0.000	0.000	0.000	0.005	0.010
16_Pakistan_IN_Pakistan-PJL-Urdu	0.001	0.000	0.002	0.000	0.000	0.001	0.005	0.008
23_S.Europe_AL-BS_Yugoslavia-Kosovo-Albania	0.001	0.000	0.002	0.000	0.000	0.000	0.005	0.010
29_W.Russia_BS-UR_Saami-Russian	0.001	0.000	0.004	0.000	0.000	0.000	0.003	0.018
43_E.India-China_TB_.	0.001	0.000	0.002	0.000	0.000	0.000	0.005	0.012
53_Krygyzstan_TK_Tubalar-Krygyzstan-Kyrgyz	0.001	0.000	0.002	0.000	0.000	0.000	0.005	0.012
44_C.India-Tibet_TB_Sherpa-Tibet-Changapa	0.000	0.000	0.001	0.000	0.000	0.000	0.003	0.005
47_Mongolia_MG_Mongolian	0.000	0.000	0.001	0.000	0.000	0.000	0.003	0.005
31_Finland_UR_FIN-Finnish	0.000	0.000	0.002	0.000	0.000	0.000	0.004	0.010
40_S.MainlandSEA_MK-TD_Cambodian-Thai	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.011
51_E.Siberia_EA-ND_Aleut-Tlingit	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.005
14_Andaman.Is_JO_Onge	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001
49_Siberia_TG-TK_Even-Yakut	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.003
50_E.Siberia_EA_Eskimo	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002

48_China_TG_Oroqen-Hezhen-Ulchi	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002
45_China_MG-TG_.	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.003
36_E.China_SN-KR_CHB-Han-Korean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
52_China_TK-UR_Mansi-Uygur	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table A.13: List of possible non-European contributors to the SAC genetic heritage and their relationship to the fS-inferred GR clusters available and if identified as a contributor. Some suggestive mentions in the literature are provided ('Numbers') but this is not a comprehensive list.

Region	Origin	Numbers (References)	fS-cluster identified	fS-clusters not identified
Southern Africa	KhoeSan		19_S.Africa_KS	Nothing else available
West Africa	Angola	170 individuals ([152]; [21]; [185])	20_Nigeria_NK	Nothing else available
West Africa	Dahomey (Benin)		20_Nigeria_NK	Nothing else available
West Africa	Guinea	228 individuals ([152])	20_Nigeria_NK	Nothing else available
East Africa	Mozambique	161 individuals ([152]) Taken to Daman and Diu by Gujarati merchants ([133])	20_Nigeria_NK	Nothing else available
East Africa	Madagascar	More than half during the late 18th century. 569 individuals from 5 batches between 1672 to 1682 ([152]) General mention ([152]; [133])	NOTHING AVAILABLE ~20_Nigeria_NK ~42_Singapore_MP	Nothing else available
South Asia West Coast	Bengal	Slave names specified ([133]) 25 individuals in 1701-1790 ([133]) General mention ([152])	08_Bangladesh_IN (exclusively Bengali/Bangladeshi)	none
South Asia West Coast	Coromandel (South West coast along tip)	Slave names specified ([133]) Examples specified; 2 individuals in 1790 ([133]) General mention ([133])	03_P.India_IN-DV 06_C-S.India_DV	01_S.India_DV 07_CW-CE.India_IE-DV (West Coast)
South Asia Sri Lanka	Ceylon (Sri Lanka)	Examples specified; 4 individuals in 1790 ([133]) General mention ([152])	03_P.India_IN-DV (includes Sri Lankan Tamil)	04_C-S.India_DV (exclusively Tamil)
South Asia East Coast	Cochin (India)	23 individuals from the East Coast in 1701-1790 ([133]) General mention ([152]; [133])	03_P.India_IN-DV 06_C-S.India_DV	01_S.India_DV (East coast including Cochin)
Castes sold at Cochin	Polia	53 individuals in 1753 ([133])	none	01_S.India_DV
	Chego	49 individuals in 1753 ([133])	none	07_CW-CE.India_IE-DV
	Bettua	10 individuals in 1753 ([133])	none	-

	Kanaka	6 individuals in 1753 ([133])	none	01_S.India_DV
	Parea	2 individuals in 1753 ([133])	none	-
	Mocqua	2 individuals in 1753 ([133])	none	-
	Moor	2 individuals in 1753 ([133])	none	-
	Marin	1 individuals in 1753 ([133])	none	-
	Oellada	1 individuals in 1753 ([133])	none	-
	Nairo	1 individuals in 1753 ([133])	none	-
South Asia East Coast	Surat	General mention ([133])	None	07_CW-CE.India_IE-DV 09_NW-C.India_IN 10_CE.India_IN 11_CE-CW.India_MN-DV 17_N.India_IN 18_N.India_IE (Includes Gujarat area, North of Surat)
South Asia East Coast	Malabar	Slave names specified ([133]) Examples specified; 29 individuals in 1790 ([133]) 45 individuals in 1701-1790 ([133]) General mention ([152]; [133])	03_P.India_IN-DV 06_C-S.India_DV	01_S.India_DV (East Coast)
South-East Asia	Moluccas Spice Islands or Banda Arc (Indonesia)	General mention ([152])	~42_Singapore_MP	Nothing else available
South-East Asia	Batavia Jakarta West Java (Indonesia)	102 individuals (in one year) ([152]) examples specified ([133]) Slave names specified ([133]) General mention ([152]; [133])	~42_Singapore_MP	Nothing else available
South-East Asia	Central Java	11 individuals in 1701-1790 ([133])	~42_Singapore_MP	Nothing else available
South-East Asia	Bali	examples specified ([133])	~42_Singapore_MP	Nothing else available
South-East Asia	Makassar (South Sulawesi/Celebes)	Slave names specified ([133]) 19 individuals in 1701-1790 ([133]) examples specified ([133])	~42_Singapore_MP	Nothing else available
South-East Asia	Mandhaar/Mandhar (south Sulawesi Indonesia)	Slave names specified ([133]) examples specified ([133])	~42_Singapore_MP	Nothing else available
South-East Asia	Timor	examples specified ([133])	~42_Singapore_MP	Nothing else available

South-East Asia	Batak-Toba (north Sumatra)		~42_Singapore_MP	Nothing else available
South-East Asia	Bugis	As slave traders around Maluku ([133]) Slave names specified ([133])	~42_Singapore_MP	Nothing else available
South-East Asia	Nias	examples specified ([133])	~42_Singapore_MP	Nothing else available
East Asia	Chinese	As slave traders around Maluku and of Balinese slaves (17-18th) ([133])	None	Eight possible sources

B

Chapter 5 Supplementary

B.1 Supplementary Figures

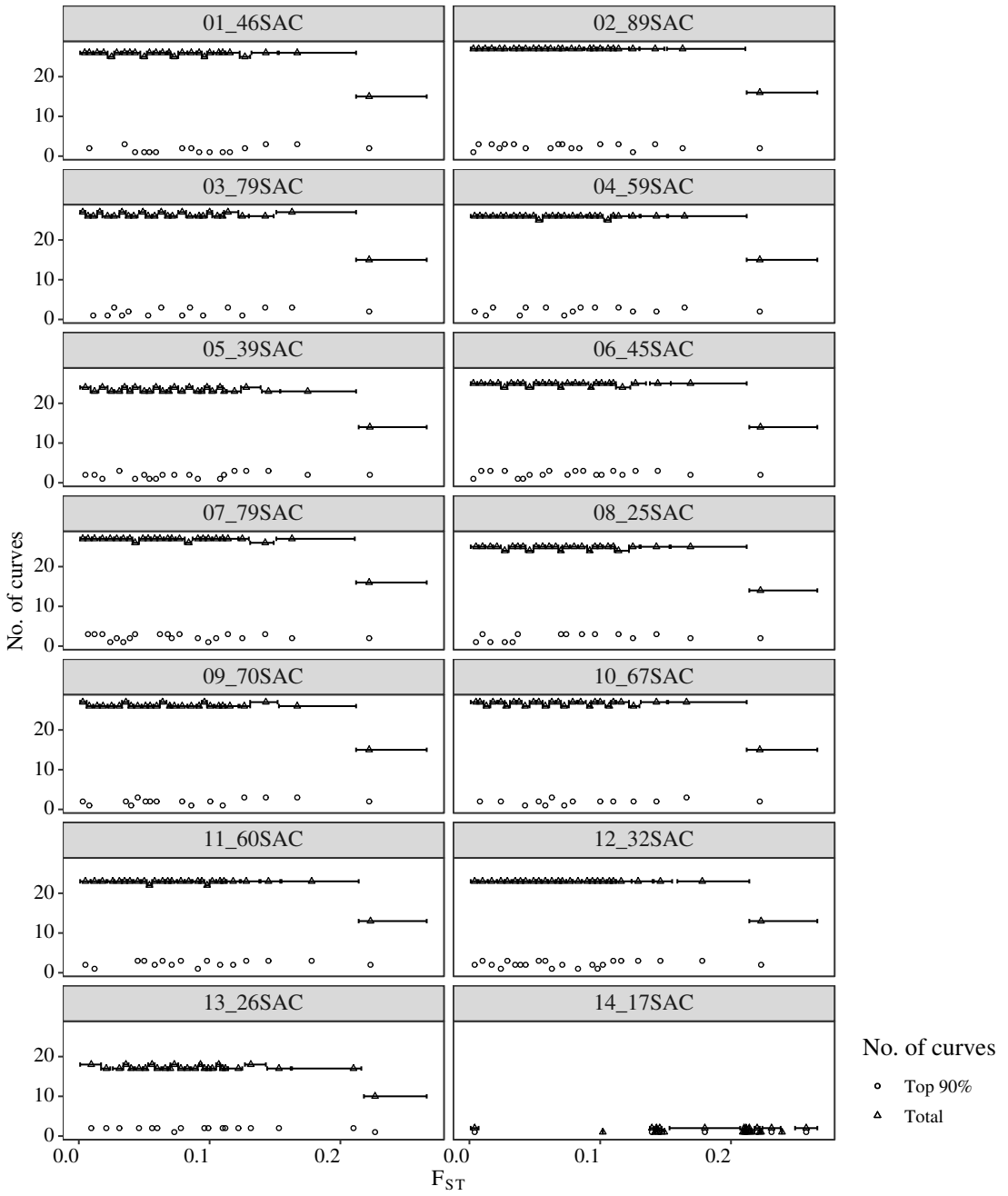


Figure B.1: Summary of the F_{ST}-based binning procedure used to identify top amplitudes for the first event. Results are displayed per fS-inferred SAC cluster. Shown are the number of curves in the dataset (Total) and in the top 90% per F_{ST} bin. Points plotted on the median F_{ST} value, bars indicate the minimum and maximum F_{ST} within the bin.

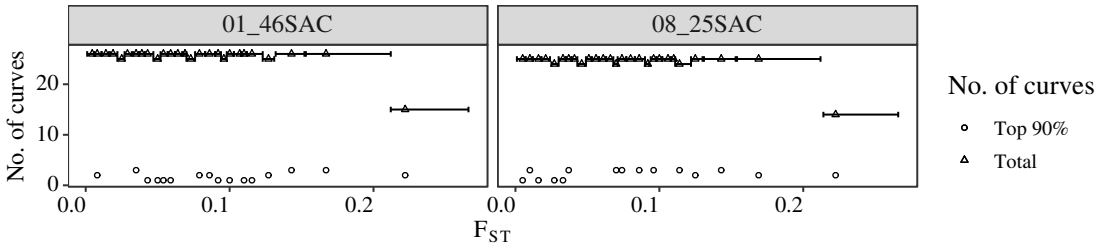
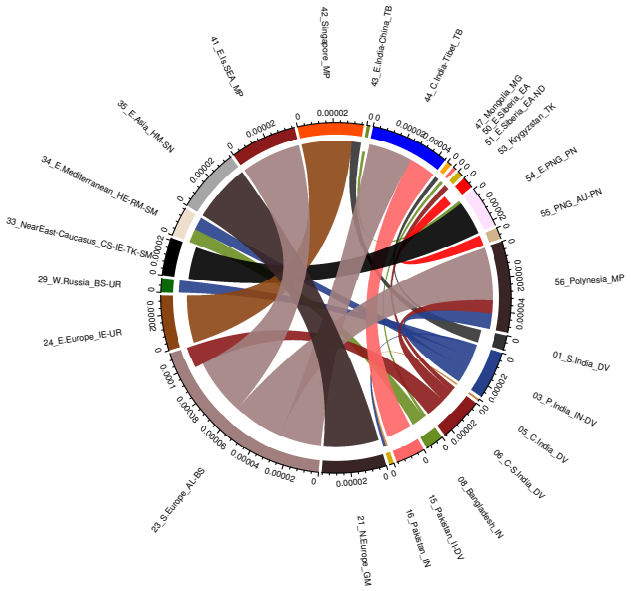
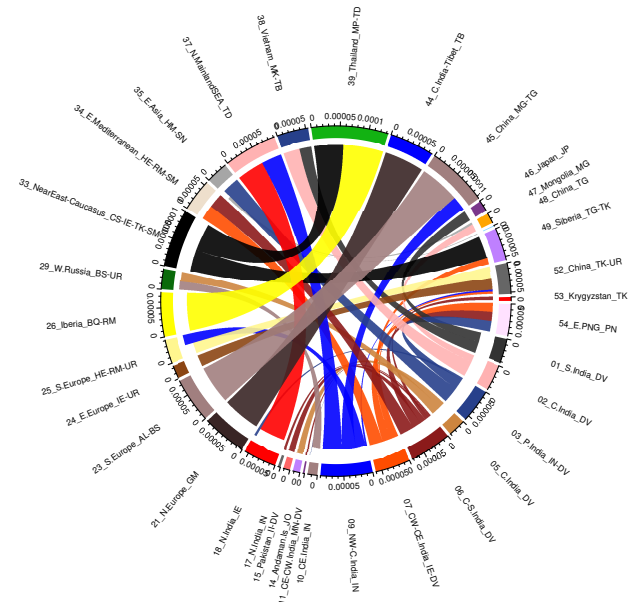


Figure B.2: Summary of the F_{ST}-based binning procedure used to identify top amplitudes for the second event. Results are displayed per fS-inferred SAC cluster. Shown are the number of curves in the dataset (Total) and in the top 90% per F_{ST} bin. Points plotted on the median F_{ST} value, bars indicate the minimum and maximum F_{ST} within the bin.

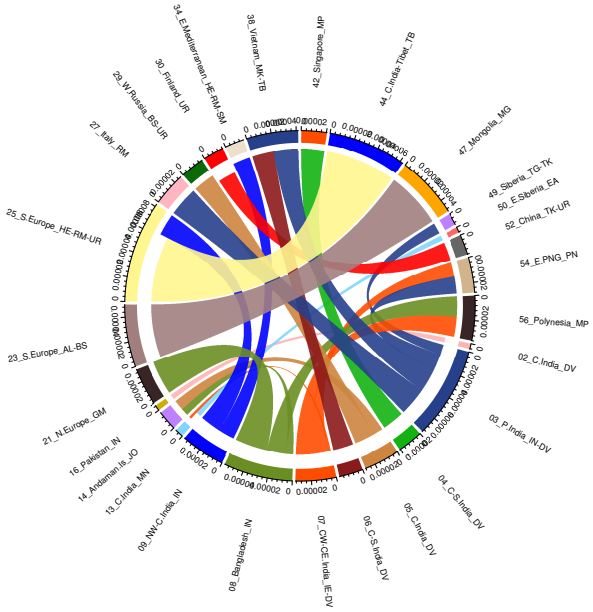


(a) 01_46SAC

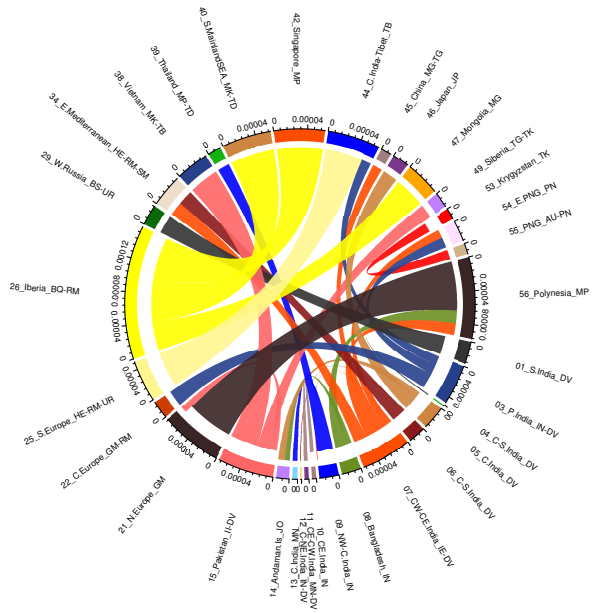


(b) 02_89SAC

Figure B.3: Chord plot visualisation of top 90% of LD decay curves ranked by amplitude for a specified F_{ST} bin (late events). Subplots are of the SAC fS-inferred clusters. Chords connecting GR sources indicate a curve recovered and the width indicates the amplitude size. Only non-African GR curves are included.

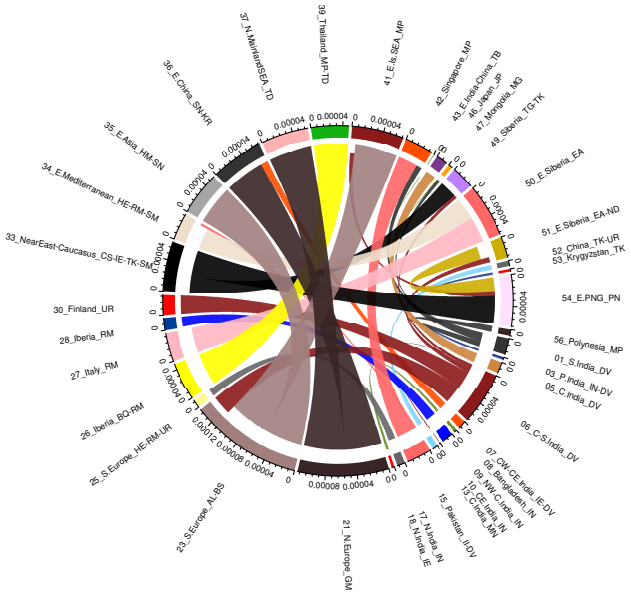


(a) 03_79SAC

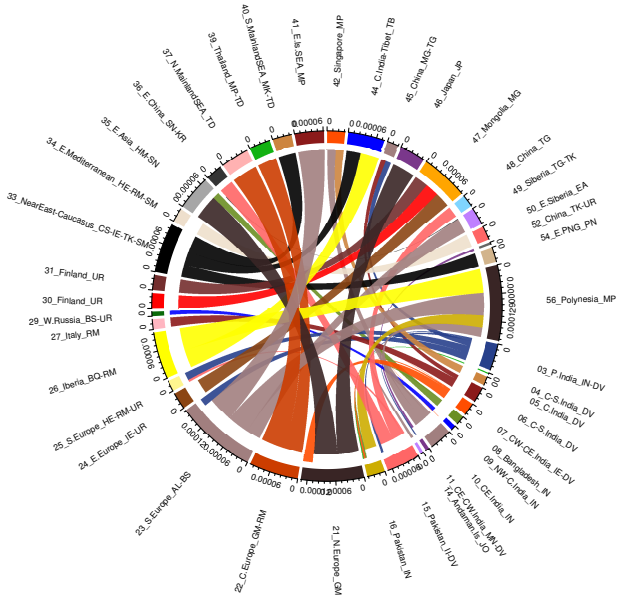


(b) 04_59SAC

Figure B.4: Chord plot visualisation of top 90% of LD decay curves ranked by amplitude for a specified F_{ST} bin (late events) (continued from previous plot...) Subplots are of the SAC fS-inferred clusters. Chords connecting GR sources indicate a curve recovered and the width indicates the amplitude size. Only non-African GR curves are included.

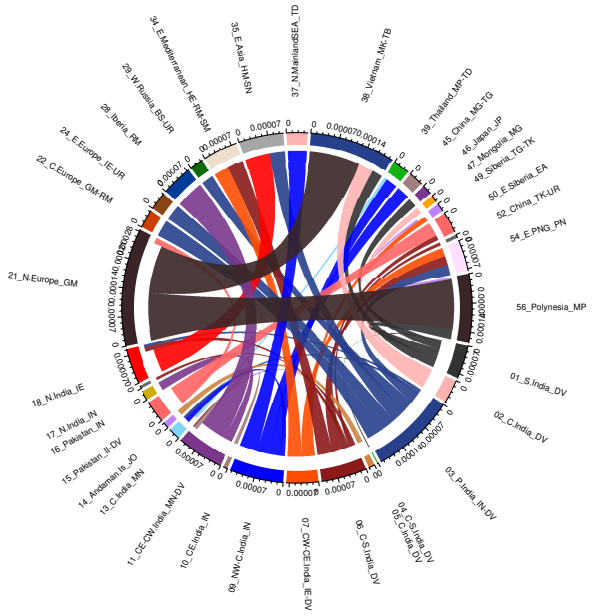


(a) 05_39SAC

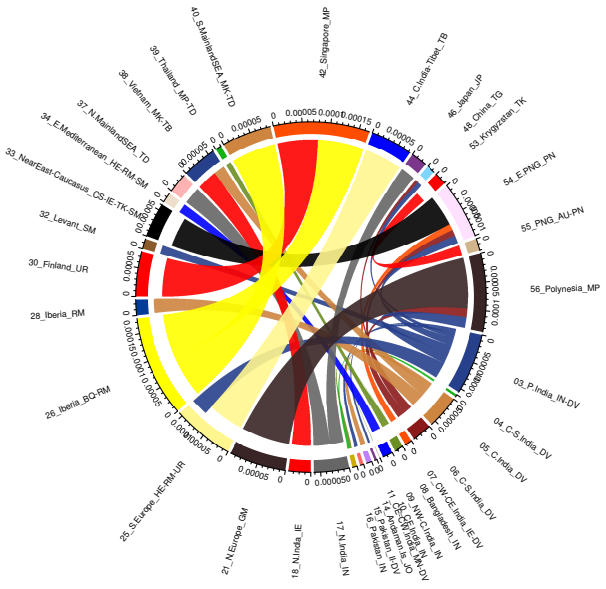


(b) 06_45SAC

Figure B.5: Chord plot visualisation of top 90% of LD decay curves ranked by amplitude for a specified F_{ST} bin (late events) (continued from previous plot...) Subplots are of the SAC fS-inferred clusters. Chords connecting GR sources indicate a curve recovered and the width indicates the amplitude size. Only non-African GR curves are included.

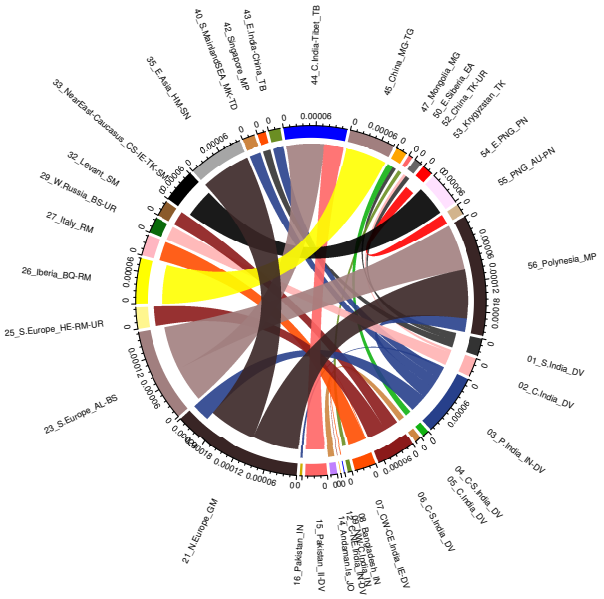


(a) 07_79SAC

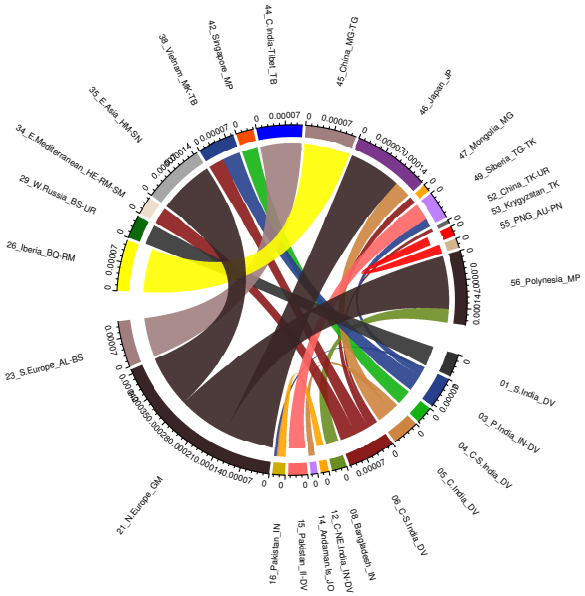


(b) 08_25SAC

Figure B.6: Chord plot visualisation of top 90% of LD decay curves ranked by amplitude for a specified F_{ST} bin (late events) (continued from previous plot...) Subplots are of the SAC fS-inferred clusters. Chords connecting GR sources indicate a curve recovered and the width indicates the amplitude size. Only non-African GR curves are included.

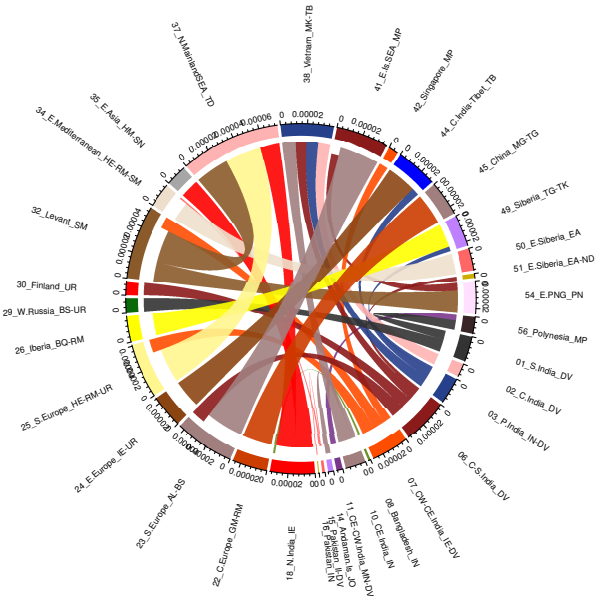


(a) 09_70SAC

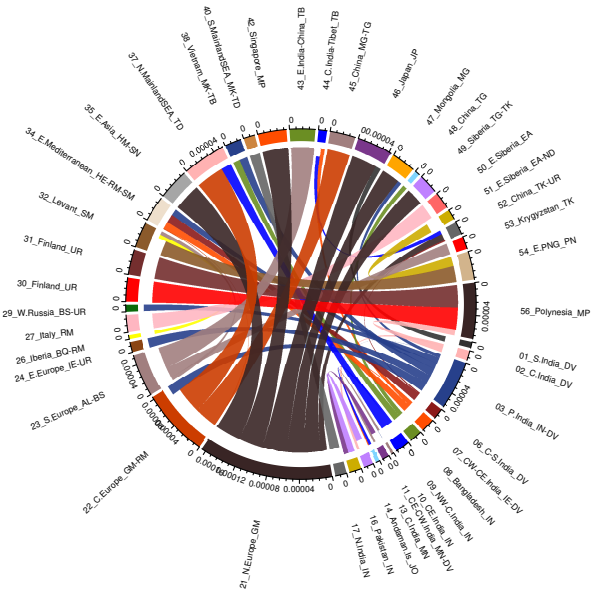


(b) 10_67SAC

Figure B.7: Chord plot visualisation of top 90% of LD decay curves ranked by amplitude for a specified F_{ST} bin (late events) (continued from previous plot...) Subplots are of the SAC fS-inferred clusters. Chords connecting GR sources indicate a curve recovered and the width indicates the amplitude size. Only non-African GR curves are included.

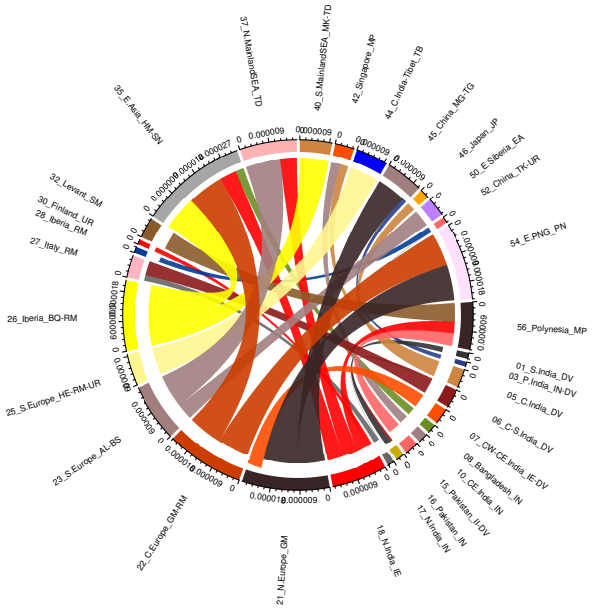


(a) 11_60SAC



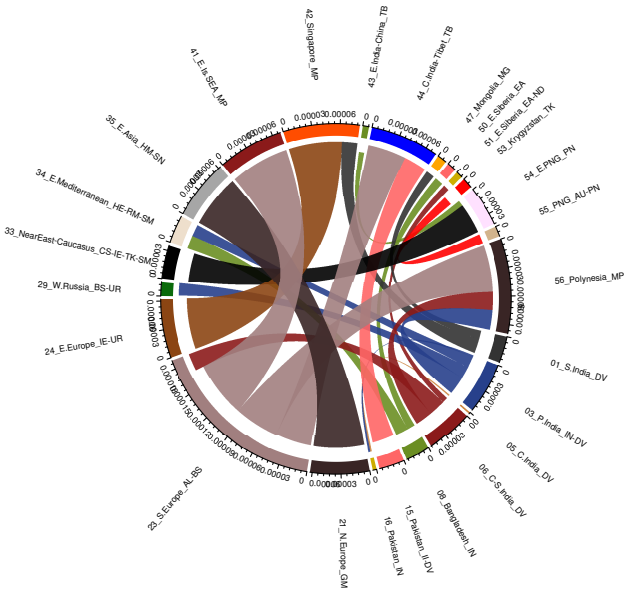
(b) 12_32SAC

Figure B.8: Chord plot visualisation of top 90% of LD decay curves ranked by amplitude for a specified F_{ST} bin (late events) (continued from previous plot...) Subplots are of the SAC fS-inferred clusters. Chords connecting GR sources indicate a curve recovered and the width indicates the amplitude size. Only non-African GR curves are included.

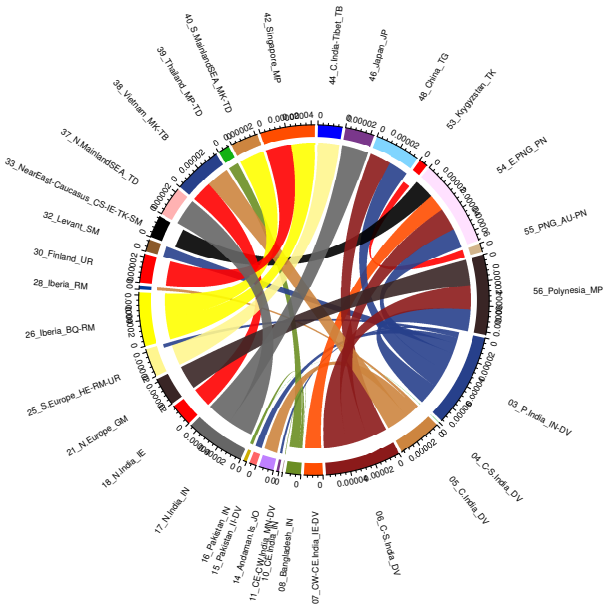


(a) 13_26SAC

Figure B.9: Chord plot visualisation of top 90% of LD decay curves ranked by amplitude for a specified F_{ST} bin (late events) (continued from previous plot...) Subplots are of the SAC fS-inferred clusters. Chords connecting GR sources indicate a curve recovered and the width indicates the amplitude size. Only non-African GR curves are included.



(a) 01_46SAC



(b) 08_25SAC

Figure B.10: Chord plot visualisation of the top 90% of LD decay curves ranked by amplitude for a specified F_{ST} bin (early event). Subplots are of the SAC fS-inferred clusters. Chords connecting GR sources indicate a curve recovered and width indicates the amplitude size. Only non-African GR curves are included.

B.2 Supplementary Tables

Table B.1: Summary of the bootstrap iterations which best match the events found from the fS-inferred SAC clusters A total of 993 iterations considered. Indicated are the total number and proportion of iterations (total and prop. iteration) which fall within the error margin for the event (= match), the minimum, maximum and mean percentage sample overlap (SO %; individuals in the SAC cluster/total individuals in iteration) with the respective fS-inferred SAC cluster (min, max, mean) for all iteration matching (match), all iterations not matching (all non-match) and for those not matching subsetted to a similar sample size (ss non-match) as for 'match'. Results of a t-test between 'match' and 'ss non-match' of the SO with the respective SAC fS-inferred cluster are indicated where possible, reporting the T statistic (T), degrees of freedom (df) and p-value (p). * indicates values significant following Šidák-Holms sequential correction [259]

Focal event		Matching iterations		Sample overlap with focal SAC fS-cluster (%)									t-test		
				match			ss non-match			all non-match					
Event	fS-cluster	total	prop.	min	max	mean	min	max	mean	min	max	mean	T	df	p
Late	01_46SAC	796	0.802	0.000	0.226	0.063	0.000	0.167	0.062	0.000	0.167	0.061			
Early	01_46SAC	107	0.108	0.000	0.148	0.061	0.000	0.155	0.062	0.000	0.226	0.063	1.200	211.299	0.232
Early	08_25SAC	1	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.235	0.036			
Late	14_17SAC	20	0.020	0.000	0.045	0.011	0.000	0.067	0.023	0.000	0.182	0.022	-2.051	32.164	0.048
Single	02_89SAC	8	0.008	0.067	0.370	0.185	0.024	0.300	0.131	0.000	0.364	0.120	-0.828	13.334	0.422
Single	03_79SAC	4	0.004	0.071	0.200	0.145	0.069	0.128	0.107	0.000	0.381	0.110	-3.434	3.443	0.033
Single	04_59SAC	1	0.001	0.133	0.133	0.133	0.058	0.058	0.058	0.000	0.355	0.081			
Single	05_39SAC	1	0.001	0.200	0.200	0.200	0.033	0.033	0.033	0.000	0.241	0.053			
Single	06_45SAC	1	0.001	0.067	0.067	0.067	0.029	0.029	0.029	0.000	0.300	0.062			
Single	07_79SAC	3	0.003	0.048	0.273	0.129	0.000	0.138	0.056	0.000	0.333	0.107	-0.512	2.279	0.654
Single	08_25SAC	11	0.011	0.000	0.235	0.060	0.000	0.091	0.032	0.000	0.152	0.036	-0.911	18.874	0.374
Single	09_70SAC	13	0.013	0.000	0.176	0.084	0.045	0.167	0.095	0.000	0.280	0.094	-4.099	17.143	0.001*
Single	10_67SAC	28	0.028	0.000	0.167	0.054	0.000	0.136	0.095	0.000	0.321	0.091	-5.225	37.759	0.000*
Single	11_60SAC	1	0.001	0.133	0.133	0.133	0.079	0.079	0.079	0.000	0.312	0.082			
Single	12_32SAC	1	0.001	0.000	0.000	0.000	0.016	0.016	0.016	0.000	0.267	0.042			
Single	13_26SAC	8	0.008	0.000	0.056	0.016	0.000	0.087	0.045	0.000	0.188	0.035	-3.819	8.262	0.005*
All	02 - 13	36	0.036	0.833	1.000	0.920	0.824	1.000	0.920	0.714	1.000	0.915	-7.555	46.947	0.000*
All	01 and 14	986	0.993	0.000	0.286	0.085	0.000	0.105	0.036	0.000	0.105	0.044			

C

Chapter 6 Supplementary

C.1 Supplementary Figures

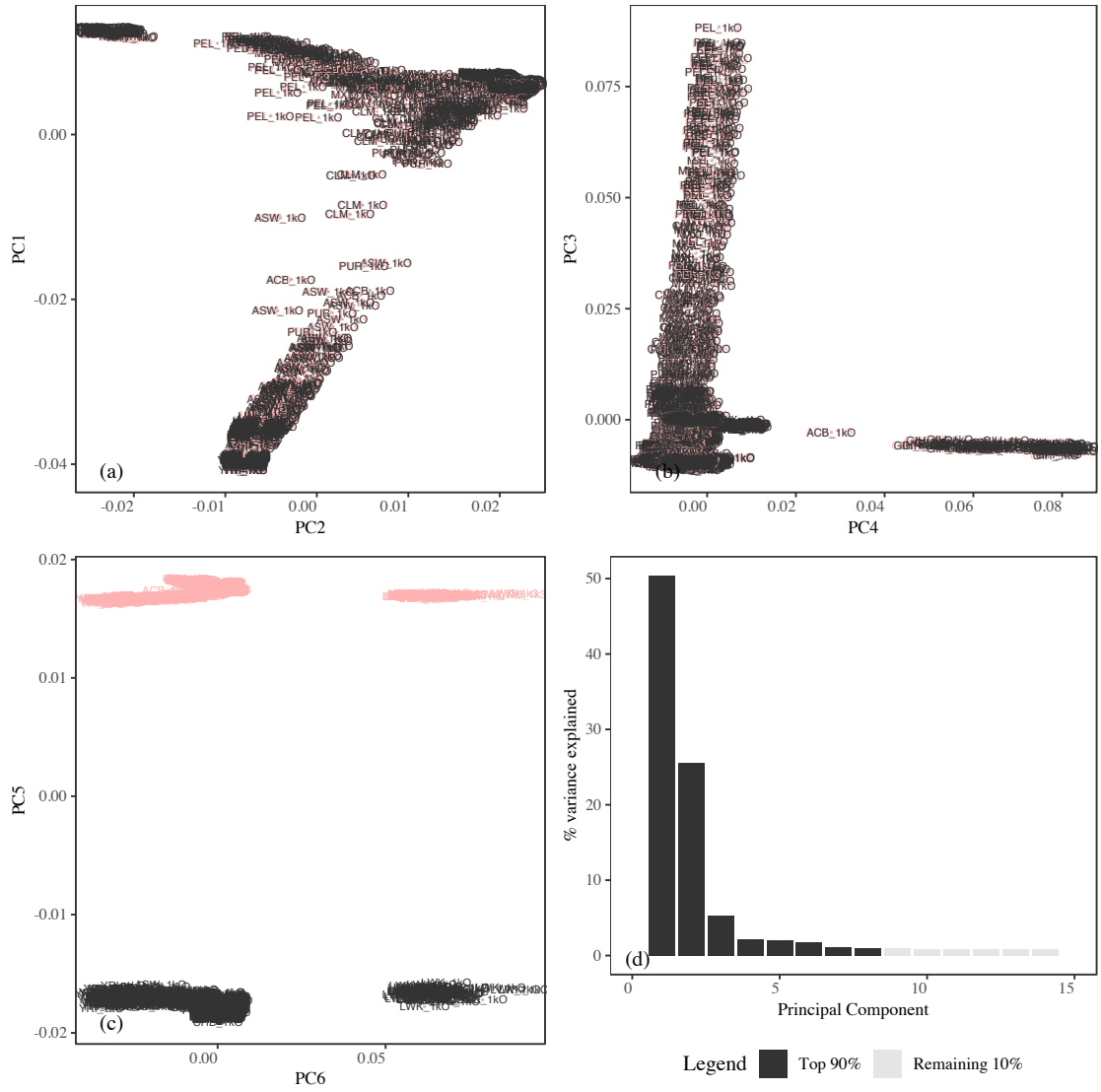


Figure C.1: Principal Component Analysis demonstrating batch effect in the 1KGP phase3 WGS data merge before removal of discordant SNPs. (a-c) Plotted are 1,623 individuals genotyped with WGS and on the Illumina Omni1M SNP chip before removal of discordant SNPs (7,956). The effect of discordant SNPs can be seen in PC5. (d) Percentage variance explained by each of the top 15 principal components.

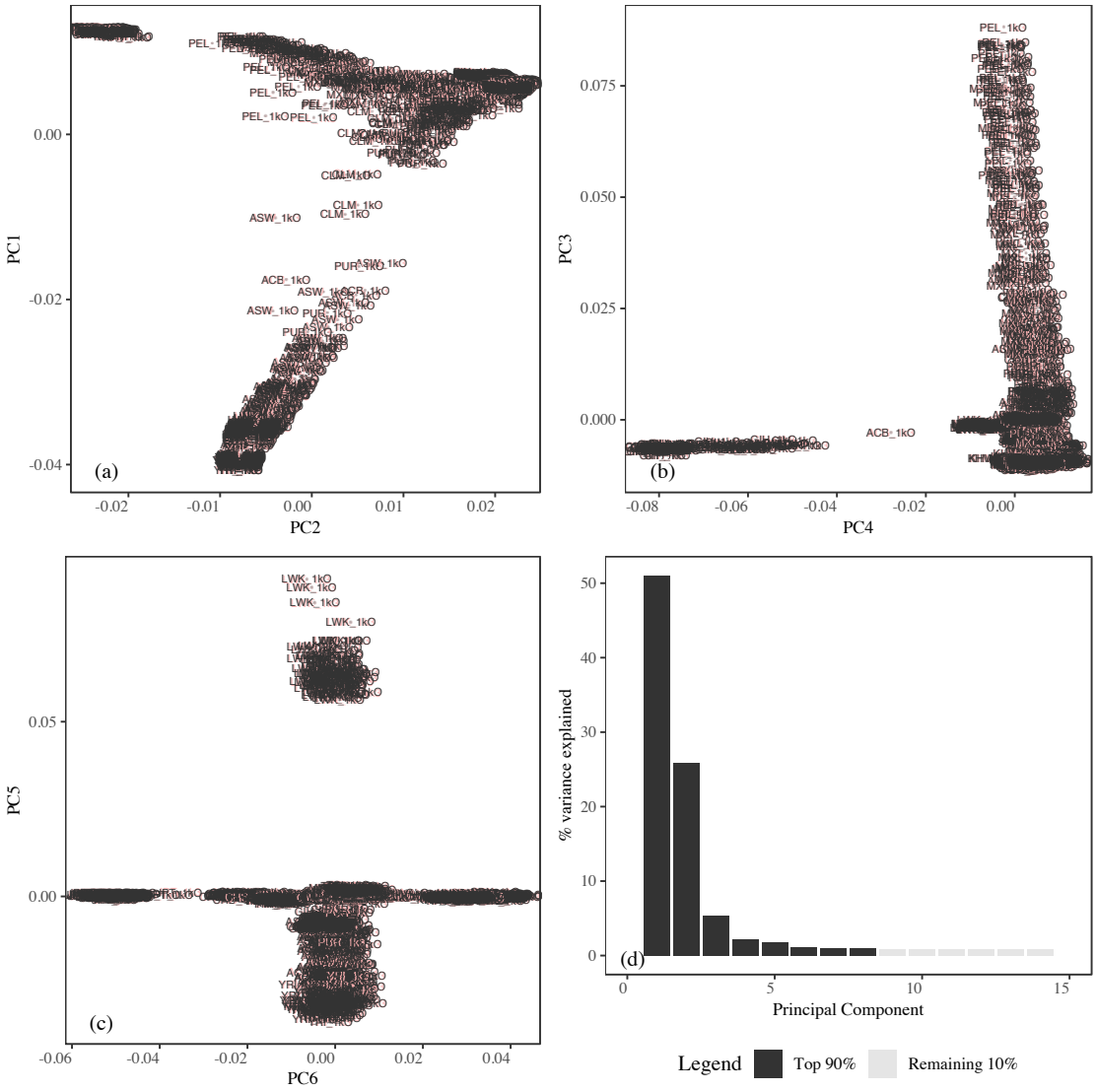


Figure C.2: Principal Component Analysis demonstrating batch effect in the 1KGP phase3 WGS data merge after the removal of discordant SNPs. (a-c) Plotted are 1,623 individuals genotyped with WGS and the on the Illumina Omni1M SNP chip after removal of discordant SNPs (7,956). The effect of discordant SNPs was removed as can be seen in PC5. (d) Percentage variance explained by each of the top 15 principal components.

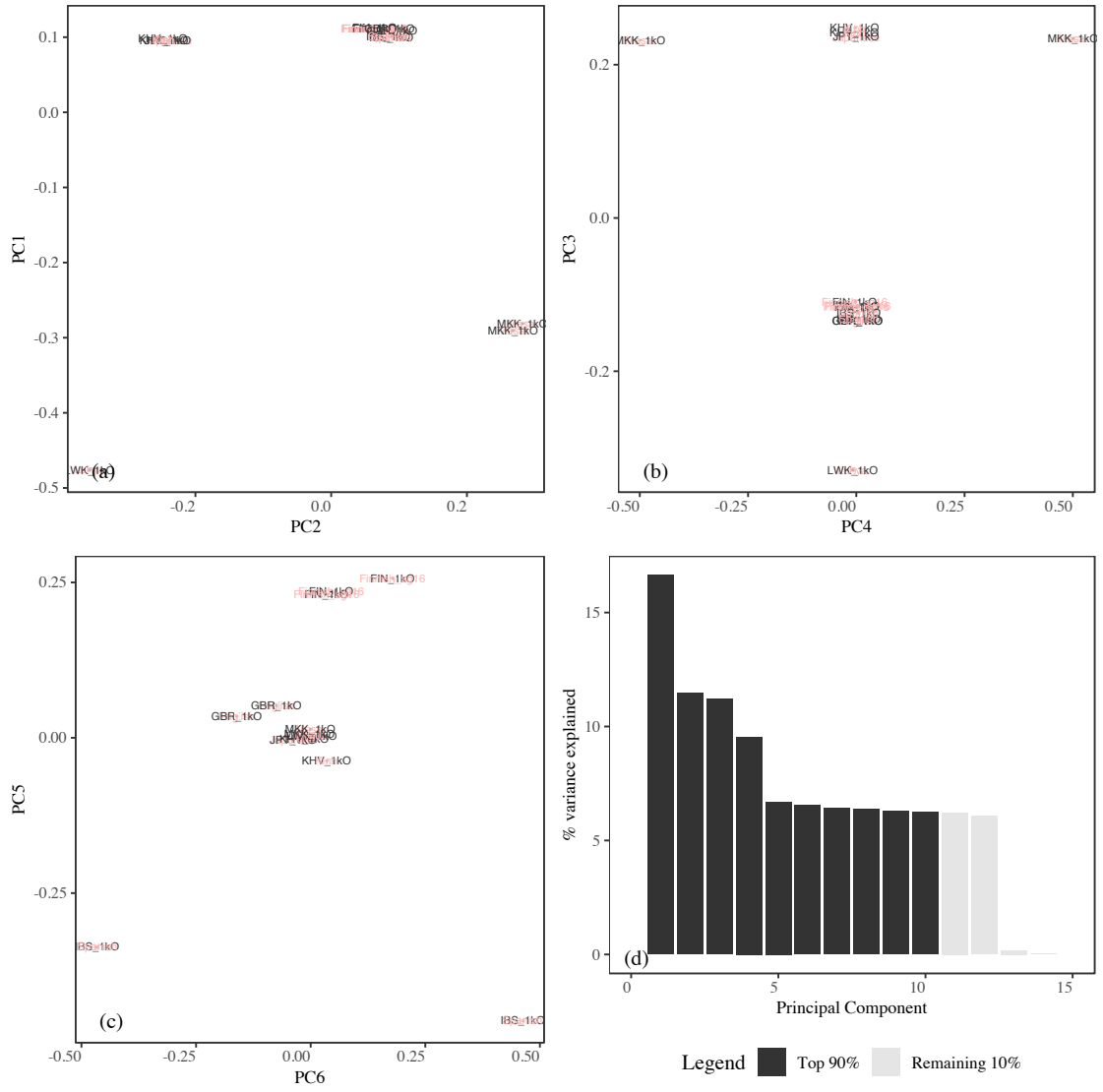


Figure C.3: Principal Component Analysis demonstrating batch effect in the Simons Genome Diversity Project WGS data merge. (a-c) Plotted are the same 13 individuals genotped in the SGDP and the 1000 Genome Project on the Illumina Omni1M SNP chip before the removal of the discordant SNPs (36,011). No effect of the discordant SNPs is detected here. (d) Percentage variance explained by each of the top 15 principal components.

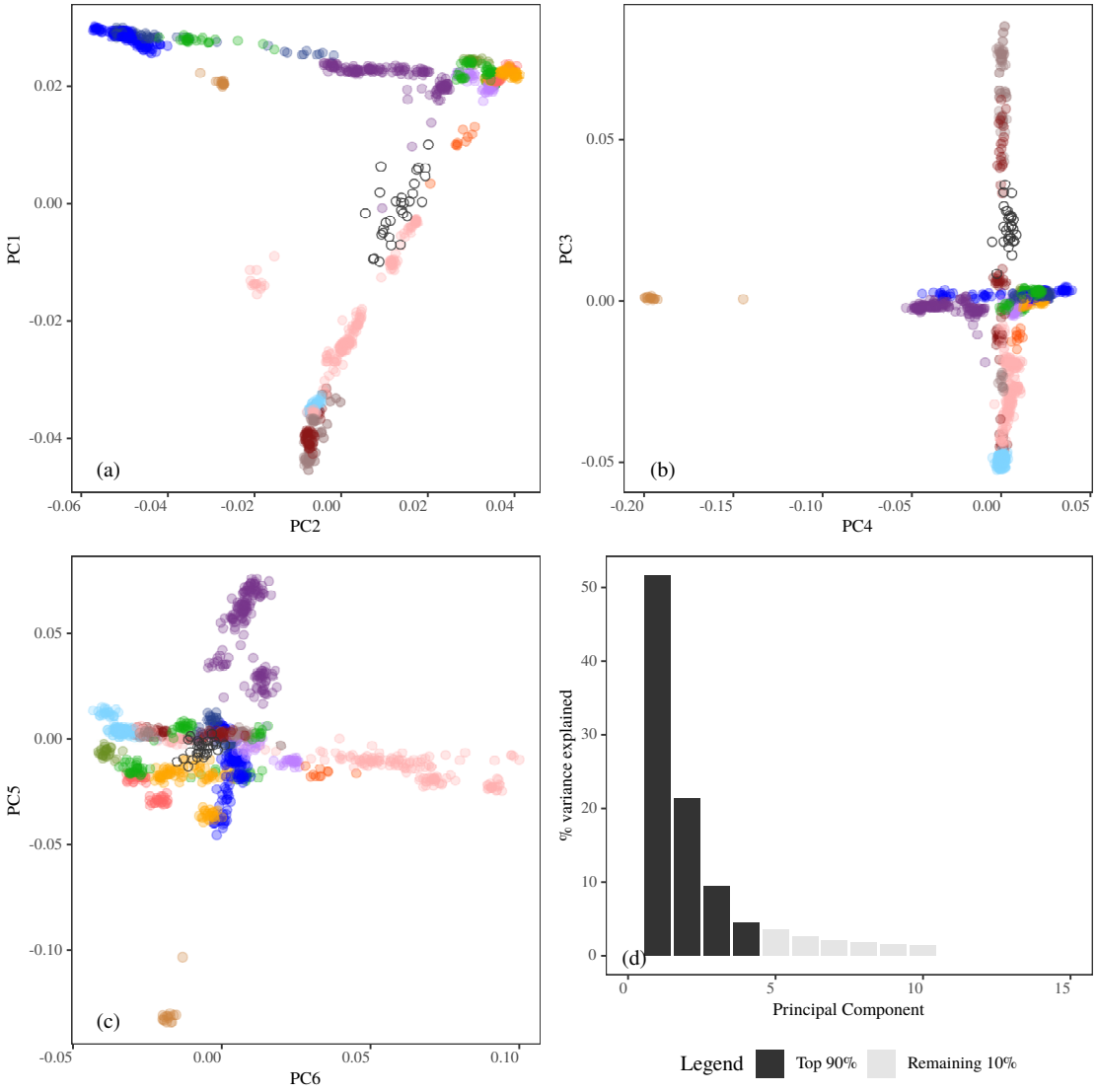


Figure C.4: PCA plots for Basters. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

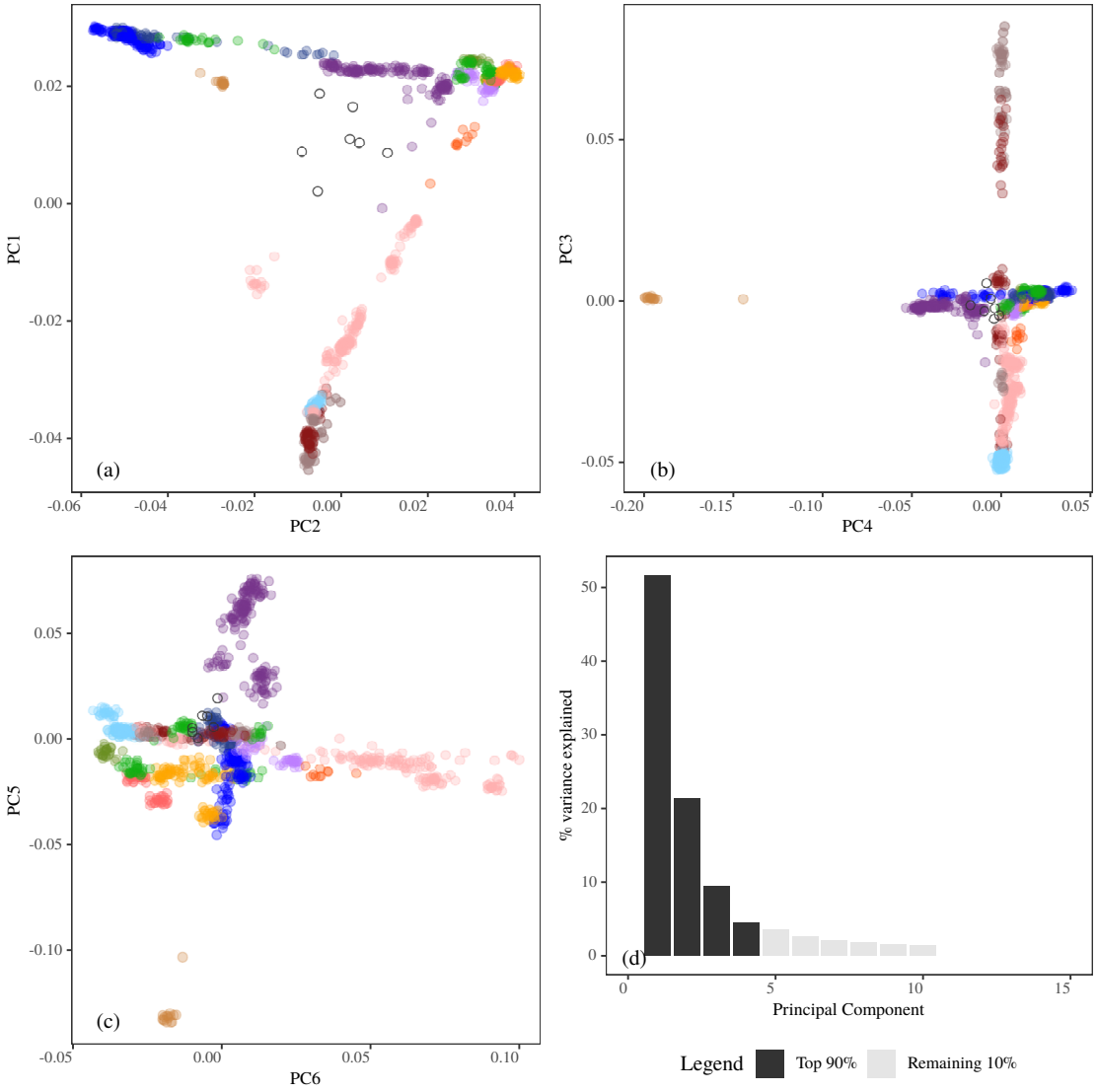


Figure C.5: PCA plots for CAPEMALAY_CPT. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

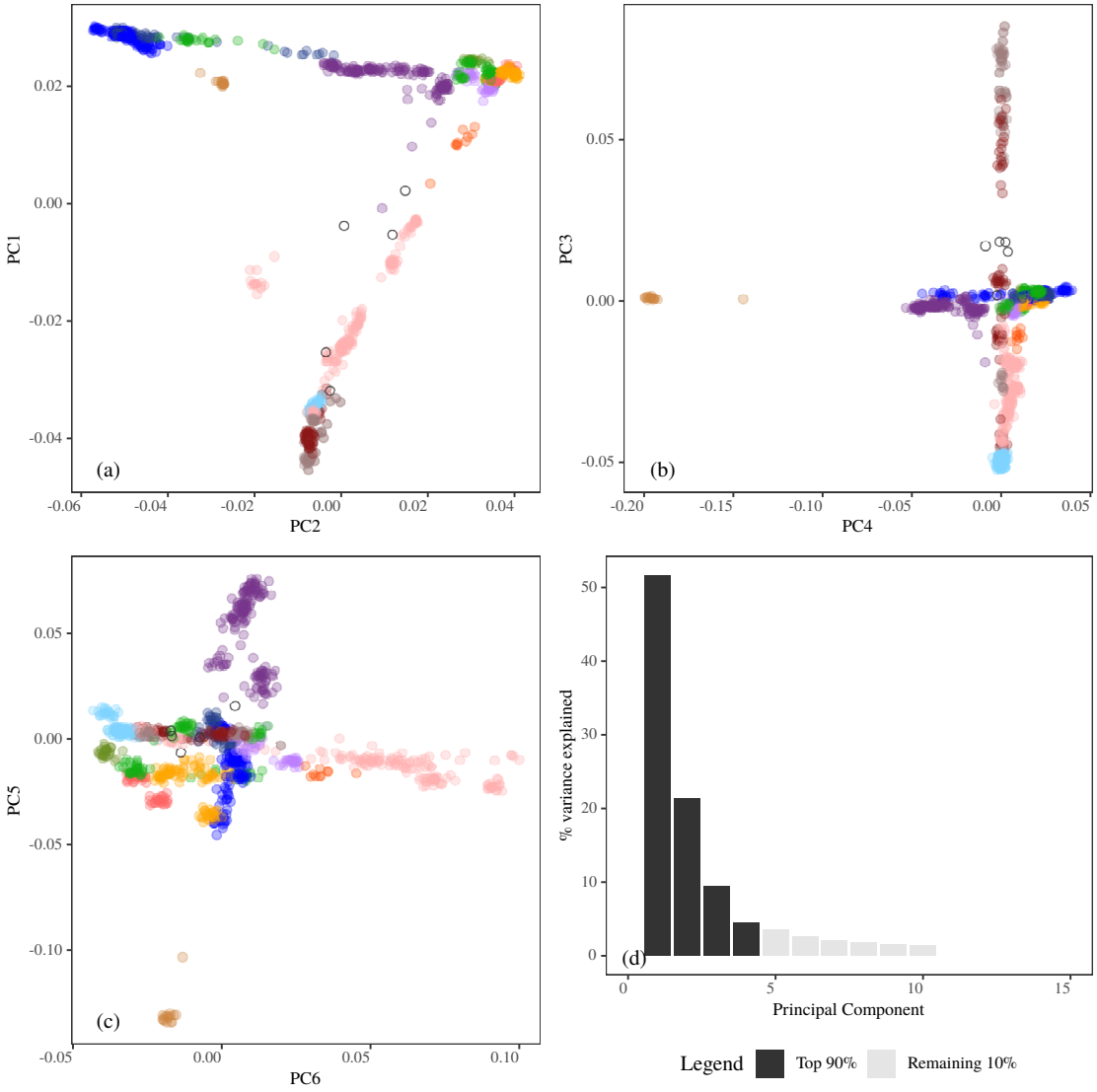


Figure C.6: PCA plots for COLOURED_BFN. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

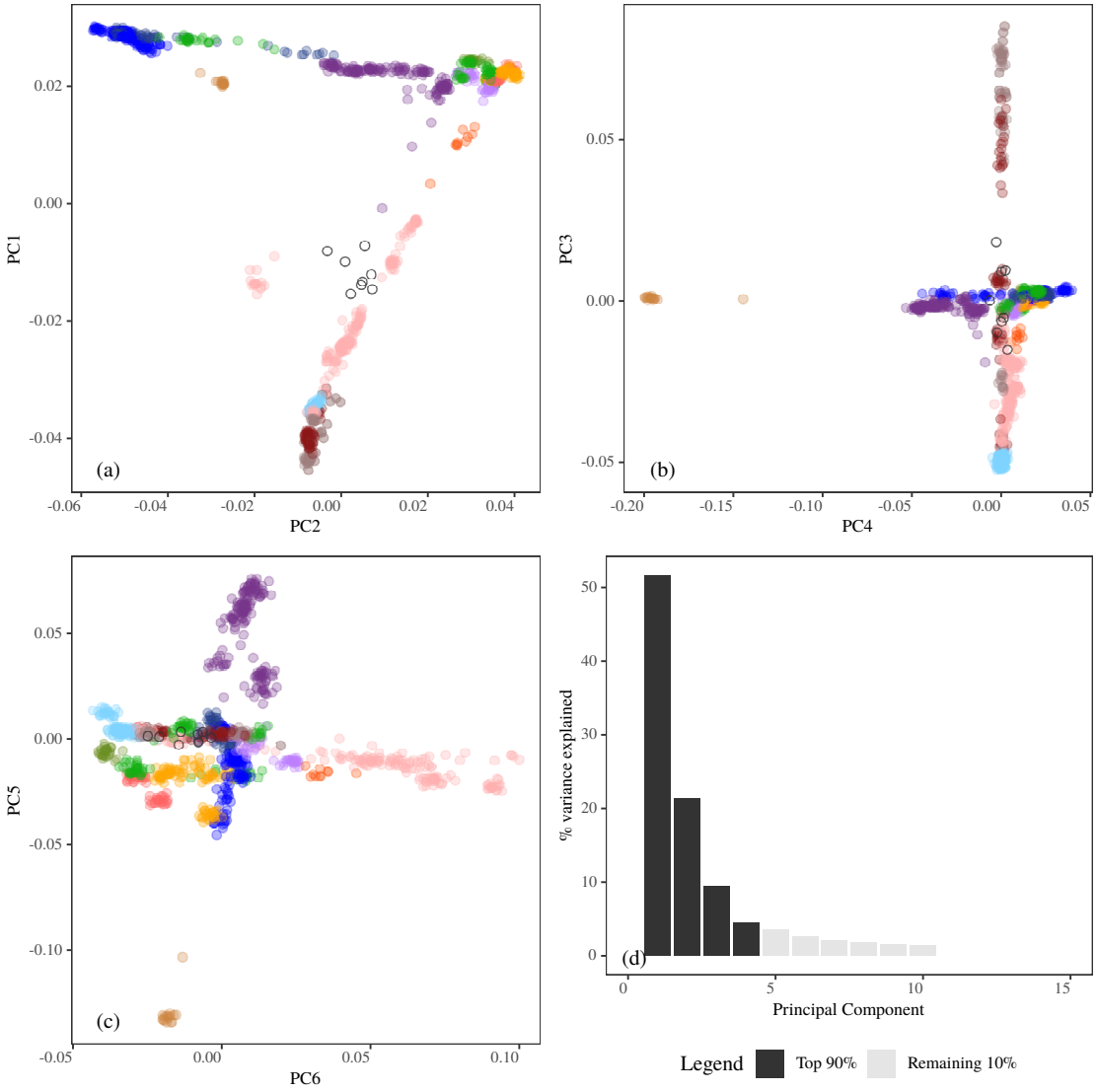


Figure C.7: PCA plots for COLOURED_BRT. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

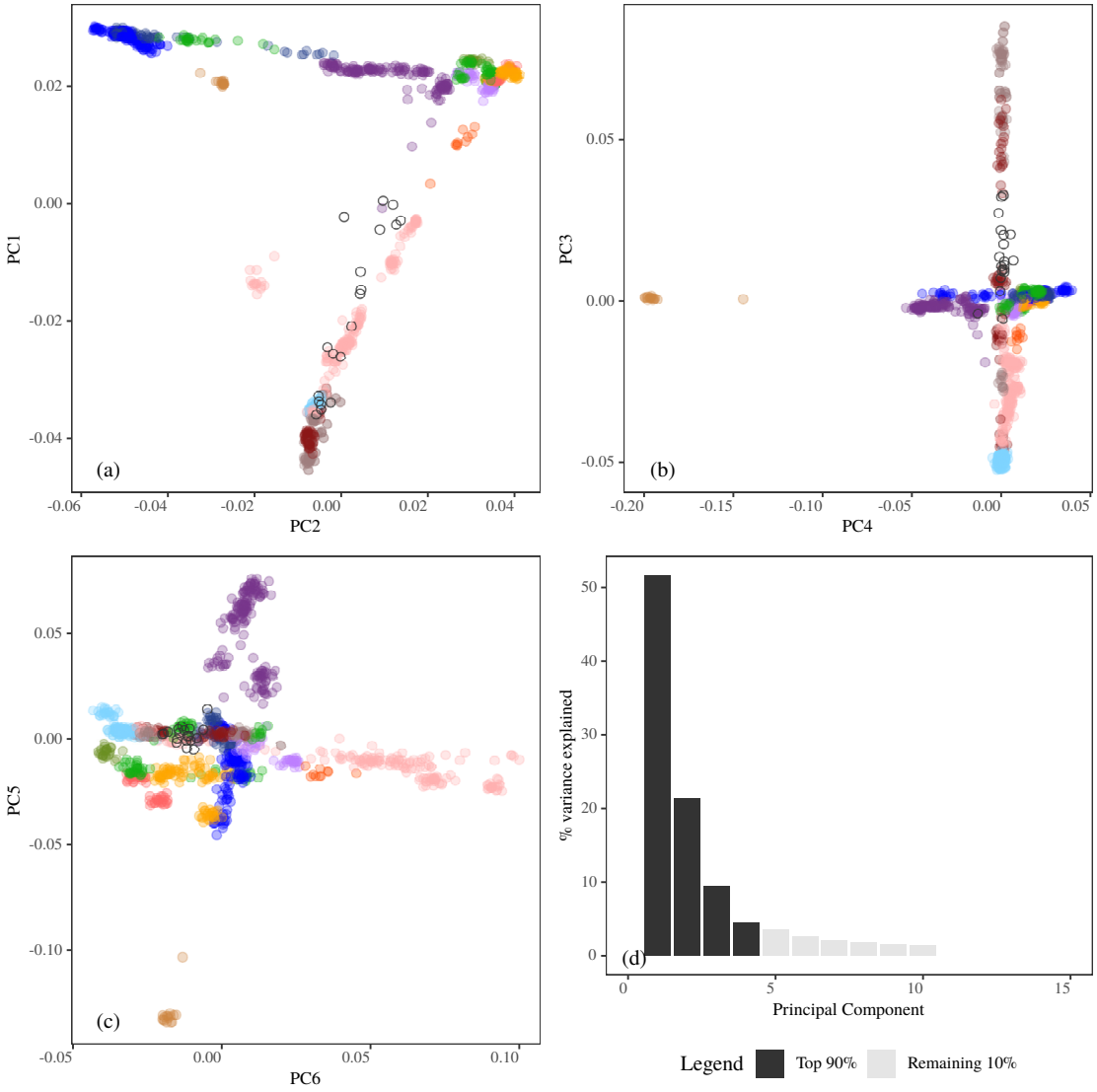


Figure C.8: PCA plots for ColouredColesberg. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

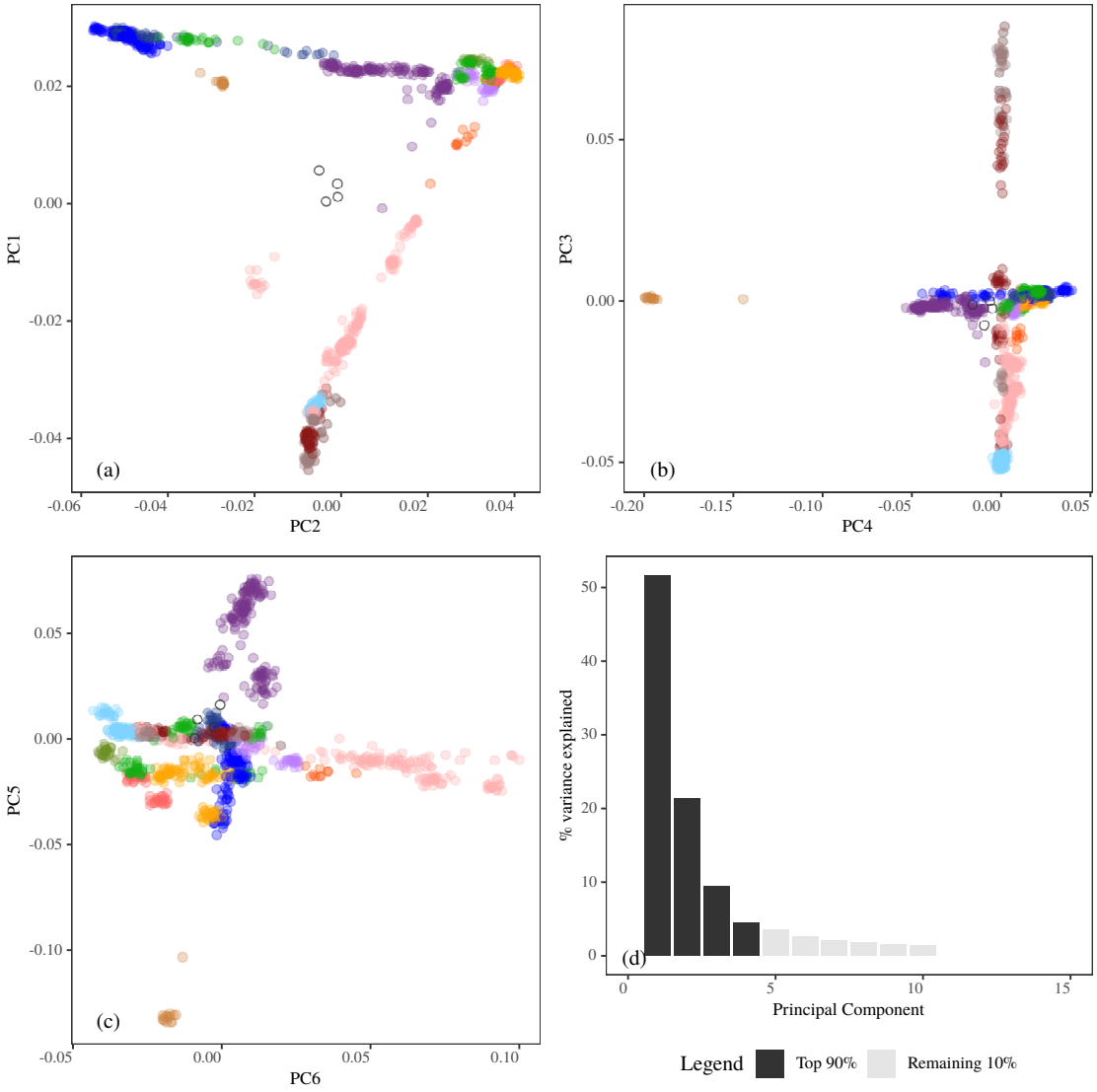


Figure C.9: PCA plots for COLOURED_CPT. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

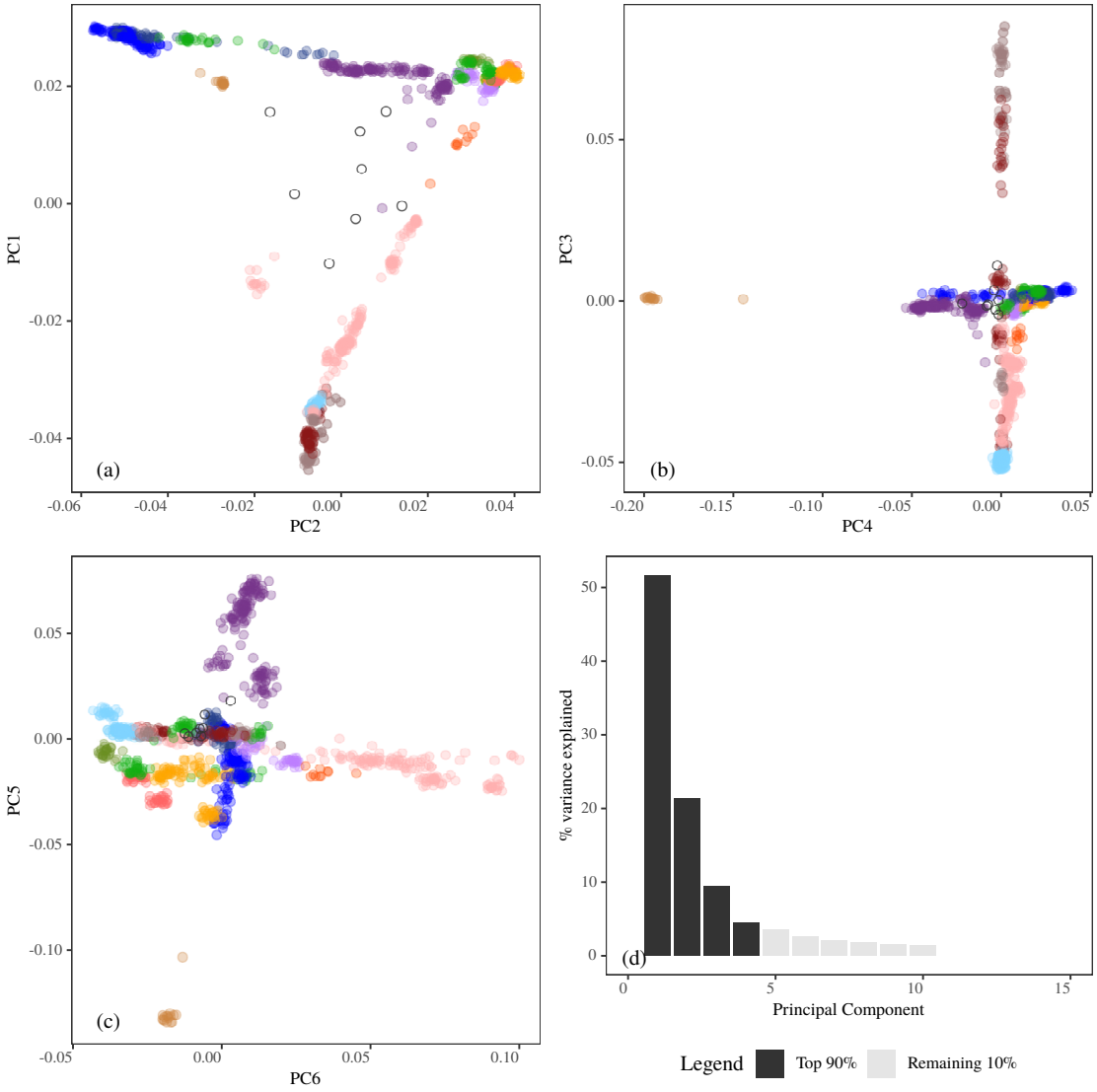


Figure C.10: PCA plots for Coloured-D6. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

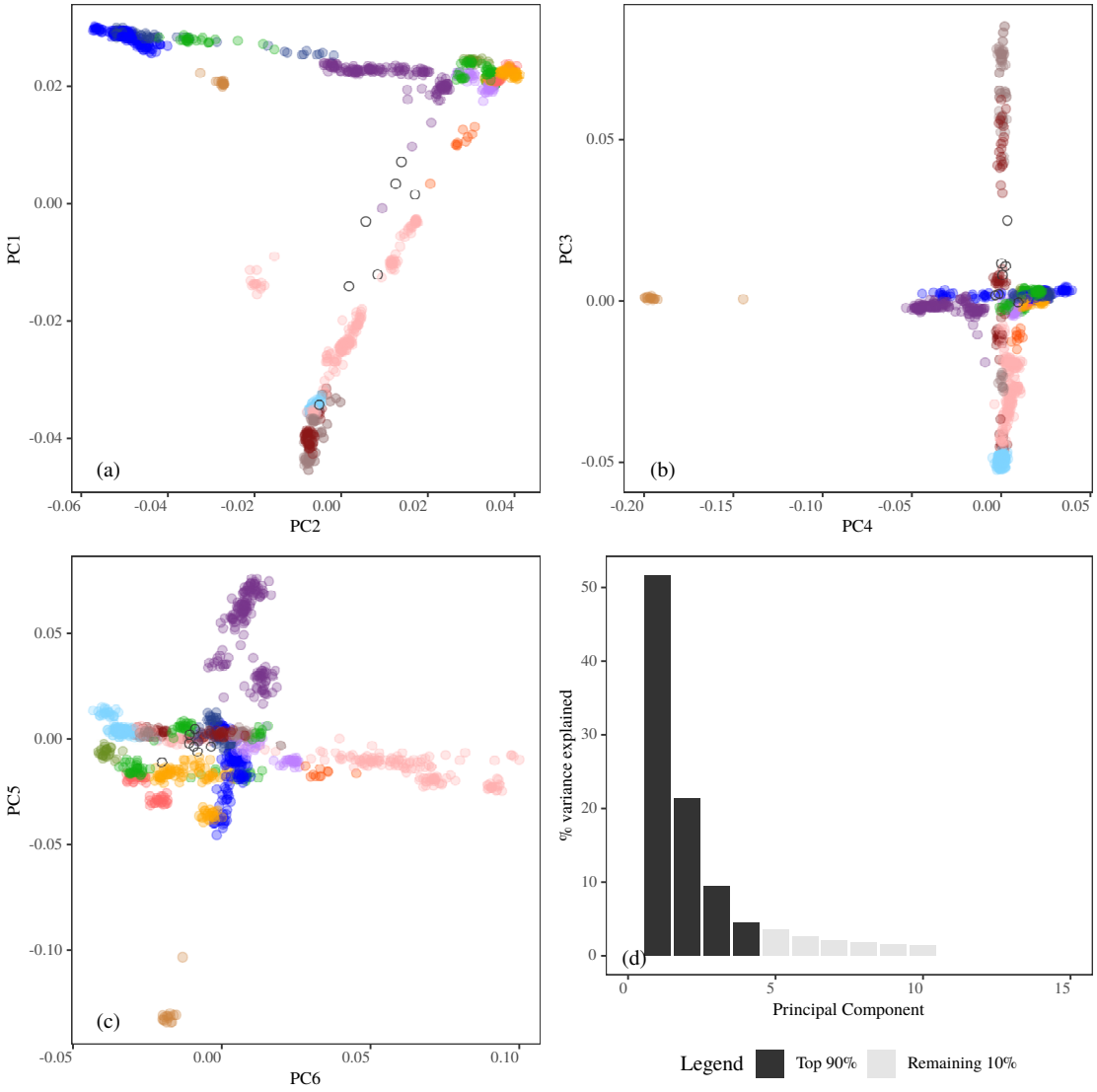


Figure C.11: PCA plots for Coloured-EC. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

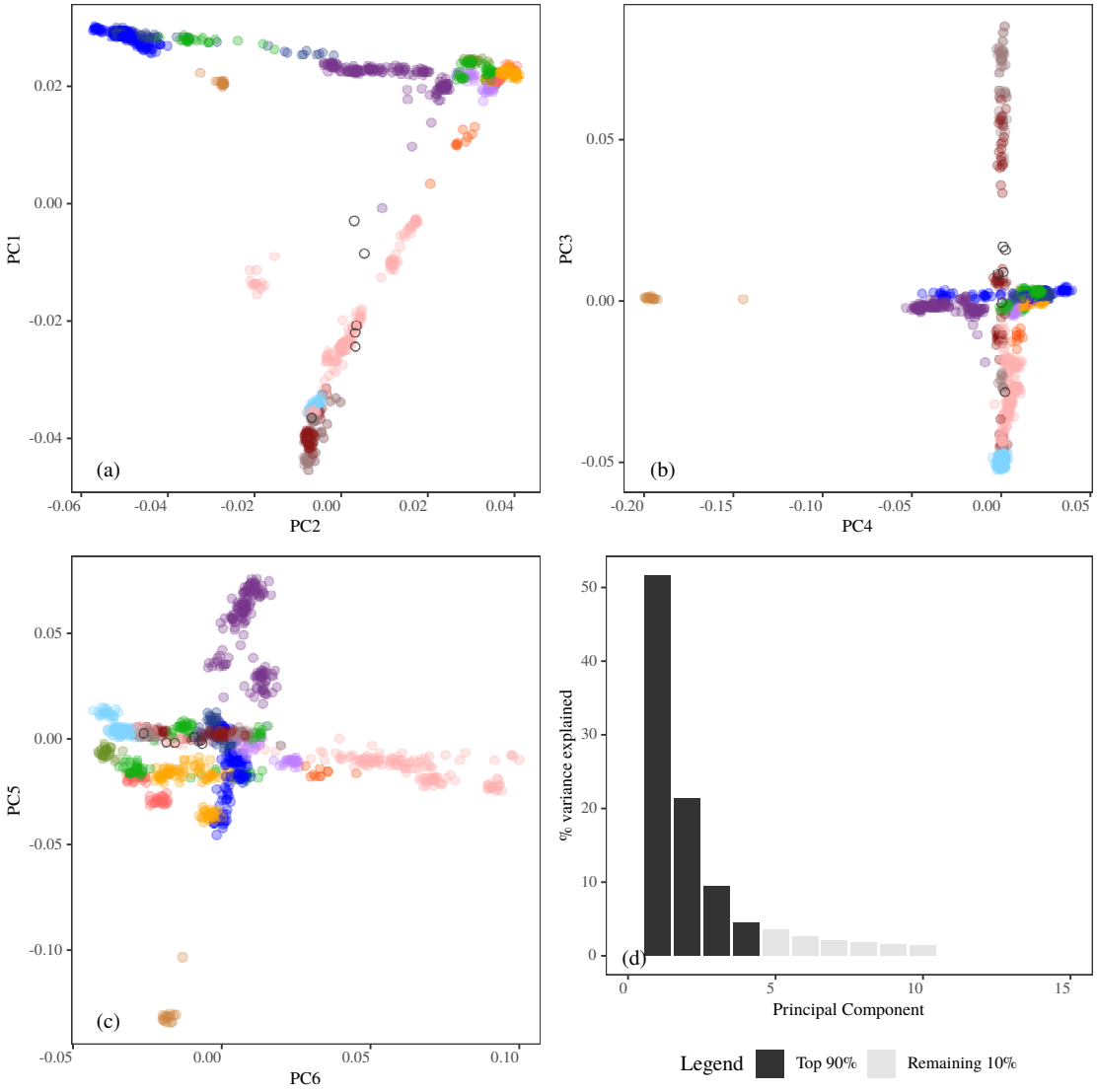


Figure C.12: PCA plots for COLOURED_EL. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

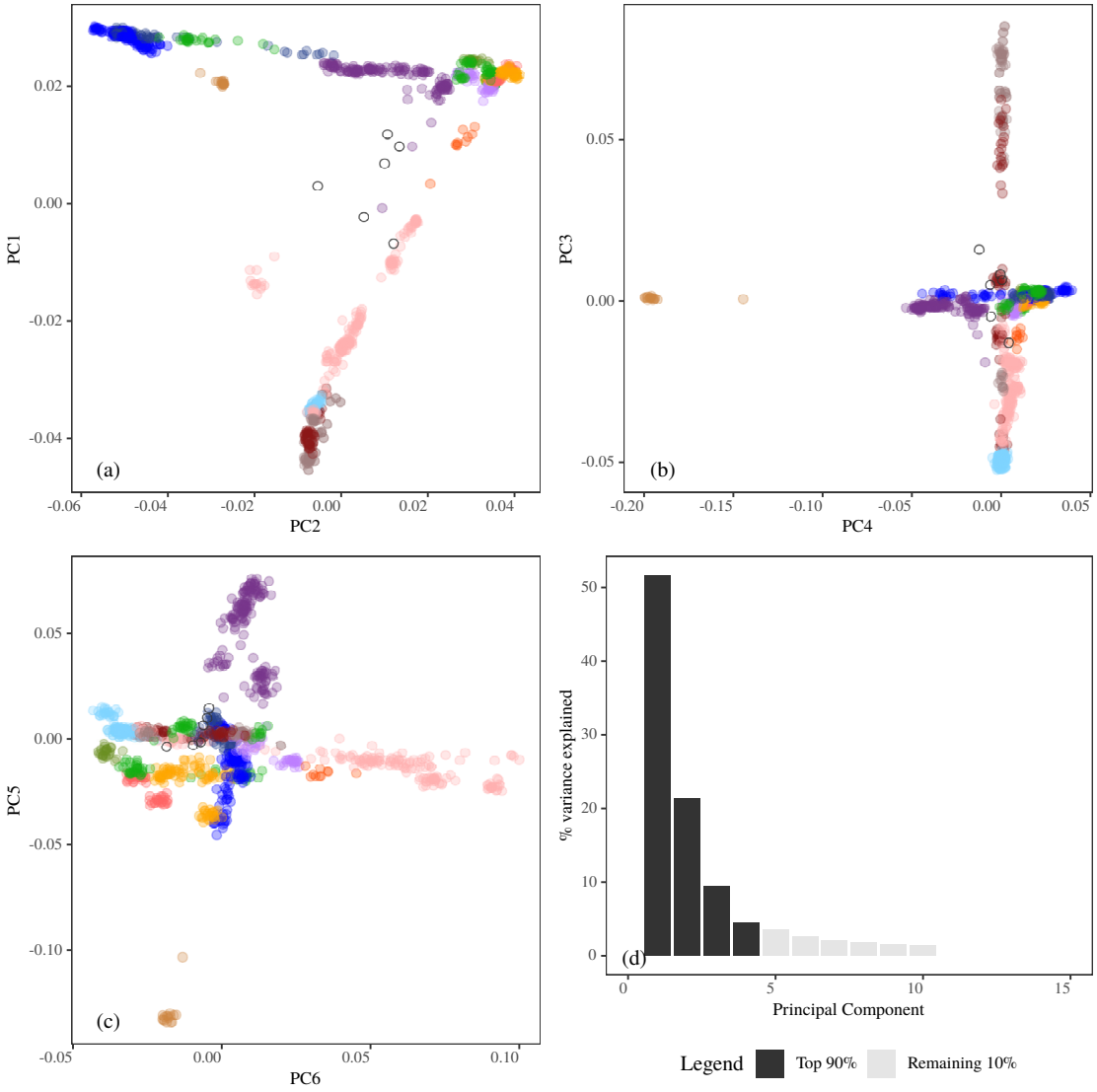


Figure C.13: PCA plots for COLOURED_JHB. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

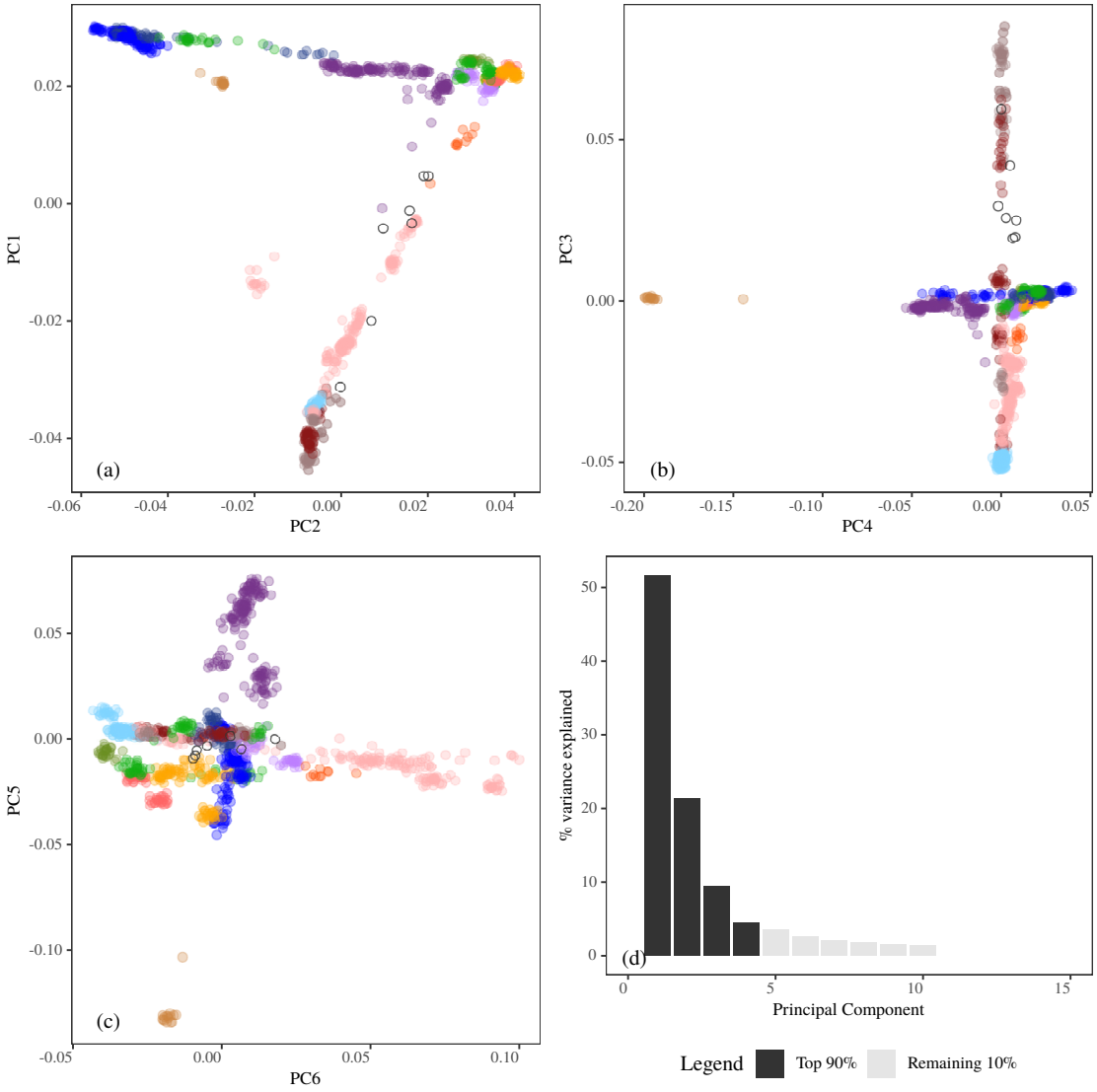


Figure C.14: PCA plots for COLOURED_KB. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

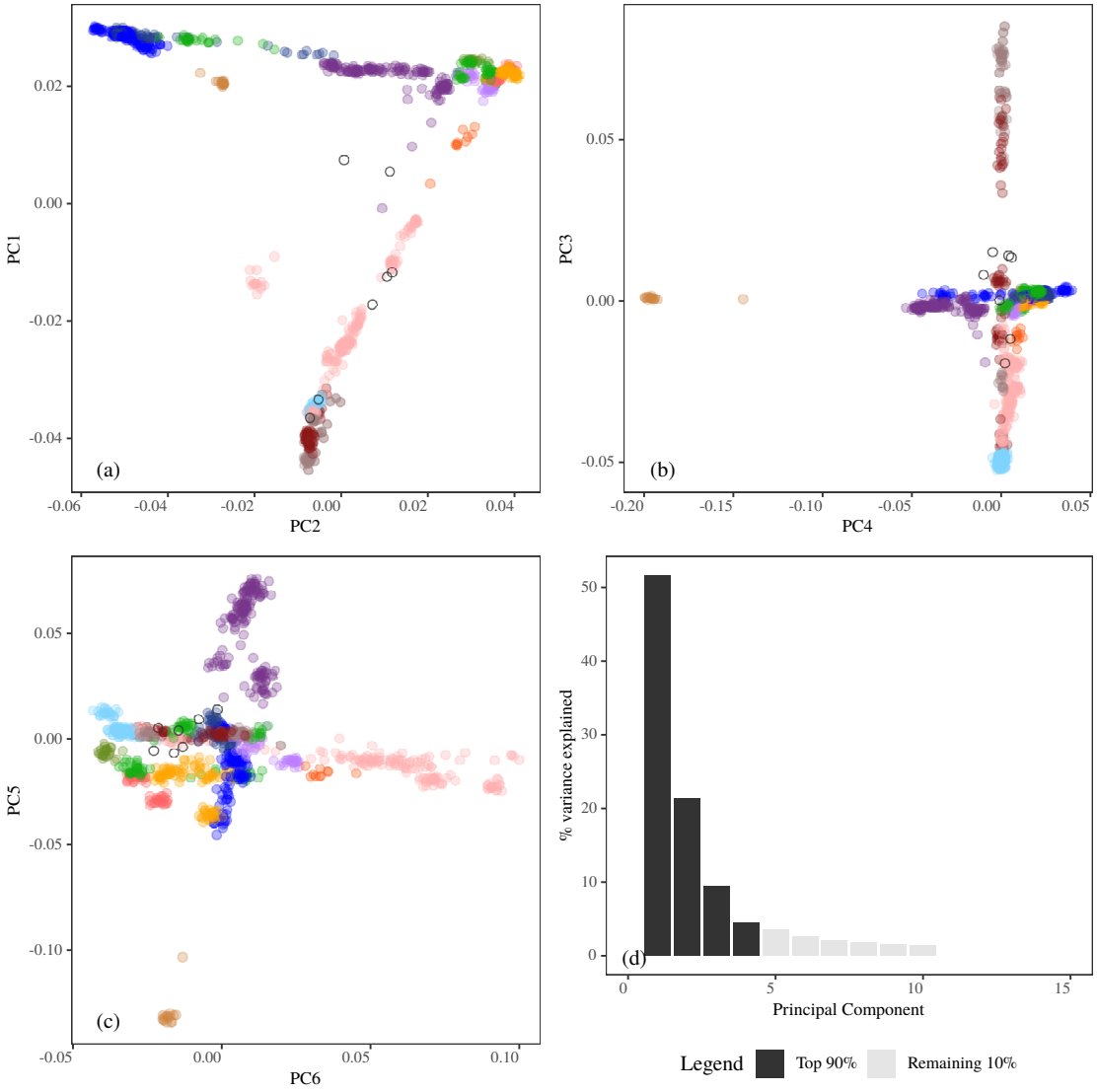


Figure C.15: PCA plots for COLOURED_KS. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

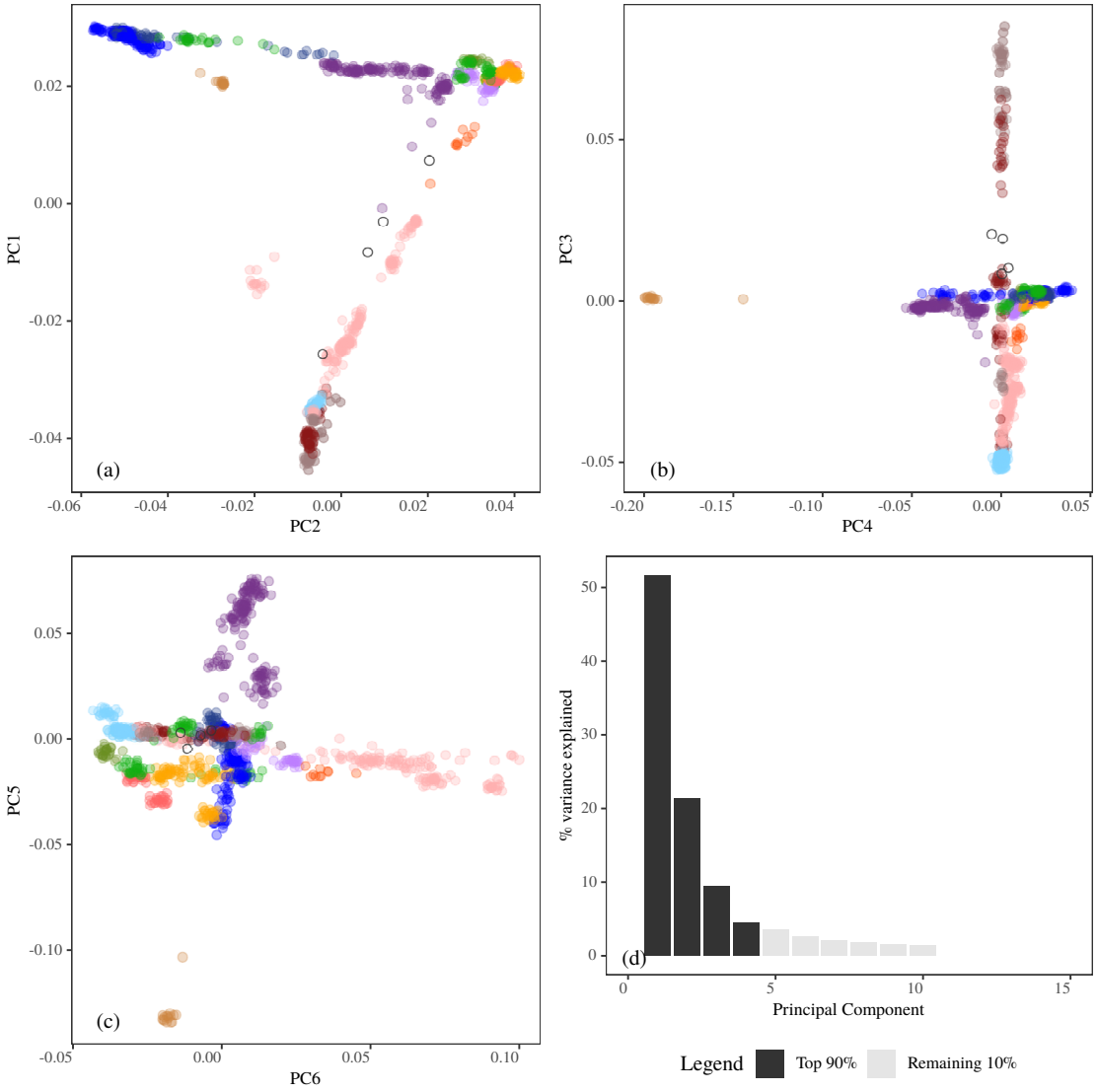


Figure C.16: PCA plots for COLOURED_KWT. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

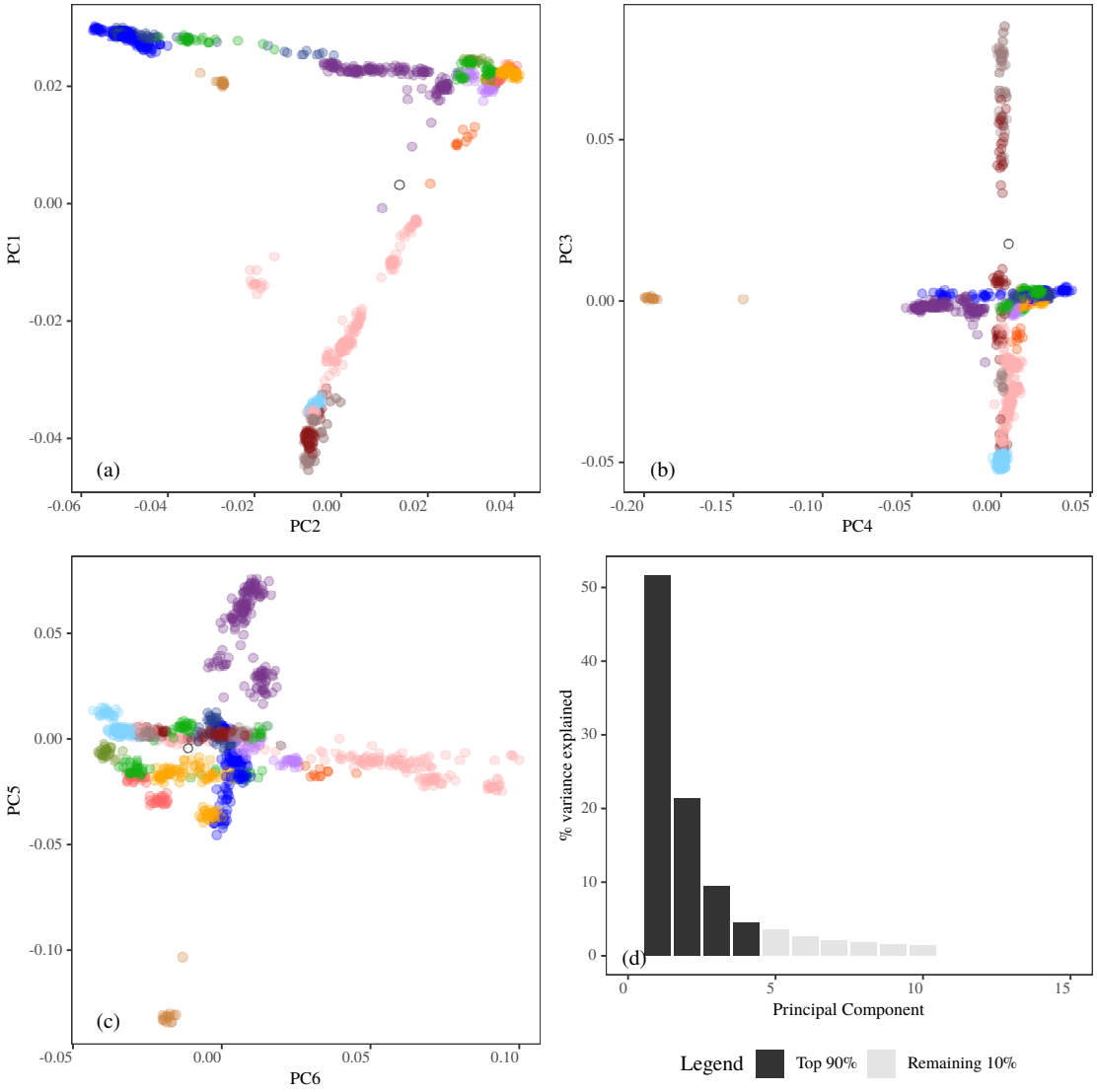


Figure C.17: PCA plots for COLOURED_MHK. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

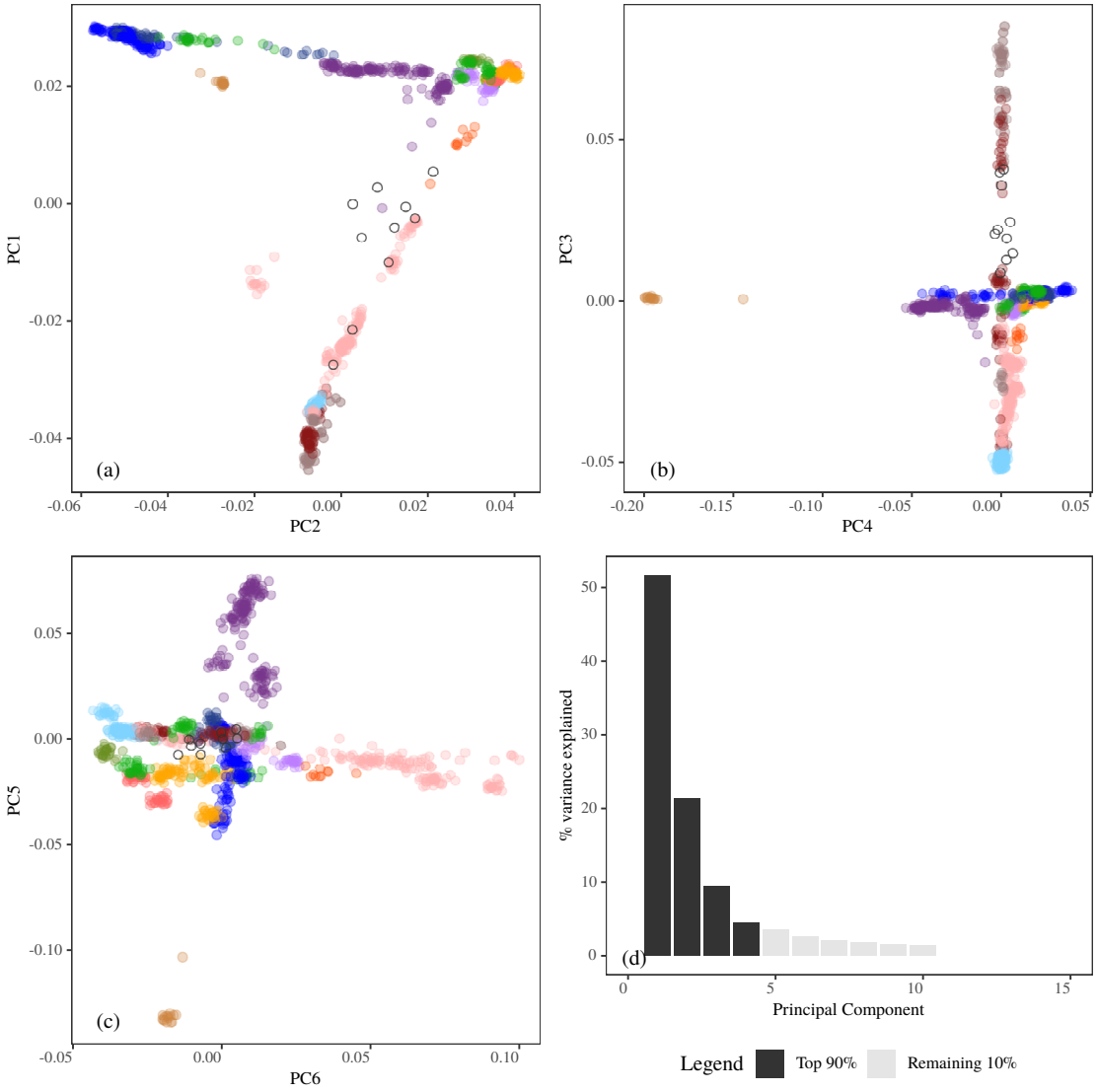


Figure C.18: PCA plots for Coloured-NC. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

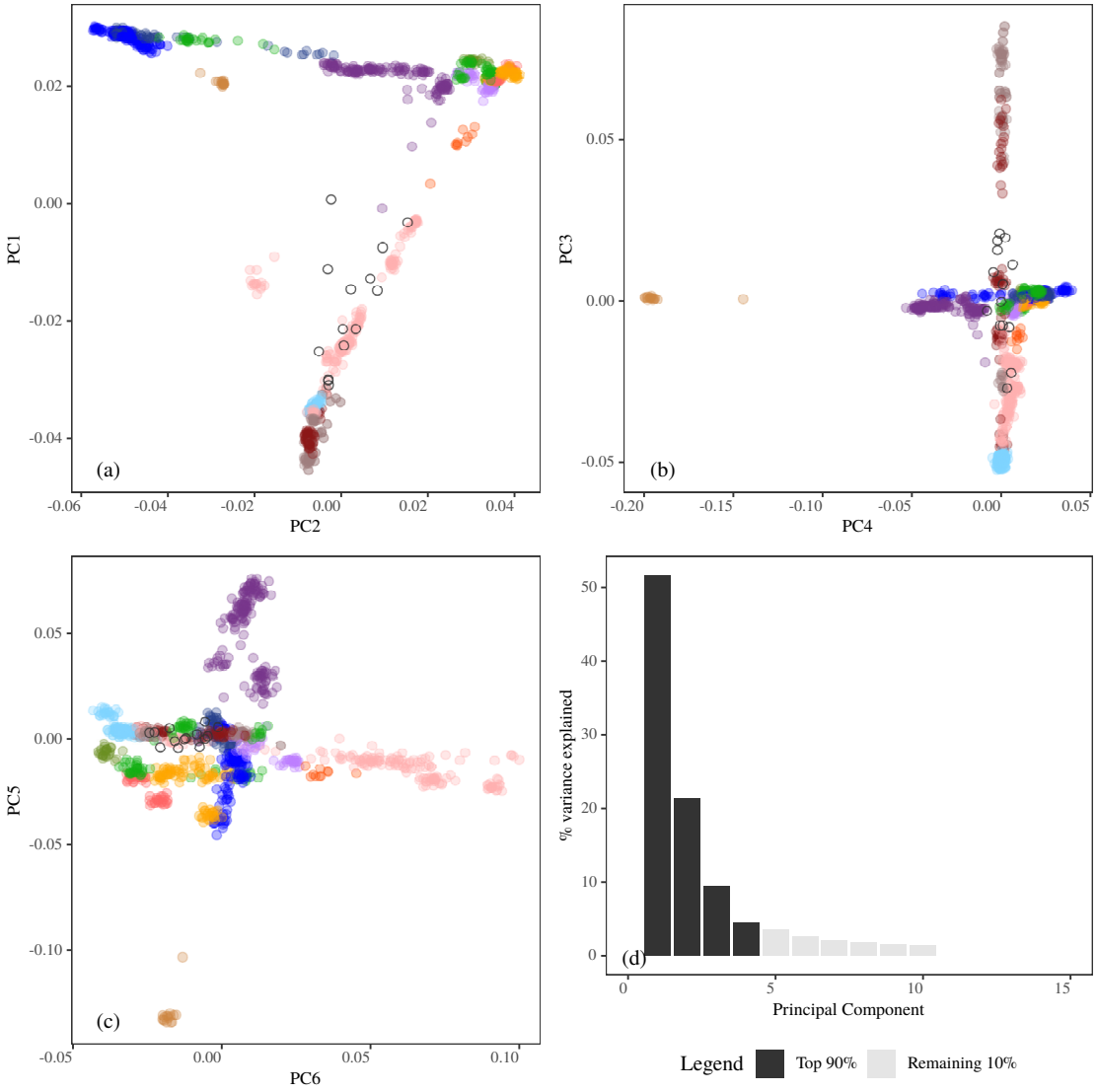


Figure C.19: PCA plots for COLOURED_PTG. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

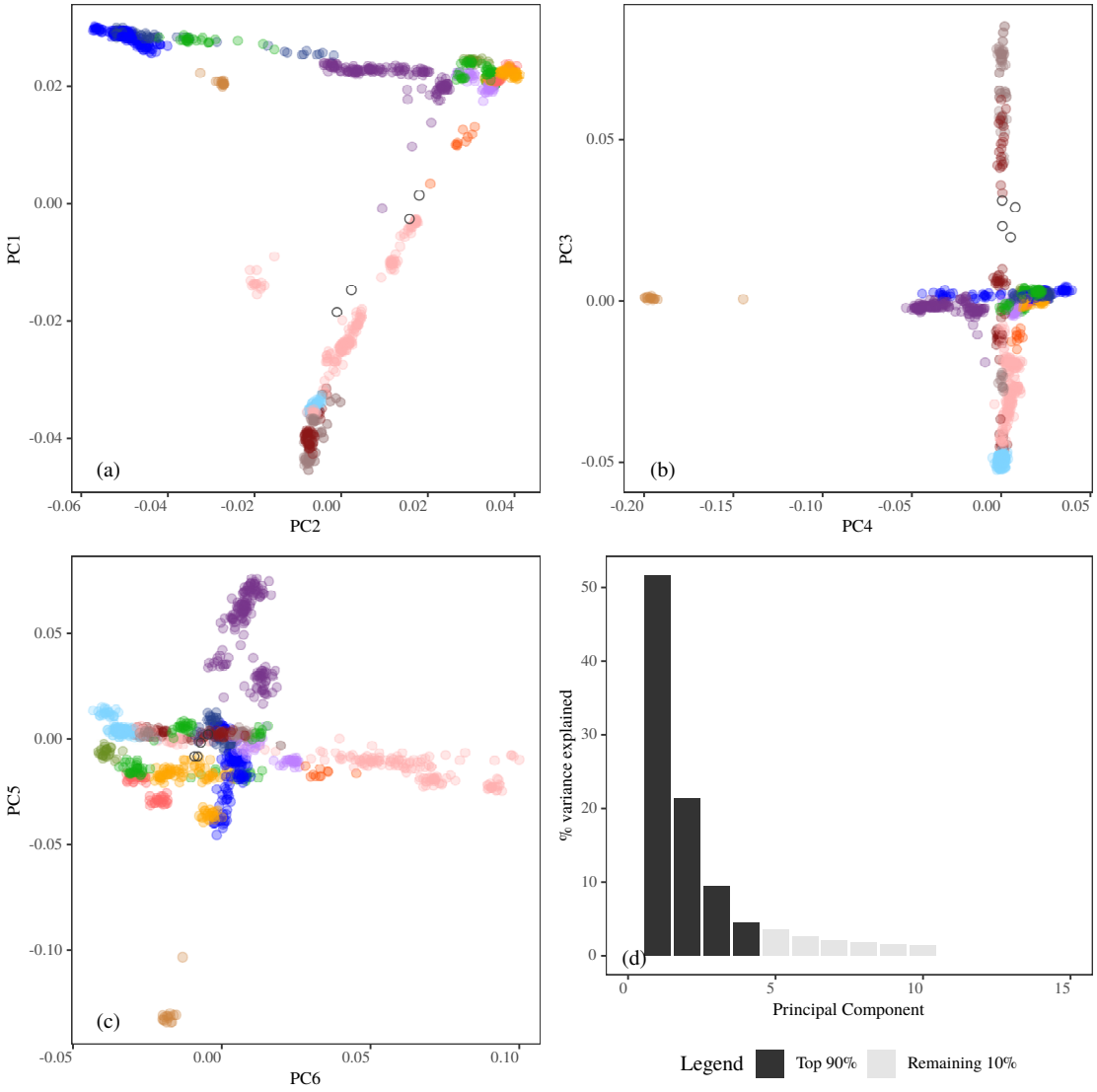


Figure C.20: PCA plots for COLOURED_TBV. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

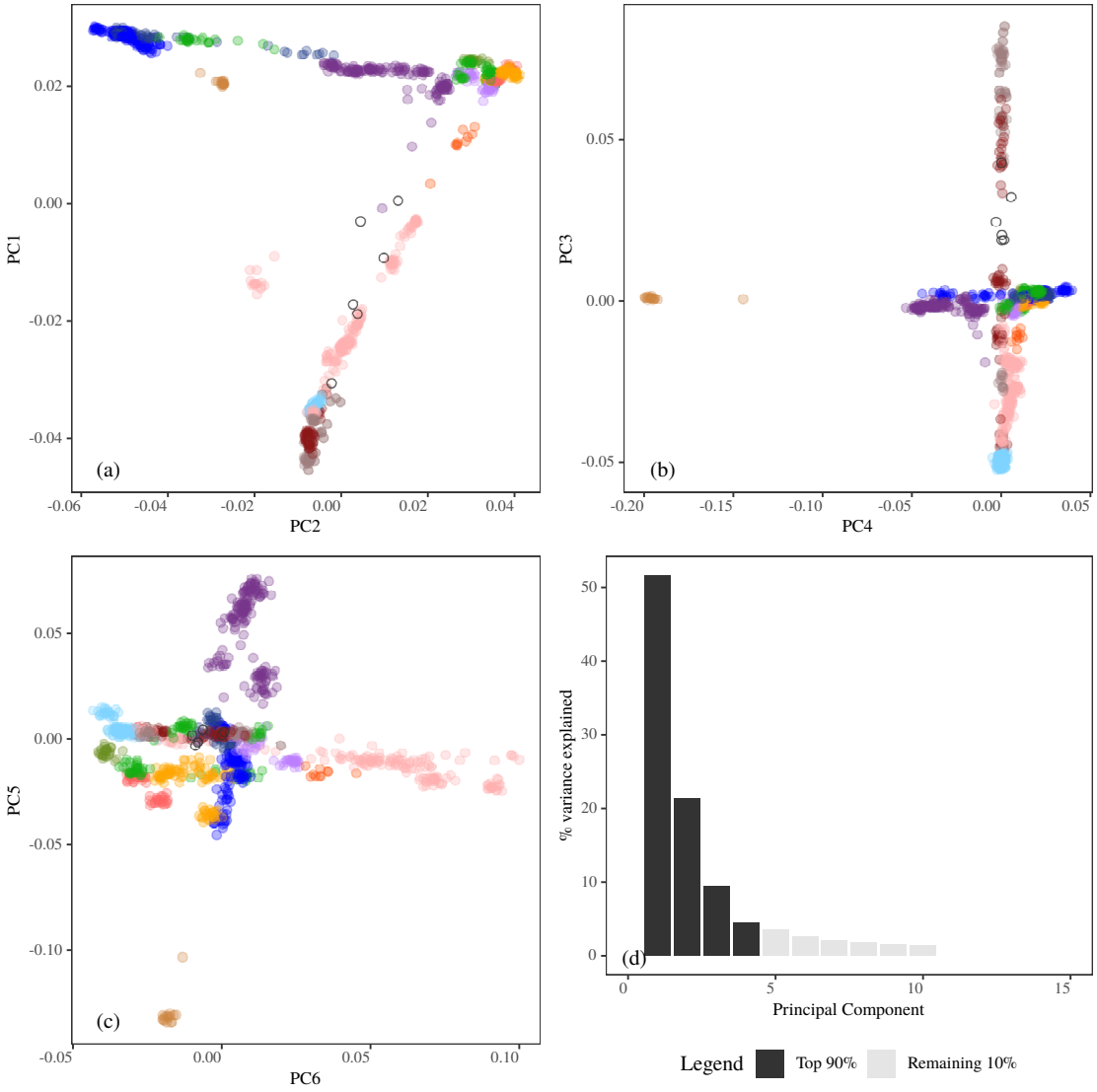


Figure C.21: PCA plots for COLOURED_UTN. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

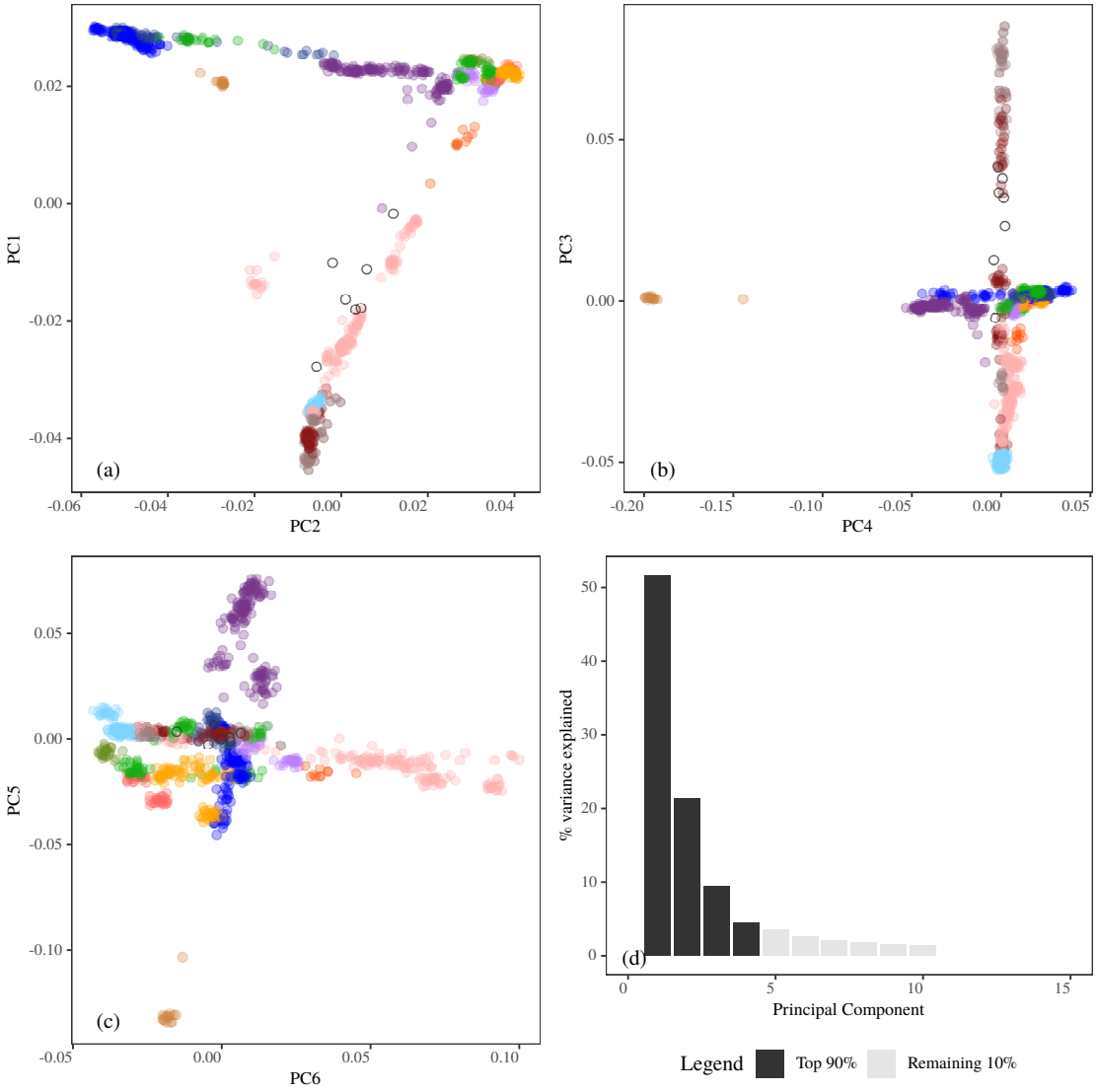


Figure C.22: PCA plots for COLOURED_VRE. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

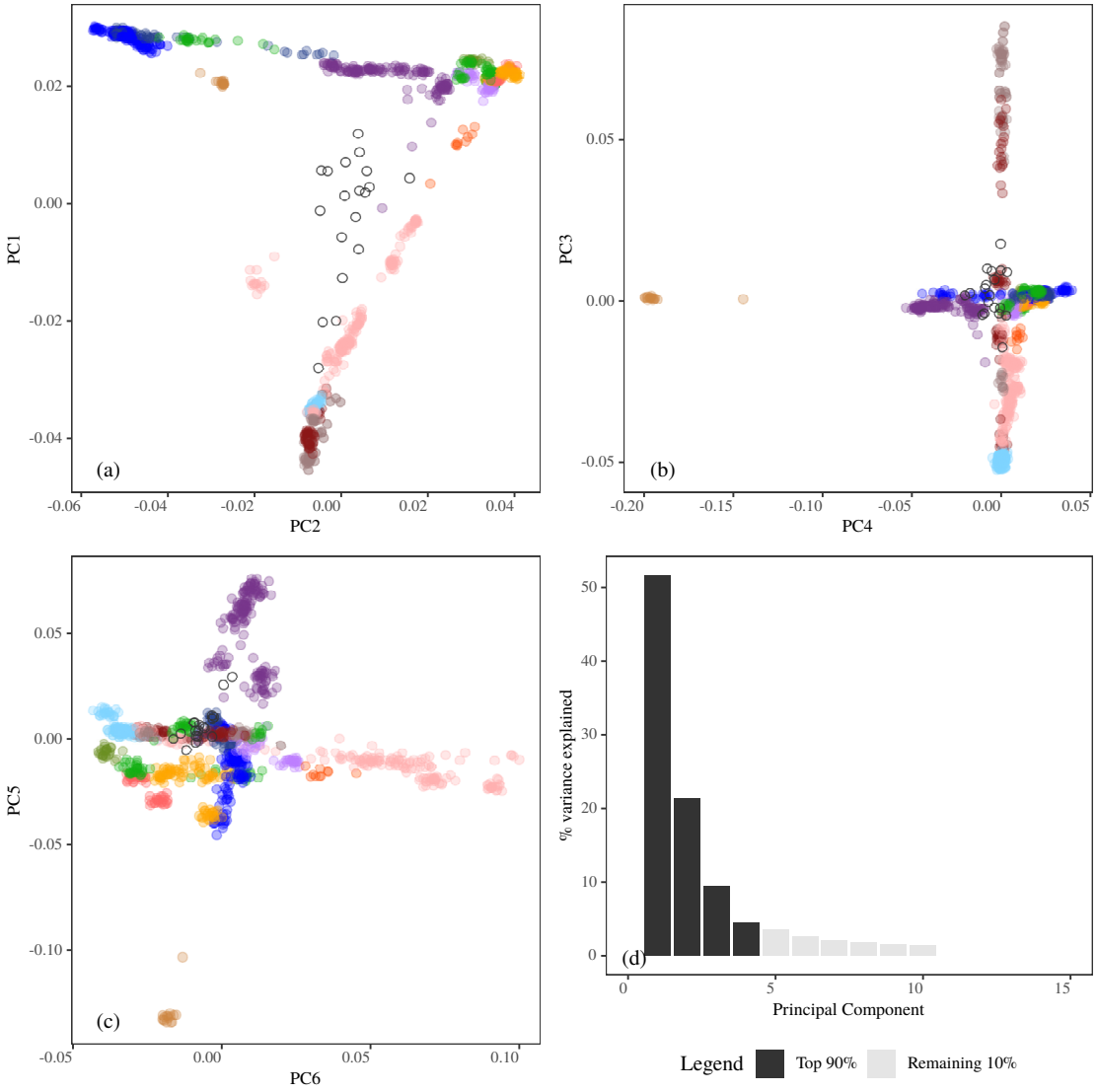


Figure C.23: PCA plots for ColouredWellington. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

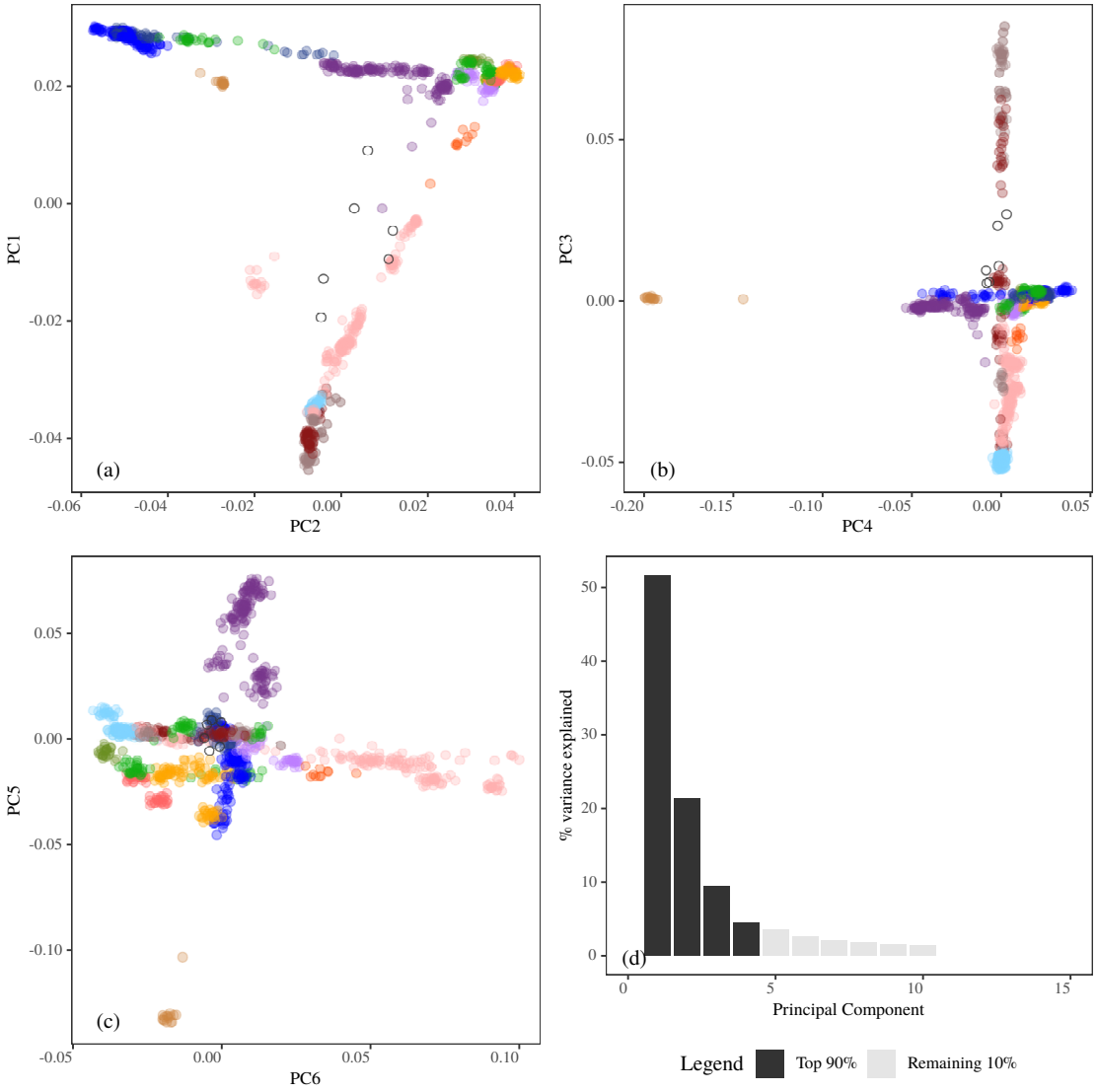


Figure C.24: PCA plots for GRIEKWA_CPT. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

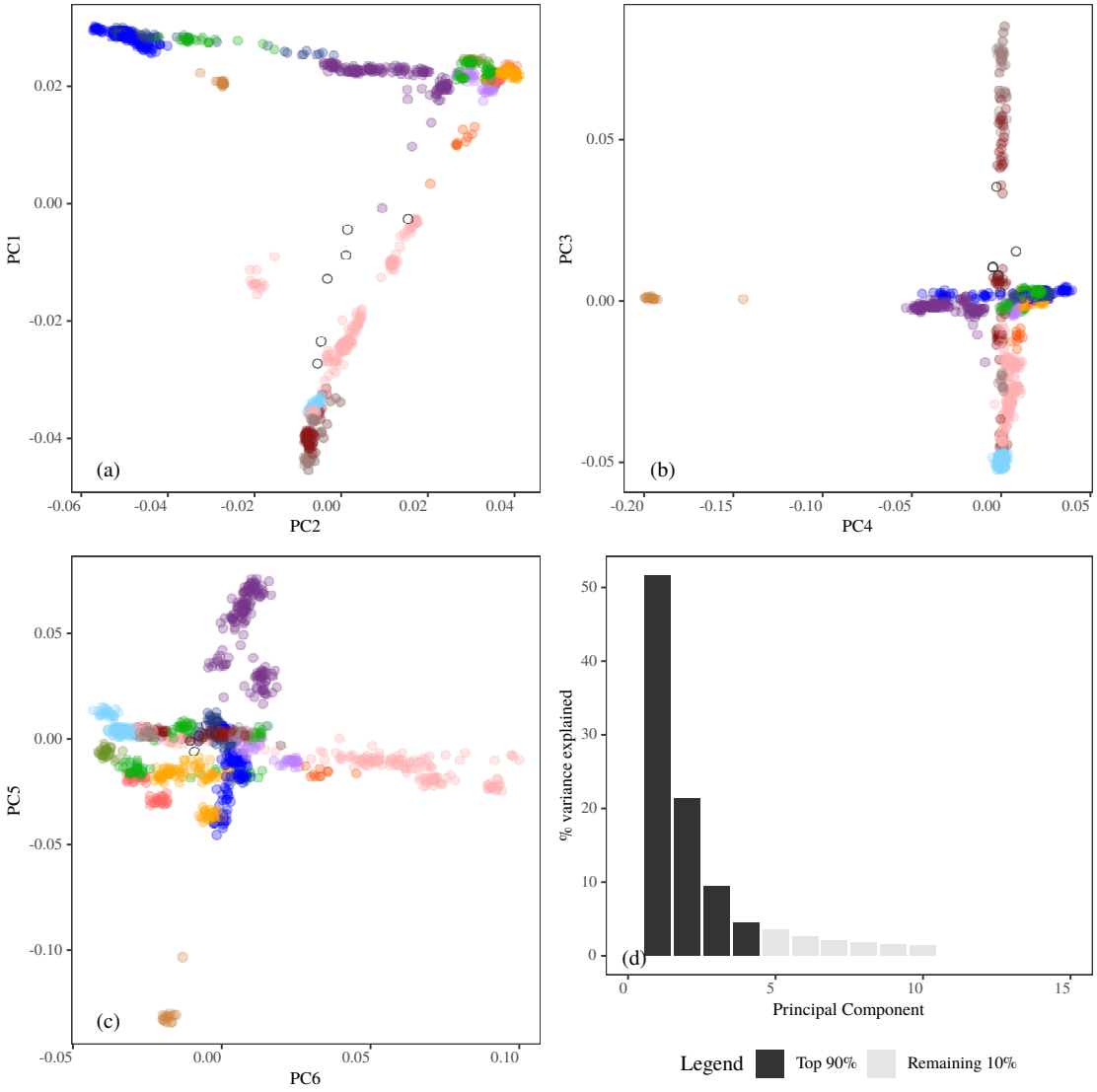


Figure C.25: PCA plots for GRIEKWA_KNY. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

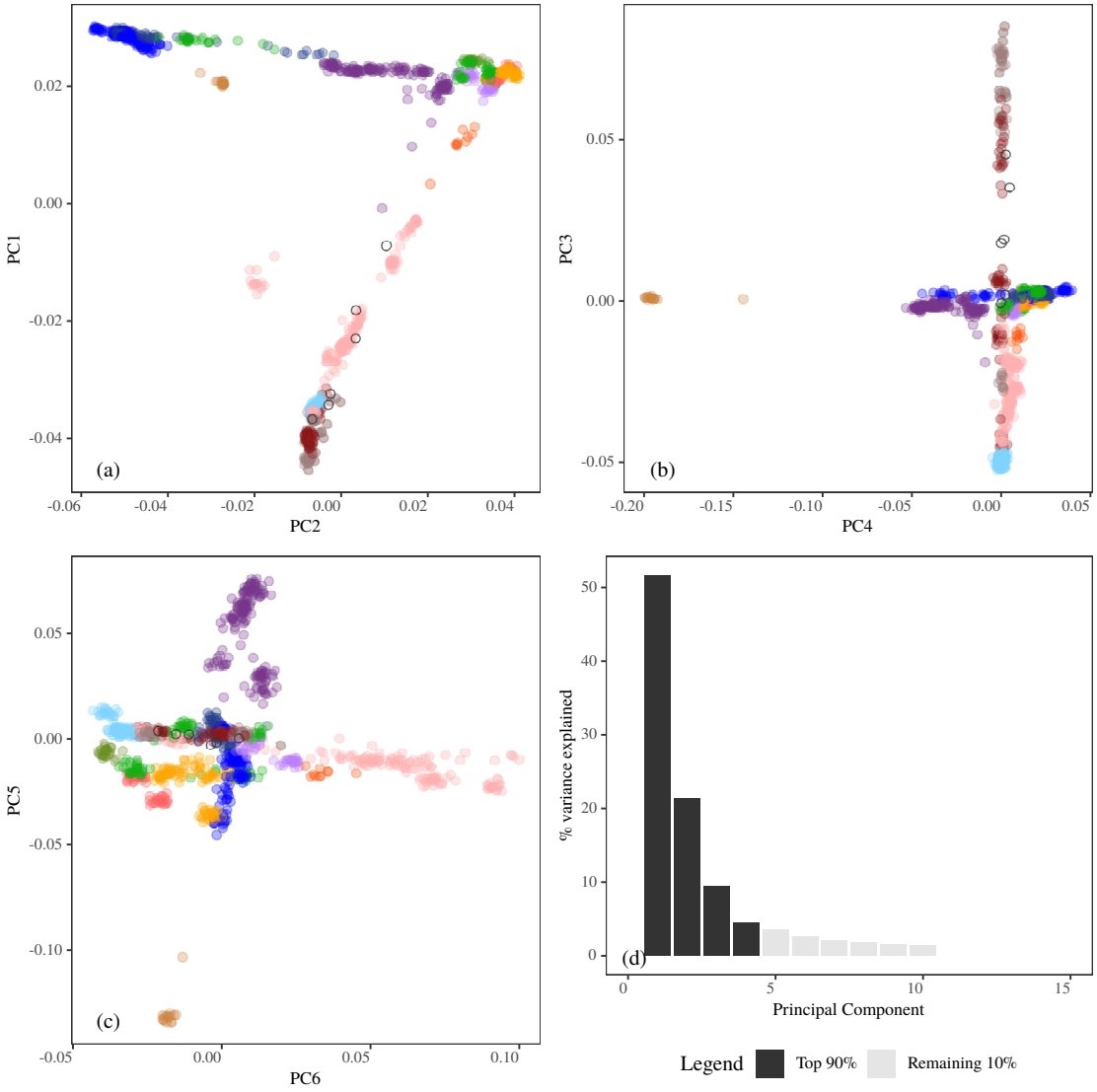


Figure C.26: PCA plots for GRIEKWA_UTN. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

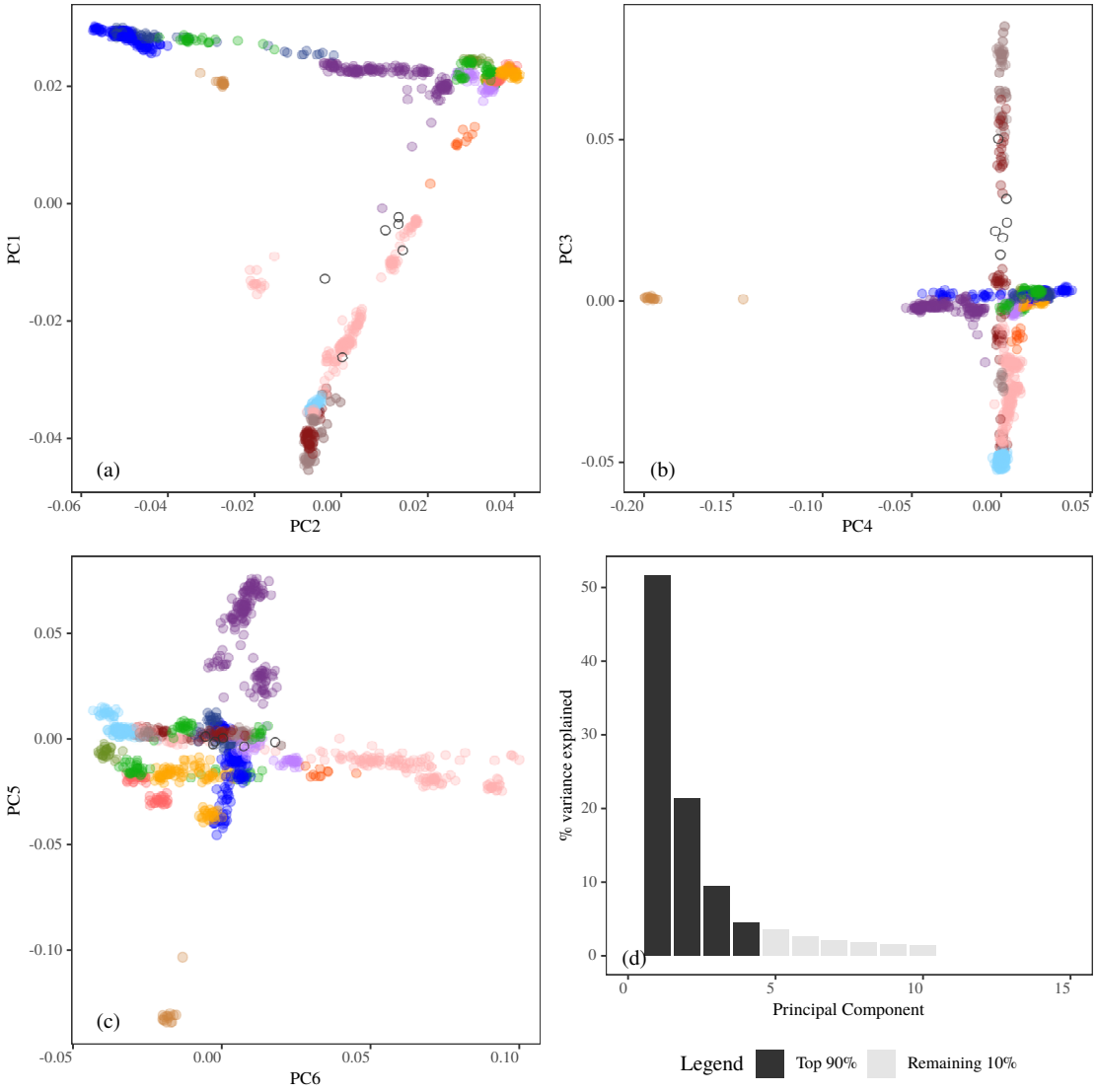


Figure C.27: PCA plots for GRIEKWA_VRE. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

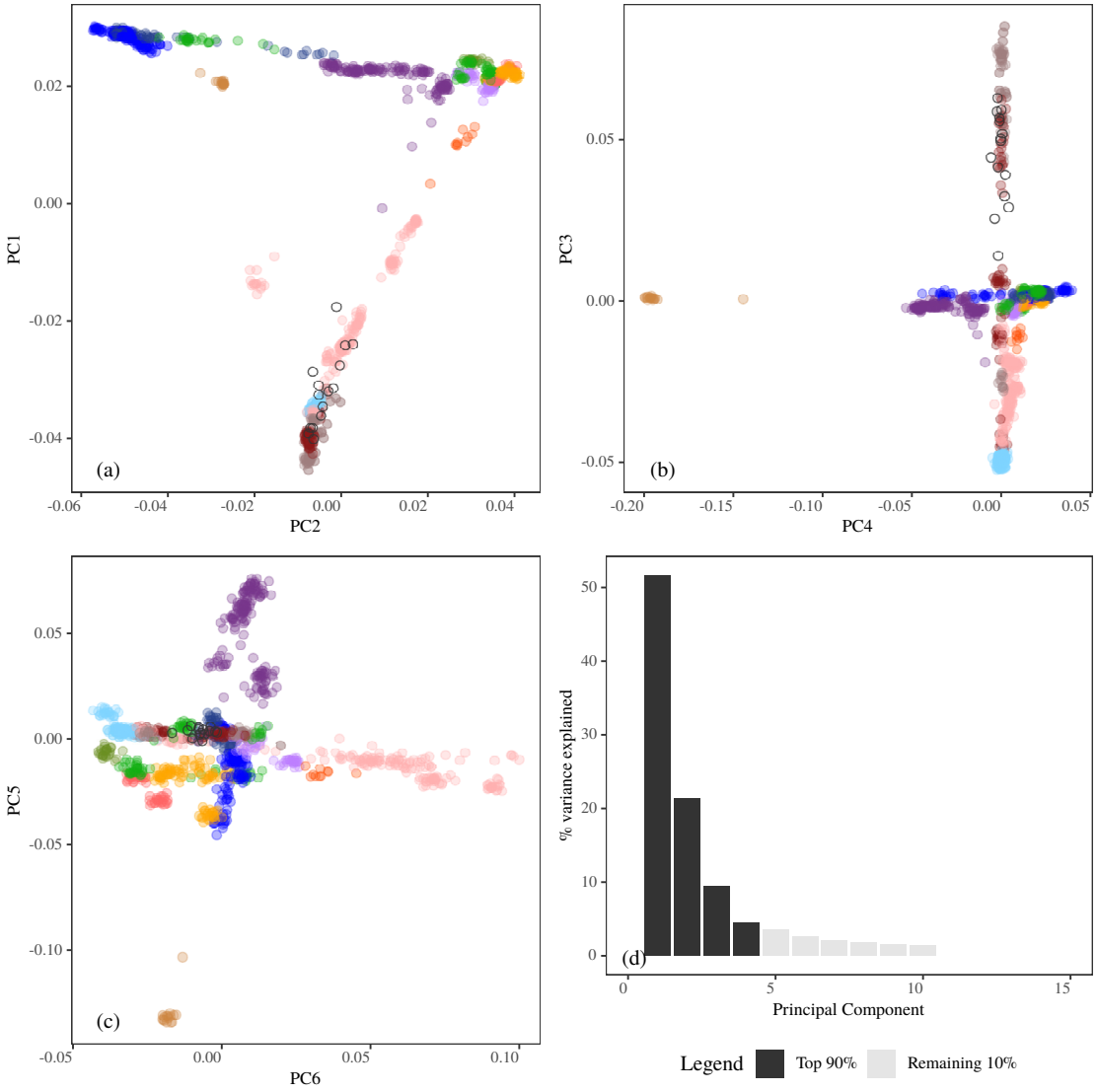


Figure C.28: PCA plots for Karretjie. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

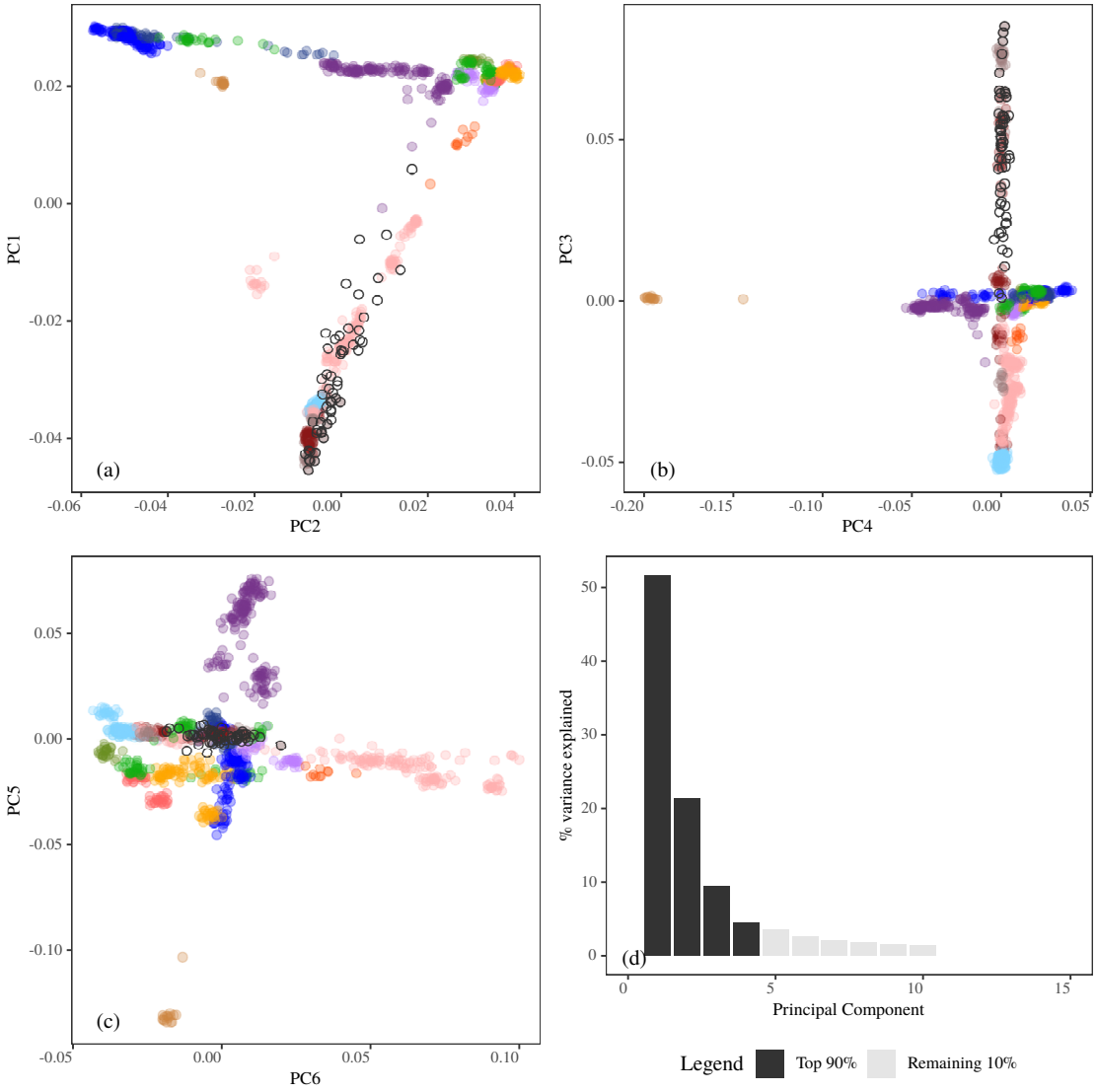


Figure C.29: PCA plots for KHM_SA. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

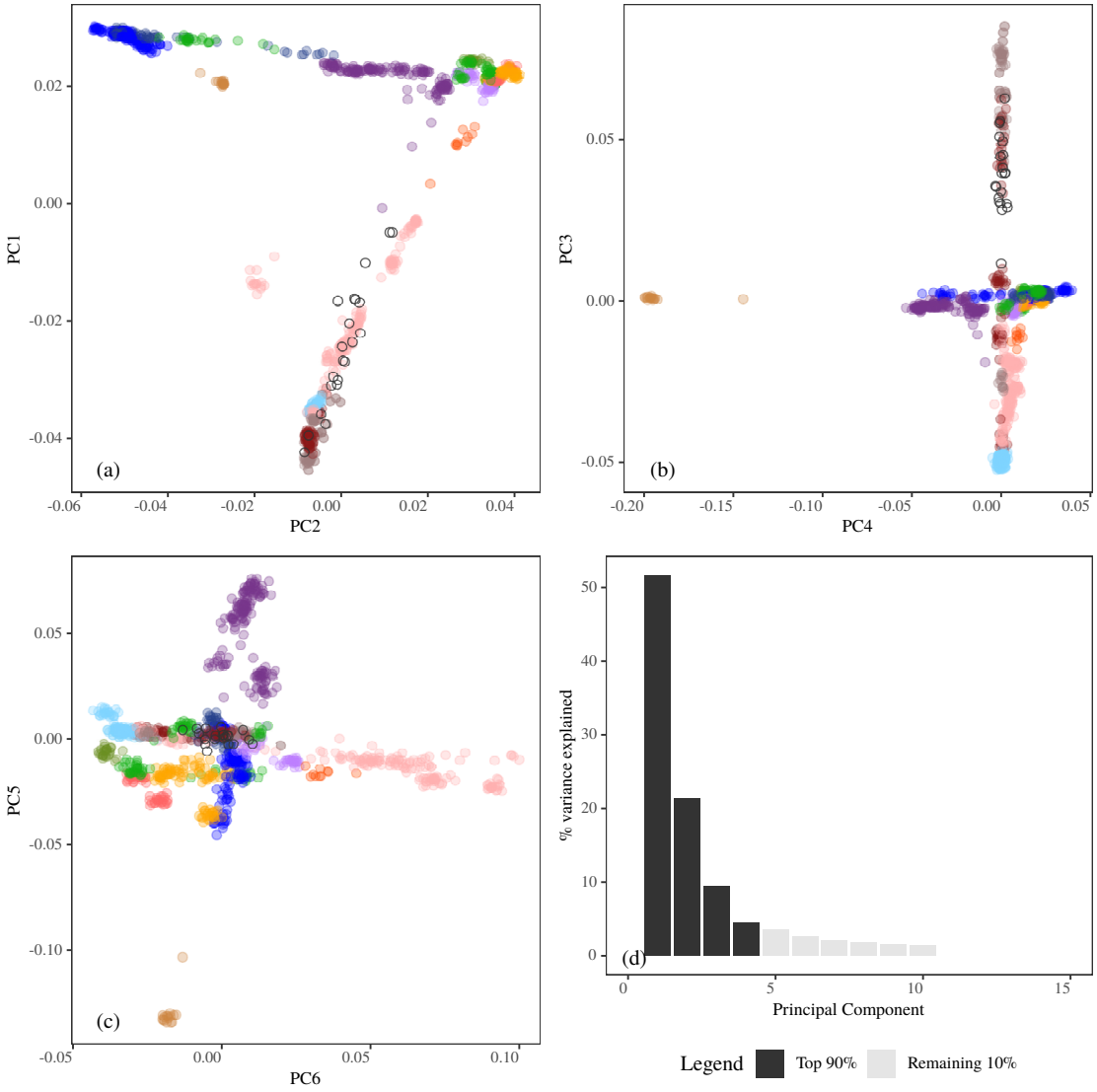


Figure C.30: PCA plots for Khomani. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

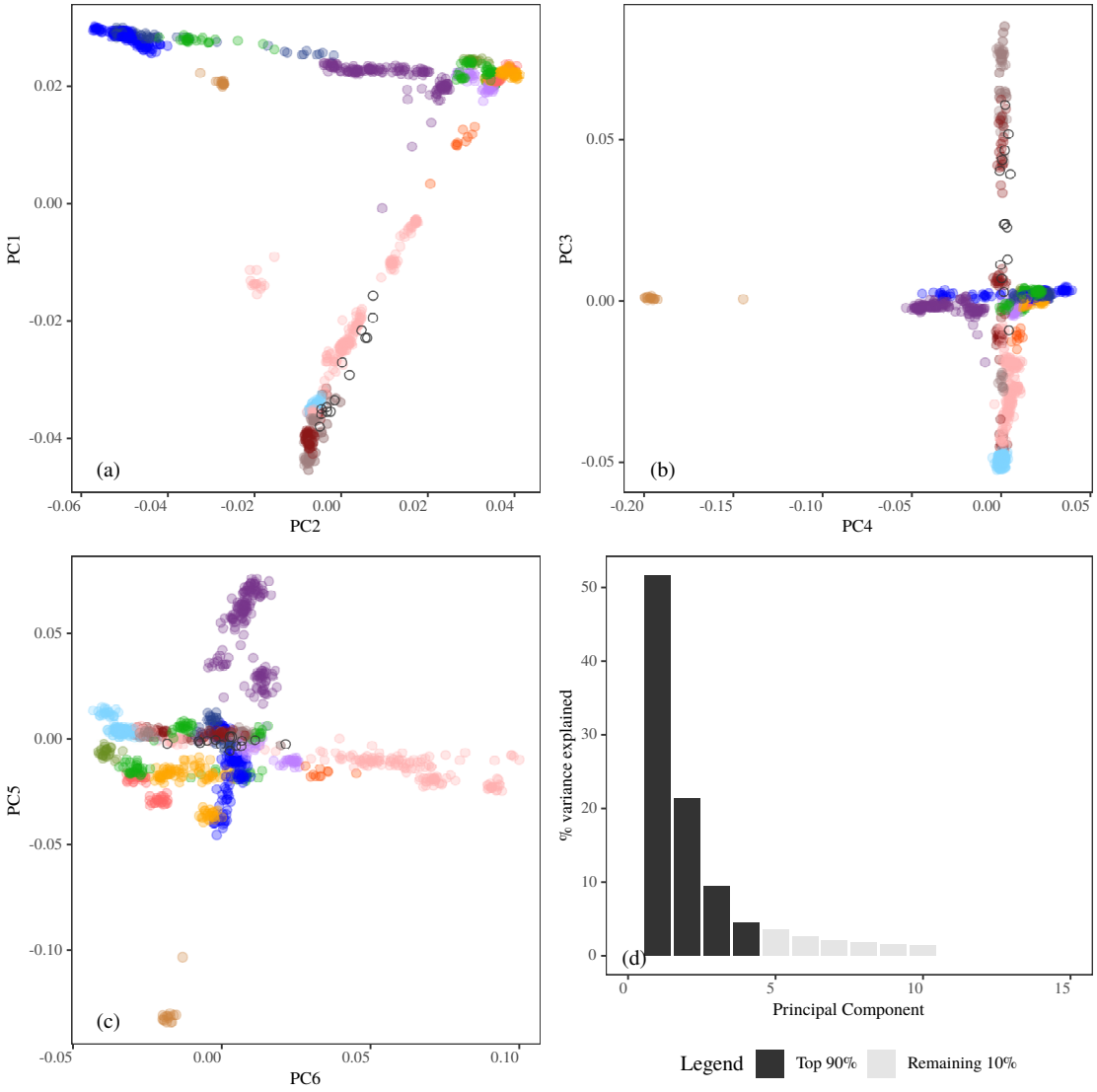


Figure C.31: PCA plots for Nama. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

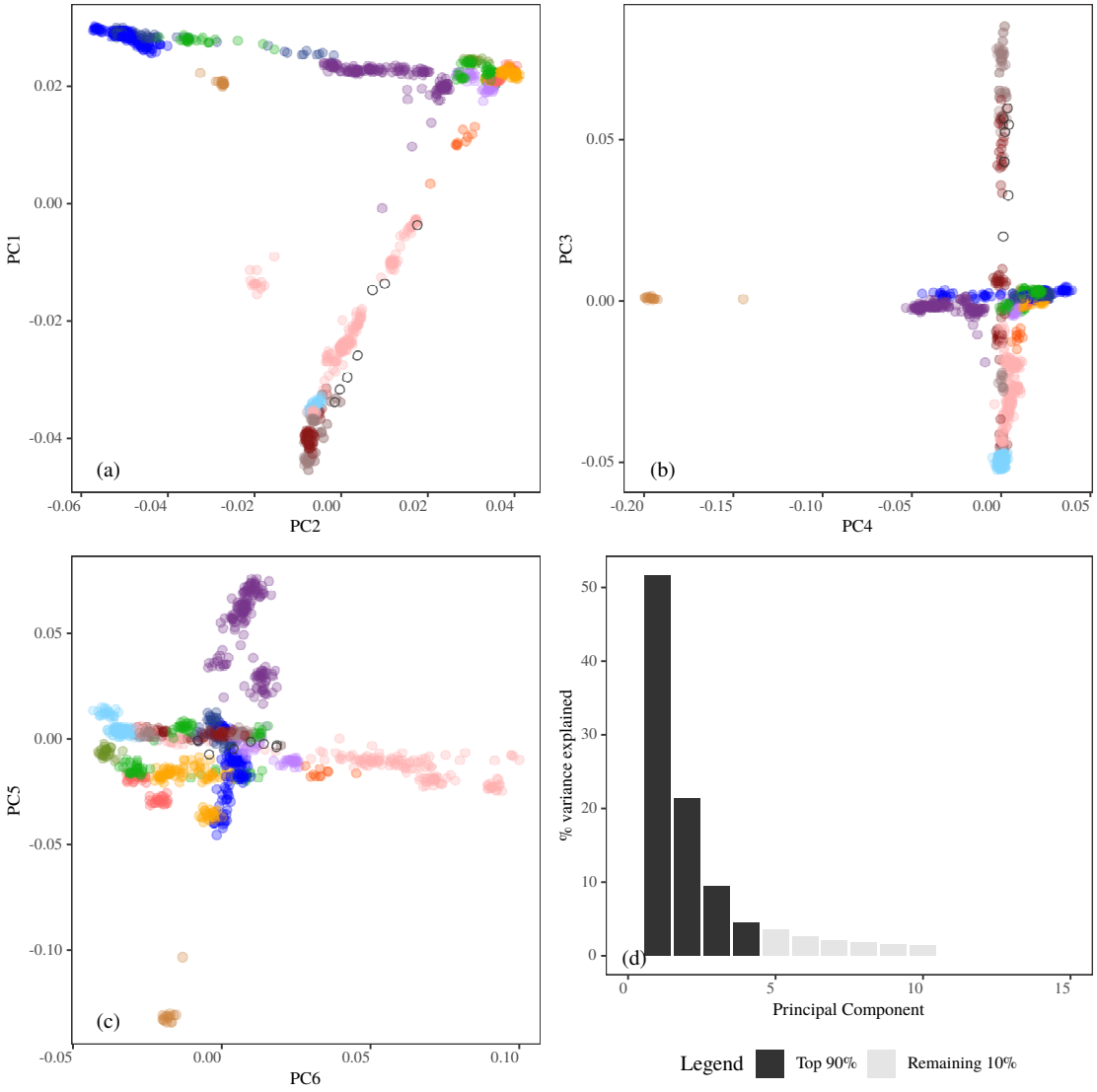


Figure C.32: PCA plots for NAMA_SA. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

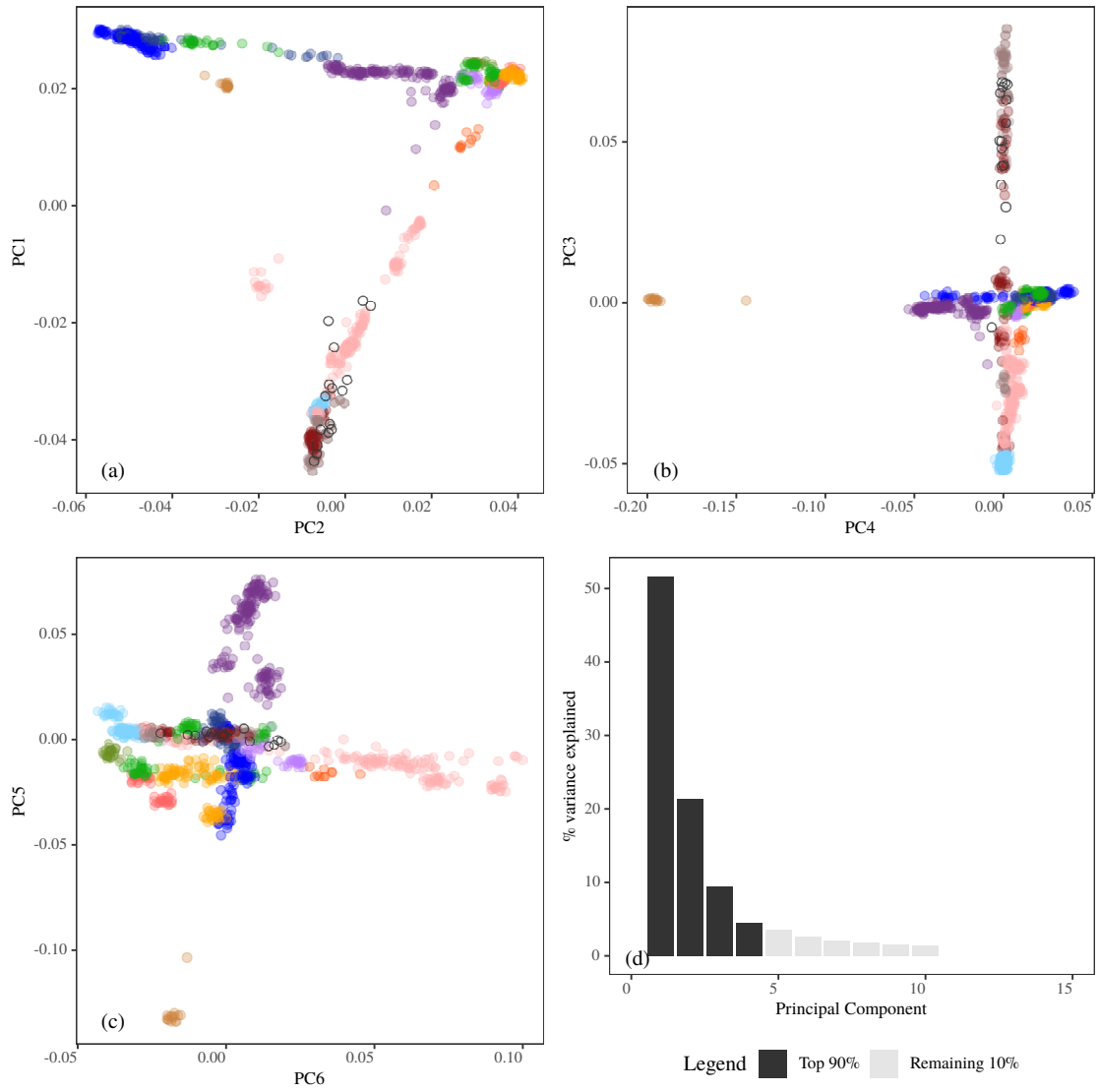


Figure C.33: PCA plots for SAN_he11. Shown is (a-c) the focal sample set (open circles) plotted onto GR data coloured by UN Region. (d) Percentage variance explained by each of the top 15 principal components.

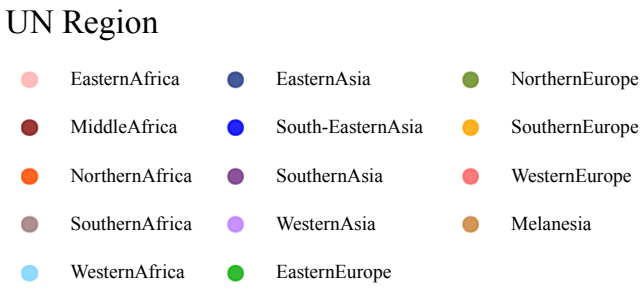


Figure C.34: Legend of GR populations colours for PCA plots. Colours refer to UN Regions for Figure C.4 C.33

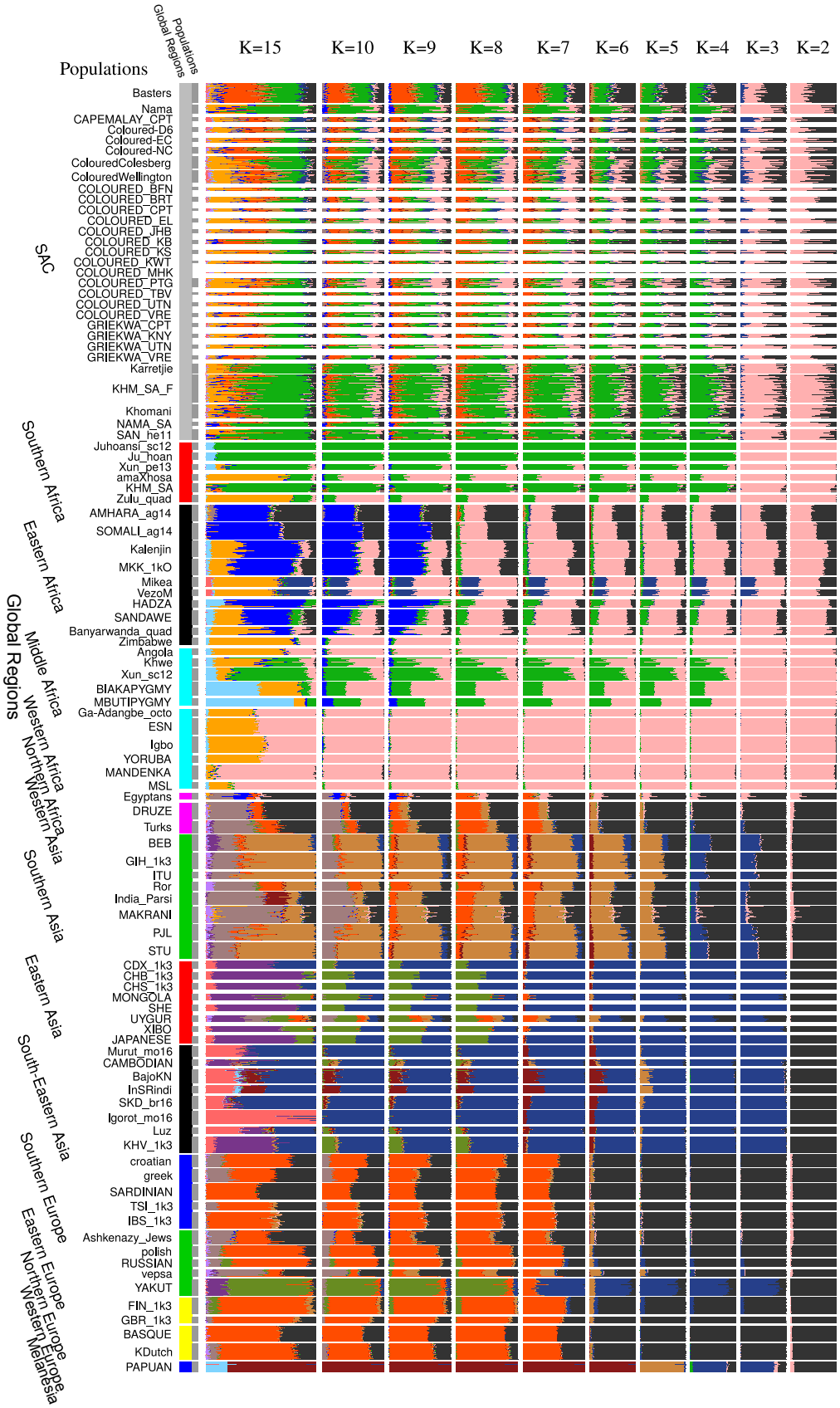


Figure C.35: ADMIXTURE profiles (K = 2...10, 15) for the SAC + GR data. Individual profiles arranged by UN global regions (labelled to the left).

C.2 Supplementary Tables

Table C.1: Summary of the *a priori* populations and the code (PopCode) used in chapter 6 . Indicated are the number of samples considered (n), their GPS coordinates in decimal degrees North and East, UN region designation and language grouping (see Table 6.4). Also indicated are the number of individuals removed due to within group kinship (WK), between group kinship (BK), as individuals removed as outliers based on Tukey criterion on IBD (I), iteratively based on position in PCA space using `smartpca` (S), and for the homogenous KhoeSan groups, as outliers based on PCA distances from the Ju|hoan (KP). Populations excluded due to non-homogeneity indicated (NHMG), and the final number of samples included as Global Reference (GR) and SAC or focal (F) indicated.

			Decimal Degrees		UN Regions			Outliers			Kinship			Dataset subdivision	
Population	PopCode	Source	N	E	Region	Ling. abb.	n	I	S	KP.	WK	BK	NHMG	F	GR
Malagasy	Mikea	[260]	-22.16	43.40	Eastern Africa	BG	21		3		3				18
Malagasy	VezoM	[260]	-23.36	43.65	Eastern Africa	BG	24							10	14
Somali	SOMALI_ag14	[4]	2.06	45.24	Eastern Africa	CU	39								25
Hadza	HADZA	[113]	-3.40	36.61	Eastern Africa	HZ	17								12
Masai	MKK_1kO	[245]	-1.75	35.82	Eastern Africa	NE	31				2				25
Kalenjin	Kalenjin	[4]	-1.04	36.51	Eastern Africa	NS	100								25
Ban-yarwanda	Banyarwanda_quad	[4]	0.31	32.53	Eastern Africa	SB	20					1			12
Zimbabwean	Zimbabwe	Montinaro *et al.* 2019	-19.70	31.10	Eastern Africa	SB	12		1						12
Sandawe	SANDAWE	[113]	-5.50	36.50	Eastern Africa	SD	28				5				23
Amhara	AMHARA_ag14	[4]	8.96	38.64	Eastern Africa	SE	42								25
Biaka	BIAKAPYGMY	[90]	6.61	20.94	Middle Africa	BI	21								21
!Xun	Xun_sc12	[115]	-14.63	17.67	Middle Africa	JU	19								19
Khwe	Khwe	[115]	-17.36	22.95	Middle Africa	KH	17		3						16
Mbuti	MBUTIPYGMY	[90]	-4.04	21.76	Middle Africa	SE	12								12
Angolan	Angola	Montinaro *et al.* 2019	-8.84	13.29	Middle Africa		10								10
Egyptians	Egyptans	[254]	26.83	26.38	Northern Africa	SE	12		2						12
Basters	Basters	[99]	-23.32	17.05	Southern Africa	GM	30							30	
Cape Malay	CAPEMALAY_CPT	This Study	-33.93	18.42	Southern Africa	GM	8				1			7	
"Coloured"	COLOURED_BFN	This Study	-29.12	26.21	Southern Africa	GM	5							5	
"Coloured"	COLOURED_BRT	This Study	-25.79	31.05	Southern Africa	GM	8							8	
"Coloured"	COLOURED_CPT	This Study	-33.93	18.42	Southern Africa	GM	4							4	
"Coloured"	COLOURED_EL	This Study	-33.03	27.85	Southern Africa	GM	6							6	
"Coloured"	COLOURED_JHB	This Study	-26.20	28.05	Southern Africa	GM	6							6	
"Coloured"	COLOURED_KB	This Study	-28.45	16.99	Southern Africa	GM	7							7	
"Coloured"	COLOURED_KS	This Study	-30.55	29.42	Southern Africa	GM	7							7	
"Coloured"	COLOURED_KWT	This Study	-32.89	27.42	Southern Africa	GM	4							4	
"Coloured"	COLOURED_MHK	This Study	-25.86	25.64	Southern Africa	GM	1							1	
"Coloured"	COLOURED_PTG	This Study	-23.90	29.45	Southern Africa	GM	14							14	
"Coloured"	COLOURED_TBV	This Study	-32.59	26.72	Southern Africa	GM	4							4	
"Coloured"	COLOURED_UTN	This Study	-28.45	21.26	Southern Africa	GM	6							6	
"Coloured"	COLOURED_VRE	This Study	-31.67	18.50	Southern Africa	GM	7							7	
"Coloured"	Coloured-D6	[99]	-33.92	18.41	Southern Africa	GM	8							8	
"Coloured"	Coloured-EC	[99]	-33.80	25.25	Southern Africa	GM	7							7	
"Coloured"	Coloured-NC	[99]	-29.36	21.04	Southern Africa	GM	10							10	
"Coloured"	ColouredColesberg	[115]	-30.71	25.10	Southern Africa	GM	20							19	
"Coloured"	ColouredWellington	[115]	-33.64	19.01	Southern Africa	GM	20							19	
Griekwa	GRIEKWA_CPT	This Study	-33.93	18.42	Southern Africa	GM	7					1		6	
Griekwa	GRIEKWA_KNY	This Study	-34.04	23.05	Southern Africa	GM	8				2			6	
Griekwa	GRIEKWA_UTN	This Study	-28.45	21.26	Southern Africa	GM	7				1			6	

Griekwa	GRIEKWA_VRE	This Study	-31.67	18.50	Southern Africa	GM	6						6	
Karretjie	Karretjie	[115]	-30.71	25.10	Southern Africa	GM	20						15	5
Ju|hoan	Ju_hoan	[99]	-19.23	20.09	Southern Africa	JU				1				12
Ju'/hoansi	Juhoansi_sc12	[115]	-19.60	20.49	Southern Africa	JU	18	1		1		4		11
!Xun	Xun_pe13	[99]	-19.04	19.15	Southern Africa	JU				1				10
Nama	Nama	[115]	-22.56	17.07	Southern Africa	KK	20						14	5
Nama	NAMA_SA	[107]	-28.50	17.00	Southern Africa	KK	13				3		7	3
amaXhosa	amaXhosa	[99]	-31.59	28.71	Southern Africa	SB	15						10	5
Zulu	Zulu_quad	[4]	-29.37	29.82	Southern Africa	SB	95							12
≠Khomani	KHM_SA	[107]	-30.56	22.94	Southern Africa	UT	86				30		42	14
≠Khomani	Khomani	[115]	-26.97	20.79	Southern Africa	UT	39				1	9	21	7
San	SAN_he11	[113]	-28.44	21.07	Southern Africa	UT	31					10	16	5
Esan	ESN	[2]	9.07	7.48	Western Africa	ED	99							25
Ga	Ga-Adangbe_octo	[4]	7.90	-3.28	Western Africa	GB	11							11
Igbo	Igbo	[4]	9.02	4.18	Western Africa	IG	99							25
Mandenka	MANDENKA	[90]	14.50	-14.45	Western Africa	ME	22							22
Mende	MSL	[2]	8.48	-13.23	Western Africa	ME	85		7				10	25
Yoruba	YORUBA	[90]	9.08	8.68	Western Africa	YB	21							12
Japanese	JAPANESE	[90]	36.20	138.25	Eastern Asia	JP	28							12
Dai Chinese	CDX_1k3	[2]	22.00	100.78	Eastern Asia	KD	93	1						12
Mongola	MONGOLA	[90]	35.86	104.20	Eastern Asia	MG	10							10
She	SHE	[90]	35.86	104.20	Eastern Asia	MH	10							10
Han Chinese	CHB_1k3	[2]	39.92	116.38	Eastern Asia	SN	103							12
Han Chinese	CHS_1k3	[2]	23.13	113.27	Eastern Asia	SN	105						10	12
Uygur	UYGUR	[90]	35.86	104.20	Eastern Asia	TC	10							10
Xibo	XIBO	[90]	35.86	104.20	Eastern Asia	TU	9							9
Cambodian	CAMBODIAN	[90]	12.57	104.99	South-Eastern Asia	AA	10							10
Bajo	BajoKN	[260]	-1.85	120.52	South-Eastern Asia	BG	22							22
Kalamantan	SKD_br16	[86]	-3.06	115.15	South-Eastern Asia	LD	42					1		20
Igorot	Igorot_mo16	[211]	17.98	121.00	South-Eastern Asia	MP	21							21
Rindi	InSRindi	[261]	-9.90	120.65	South-Eastern Asia	MP	24							12
Luz	Luz	[211]	14.60	120.94	South-Eastern Asia	MP	12		2					12
Murut	Murut_mo16	[211]	4.55	114.16	South-Eastern Asia	NB	17							17
Vietnamese	KHV_1k3	[2]	10.78	106.68	South-Eastern Asia	VM	99							25
Tamil	STU	[2]	6.90	79.90	Southern Asia	DS1	102		4		2	1		25
Telugu	ITU	[2]	17.37	78.48	Southern Asia	DS2	102	2					10	25
Ror	Ror	[262]	27.66	76.90	Southern Asia	II	15							14
Bengali	BEB	[2]	23.70	90.35	Southern Asia	IN	86							25
Gujurati	GIH_1k3	[2]	23.22	72.68	Southern Asia	IN	103				2			25
Punjabi	PJL	[2]	31.55	74.36	Southern Asia	IN	96							25
Parsi	India_Parsi	[263]	25.19	66.59	Southern Asia	IR	19							19
Makrani	MAKRANI	[90]	30.38	69.35	Southern Asia	IR	25							25
Druze	DRUZE	[90]	32.74	35.05	Western Asia	SE	42							25
Turks	Turks	[254]	39.00	30.69	Western Asia	TC	19							19
Polish	polish	[82]	51.77	19.34	Eastern Europe	BS	33				16			17
Russian	RUSSIAN	[90]	55.75	37.60	Eastern Europe	BS	25							12
Vepsa	vepsa	[264]	63.49	33.00	Eastern Europe	FN	11							11
Yakut	YAKUT	[90]	60.83	99.13	Eastern Europe	TC	25							25
Ashkenazi Jew	Ashkenazy_Jews	[254]	51.87	14.65	Eastern Europe		21							21
Finnish	FIN_1k3	[2]	60.17	24.93	Northern Europe	FN	99							25
British	GBR_1k3	[2]	52.49	-1.89	Northern Europe	GM	91		2				10	25
Croatian	croatian	[171]	44.42	14.16	Southern Europe	BS	19							19
Greek	greek	[82]	38.06	19.99	Southern Europe	HE	40				20			20
Iberian	IBS_1k3	[2]	40.38	-3.72	Southern Europe	RM	107							25
Sardinian	SARDINIAN	[90]	40.12	9.01	Southern Europe	RM	28							25
Italian	TSI_1k3	[2]	42.10	12.00	Southern Europe	RM	107							12
Basque	BASQUE	[90]	43.43	-1.37	Western Europe	BQ	24							24

Dutch	KDutch	Kayser Unpub.	52.13	5.29	Western Europe	GM	91	1	6	25
Papuan	PAPUAN	[90]	-6.31	143.96	Melanesia	PN	16			16

Table C.2: Summary of the *a priori* populations and the code (PopCode) used in chapter 6 MALDER analysis . Indicated are the GPS coordinates in decimal degrees North and East, UN region designation and language grouping (see Table 6.4).

Population	PopCode	Source	Decimal Degrees		Region	Ling. abb.
			N	E		
Ju'/hoansi	Juhoansi_sc12	[115]	-19.60	20.49	Southern Africa	JU
Amhara	AMHARA_ag14	[4]	8.96	38.64	Eastern Africa	SE
Hadza	HADZA	[113]	-3.40	36.61	Eastern Africa	HZ
Mozambican	Mozambique	Montinaro *et al.* 2019	-25.89	32.61	Eastern Africa	
Biaka	BIAKAPYGMY	[90]	6.61	20.94	Middle Africa	BI
!Xun	Xun_sc12	[115]	-14.63	17.67	Middle Africa	JU
Mbuti	MBUTIPYGMY	[90]	-4.04	21.76	Middle Africa	SE
Esan	ESN	[2]	9.07	7.48	Western Africa	ED
Dutch	KDutch	Kayser Unpub.	52.13	5.29	Western Europe	GM
French	FRENCH	[90]	46.23	2.21	Western Europe	RM
British	GBR_1k3	[2]	52.49	-1.89	Northern Europe	GM
Iberian	IBS_1k3	[2]	40.38	-3.72	Southern Europe	RM
Bedouin	BEDOUIN	[90]	31.05	34.85	Western Asia	SE
Bengali	BEB	[2]	23.70	90.35	Southern Asia	IN
Balochi	BALOCHI	[90]	30.38	69.35	Southern Asia	IR
Gujurati	GIH_1k3	[2]	23.22	72.68	Southern Asia	IN
Tamil	STU	[2]	6.90	79.90	Southern Asia	DS1
Malay	MAS	[13]	103.82	1.35	Southern Asia	MS
Bajo	BajoKN	[260]	-1.85	120.52	South-Eastern Asia	BG
Igorot	Igorot_mo16	[211]	17.98	121.00	South-Eastern Asia	MP
Dusun	Dusun_mo16	[211]	4.55	114.16	South-Eastern Asia	NB
Malay	Malay_peninsular	[265]	1.49	103.76	South-Eastern Asia	MS
lebbo	lebbo	[260]	0.97	114.38	South-Eastern Asia	
Han Chinese	CHS_1k3	[2]	23.13	113.27	Eastern Asia	SN

Table C.3: Post-hoc Tukey pairwise comparisons of the inter-individual distances within the focal admixed populations including the Southern African KhoeSan. Difference in mean PCA position (Diff.) ordered and only values ≤ 0.05 shown. P-values are adjusted for multiple comparisons. Family-wise 95% confidence intervals indicated (CI).

Comparison	Diff.	Lower 95% CI	Upper 95% CI	Adjusted p-value
KHM_SA-Basters	0.01	0.01	0.02	0.00
Nama-Basters	0.01	0.01	0.02	0.00
ColouredColesberg-Basters	0.01	0.01	0.02	0.00
COLOURED_KS-Basters	0.02	0.01	0.03	0.00
COLOURED_PTG-Basters	0.01	0.01	0.02	0.00
SAN_he11-Basters	0.01	0.00	0.02	0.00
GRIEKWA_UTN-Basters	0.02	0.00	0.03	0.00
COLOURED_KS-COLOURED_CPT	0.02	0.00	0.04	0.00
Khomani-Basters	0.01	0.00	0.02	0.00
Nama-COLOURED_CPT	0.02	0.00	0.03	0.00
KHM_SA-COLOURED_CPT	0.02	0.00	0.03	0.01
COLOURED_PTG-COLOURED_CPT	0.02	0.00	0.03	0.01
ColouredColesberg-COLOURED_CPT	0.02	0.00	0.03	0.01
COLOURED_KS-CAPEMALAY_CPT	0.02	0.00	0.03	0.02
GRIEKWA_UTN-COLOURED_CPT	0.02	0.00	0.04	0.02
SAN_he11-COLOURED_CPT	0.02	0.00	0.03	0.02
COLOURED_KS-COLOURED_BRT	0.01	0.00	0.03	0.05
COLOURED_MHK-COLOURED_KS	-0.03	-0.06	-0.00	0.05