Human Genetic Variation and the Dispersal of Bantu-Speaking Populations

A Thesis submitted for the Degree of *Doctor of Philosophy*

Miguel González dos Santos

St. Cross College and Department of Zoology

University of Oxford

Hilary Term 2018

The DPhil studentship to Miguel González dos Santos (SFRH/BD/90648/2012) was funded by the Portuguese Foundation of Science and Technology (FCT) co-financed by the European Social Fund through the POPH Programme.

ABSTRACT

Human Genetic Variation and the Dispersal of Bantu-Speaking Populations

Miguel González dos Santos

St. Cross College and the Department of Zoology, University of Oxford

A Thesis submitted for the Degree of *Doctor of Philosophy*

Hilary Term 2018

The expansion of Bantu-speaking agropastoralist populations had a significant impact on the genetic, linguistic, and cultural variation of sub-Saharan Africa. Bantu languages originated in an area close to the present-day border between Cameroon and Nigeria not earlier than 5000 years ago. From their homeland, these languages spread south and east through most of sub-Saharan Africa becoming by far the largest African linguistic family. Even though it is generally accepted that this dispersal was the result of the migration of people, there are still many unanswered questions about the dynamics of this process. The demic consequences of this dispersal are reflected in the relatively high genetic homogeneity observed across most of sub-Saharan Africa populations but its implications on the genetic diversity of the continent are not fully appreciated yet. Incorporating African populations' genome-wide data available from Prof. Capelli's research group with previously published data, this thesis addresses and tries to answer some of the questions related to the genetic impact of this dispersal.

I start by addressing the process of admixture between the migrating Bantu speakers and previous inhabitants in south-east Africa.

Rejecting the hypothesis of a potentially pre-Bantu component found in south-east African Bantu-speaking populations as the result of extensive gene flow, I highlight the importance of the demographic dynamics associated with the Bantu dispersal in shaping the genetic diversity of these communities. I then focus on the analysis of the genetic variation observed in sub-Saharan Niger-Congo populations, highlighting the impact of admixture events and the existence of population structure within Bantu speakers, often considered as a genetically homogeneous group. Lastly, I attempt to mask the admixture events I found and reconcile genetics, geography, and linguistics in order to reconstruct the relationships among the Bantu-speaking populations across the continent.

Overall, my results emphasize the diversity observed in Bantu-speaking populations and the impact their rapid dispersal and contact with other communities had on the genetic variation they show.

CONTENTS

ABSTRACT IV

CONTENTS VI

LIST OF ABBREVIATIONS X

CHAPTER 1 GENERAL INTRODUCTION 1

1.1 HUMAN EVOLUTIONARY GENETICS 2

1.1.1 Brief history of the field *2*

1.1.2 Human genome *4*

1.2 THE ORIGINS OF MODERN HUMANS IN AFRICA 8

1.3 HUMAN POPULATION MOVEMENTS 11

1.4 AFRICAN LINGUISTICS 13

1.4.1 Language groups in Africa *13*

1.4.2 Bantu languages *15*

1.5 AFRICAN ARCHAEOLOGY 20

1.6 AFRICAN GENETICS 24

1.7 OVERVIEW AND THESIS AIM 28

CHAPTER 2 MATERIALS AND METHODS 30

2.1 GENETIC DATASETS 31

2.1.1 Novel Samples *31*

2.1.1.1 Collection of the samples 31

2.1.1.2 DNA extraction and genotyping 31

2.1.2 Genetic dataset assembly *33*

2.1.3 Quality Control *37*

2.2 LINGUISTIC DATASET 39

2.3 GEOGRAPHICAL DISTANCES 40

2.3.1 Great-circle line distance (GCD) 40

2.3.2 Model-based geographic distance 41

2.4 STATISTICAL ANALYSIS 44

2.4.1 Admixture analysis 44

2.4.2 F_{ST} analysis 45

2.4.3 Formal tests of admixture (three-population tests) 47

2.4.4 Correlations 48

2.4.5 Correction to genetic distances 49

CHAPTER 3 GENOME-WIDE SNP ANALYSIS OF SOUTHERN AFRICAN POPULATIONS PROVIDES NEW INSIGHTS INTO THE DISPERSAL OF BANTU SPEAKING GROUPS 51

3.1 INTRODUCTION 52

3.2 RESULTS 55

3.2.1 Principal Component Analysis 55

3.2.2 ADMIXTURE 57

3.2.3 F_{ST} components 58

3.2.4 Three-population tests 59

3.2.5 Three-population tests resampling 60

3.2.6 Hierarchical pairwise F_{ST} tree 61

3.3 DISCUSSION 62

CHAPTER 4 CHARACTERIZATION OF THE VARIATION IN BANTU-SPEAKING POPULATIONS 67

4.1 INTRODUCTION 68

4.2 RESULTS 70

4.2.1 Datasets 70

4.2.2 Principal Component Analysis 73

4.2.3 ADMIXTURE 78

4.2.4 Three-population tests (f3) 84

4.2.4.1 Somalia 84

4.2.4.2 Europe 89

4.2.4.3 Mbuti RHG 89

4.2.4.4 'Khoisan' 89

4.2.5 F_{ST} population tree 90

4.2.6 Consolidation of the identified admixture events and the population tree 94

4.3 DISCUSSION 97

CHAPTER 5 AN ATTEMPT TO RECONCILE THE GENETICS, GEOGRAPHY, AND LANGUAGES OF BANTU-SPEAKING POPULATIONS 103

5.1 INTRODUCTION 104

5.2 RESULTS 107

5.2.1 Genetic datasets 107

5.2.2 Geographical clustering of the samples 109

5.2.3 Choice of F_{ST} estimator 111

5.2.4 Analysis of the correlation between pairwise distances 114

5.2.5 F_{ST} Population Tree 117

5.2.6 Correction of genetic distances 119

5.2.7 Model-based geographic distances 127

5.2.7.1 Model description 129

5.2.7.2 Model testing 130

5.2.7.3 Model refinement 133

5.2.7.4 Testing the refined models 136

5.2.7.5 Geographical distribution of the linguistic and genetic distances 140

5.2.7.6 Modifications to the geographical models 142

5.3 DISCUSSION 149

CHAPTER 6 FINAL REMARKS 158

REFERENCES 164

CONTRIBUTORS 186

SUPPLEMENTARY MATERIAL 188

LIST OF ABBREVIATIONS

BCE	Before Common Era
BG	Bantu Genetic (Dataset)
BLG	Bantu Languages Genetic (Dataset)
BP	Before Present
CV	Cross-Validation
DNA	Deoxyribonucleic Acid
EB	Eastern Bantu
ES	Early Split (Model)
GCD	Great-Circle Distance
IBD	Identity by Descent
LS	Late Split (Model)
MG	Mozambique Genetic (Dataset)
MB	Mozambique Bantu
mtDNA	Mitochondrial DNA
NC	Niger-Congo
NCG	Niger-Congo Genetic (Dataset)
NLIA	(Populations with) No Linguistic Information Available
NRY	Nonrecombining Region of the Y chromosome
PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
RHG	Rainforest Hunter-Gatherer
SNP	Single Nucleotide Polymorphism
WB	Western-Bantu

Chapter 1

GENERAL INTRODUCTION

Homo sapiens is widespread across almost the entire world, having settled and adapted to numerous different habitats. The current distribution of human populations, as well as their genetic composition and cultural and technological innovations, are the result of countless movements and migrations of people around the world.

1.1 Human Evolutionary Genetics

All living and extinct organisms share a common ancestor. The evolution of each species from this most recent common ancestor has left a footprint in their genome, a genetic record of the path through which all organisms are related to each other. Evolutionary Genetics as a discipline, and specifically Human Evolutionary Genetics, is based on this principle and involves the study of this genetic record, the differences between genomes, and the mechanisms behind this process.

1.1.1 Brief history of the field

The investigation of the origin of modern humans started much earlier than the establishment of genetics as a field, with records dating back to Ancient Greece (Wood 2005). In the 1700s, Linnaeus recognized for the first time in his book *Systema Naturae* that humans could have biological connections with other animals by including humans and other primates in a single order (first named Anthropomorpha and later Primates) (Willoughby 2007). However, it was only in the 1860s that one of the first

systematic reviews of the differences between modern humans and other primates was performed, by T.H. Huxley (Huxley 1863; Wood 2005). In his work, a milestone in Human Evolutionary studies, Huxley established that modern humans were morphologically closer to chimpanzees and gorillas than these were to Asia's great ape (the orangutan). This morphological evidence led many researchers (including Charles Darwin) to situate Africa as the most likely place where to find the ancestors of modern humans (Darwin 1871; Wood 2005).

The ABO blood group system, discovered in 1900 by Karl Landsteiner, is considered the first defined human genetic polymorphism (Jobling et al. 2013). By mixing red blood cells and serum from different blood samples, Landsteiner discovered that each individual's blood could be divided into one of four groups (A, B, AB, and O) on the basis of two antigens (A and B) that can be either present (one or both) or absent in the red blood cells. Moreover, the subsequent finding that these blood groups are inherited and that their frequencies differ between populations led to the realization that this information could be used to infer the ancestry of these populations, marking the beginning of the study of human genetic diversity.

Since the discovery of the ABO blood group system, the methods used in this discipline have been evolving. Other immunological indirect methods (based on the products of genes) like protein electrophoresis allowed the identification of an increasing number of protein polymorphisms

and led the way to the use of DNA-based genetic methods to study human diversity and history.

Methods for the direct characterization of DNA variation have been used in Human Evolutionary Genetics since the 1970s. However, the analysis of human genetic diversity was truly revolutionized by the development of the Polymerase Chain Reaction (PCR) in the mid-1980s (Mullis and Faloona 1987). This technique allows the amplification of copies of a defined DNA sequence and became an essential step for DNA sequencing and other molecular analyses. The development of this technique opened the door for new genetic markers and direct methods for studying the human genetic variation.

1.1.2 Human genome

The human nuclear genome is diploid, which means that the 46 chromosomes in which the DNA is organized are divided into 23 pairs, one inherited from each parent. Of these, the first 22 pairs are autosomes and equal in both sexes. The last pair represents the sex chromosomes: females have two copies of the X chromosome, and males have one X chromosome and one Y chromosome. Outside the nucleus, in the mitochondria, a single double-stranded circular DNA molecule of approximately 16.5 kb constitutes the mitochondrial DNA (mtDNA), present in multiple copies in each mitochondrion.

The organization of the human genome plays a key role for the field of Human Evolutionary Genetics due to its implications in the genetic transmission across generations (Figure 1.1). During the formation of the gametes the autosomes undergo a process called 'genetic recombination', which means that the genetic information on the autosomes is reshuffled in each generation. On the other hand, two regions of the genome fail to follow this mode of transmission: the nonrecombining region of the Y chromosome (NRY, comprising the great majority of its length, ~60 Mb) and the mtDNA. While the NRY is only passed from fathers to male descendants, mtDNA is passed from mothers to descendants of both sexes. Neither marker

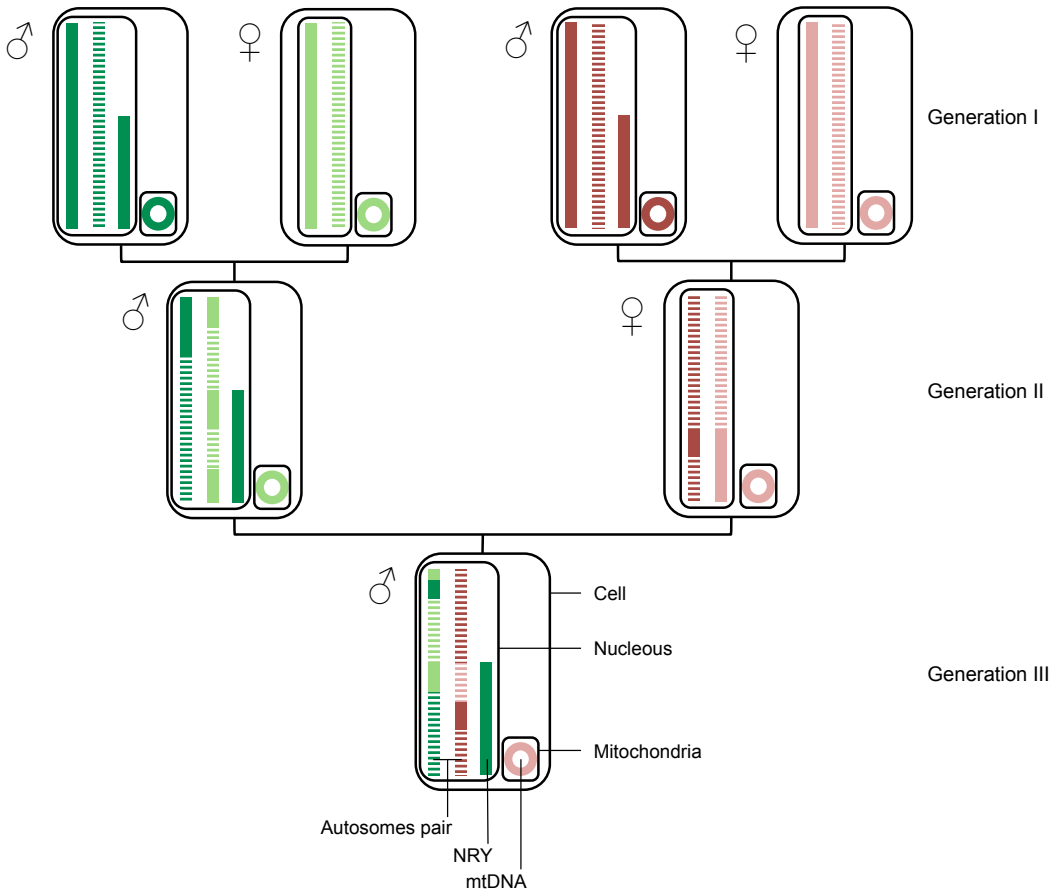


Figure 1.1 – Inheritance patterns of recombining (autosomes) and non-recombining (NRY and mtDNA) segments of the human genome across generations (Adapted from Jobling et al. 2013).

experiences recombination at any stage of life; in this way, these two markers define unaltered (save for mutation) uniparental lineages (Figure 1.1).

Genetic variation in the human genome is present as different types of polymorphisms. The most basic and most abundant source of variation in a DNA sequence is the alteration of a single nucleotide base – a single nucleotide polymorphism (SNP) (Sobrino et al. 2005). One of the great advantages of this kind of marker for Human Evolutionary Genetics is its low mutation rate, in the order of 10^{-8} mutations/generation (Kong et al. 2012; Nachman and Crowell 2000). The low mutation rate observed means that a polymorphism in a particular position in two different sources is normally inherited from a common ancestor – ie, showing identity by descent (IBD). In the case of uniparental markers, the low mutation rate of SNPs combined with the lack of recombination and the uniparental inheritance enabled the sequential accumulation of mutations in these DNA sequences forming haplogroups, lineages of related DNA sequences with a common ancestor.

Recently there has been an explosion of available genetic information through the development of new techniques that allow the analysis of genome-wide markers and the sequencing of whole genomes in a fast and inexpensive way (eg, microarrays and next-generation sequencing) (Mardis 2008). These break-throughs, together with the development of new bio-informatics and bio-statistic methods capable of

processing increasing amounts of genetic data, have allowed us to reach an unparalleled understanding of our evolutionary history.

One of the most successful applications of these increasing amounts of available data are Genome-Wide Association Studies (GWAS). GWAS aim to investigate a very high number of markers over a large population set in order to identify overrepresented alleles in groups of samples. As such, this method plays an essential role in the identification of the underlying genetic variation causing phenotypic manifestations, including diseases. Therefore, in order to correctly interpret the data from GWAS, it is of vital importance to have well characterized patterns of genetic population structure to control for spurious associations due to general differences among groups.

Through the fast evolution of the field we have generated an increasingly refined picture of the history of human populations. However, as no single record is enough to fully explain our past and each presents uniquely different features, it is important to complement the information obtained through the analyses of genetic diversity with that obtained by other disciplines, including archaeology, palaeoanthropology, and linguistics, among others.

1.2 The Origins of Modern Humans in Africa

The origin of the human species is a complex and fascinating topic. Most of our understanding in this area comes from fossils and archaeological findings and, more recently, the investigation of the genetic record in modern samples and ancient DNA. Even though the split between humans and our closest living relatives (the chimpanzees and the bonobos) occurred several million years ago, anatomically modern *Homo sapiens* is a relatively much more recent species. However, the origins of our species still presents several gaps in its records and existing knowledge about the exact time and place for the split from other species within the *Homo* genus is still limited.

Until recently, fossil remains and genetic data of anatomically modern *H. sapiens* in Africa did not show a complete overlap (Meyer et al. 2016; Stringer 2002; Stringer and Galway-Witham 2017). Genetic data suggested that the divergence date to other hominin forms – the Neanderthals and Denisovans – more than 500,000 years ago was much older than the oldest fossil remains of anatomically modern humans (Meyer et al. 2016; Stringer and Galway-Witham 2017). However, recently, fossils recovered from the Jebel Irhoud excavations, in Morocco, were dated and identified as the earliest known record of modern *H. sapiens* at approximately 300,000 years ago (Hublin et al. 2017; Stringer and Galway-Witham 2017). This discovery, combined with the recent finding of a similarly dated *Homo naledi* in South Africa, shows us that the full

understanding of the origins of our species is still a developing investigation (Berger et al. 2017; Stringer and Galway-Witham 2017).

Genetic and archaeological studies point to Africa as the birthplace of *H. sapiens* (Hublin et al. 2017; Manica et al. 2007; Ramachandran et al. 2005; Ray et al. 2005; Stringer and Galway-Witham 2017). Since their origin, humans have expanded through the whole planet, developing different languages and lifestyles. A fast expansion out of Africa and the colonization of the rest of the world would imply a series of bottleneck events (founder events) resulting in a progressive loss of genetic diversity (Manica et al. 2007). Indeed, Africa is where the highest levels of genetic diversity are found today, as well as being one of the most culturally and linguistically diverse regions in the world (Henn et al. 2011; Jorde et al. 2000; Tishkoff and Williams 2002). Nearly a third of the world's living languages (more than 2000 languages) exist in the continent (Simons and Fennig 2018), which is also home to the earliest known modern human remains (Berger et al. 2017; Hublin et al. 2017; Richter et al. 2017; Stringer and Galway-Witham 2017). Moreover, a cline on the loss of diversity with distances from Africa is observed for both cranial phenotypic and genetic data (Figure 1.2) (Manica et al. 2007).

The diversity found in the African continent is not just restricted to the human populations. The topography and climates found in this region are extremely diverse, sometimes even over small distances. Additionally, African populations live in very disparate climates and still exhibit to this

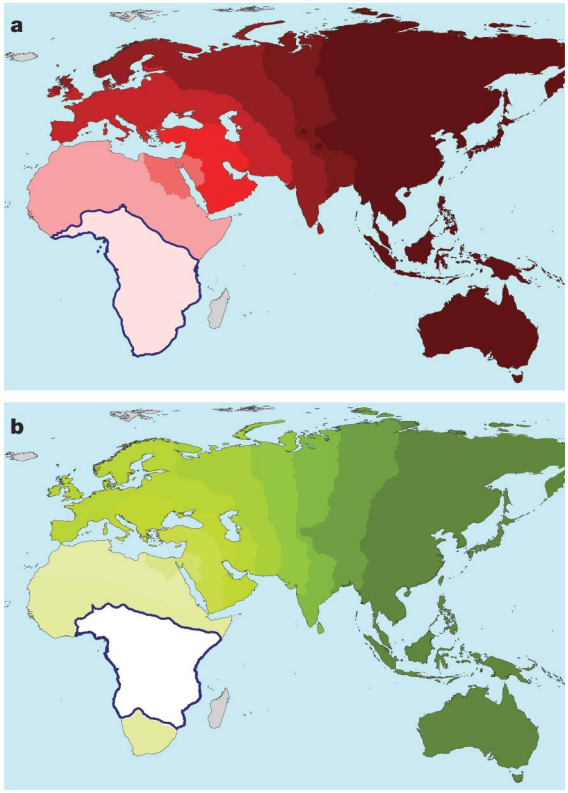


Figure 1.2 – Maps showing the loss of diversity with the distance from an origin location, for (a) phenotypic and (b) genetic data.
The area enclosed by a blue line indicates the most likely origin for each data (From Manica et al. 2007).

date very different subsistence patterns, from agriculture to some of the last examples of hunting and foraging (Mitchell 2002).

All this evidence point to a single African origin for the birthplace of modern humans. The study of African populations is thus an essential step in order to better understand our past and our origins. However, despite its importance, Africa is still underrepresented in human evolutionary genetics studies.

1.3 Human Population Movements

Since the origin of the first hominins, Africa has been the stage for several major population movements, both within the continent as well as out of it. These movements and subsequent episodes of isolation, differentiation, and admixture (both ancient and more recent) had a crucial role in shaping the genetic landscape that can be observed today.

Homo erectus, generally considered the first species of the genus *Homo*, is the earliest hominin found outside Africa (Jobling et al. 2013; Tattersall 2009; Wood and Collard 1999). This genus witnessed several population movements out of the African continent during its history. However, one of these expansions is especially worth mentioning due to its overwhelming impact on the evolutionary history of our species. This 'Out of Africa' migration of anatomically modern humans occurred 50-100 thousand years ago and eventually led to the colonization of almost the entire world with the extinction of all earlier archaic forms (although with some interbreeding) (Jobling et al. 2013; H. Li and Durbin 2011; Lohse and Frantz 2014). As a result of this expansion, the genetic diversity of non-African human populations is a subset of the genetic diversity present in African groups. Moreover, the genetic diversity in the rest of the world shows a decrease as the geographic distance from Africa increases (Manica et al. 2005; Manica et al. 2007; Prugnolle et al. 2005). This is consistent with serial colonization events during the expansion and an expected associated

loss of diversity, a process known as serial founder effect (Figure 1.3) (Henn et al. 2012b; H. Li and Durbin 2011; Ramachandran et al. 2005).

Along the journey of colonization across the world, one of the most significant developments in human evolution was the shift to food production based on the domestication of plants and animals, which happened independently on several places across the globe (Diamond and Bellwood 2003). The development of food production as opposed to hunter-gathering conferred many advantages to these populations, leading to higher population densities that allowed an expansion of these peoples and their lifestyle and culture. As such, it was hypothesized that the farmers and their culture would have replaced hunter-gatherer inhabitants of the settled regions, and therefore a relationship between the spread of agriculture and the distribution of languages would be expected (Diamond and Bellwood 2003). Even though these relationships are disputed and controversial for

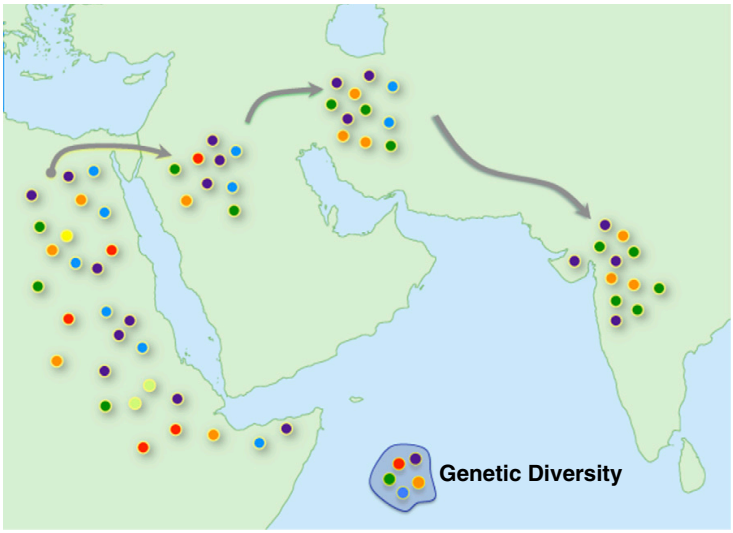


Figure 1.3 – Schematic representation of the loss of genetic diversity associated with the 'Out of Africa' migration and colonization of the rest of the world (From Henn et al. 2012).

other centres of agriculture origins, this has been subject to intense investigation and is well supported in particular for the spread of farmers speaking Bantu languages in sub-Saharan Africa.

1.4 African Linguistics

1.4.1 Language groups in Africa

Languages in the African continent are a key element of most individuals' self-identity and identification as members of a population is deeply rooted in the languages that someone speaks (Phillipson 2005). The actual number of languages spoken today in Africa is very difficult to establish, as it is not always easy to agree on where to draw the line that separates differentiated languages. Notwithstanding, despite some fluctuations, this number has been proposed to be higher than 2000 different languages, amounting to approximately 30% of the world's languages (Heine and Nurse 2000; D. Nurse and Tucker 2002; Simons and Fennig 2018). Excluding languages non-native to the continent and introduced later in its history (namely Indo-European languages introduced during the exploration of parts of the continent by European colonists), the languages in continental Africa can be divided into four main groups: the Niger-Congo, Afroasiatic, Nilo-Saharan, and 'Khoisan' languages (Heine and Nurse 2000) (Figure 1.4).

The Niger-Congo phylum is not only the largest in Africa but also in the world, comprising more than 1400 languages (D. Nurse and Tucker 2002). The distribution of this phylum across the continent is strikingly dominant in comparison to the other phyla present (Figure 1.4). A specific group of Niger-Congo languages especially contributes to this overwhelming distribution – the Narrow Bantu family, which will be explored in further detail in the next section. On the other hand, Afroasiatic and Nilo-Saharan languages are mostly located above the area of distribution of the Niger-Congo phylum, in north and east Africa.

Finally, 'Khoisan' languages represent the smallest of the four groups. This group was once proposed as a family of languages from

Figure 1.4 – Geographic distribution of the main continental African language groups (From Jobling et al. 2013).

southeast Africa sharing the linguistic feature of phonemic clicks (Greenberg 1963; Güldemann and Fehn 2014). However, current evidence points to the fact that the concept 'Khoisan' groups southern and eastern African languages from three isolated families[1]: Tuu, Kx'a, and Khoe-Kwadi (Mitchell 2010; Simons and Fennig 2018). From a genetic perspective, populations from these three families show some of the oldest branches of both male and female lineages (Barbieri et al. 2013b; Batini et al. 2011). Moreover, genome-wide data has allowed the identification of genetic structure among these 'Khoisan' populations. However, this genetic structure has been shown not to correlate with linguistic affiliation or subsistence strategy, correlating instead with geography (Montinaro et al. 2017). The term 'Khoisan' will still be used in parts of this thesis as a loose grouping of the hunter-gatherer and pastoralist populations from southern Africa. Still, this does not mean to suggest any close relationship between these populations and only as a proxy to describe them.

1.4.2 Bantu languages

As stated above, the Narrow-Bantu language family, within African's largest phylum – the Niger-Congo –, shows a clear dominance over the southern half of the continent (Figure 1.4). The close similarity

[1] Different authors have named the three families differently – Tuu, Ju, and Khoe; Southern, Northern, and Central Khoisan (or Khoesan); !Ui-Taa, Ju, and Khoe (Mitchell 2010). In this thesis I used the nomenclature used in Ethnologue (Simons and Fennig 2018).

among Bantu languages was noted as early as the 15th century, when Portuguese ships on their way to India reached the Limpopo estuary on the east African coast and one of Vasco da Gama's sailors, who could speak several languages from the west coast, was able to communicate with some of the inhabitants (de Maret 2013; Vansina 1979). The term 'Bantu' – after the Zulu term 'aba-ntu' (meaning 'men'), also found in related forms in many other Bantu languages – and its use to describe these closely related languages was coined by German linguist Wilhelm Bleek in the 1850s (Eggert 2005).

As expected, the continuous gradient of linguistic differences observed among Bantu languages hinders the task of determining where the boundaries between languages and dialects are, and thus, the task of defining the total number of languages. Even though the figures on the number of Bantu languages vary greatly between authors and depending on the criteria used, several hundreds of Bantu languages exist today, spoken by more than 300 million people in Africa (around one-fourth of the African population) (de Maret 2013; D. Nurse and Tucker 2002; Vansina 1979).

The overwhelming dominance of Bantu languages over an area of more than 11 million km^2 and their close relatedness generated extensive discussions throughout the years about the origin, evolutionary history, and relationships between these languages (Barbieri et al. 2013a; Busby et al. 2016; Choudhury et al. 2017; Currie et al. 2013; de Filippo et al. 2012; de Maret 2013; Grollemund et al. 2015; Lane et al. 2002; Mitchell 2013; Patin

et al. 2017). The vast area presently occupied combined with their low differentiation suggests that the differentiation from a common ancestral language was possibly the result of a relatively recent and fast expansion from a single location – an hypothesis first proposed by Harry Johnston and commonly labelled as the 'Bantu expansion' (de Maret 2013; Johnston 1886; Vansina 1979). Still, several questions about the details of this dispersal can be addressed, some still debatable even after years of study on these languages and their relationships.

The place of origin for the Bantu languages family has been suggested as the Grassfields region of present-day Cameroon, near the border with Nigeria, the place where Bantu languages show their highest levels of diversity and they coincide with the greatest number of related non-Bantu languages (de Maret 2013; Greenberg 1955). From here, the start of the dispersal of Bantu languages across sub-Saharan Africa has been dated from linguistic data to the first or second millennium BCE (de Maret 2013).

Due to limitations on the availability and quality of the linguistic records, the study of the relationships among Bantu languages relies mostly on the field of lexicostatistics, a statistical comparison analysis of lists of words from different languages. This method allows for a general classification of large numbers of languages and the identification of subgroups splitting from the original ancestral language (Blench 2013; de Maret 2013). Through this analysis, four main groups of Bantu languages have been identified: two in the northwestern region of the Bantu distribution

area, near the Bantu homeland and, thus, where the highest levels of linguistic diversity are found; one West/Central Bantu group; and, finally, one East/southern Bantu group (Bastin et al. 1999; de Maret 2013). On a finer note, Malcolm Guthrie proposed what is arguably the most influential division of Bantu languages, initially establishing 15 Bantu zones[2] (Figure 1.5) (Guthrie 1967; D. Nurse and Tucker 2002). Each of Guthrie's Bantu zones can be further subdivided into numbered groups (10, 20, etc.) of

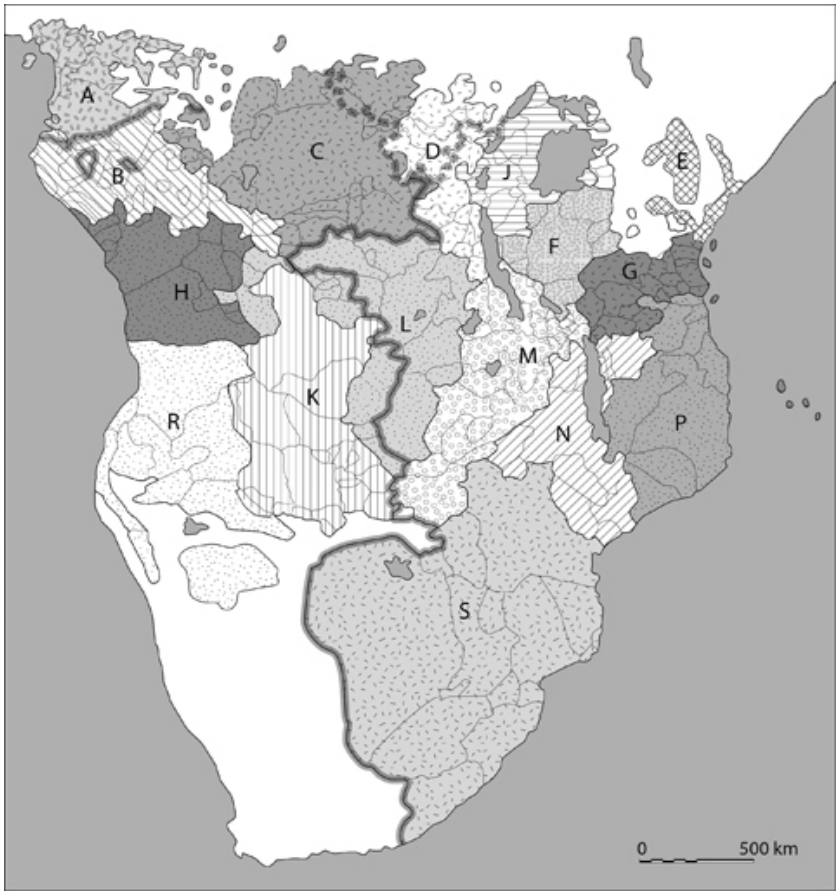


Figure 1.5 – Map with location of Guthrie's division of Bantu languages in geographical zones. The bold line separates the northwest region, the West/Central and the East/Southern Bantu languages (From de Maret 2013).

[2] Guthrie's Bantu zones were labelled with the letters A, B, C, D, E, F, G, H, K, L, M, N, P, R, and S. Later, a new zone 'J' was added by combining parts of zones D and E (D. Nurse and Tucker 2002).

several languages, each individually numbered.

While the place and timings of the beginning of this expansion gather general consensus, the way Bantu languages split between the East and West groups causes much more debate. Two main hypotheses address this issue: the so-called 'early-split' and 'late-split' models (Figure 1.6). Both models admit a southward movement from the Bantu homeland in central-west Africa along the Atlantic coast, giving origin to the West branch of Bantu languages. The major difference between the two hypotheses is then related to the way languages that would originate the East branch travelled east from their homeland. The 'early-split' model supports that the split between the West and East branches would have occurred north of the rainforest, early in the dispersal process. According to this model, an eastern branch would already have been established before heading south (Currie et al. 2013; de Filippo et al. 2012; Holden 2002). On the other hand, the 'late-split' model proposes that the split between these two branches would occur only later in the dispersal process and south of the rainforest.

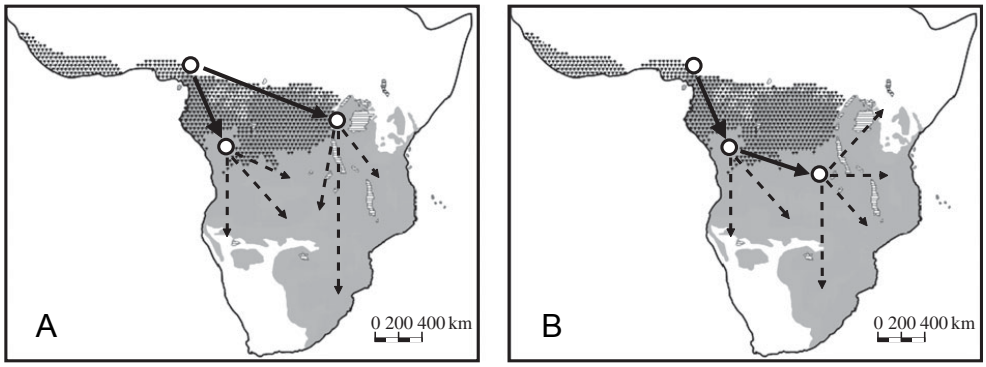


Figure 1.6 – The routes of the Bantu expansion according to the (A) 'early-split' and (B) 'late-split' models (From de Filippo et al. 2012).

After the split, both a movement north (in the direction of the Great Lakes) and a movement south to other areas of southeastern Africa would occur (Ehret 2001; Rexová et al. 2006). Recent studies on this issue have shown that the 'late-split' hypothesis shows a better match with the current distribution of the available linguistic data (Currie et al. 2013; Grollemund et al. 2015).

1.5 African Archaeology

Even though the first studies on the Bantu expansion focused, not surprisingly, on their linguistic aspects, archaeology has played an ever-increasing role in the field. The application of the term 'Bantu' was broadened from its original linguistic meaning and was associated with a socio-economic, cultural, and even ethnographic meaning (de Maret 2013; Eggert 2005).

In the 1960s, archaeological publications stressed the importance of some cultivars for the Bantu-speaking populations and the significance of ironworking to the Bantu dispersal (Eggert 2005; Posnansky 1961). Iron was indeed the first metal worked in sub-Saharan Africa, differently to other parts of the world where the Stone Age was followed by the Copper or Bronze Age (Jobling et al. 2013). Even though some older dates may still remain controversial, there is consensus over evidence for the production of iron tools south of the Sahara around 2300 BP (de Maret 2013).

Ceramics are the main archaeological record left by these populations and similarities in their design allow the establishment of links between archaeological sites, which, in the absence of significant trade, would allow the identification of populations (Mitchell 2013). Archaeological evidence for iron-working can be easily found in sub-Saharan Africa, with examples including the Nok (in Nigeria) and Urewe (in the Great Lakes) centres (Jobling et al. 2013). However, other areas – namely near the homeland of the Bantu expansion – show very sparse archaeological findings hindering the interpretations about the spread of this technology across the continent (de Maret 2013; Jobling et al. 2013). However, where available, the archaeological data on iron-working communities show high levels of homogeneity. The presence of 'dimple-based' and 'channeled' pottery in central, east, and southeast Africa was identified as an indication of a rapid dispersal of people across the continent (Eggert 2005). These communities in southern and eastern Africa are then generally considered a single archaeological culture – the Chifumbaze complex (Jobling et al. 2013; Phillipson 2005). In fact, the Chifumbaze complex comprises two different traditions (Kalundu and Urewe) and present-day Eastern Bantu-speakers as the Shona and the Swahili can be traced to both of them (Mitchell 2013). This connection to present-day Bantu speakers and the widespread distribution of the Chifumbaze complex was interpreted as the result of a dispersal of early farming communities through different dispersal routes in western and eastern Africa (Eggert 2005; Mitchell 2013) (Figure 1.7). In addition, this complex's distribution greatly coincides with the widespread

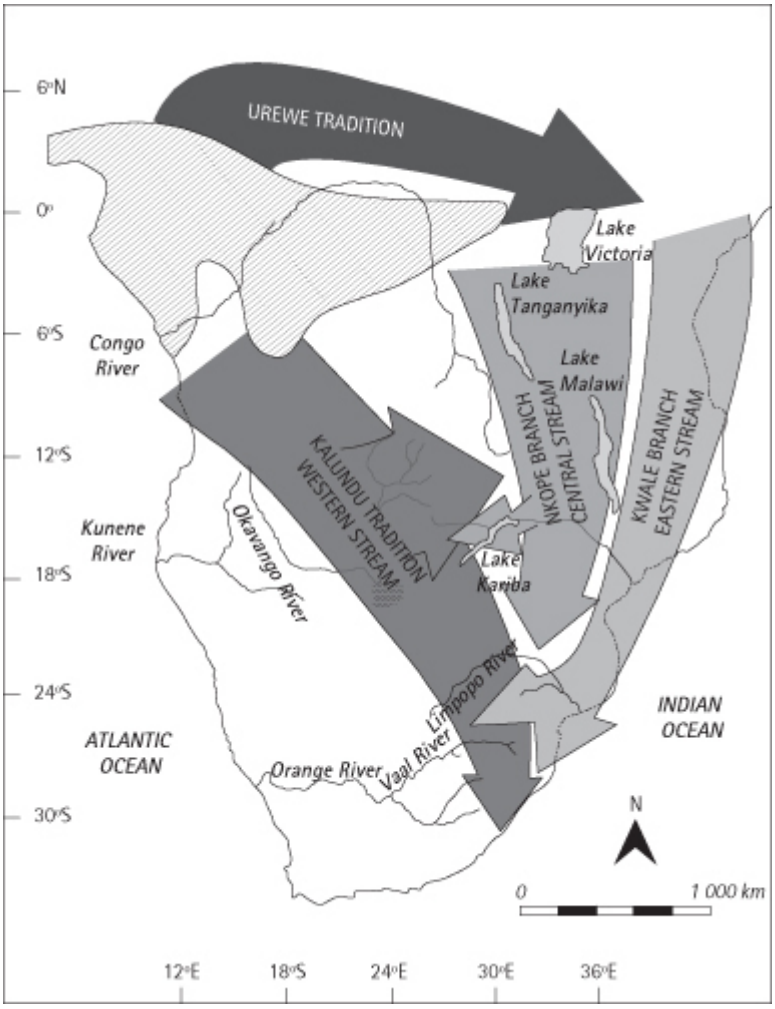


Figure 1.7 – Likely routes of dispersal of Early Farming Communities based on their archaeological traditions (From Mitchell 2013).

distribution of Bantu languages across sub-Saharan Africa discussed in the previous section. This has been interpreted as an indication that the distribution of present-day Bantu speakers was the result of large-scale population movements of iron-working agriculturalists speaking Bantu languages (Eggert 2005; Jobling et al. 2013).

Even though most of the subsistence information about these populations is about their livestock (mostly cattle, sheep, and goats; but also chickens and dogs), there is evidence that the principal crops of these

communities were pearl millet, sorghum, and finger millet (Mitchell 2013). The climatic requirements of the crops deeply restricted the biomes that could be settled by these farming communities. These requirements were met in miombo and savannah biomes, which meant that biomes such as the Kalahari Desert acted as actual geographical barriers to their expansion (Mitchell 2013).

Together with agriculture, the emergence and advantages of iron technologies for land cultivation are thought to have stimulated population growth and movements (Mitchell 2002; Phillipson 2005). The expansion of agriculture alongside the Bantu languages led to an almost complete replacement of previously inhabitant populations in some regions. However, the degree of interaction with the groups inhabiting the regions that were being settled is also still the subject of much debate, with different dynamics of interaction being witnessed in different places. The arrival of Bantu-speaking agriculturalists could have resulted in different degrees of admixture with the local populations of pastoralists and hunter-gatherers leading to the isolation and/or replacement of these autochthonous groups (de Filippo et al. 2012; Destro-Bisol et al. 2004; Mitchell et al. 2008; Mitchell 2009; Patin et al. 2014).

To study this process is then of major importance to gain a better understanding of our own origins and of the individual variation observed in the present and its implications.

1.6 African Genetics

More recently, human genetic studies have increasingly complemented the findings observed for linguistics and archaeology on the expansion of Bantu-speaking populations. Linguistic and archaeological data alone could not answer the question of whether the Bantu expansion was an actual movement of people or the result of a linguistic shift and cultural diffusion. However, the present genetic composition of sub-Saharan populations should be deeply linked to whichever of these two models (or combination of models) was the basis for the Bantu expansion. Thus, genetic data played a key role in the disentangling of the two models of dispersal.

On the one hand, if the expansion of Bantu languages did not involve the actual movement of Bantu speakers and was only the result of cultural diffusion, it would not be expected that the present-day Bantu communities would be genetically closer among themselves than to other neighbouring populations. Alternatively, if the Bantu expansion was the result of a southwards and eastwards migration of Bantu speakers from their homeland in central-west Africa and the settlement of these people on most of sub-Saharan Africa, the opposite would be true and Bantu communities would be genetically closer to each other than to other neighbouring populations. The available genetic data clearly support the latter option, showing that an actual movement of people (a demic diffusion) was involved

in this expansion (Busby et al. 2016; de Filippo et al. 2012; S. Li et al. 2014; Pakendorf et al. 2011; Tishkoff et al. 2009).

It is important to note that some exceptions are observed and this relative genetic homogeneity is not observed for all Bantu-speaking populations in sub-Saharan Africa. In this context, the most striking example is the case of some Bantu-speaking Pygmy populations from central Africa. Even though these Pygmy populations share linguistic affinities with other neighbouring Bantu groups, they are genetically and culturally well differentiated. However, evidence points to the fact that Bantu-speaking Pygmies obtained the language through language shift, promoted by interactions with populations arriving in the area following the Bantu expansion (Bahuchet 2012; Rocha and Fehn 2016).

Even though the genetic perspectives on the Bantu expansion are still fairly recent, there are already a variety of approaches and methodologies used by the field providing different perspectives on the topic. The study of different genetic markers, for example, allowed the addressing of different dynamics experienced by these populations during their expansion.

The analyses of uniparental SNP markers (NRY and mtDNA) data reflected the already mentioned close relationship among the different Bantu communities, with the genetic distances measured on both these markers among Bantu populations being significantly lower than those between

Bantu and other African populations (de Filippo et al. 2012). Moreover, the close relatedness of Bantu-speaking populations can also be witnessed in their association with a small number of uniparental lineages. In the case of the Y chromosome paternal lineages, this association is mainly dominated by a single haplogroup, E1b1a (and, in a smaller scale, others as B2a and E2) (Berniell-Lee et al. 2009; de Filippo et al. 2011). In the case of mitochondrial DNA maternal lineages, a comparatively higher number of lineages seem to be associated with Bantu communities, namely the haplogroups L0a, L1c, L2a, L3b, and L3e (Coelho et al. 2009; Rocha and Fehn 2016; Salas et al. 2002).

More recently, genome-wide studies started bringing important insights into the genetic variation of Bantu-speaking populations, analysing the history of these communities and their genetic relationships with other African groups. The results from genome-wide studies corroborate the relative homogeneity found in Bantu-speaking populations, reflecting a recent separation of different Bantu groups following an expansion through sub-Saharan Africa (Patin et al. 2017). Moreover, these studies allowed the identification of complex dynamics of admixture between the migrating Bantu agriculturalists and local inhabitants of the settled regions (Busby et al. 2016; Marks et al. 2015; Montinaro et al. 2017; Patin et al. 2017; Schlebusch et al. 2012). The combination of genome-wide and uniparental data allowed to disentangle some of these complex dynamics and the identification of sex-mediated admixture events throughout the African continent. While in the rainforest regions of central Africa a male-driven

Bantu to Pygmy admixture was identified (Patin et al. 2014), the admixture dynamics in southern Africa seem to be very different. Here, the identification of high frequencies of maternal lineages related to ‘Khoisan’ populations (mtDNA haplogroups L0d and L0k) without the corresponding ‘Khoisan’-related NRY lineages in similar proportions point to a female-driven ‘Khoisan’ to Bantu admixture (Barbieri et al. 2013b; Schlebusch et al. 2013).

The use of haplotype-based admixture inference methods also allows for the exploration of modes of dispersal of Bantu-speaking populations. These methods, together with evidence provided by linguistic data, support a mode of dispersal matching the late-split model, with a first southwards movement followed by differentiation south of the rainforest and subsequent movements eastwards and farther southwards (Busby et al. 2016; Grollemund et al. 2015; Patin et al. 2017).

1.7 Overview and Thesis Aim

This DPhil thesis is divided into a total of six chapters:

The current chapter, Chapter 1, is an introduction to the theme and relevant general concepts aiming to provide the context and relevant information to the field.

In Chapter 2 I have included an overview of all the data used in the different stages of the subsequent three data chapters. I have included as well a detailed description of all the methodology followed.

Chapter 3 focuses on the process of admixture during the Bantu expansion between incoming farmers and foraging residents. I address the hypothesis of an early pre-Bantu component present in south-east African populations as previously proposed (Sikora et al. 2011). This work was the basis of a paper that was published in the journal *Genome Biology and Evolution* and it is presented in its final published form in the Supplementary Material.

Chapter 4, following insights from the results of the previous chapter, focuses on the investigation of the genetic variation observed in sub-Saharan Niger-Congo populations and the consequences of the interactions of these populations with earlier inhabitants of the area.

On Chapter 5 I test the fit of different models to the available data, trying to reconcile genetics, geography, and linguistics on the spread of Bantu-speaking populations and to reconstruct the patterns of populations' splits and their implications on the genetic diversity observed today.

Finally, Chapter 6 consists of a general discussion of the preceding chapters and analyses the future prospects of the work.

The present-day widespread distribution of Bantu-speaking populations across sub-Saharan Africa is still a topic with several unanswered questions, namely regarding the details of their expansion and its implications on the genetic variation of the continent.

Altogether, my DPhil thesis aims to obtain a better understanding and characterization of the human genetic variation found in sub-Saharan Niger-Congo populations. Through the use of genome-wide data and a better data coverage, I aim to analyse the mode and genetic implications of the dispersal of Bantu-speaking populations and the routes taken by them.

Chapter 2

MATERIALS AND METHODS

2.1 Genetic Datasets

2.1.1 Novel Samples

2.1.1.1 Collection of the samples

Saliva samples from unrelated individuals were collected either by me, by members of the group, or by collaborators in various field trips using the Oragene® DNA collection kits (DNA Genotek, Inc., Ottawa, Ontario, Canada). The samples were collected after ethical approval by the appropriate Ethics Committees for each trip was obtained and all the participants are healthy adults from whom informed consent was obtained at the time of the sampling. A questionnaire regarding the ethnic and linguistic information about the donors, as well as from their parents and grandparents, was given to the participants and this information was obtained when known.

2.1.1.2 DNA extraction and genotyping

DNA was extracted according to the manufacturer's protocols. The samples collected with these kits are properly mixed after collection and incubated at 50ºC in order to inactivate the nucleases present and so that the DNA is released prior to the extraction. Part of this mixture is mixed with PT-L2P and incubated in ice to precipitate inhibitors and other impurities. The samples are then centrifuged so that the impurities form a pellet in the bottom of the tube and the DNA remains in the supernatant. In order to

precipitate the DNA in the sample 95% ethanol is added at room temperature. The tubes are again centrifuged and, this time, the supernatant is discarded as the DNA is precipitated in the pellet. A final 70% ethanol wash is performed at room temperature and the supernatant is again discarded. The DNA samples were diluted in TE solution (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) storage buffer and incubated to at 50ºC to facilitate the rehydration of the DNA.

Two Illumina SNP arrays (Illumina, San Diego, CA, USA) were used to genotype the samples. Basotho and Owambo samples were genotyped with the Illumina Human 610-Quad BeadChip and all the others were genotyped with the Human Omni5-Quad BeadChip. Information about the samples genotyped by me or other members of Prof. Capelli's research group included in this thesis is available in Table 2.1.

Table 2.1 – Population samples genotyped by members of Prof. Capelli's research group and included in this thesis.

Population	n	Country of Origin	Provider
Angola	10	Angola	Sampled by myself in Portugal, 2016
Basotho	8	Lesotho	Sampled by members of the group, 2009
Ndebele	10	Lesotho	Sampled by members of the group, 2009
Owambo	10	Namibia	Sampled by members of the group, 2010
Himba	10	Namibia	Sampled by members of the group, 2010
Mbukushu	8	Namibia	Sampled by members of the group, 2012
Kwangali	7	Namibia	Sampled by members of the group, 2012
Damara	8	Namibia	Sampled by members of the group, 2012
Zimbabwe	12	Zimbabwe	Celia May (University of Leicester)
Zambia	11	Zambia	Masharip Atadzhanov (University of Zambia)
Mozambique	13	Mozambique	Nedio Mabunda (Instituto Nacional de Saúde, Mozambique)
Yansi	4	Democratic Republic of the Congo	Pierre de Maret (Université Libre de Bruxelles)
Kongo	6	Democratic Republic of the Congo	Pierre de Maret (Université Libre de Bruxelles)
Fwe	8	Zambia	Brigitte Pakendorf (CNRS)
Tswana	8	Botswana	Brigitte Pakendorf (CNRS)
Kgalagadi	9	Botswana	Brigitte Pakendorf (CNRS)

2.1.2 Genetic dataset assembly

In order to perform comparative studies with other populations, I used the software PLINK versions 1.07 and 1.9 (Chang et al. 2015; Purcell et al. 2007) to merge my data with available genome-wide SNP data from different African populations genotyped on different Illumina and Affymetrix platforms (Table 2.2). Four datasets were assembled and used in different parts of this thesis (Table 2.3):

- **'Mozambique Genetic Dataset'** (Chapter 3);
- **'Niger-Congo Genetic Dataset'** (Chapter 4);
- **'Bantu Genetic Dataset'** (Chapters 4 and 5); and
- **'Bantu Languages Genetic Dataset'** (Chapter 5).

The **'Mozambique Genetic Dataset'** (MG) was the first dataset I assembled for this thesis and consists of genome-wide SNP data from 35 sub-Saharan Africa populations, both from Niger-Congo and other linguistic groups. This dataset was used in Chapter 3 to investigate the presence of a reported 'non-Bantu' genetic component in Mozambique.

The **'Niger-Congo Genetic Dataset'** (NCG) was then assembled to further investigate the findings from Chapter 3 and to better explore the genetic relationships among different Bantu-speaking populations by increasing both the range of populations and the number of markers used. This dataset consists of several sub-Saharan African populations, mainly

Table 2.2 – Populations from other studies included in this thesis. NC, Niger-Congo; AA, Afroasiatic; NS, Nilo-Saharan; NLIA, no linguistic information available; K, 'Khoisan'

Population	Country of Origin	Language Family	Source
Bamoun	Cameroon	NC, Bantoid	Bryc et al. 2010
Brong	Ghana	NC, non-Bantu	Bryc et al. 2010
Fang	Cameroon	NC, Narrow Bantu	Bryc et al. 2010
Kongo	Democratic Republic of the Congo	NC, Narrow Bantu	Bryc et al. 2010
MbororoFulani	Nigeria	NC, non-Bantu	Bryc et al. 2010
Xhosa	South Africa	NC, Narrow Bantu	Bryc et al. 2010; Petersen et al. 2013
Lhuya	Kenya	NC, Narrow Bantu	Frazer et al. 2007
Baganda	Uganda	NC, Narrow Bantu	Gurdasani et al. 2014
Banyarwanda	Uganda	NC, Narrow Bantu	Gurdasani et al. 2014
Barundi	Uganda	NC, Narrow Bantu	Gurdasani et al. 2014
Fula	Gambia	NC, Senegambian	Gurdasani et al. 2014
GaAdangbe	Ghana	NC, Kwa	Gurdasani et al. 2014
Jola	Gambia	NC, Bak	Gurdasani et al. 2014
Kikuyu	Kenya	NC, Narrow Bantu	Gurdasani et al. 2014
Somali	Ethiopia/Somalia	AA, Cushitic	Gurdasani et al. 2014
Sotho	South Africa	NC, Narrow Bantu	Gurdasani et al. 2014
Wolof	Gambia	NC, Senegambian	Gurdasani et al. 2014
Zulu	South Africa	NC, Narrow Bantu	Gurdasani et al. 2014
Igbo	Nigeria	NC, Igboid	Gurdasani et al. 2014; Bryc et al. 2010
Mandenka	Gambia/Senegal	NC, Mande	Gurdasani et al. 2014; Li et al. 2008
Yoruba	Nigeria	NC	Gurdasani et al. 2014; Li et al. 2008
Maasai	Kenya	Language isolate	HapMap
Hadza	Tanzania	Language isolate	Henn et al. 2011
Sandawe	Tanzania	Language isolate	Henn et al. 2011
Lemande	Cameroon	NC, Narrow Bantu	Jarvis et al. 2012
Ngumba	Cameroon	NC, Narrow Bantu	Jarvis et al. 2012
TikarSouth	Cameroon	NC, Bantoid	Jarvis et al. 2012
BantuKenya	Kenya	NC, Narrow Bantu	Li et al. 2008
BantuSouthAfrica	South Africa	NC, Narrow Bantu	Li et al. 2008
BiakaPygmy	Central African Republic	NS	Li et al. 2008
PygmyMbuti	Democratic Republic of the Congo	NS and NC	Li et al. 2008
BurkinaFaso	Burkina Faso	NLIA	MalariaGEN
Cameroon	Cameroon	NLIA	MalariaGEN
Gambia	Gambia	NLIA	MalariaGEN
Ghana	Ghana	NLIA	MalariaGEN
Kenya	Kenya	NLIA	MalariaGEN
Malawi	Malawi	NLIA	MalariaGEN
Mali	Mali	NLIA	MalariaGEN
Tanzania	Tanzania	NLIA	MalariaGEN
Soweto	South Africa	NC, Narrow Bantu	May et al. 2013
Anuak	Ethiopia	NS	Pagani et al. 2012
AriBlacksmith	Ethiopia	AA	Pagani et al. 2012
AriCultivator	Ethiopia	AA	Pagani et al. 2012
Gumuz	Ethiopia	NS	Pagani et al. 2012
SouthSudan	South Sudan	NS	Pagani et al. 2012
Bakiga	Uganda	NC, Narrow Bantu	Patin et al. 2014
Nzebi	Gabon	NC, Narrow Bantu	Patin et al. 2014
Nzime	Cameroon	NC, Narrow Bantu	Patin et al. 2014
Himba	Namibia	NC, Narrow Bantu	Pickrell et al. 2012
Mbukushu	Namibia	NC, Narrow Bantu	Pickrell et al. 2012
Owambo	Namibia	NC, Narrow Bantu	Pickrell et al. 2012
Tswana	Botswana	NC, Narrow Bantu	Pickrell et al. 2012
G/wi-G//ana	Botswana	K	Schlebusch et al. 2012
Herero	Namibia	NC, Narrow Bantu	Schlebusch et al. 2012
Ju/'hoansi	Namibia	K, K'xa	Schlebusch et al. 2012
Karretjie	South Africa	K	Schlebusch et al. 2012
≠Khomani	South Africa	K	Schlebusch et al. 2012
Khwe	Angola	K	Schlebusch et al. 2012
Nama	Namibia	K	Schlebusch et al. 2012
SEBantu	South Africa	NC, Narrow Bantu	Schlebusch et al. 2012
!Xũ	Namibia	K	Schlebusch et al. 2012; Petersen et al. 2013
MozambiqueBantu	Mozambique	NC, Narrow Bantu	Sikora et al. 2011

Table 2.3 – Number of individuals included in each populations for each dataset analysed in this thesis.

		Mozambique Genetic Dataset	*Niger-Congo Genetic Dataset*	*Bantu Genetic Dataset*	*Bantu Languages Genetic Dataset*
Non-Niger-Congo					
	CEU	*Note 1**	60	-	-
	Ju/'hoansi	42	18	-	-
	Damara	-	8	-	-
	G/wi-G//ana	14	-	-	-
	Karretjie	11	-	-	-
	≠Khomani	16	-	-	-
	Khwe	17	-	-	-
	Nama	12	-	-	-
	!Xũ	31	-	-	-
	PygmyMbuti	13	13	-	-
	PygmyBiaka	21	-	-	-
	AriBlacksmith	15	-	-	-
	AriCultivator	22	-	-	-
	Hadza	3	-	-	-
	Maasai	96	-	-	-
	Sandawe	23	-	-	-
	Anuak	21	-	-	-
	Gumuz	13	-	-	-
	SouthSudan	21	-	-	-
	Somali	-	39	-	-
Populations with no linguistic information available					
	Angola	-	10	-	-
	Mozambique	-	13	-	-
	Tanzania	-	147	-	-
	Mali	-	99	-	-
	Malawi	-	100	-	-
	Kenya	-	232	-	-
	Ghana	-	150	-	-
	Gambia	-	387	-	-
	Cameroon	-	100	-	-
	BurkinaFaso	-	50	-	-
Niger-Congo non-Narrow-Bantu					
	MbororoFulani	-	11	-	-
	Wolof	-	78	-	-
	Fula	-	74	-	-
	Brong	-	6	-	-
	TikarSouth	-	19	-	-
	Yoruba	122	120	-	-
	Mandenka	22	110	-	-
	GaAdangbe	-	100	-	-
	Igbo	-	112	-	-
	Bamoun	-	18	-	-
	Jola	-	79	-	-
Niger-Congo Narrow-Bantu					
	MozambiqueBantu	69	-	-	-
	Lemande	-	19	19	19
	Fang	-	15	15	15
	Ngumba	-	20	20	20
	Nzime	-	50	50	50
	Nzebi	-	20	20	20
	Yansi	-	4	4	4
	Kongo	-	15	15	15
	Baganda	-	100	100	100
	Bakiga	-	32	32	32
	Banyarwanda	-	100	100	100
	Barundi	-	97	97	97
	Kikuyu	-	99	99	99
	Luhya	81	86	86	86
	BantuKenya	8	10	10	-
	Herero	9 (*Note 2**)	12	12	12
	Kwangali	7	7	7	7
	Owambo	10	15	15	15
	Fwe	-	8	8	-
	Himba	-	15	15	-
	Mbukushu	8	12	12	-
	Zambia	-	11	11	-
	Zimbabwe	-	12	12	-
	Xhosa	15	18	18	18
	Basotho	8	8	8	8
	Kgalagadi	-	9	9	9
	Ndebele	-	10	10	10
	Tswana	-	12	12	12
	Zulu	-	100	100	100
	BantuSouthAfrica	8	8	8	-
	SEBantu	19	20	20	-
	Sotho	-	86	86	-
	Soweto	91	94	94	-

**Note 1:* In the 'Mozambique Genetic Dataset', the CEU data was included in a first stage. After performing an ADMIXTURE analysis to detect individuals in the other samples characterized by more than 10% of the non sub-Saharan Africa coponent, I removed CEU (and the detected individuals across samples) from the final dataset.

**Note 2:* In the 'Mozambique Genetic Dataset', the Herero sample is coded as 'SWBantu'.

from the Niger-Congo linguistic family[3]. Additionally I also included, for reference, a European population, an east African non-Niger-Congo population (from Somalia, speaking a Cushitic language), a Khoisan population[4] (the Ju/'hoansi from the Kx'a language family), and a rainforest hunter-gatherer (RHG, a.k.a. Pygmies). The Damara, a Khoe-speaking population from Namibia, were also included in this dataset, as they have been shown to cluster with Niger-Congo populations before (Pickrell et al. 2012).

The **'Bantu Genetic Dataset'** (BG) is a subset of the 'Niger-Congo Genetic Dataset' comprising only the populations speaking Narrow Bantu languages. Even though some of the populations from the 'Niger-Congo Genetic Dataset' were included in that dataset without linguistic affiliation – and notwithstanding the fact that some of those populations showed similar patterns of variation and clustered with Bantu populations in some of the subsequent analyses in the following chapters – in this subset I decided to only keep populations with language information to avoid inaccurate assessments. As even when the language affiliation of a population was known linguistic data to build the linguistic distance matrix were not always available, a subset of this dataset was assembled to include only the

[3] Even though linguistic data were not available for all populations some were still included in the analyses as they were from predominantly Niger-Congo-speaking countries and they are shown to cluster with other Niger-Congo populations in the subsequent analyses.

[4] 'Khoisan' is here used as a term for non-Bantu southern Africa populations speaking click-rich languages (Güldemann and Fehn 2014).

populations for which linguistic data made it possible for them to be included in the linguistic distance matrix used in Chapter 5 and subsequent analyses – the **'Bantu Languages Genetic Dataset'** (BLG).

2.1.3 Quality Control

The datasets went through a similar quality control (QC) process before any analysis was performed. Before the assembly, each individual dataset was first processed using the software PLINK versions 1.07 or 1.9 (Chang et al. 2015; Purcell et al. 2007) to remove markers and then individuals with a missing call rate higher than 10%. The variant positions for each individual dataset were also lifted to build 37 human genetic maps using data provided by Illumina and Affymetrix.

After this first QC step, all the populations were merged one at a time according to the marker's position to avoid differences in nomenclature between genotyping platforms. When all population data were merged an additional QC step was performed, this time removing variants and afterwards individuals with a missing call rate higher than 2%. In order to overcome the effects of markers in strong Linkage Disequilibrium (LD), I also removed all markers with a correlation (r^2) greater than 0.4 using a sliding window of 200 SNPs, shifted at 25 SNPs intervals (Behar et al. 2010).

After the datasets were assembled I used KING software (Manichaikul et al. 2010) to infer kinship between samples and randomly

removed one individual from pairs showing a kinship rate higher than 0.0884 (up to second degree relationship).

Merging genotype data produced with both Illumina and Affymetrix platforms has been previously employed with success, with no evidence of errors or biases (Henn et al. 2012a; Montinaro et al. 2017; Reich et al. 2009). The analyses performed on this thesis on datasets with data from both manufacturers also did not show any significant differences between related groups genotyped with different platforms.

An additional QC step was performed in the 'Mozambique Genetic Dataset', in order to detect (and subsequently remove) samples influenced by strong recent European admixture. The CEU data from HapMap Phase 3[5] (International HapMap Consortium 2003) was included in the dataset (after following the same QC steps described above) and I performed an ADMIXTURE analysis (D. H. Alexander et al. 2009) with the default options and removed all individuals characterized by more than 10% of the non-sub-Saharan Africa component. In the 'Niger-Congo Genetic Dataset' I include instead the CEU population sample in all the analyses (except were noted) as a reference so this additional QC step was not performed[6].

[5] Data available at the website http://hapmap.ncbi.nlm.nih.gov/.

[6] Even though the additional QC step to remove samples with recent European admixture was not performed in the 'Niger-Congo Genetic Dataset', after the first Principal Component analysis two samples (both from the Mozambique population) clustered with the rest of the European samples and away from all

Table 2.4 shows the final number of individuals, populations, and markers that composed the final four datasets.

Table 2.4 – Summary of the datasets used in this thesis.

	Mozambique Genetic Dataset	*Niger-Congo Genetic Dataset*	*Bantu Genetic Dataset*	*Bantu Languages Genetic Dataset*
Number of populations	35	59	32	22
Number of individuals	868	3277	1124	848
Number of markers	1755	10809	10809	10809

2.2 Linguistic dataset

The linguistic dataset I use in Chapter 5 of this thesis was assembled by a collaborator from the Departments of English and Anthropology from the University of Missouri, Rebecca Grollemund, based on the languages spoken by the Niger-Congo-speaking populations present on the 'Niger-Congo Genetic Dataset'. Briefly, to obtain this dataset 100 basic vocabulary words[7] were considered and cognate sets (similar words with similar meanings, and thus indicating that those languages might be related) for each of these words were identified and coded as discrete

the African ones. Being so, these two samples seem to represent some very recent non-African influence not noted during the sampling process and thus I removed them from the final dataset and all the analyses showed in this thesis.

[7] The 100 words analysed were, in English: *all*, *animal*, *arm*, *ashes*, *bark*, *bed*, *belly*, *big*, *bird*, *bite*, *blood*, *bone*, *breast* (woman), *burn*, *child*, *cloud*, *come*, *count*, *die*, *dog*, *drink*, *ear*, *eat*, *egg*, *elephant*, *eye*, *face*, *fall*, *fat*, *feather*, *fingernail*, *fire*, *fire-wood*, *fish*, *five*, *fly*, *four*, *give*, *goat*, *ground*, *soil*, *hair* (on head), *head*, *hear*, *heart*, *horn*, *house*, *hunger*, *intestine*, *kill*, *knee*, *knife*, *know*, *leaf*, *leg*, *liver*, *louse*, *man*, *moon*, *mountain*, *hill*, *mouth*, *name*, *navel*, *neck*, *night*, *nose*, *oil*, *one*, *person*, *rain*, *road/path*, *root* (big), *salt*, *sand*, *see*, *send*, *shame*, *sing*, *skin*, *sky*, *sleep*, *smoke*, *snake*, *spear*, *steal*, *stone*, *sun*, *tail*, *ten*, *three*, *tongue*, *tooth*, *tree*, *two*, *urine*, *village*, *vomit*, *walk*, *war*, *water*, *wind*, and *woman*.

binary characters. The distance matrix between languages was estimated based on the Hamming distance, counting the number of differences between pairs of sequences/characters. A neighbour-joining algorithm with sequential agglomeration (data were combined into progressively larger overlapping clusters) was used to construct and subset a phylogenetic network based on the linguistic distance matrix of the Niger-Congo-speaking populations included in this thesis.

2.3 Geographical distances

The geographical coordinates of the different populations used in this thesis were registered during fieldwork or retrieved from the original study. In the case of populations not sampled on the country of origin or from which no geographical information was available I used the coordinates of the capital city from the country of origin as a proxy. The geographical distance between all pairs of populations in the BG dataset based on their geographical coordinates was calculated through two different approaches: the great-circle line distance and a model-based approach (explained in detail in the corresponding chapter, Chapter 5).

2.3.1 Great-circle line distance (GCD)

In a first step the great-circle line distance (GCD) between each pair of populations was calculated. The great-circle line distance between two populations is the shortest distance between them taking into account

the curvature of the surface of the Earth and, therefore, ignores the existence of any kind of geographical barrier along its path. However, it is known that geographical barriers played a major factor in the dispersal of human populations and thus the great-circle line distance is probably not a true representation of the real distance that populations had to travel during the population movements that characterized the Bantu expansion.

2.3.2 Model-based geographic distance

I developed and calculated model-based geographical distances based on clusters of different neighbouring populations connected to each other through waypoints. I first performed a hierarchical clustering of the great-circle distances between pairs of populations in the BG dataset in order to identify clusters of populations (Figure 2.1). I identified four geographical clusters in my dataset – '*cluster A*', with populations from west Africa; '*cluster B*', with populations from east Africa; '*cluster C*', with populations from Namibia, Zambia, and Zimbabwe; and '*cluster D*', with populations from the rest of southern Africa.[8] This clustering is solely based on geography and does not take into consideration any other factors. However, *cluster C* includes populations speaking languages from both the East (Zambia and Zimbabwe) and West (Herero, Kwangali, Owambo, Fwe,

[8] The geographical clustering of the samples was performed on Chapter 5 in order to implement the models of geographical dispersal. However, for clarity and consistency this division was also used on Chapter 4 for the interpretation of the results regarding the BG dataset.

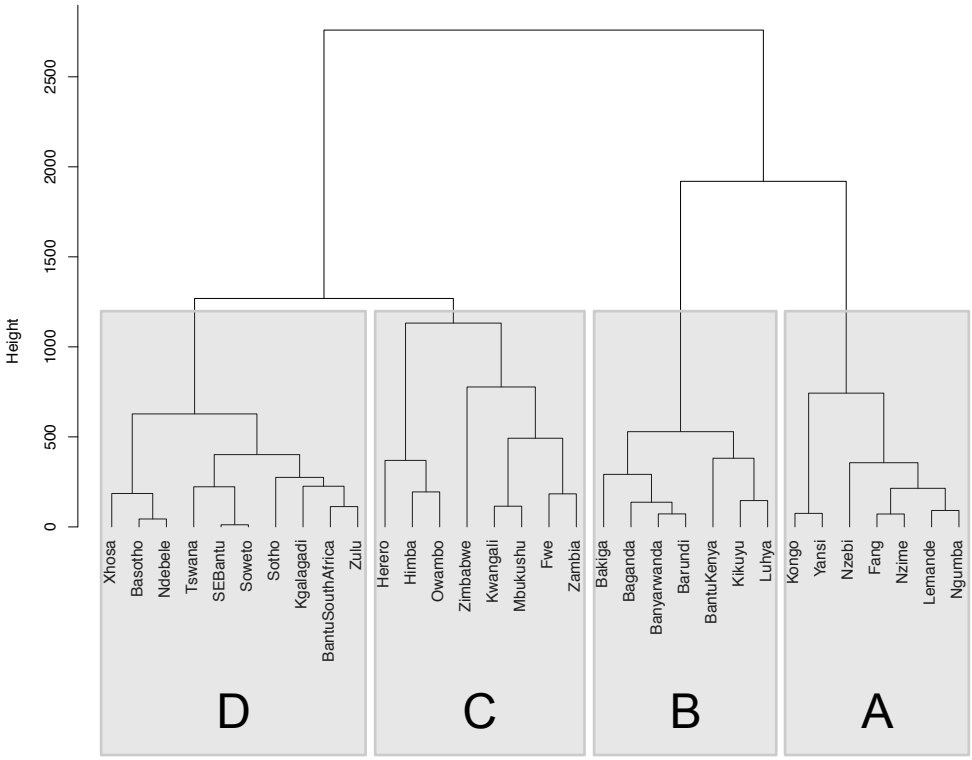


Figure 2.1 – Hierarchical clustering of the great-circle distances in the 'Bantu Genetic Dataset', identifying four Bantu geographical clusters (A to D).

Himba, and Mbukushu) branches of the Narrow Bantu linguistic family. By incorporating the linguistic information into the original clustering, *cluster C* can then be further divided into two subgroups: '*cluster C East*' and '*cluster C West*'. Given the geographical location of these clusters in the African continent I also identified one (for *clusters A*, *B*, and *D*) or more (for *cluster C*) waypoints – waypoints *mA*, *mB*, *mC1*, *mC2*, *mC3*, and *mD* – to represent the borders of each cluster, corresponding, as appropriate, to the combination of either the northernmost or the southernmost population coordinate with the easternmost or the westernmost coordinate in that cluster (Figure 2.2). An alternative waypoint – waypoint *mAB* – was also included as the midpoint between the waypoints *mA* and *mB* to be used in some of the geographical models tested (see Chapter 5).

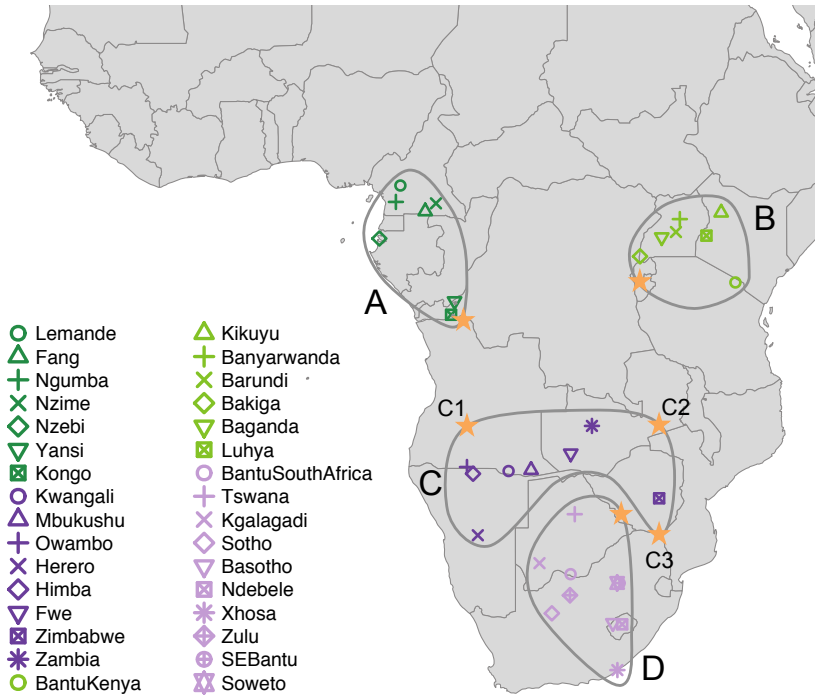


Figure 2.2 – Map with the location and populations comprising each of the four Bantu geographical clusters identified for the 'Bantu Genetic Dataset'.

After this, the distances were calculated based on different models tested as described in Chapter 5. Distances between pairs of populations belonging to the same cluster were calculated as the great-circle distance between them for all the models. The different models differ in the way that pairs of populations in different clusters are connected, the distance between them being the sum of the distance between each population and the corresponding waypoint for its cluster and the distance between waypoints (directly or through a third or more waypoints). In the specific case of *cluster C* (the only one with more than one waypoint) the waypoint to which its populations are connected and through which the other clusters are linked depends on the model being tested.

The different models used to calculate the distances based on this model-based method are explained with more detail in the corresponding chapter (Chapter 5). All distances were calculated based on the population sampling location coordinates with the function *rdist.earth* of the package *fields* (Nychka et al. 2015).

2.4 Statistical analysis

2.4.1 Admixture analysis

The population genetic structure was initially explored via Principal Component (PC) analysis on the SNP data using the software PLINK version 1.9 (Chang et al. 2015).

To further explore the genetic variation found among the populations in the study I used the software ADMIXTURE (version 1.23) (D. H. Alexander et al. 2009). This method allows for a model-based estimation of ancestry by implementing a maximum likelihood algorithm that defines a set number of ancestral populations (*K*) and assigns the individuals in the dataset to those populations (D. H. Alexander et al. 2009). The cross-validation (CV) procedure implemented in the software – which uses the consistency between different runs of the algorithm at any given value of *K* as an indicator of the validity of that value of *K* – was used as an indication of the most supported value of ancestral sources (D. H. Alexander et al. 2009; D. H. Alexander and Lange 2011). The genetic distance between

components was also calculated with the same software to assess the genetic differentiation between them using the F_{ST} measure (D. H. Alexander et al. 2009).

The spatial distribution of the ADMIXTURE components (for *K*=6 in the 'Mozambique Genetic Dataset' and *K*=9 in the 'Niger-Congo Genetic Dataset', see Chapters 3 and 4, respectively) was visualized by plotting the fraction of each component in the analyzed populations on a map with the interpolation plugin of the QGIS software, using the Inverse Distance Weighting method with a distance coefficient of 3 (QGIS Development Team 2014).

2.4.2 F_{ST} analysis

Pairwise F_{ST} between all pairs of populations was calculated for all datasets. For the 'Mozambique Genetic Dataset' the Weir and Cockerham's F_{ST} estimator was calculated using the *R* package *StAMPP* (Pembleton et al. 2013; Weir and Cockerham 1984). On the other hand, for the 'Niger-Congo Genetic Dataset' (and consequently for its subset, the 'Bantu Genetic Dataset') I wrote a script to calculate the Hudson estimator of the pairwise F_{ST}, following the formula (1) proposed by Bhatia and collaborators (Bhatia et al. 2013; Hudson et al. 1992).

$$F_{ST}^{Hudson} = \frac{(p_1 - p_2)^2 - \frac{p_1(1-p_1)}{n_1 - 1} - \frac{p_2(1-p_2)}{n_2 - 1}}{p_1(1-p_2) + p_2(1-p_1)}, \qquad (1)$$

Bhatia and collaborators advise the use of Hudson's F_{ST} estimator as it is not sensitive to the ratio of sample sizes between pairs of populations and does not tend to overestimate F_{ST} estimations (Bhatia et al. 2013). As the index value is calculated for each individual marker, in order to combine the estimates across multiple SNPs and estimate the genome-wide F_{ST} value for each pair of populations I used the same approach as Weir and Cockerham (Bhatia et al. 2013; Weir and Cockerham 1984). In this approach, F_{ST} is calculated as a 'ratio of averages', meaning that the variance components (*i.e.*, the numerator and denominator of the estimator equation) are averaged separately across the multiple SNPs. However, this is an estimator of the theoretical index and it allows for negative values for closely related populations with a small number of individuals. As a negative genetic distance between two populations does not have any biological meaning negative values obtained using this method were transformed into zero. As further explained in more detail in Chapter 5 (and the reasoning behind it), I also measured F_{ST} by using Wright's ratio of variances (2) on that chapter (Wright 1949).

$$F_{ST}^{Wright} = \frac{var(p)}{p(1-p)}, \qquad (2)$$

The *stats* and *ape R* packages were used to build and visualize, respectively, a hierarchical tree with the pairwise F_{ST} values using the complete linkage method (Legendre and Legendre 2012; Paradis et al. 2004; R Core Team 2016).

2.4.3 Formal tests of admixture (three-population tests)

Given that *ADMIXTURE* is not a formal test of gene-flow between populations, I performed three-population admixture tests (*f3 statistics*) that are based on the concept that shared genetic drift between populations implies a shared evolutionary history (Reich et al. 2009). Briefly, in a *f3 statistic* with the form *f3(X; Pop A, Pop B)* a significantly negative value of the statistic (*Z-score* < −3) highlights a complex phylogeny of the target ('*X*') population, that has a certain amount of ancestry from populations related to *A* and *B*.

However, it is important to note that a positive *f3 test* value does not necessarily imply the absence of admixture. High levels of genetic drift in an admixed population can mask the signals of admixture and the corresponding *f3 test* might fail to be negative (Patterson et al. 2012). For this reason, and given the low number of markers used in the 'Mozambique Genetic Dataset' on Chapter 3, I performed a cross-validation analysis of *f3(X; Yoruba,* '*Source') tests* on 100 simulated samples. The simulated samples were composed both by Yoruba and either Ju/'hoansi, Sandawe, or Mbuti as '*Source'* in different proportions. These proportions, alpha and 1-alpha, respectively, ranged from 5% to 95% in increasing intervals of 5 percentage points.

All the *f3 tests* in this thesis were performed for windows of 100 markers using the *threepop* companion software in the *TreeMix* suite (Pickrell and Pritchard 2012).

2.4.4 Correlations

In order to test for correlations between the genetic, geographical, and linguistic distances between pairs of Bantu populations, the 'Bantu Genetic Dataset' and the 'Bantu Languages Genetic Dataset' were used (Chapter 5).

In a first stage, a Shapiro-Wilk test was performed in all distance matrices to test them for normality as this affects the correlation coefficient that should be used subsequently – a Pearson's correlation coefficient can be used in the case of a normal distribution from the data, while a Spearman's rank correlation coefficient should be used when a normal distribution is not observed.

A Mantel test was performed between all the distance matrices produced (genetic distances – F_{ST} –, linguistic distances, and geographical distances) with 10,000 permutations and using the *spearman method* though the *R* package *vegan* (Oksanen et al. 2017).

When comparing the correlation between the distances for the genetic correction attempted and for the different geographical models tested, a linear regression analysis was fitted to the distance data observed

and the coefficient of determination was calculated. The coefficient of determination – r^2 – explains the proportion of variation in one of the variables that is explained by the other variable analysed. This coefficient was then used to evaluate the fit of the correction/model tested to the data, assuming that when a higher proportion of the variation was explained the model tested should be closer to the actual conditions underwent by the Bantu-speaking populations in their spread through the African continent.

2.4.5 Correction to genetic distances

In chapter 5, to overcome the effects of recent admixture in Bantu-speaking populations, I made corrections to the allele frequencies used to calculate the genetic distance between pairs of populations. The basis for this correction was the ADMIXTURE analysis result obtained earlier, in which the admixture between populations can be easily observed.

Using the ADMIXTURE analysis results for *K*=9 (see Chapter 4), I developed a method to correct the genetic distances between pairs of populations. The ADMIXTURE software output comprises two files: i) a *Q-file* with the ancestry fractions for each individual in the dataset, and ii) a *P-file* with the allele frequencies of the inferred ancestral populations.

The ADMIXTURE plot presented on Chapter 4 is then based on the Q-file obtained through ADMIXTURE for *K*=9. Figure 2.3A shows a schematic representation of a similar plot, with 5 non-Niger-Congo sources (with modal components in different colours), 4 Niger-Congo sources (also

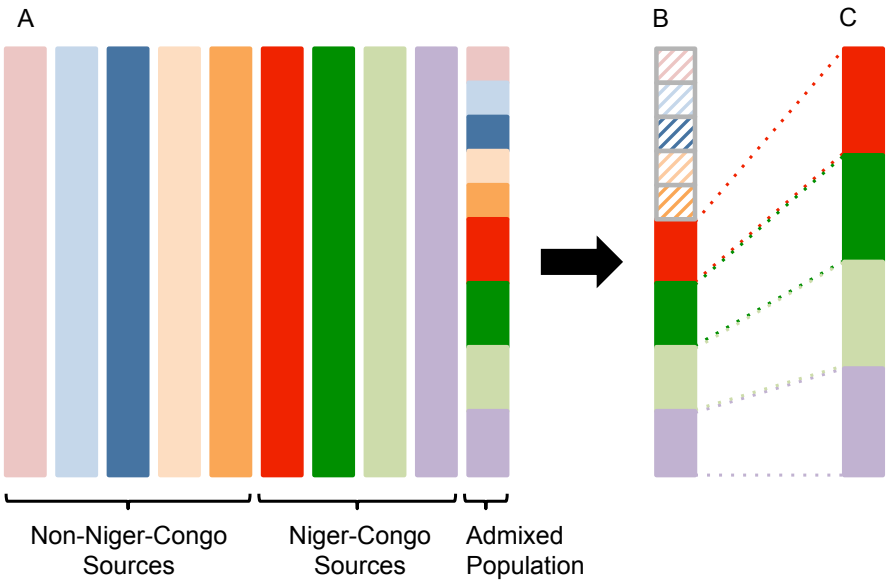


Figure 2.3 – Schematic representation of the first part of the method for correction of genetic distances performed in Chapter 5. (A) ADMIXTURE plot for *K*=9 with 5 non-Niger-Congo sources, 4 Niger-Congo sources, and 1 admixed population; (B) isolation of the Niger-Congo-specific components for the admixed population and (C) recalculation of the frequencies so that the total sum of the component retained is 1.

with modal components), and 1 putative admixed Bantu population. From the *Q-file* (and the plot) I isolate the Niger-Congo-specific components for each Bantu-speaking population (Figure 2.3B). In order to normalize the different contributions in each population I then recalculate these frequencies so that the total sum of the components retained is again 1 (Figure 2.3C). At this point I should have Bantu-speaking populations only comprised by the different Niger-Congo groups that originated them, having removed all non-Niger-Congo influences and reconstructed the ancestral 'unadmixed' population. In order to obtain the putative allelic frequencies in this original undamixed population I can then multiply the frequency of each ancestral component by the frequency of the each allele in the corresponding ancestral population (based on the information on the *P-file*).

Chapter 3

GENOME-WIDE SNP ANALYSIS OF SOUTHERN AFRICAN POPULATIONS PROVIDES NEW INSIGHTS INTO THE DISPERSAL OF BANTU-SPEAKING GROUPS

3.1 Introduction

The genetic structure of African populations is the result of both ancient and more recent episodes of migration and admixture (Hellenthal et al. 2014; Pickrell et al. 2012; Pickrell et al. 2014; Tishkoff et al. 2009). Among the historical events, the expansion of Bantu-speaking agropastoralist populations had a substantial impact on the continental distribution of genetic diversity. Bantu languages (part of the Niger-Congo linguistic group, NC) are thought to have originated near the present border between Cameroon and Nigeria approximately 5000 years before present (BP) (Diamond and Bellwood 2003; Newman 1995). From its place of origin, this branch of the NC linguistic phylum spread east and south across sub-Saharan Africa, together with agricultural techniques and the use of iron (Diamond and Bellwood 2003; Newman 1995). Today, Bantu is the largest African linguistic family both for territory extension and number of speakers, indicative of the impact this migration had on the continent (de Filippo et al. 2012). However, while there is a general consensus on the place and time of the origin of this movement, there is more debate about the actual routes used in this migration, with two main hypotheses having recently been tested. Briefly, these two main hypotheses differ on the time and place (north or south of the rainforest) of the split between East and West Bantu-speaking groups (Currie et al. 2013; de Filippo et al. 2012). Together with the route followed, another important aspect to be considered is the degree of interaction that populations involved in this expansion had with the groups

inhabiting the regions they were moving into. For example, the arrival of Bantu-speaking people in the southern regions of the continent could have led either to isolation or to admixture with the pastoralists and hunter-gatherers already present there (de Filippo et al. 2012; Destro-Bisol et al. 2004; Mitchell et al. 2008; Mitchell 2009; Patin et al. 2014).

Previous investigations have focused on the demographic dynamics of the Bantu dispersal (Barbieri et al. 2013b; Barbieri et al. 2014; Patin et al. 2014; Pickrell et al. 2012; Pickrell et al. 2014; Schlebusch et al. 2013). Within this context, the differentiation of a population sample from Mozambique from other African populations suggested the presence of a specific south-eastern non-Bantu component within the continent. This component, found at a proportion higher than 50%, has been interpreted as derived from an ancestral population (presumably related to hunter-gatherer populations) inhabiting the area before the arrival of expanding Bantu speakers (Sikora et al. 2011). The predominance of such a component in extant populations could be explained by the cultural shift of a foraging community (Figure 3.1A) or substantial gene flow into the arriving Bantu-speaking groups (Figure 3.1B). Alternatively, the reported Mozambique uniqueness could simply be the result of the dispersal process whose genetic signature would be expected to be shared with populations related to those living in Mozambique. This scenario might or might not include some degree of gene flow from hunter-gatherer populations (Figure 3.1C).

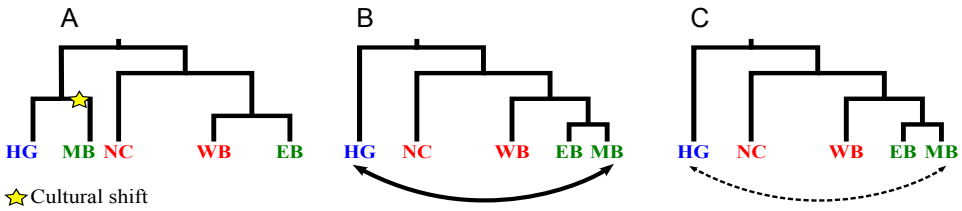


Figure 3.1 – Alternative scenarios for the Mozambique Bantu (MB). (A) Cultural shift after a split from a non-Bantu hunter-gatherer population (HG); (B) substantial gene flow into a Bantu-speaking group; (C) differentiation of an 'Eastern' Bantu component (EB, including the MB population), coupled or not with limited gene flow. NC, Niger-Congo; WB, Western-Bantu.

In order to test this hypothesis and further characterise the reported Mozambican component, I analysed genome-wide single nucleotide polymorphism (SNP) array data from 35 sub-Saharan African populations from 868 individuals (including 33 novel genotypes from four southern African populations from Namibia and Lesotho). This chapter's results place the so-called 'south-eastern component' within the broader genetic variation of southern Africa, rejecting the hypothesis that it reflects extensive gene flow between incoming agriculturalist and resident hunter-gatherer communities. Additionally, no evidence was found in the Mozambican population for substantial gene flow between Bantu and hunter-gatherer populations from this region.

3.2 Results

3.2.1 Principal Component Analysis

Relationships among individuals in the 'Mozambique Genetic Dataset' (Figure 3.2 – see Chapter 2, Methods) were explored using PCA (Figure 3.3). The first component (PC1) shows a separation of 'Khoisan'-speaking groups from the other populations, with east non-NC populations separating along the second component (PC2) (Figure 3.3A). In order to further analyse these results, a second PCA was performed removing the

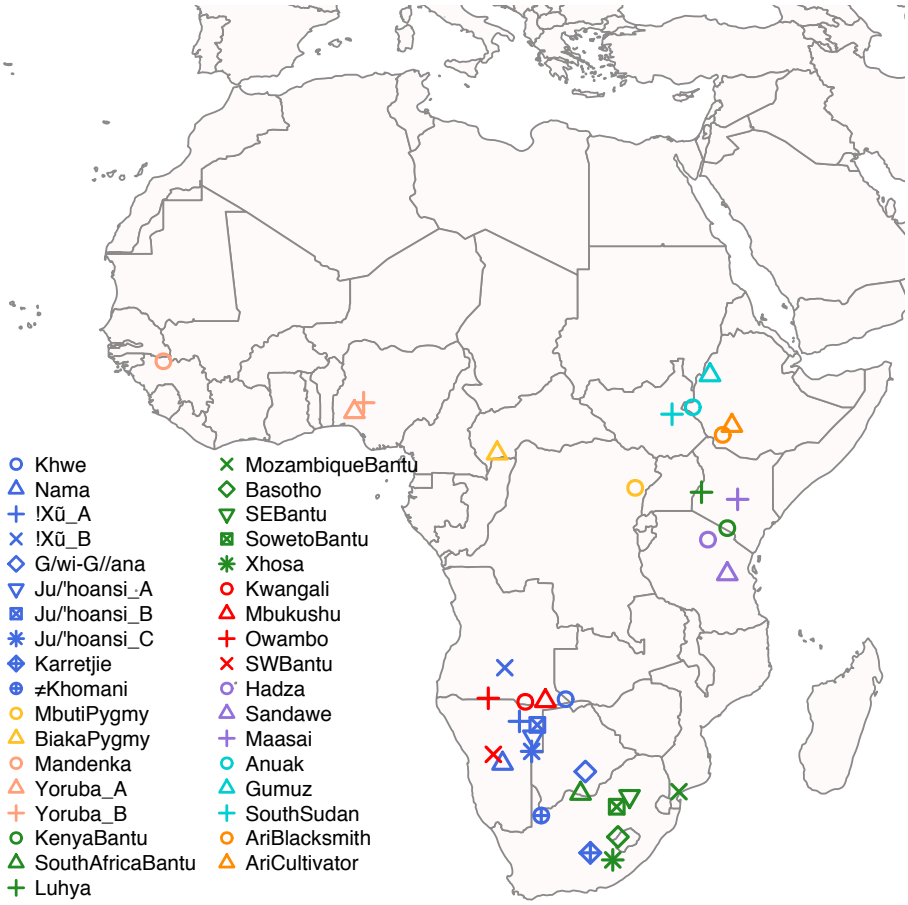


Figure 3.2 – Map of the populations analysed. The colours indicate different population groups: blue, 'Khoisan' speakers; yellow, Pygmies; salmon, non-Bantu Niger-Congo speakers; green, East-Bantu speakers; red, West-Bantu speakers; purple, language isolates; turquoise, Nilo-Saharan; and orange, Afroasiatic.

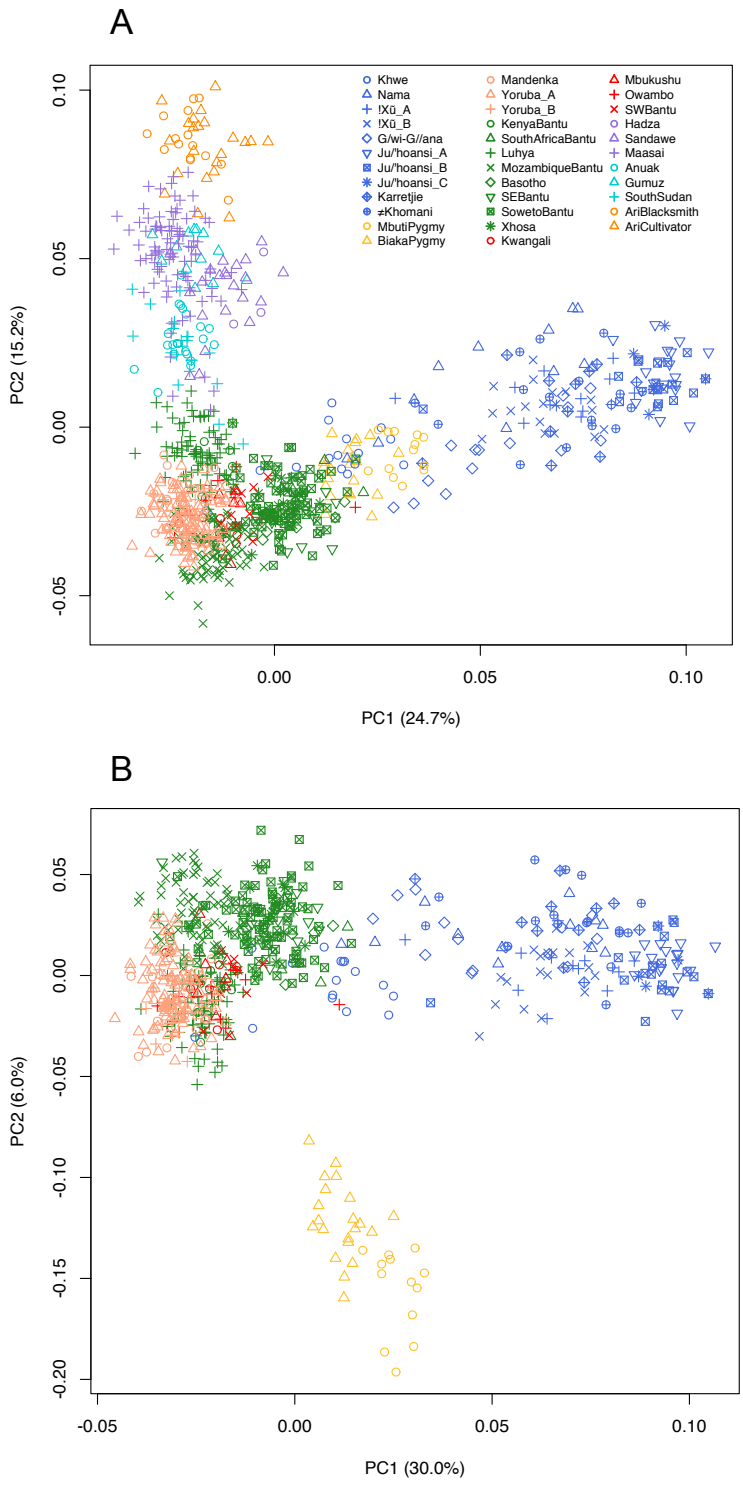


Figure 3.3 – Principal Components (PC) plots for the first and second components. (A) Full 'Mozambique Genetic Dataset'; (B) dataset excluding Afroasiatic, Nilo-Saharan, and language isolates. Symbols and colours as in Figure 3.2: blue, 'Khoisan' speakers; yellow, Pygmies; salmon, non-Bantu Niger-Congo speakers; green, East-Bantu speakers; red, West-Bantu speakers; purple, language isolates; turquoise, Nilo-Saharan; and orange, Afroasiatic.

east non-NC (Afro-Asiatic, Nilo-Saharan and linguistic isolates) populations. In this, the 'Khoisan'-speaking groups and the NC speaking populations are separated along the first component (PC1), the rainforest hunter-gatherers (RHG, a.k.a. Pygmies) located among the two (Figure 3.3B). RHG are separated from the other populations along the second component (PC2). The NC speakers are spread along the same axis, the Mandenka and Yoruba individuals on one side and various south-eastern Bantu speakers on the other (Figure 3.3B).

3.2.2 ADMIXTURE

I then performed an ADMIXTURE analysis, and obtained for *K*=6 the most suitable number of ancestral components (Supplementary Material 1). Five groups emerge, that I associate for simplicity to the linguistic group spoken by the majority of the populations characterised by the indicated component: 'Khoisan' (here considered as southern African populations speaking click-rich languages; Güldemann and Fehn 2014) (blue), Nilo-Saharan (East-1; purple) and Afro-Asiatic (East-2; orange) in east Africa, RHG (yellow), and NC speakers, the latter being characterised by two major components (red and green in Figure 3.4). These two components are present in all NC populations but their amount varies among groups. One of these components (red) dominates populations speaking West Bantu and other NC languages, while the other one (green) is more common among south-eastern Bantu speakers (Figure 3.4). The main component of the Mozambican sample (green) is similarly present in populations from Lesotho

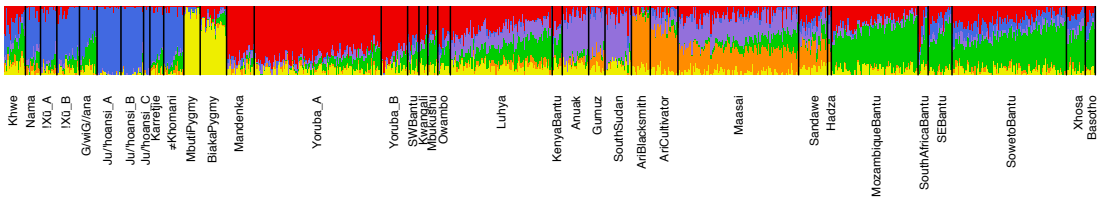


Figure 3.4 – ADMIXTURE plot for *K*=6.

and South Africa (even for higher values of *K*, Supplementary Material 2). Notably, the Basotho from Lesotho and the Xhosa/SE Bantu/SowetoBantu from South Africa also show a non-Bantu component ('Khoisan') but additional components ('Khoisan' or others) are virtually absent in samples from Mozambique. The spatial distribution of these components was visualised by interpolation maps (Figure 3.5A-F). The two main components present in NC populations mapped on to different geographic regions, so that these two can be broadly identified as NC-West and NC-Southeast (Figure 3.5E and F, respectively). The Afroasiatic and Nilo-Saharan components are mostly restricted to populations from central-east Africa (Figure 3.5B and C) and the RHG component was geographically restricted to Pygmy populations (Figure 3.5D).

3.2.3 F_{ST} components

The genetic distances among the six components were estimated via F_{ST}. The largest values are observed when the 'Khoisan' component is compared to the east Africa ones, followed by the comparisons of the first with both components from NC populations. The two NC components (West and East) are the most genetically close, their F_{ST} value (0.042) at least two times smaller than any other comparison (Table 3.1).

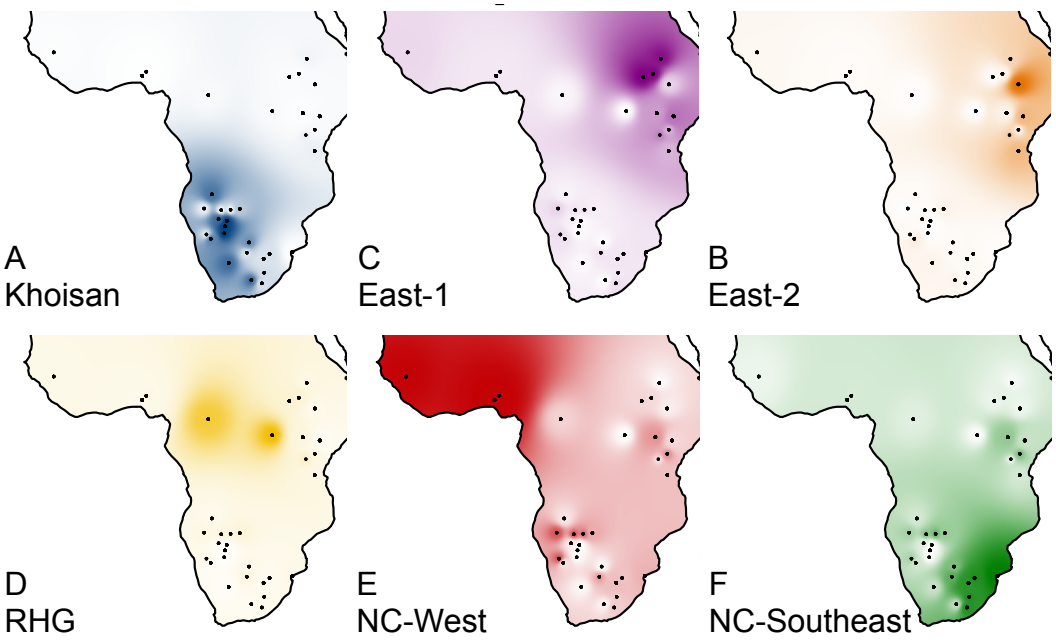


Figure 3.5 – Interpolation maps of the ADMIXTURE components identified in the 'Mozambique Genetic Dataset' for *K*=6. Colours of the components on the maps are the same as in the ADMIXTURE plot in Figure 3.4.

Table 3.1 – F_{ST} between ADMIXTURE components at *K*=6.

	NC-East	NC-West	East-1	East-2	Pygmies	Khoisan
NC-East	-					
NC-West	0.042	-				
East-1	0.099	0.081	-			
East-2	0.123	0.109	0.110	-		
Pygmies	0.097	0.095	0.132	0.137	-	
Khoisan	0.129	0.132	0.165	0.158	0.102	-

3.2.4 Three-population tests

Although STRUCTURE-like analyses are useful to identify patterns of gene flow between groups they are not a formal test of admixture since similar profiles can be the result of several different population histories. For this reason I carried out *f3* tests (Reich et al. 2009) in the 'Mozambique Genetic Dataset' for windows of 100 SNPs between all the possible trios of populations (Supplementary Material 3). As expected, this analysis identified

most of the known admixture events that characterized African populations, involving 'Khoisan' and Bantu populations in southern Africa. Interestingly, despite the low number of markers used, the analysis identified east African ancestry in 6 (out of 7) 'Khoisan' populations, as previously reported (Pickrell et al. 2012). On the other side, the analysis did not find signature of admixture in Karretjie, Owambo and Kwangali as significant values of the statistics were not observed. However, even though not significant, for these three populations negative values of *f3* statistics were observed 20, 16, and 38 times, respectively. No signature of admixture was found in Mozambique as no statistically significant values of *f3* statistics were observed. In addition, contrary to what has been reported for Karretjie, Owambo and Kwangali none of the *f3* involving Mozambique as target have non-significant negative values.

3.2.5 Three-population tests resampling

Given the low number of markers analysed, a cross-validation *f3* analysis was performed in 100 resampled samples in the form *f3(Target; Yoruba, Source)* composed by different ancestry proportions of Yoruba and *Source* (either Ju/'hoansi, Sandawe, or Mbuti), as described in Chapter 2. The *f3* statistics for all the simulated samples were always significant, with the exception of one sample for Yoruba 95%–Ju/'hoansi 5% and two for Yoruba 95%–Mbuti 5%. However it should be noted that even for these non-significant tests the Z-scores were always below −2 (Ju/'hoansi: −2.19; Mbuti: −2.98, and −2.90).

3.2.6 Hierarchical pairwise F_{ST} tree

The pairwise F_{ST} population tree shows a separation of the main groups present in the 'Mozambique Genetic Dataset', first with a group of Afro-Asiatic, Nilo-Saharan and linguistic isolates, then the 'Khoisan' speakers (except the Khwe, a highly admixed population as can be seen in Figure 3.4) and the NC speaking populations (Figure 3.6), with the RHG located between the last two. In line with what is observed for the ADMIXTURE components, the south-eastern populations grouped with the other NC populations, the internal branching order mirroring the West-East-Southeast linguistic subdivisions (Currie et al. 2013).

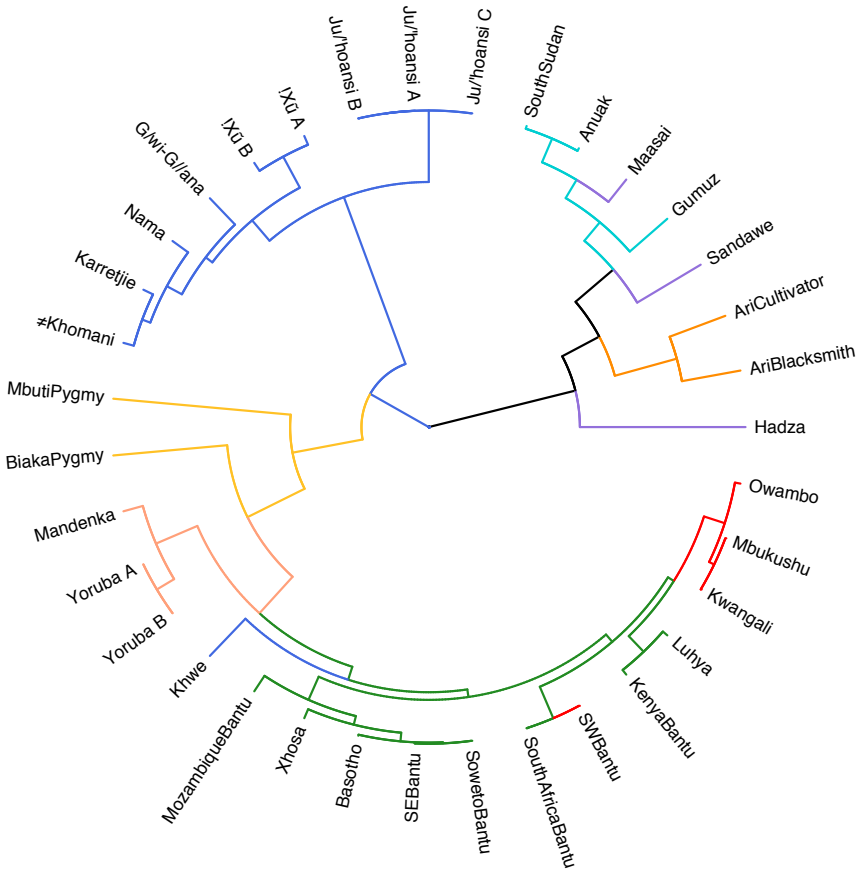


Figure 3.6 – Hierarchical population tree based on pairwise F_{ST} values. Colour of the branches corresponds to the colour of the populations in Figure 3.2.

3.3 Discussion

The spread of Bantu-speaking groups across southern Africa has significantly shaped the distribution of genetic variation of sub-Saharan Africa (Tishkoff et al. 2009). However, such dispersal did not occur in a void, as hunter-gatherers and pastoralists were already present in the areas that Bantu-speaking farmers were entering. The interactions between residents and newcomers were different in different areas, resulting in different degrees of integration and admixture between these groups (Barbieri et al. 2013a; Barbieri et al. 2014; Marks et al. 2015; Mitchell 2002; Patin et al. 2014; Pickrell et al. 2012; Pickrell et al. 2014). Archaeological and genetic data suggest that foraging groups were once present all across southern Africa, with a much larger distribution than the one currently observed (Barbieri et al. 2013b; Mitchell 2010). Their former presence over a wide geographic area has raised the possibility of the assimilation of these non-Bantu groups into farming communities via complex integration dynamics (Mitchell 2002), with the legacy of these events being recovered in contemporary Bantu-speaking populations (Barbieri et al. 2013a; Marks et al. 2015; Sikora et al. 2011). It is within this context that the previously reported 'south-eastern' African component had been interpreted: the genetic signature of a pre-Bantu community assimilated into the Bantu speakers of Mozambique (Sikora et al. 2011). Archaeological evidence supports the idea that the dispersal of Bantu speakers was relatively fast along the eastern part of Africa, which suggested reduced interactions

between the Bantu-speaking populations and the ones previously present in the occupied areas (J. A. Alexander 1984). The change in the pace of migration occurred much further south, around Lesotho and central South Africa, where the ecological conditions encountered were inappropriate for the crops used by these farming communities. The slowing of the process favoured more extended interaction between communities and possibly facilitated gene-flow (J. A. Alexander 1984; Bohannan and Plog 1967). The different admixture dynamics experienced by Bantu-speaking communities in different areas of south-eastern Africa resulted in different degree of assimilation of hunter-gatherers and pastoralist communities (Marks et al. 2015).

The previously reported differentiation of the Mozambique sample from other Bantu-speaking populations was originally interpreted as evidence for a pre-Bantu genetic component in south-eastern Africa (Sikora et al. 2011). When the Mozambican data was reanalysed here within a dataset comprising several southern African Bantu-speaking populations, the reported uniqueness of this population disappeared. In the second component of the PC analysis, Mozambique appears as part of a cluster of several sub-Saharan African populations speaking NC languages and distributed along a northeast-southwest continuum (Figure 3.3B). The ADMIXTURE components present in the Mozambican samples are the same as those of other east and south-eastern African populations from Kenya, South Africa and Lesotho (Figure 3.4). The major component in Mozambique is also the major component in other populations from the

south-eastern region (green in Figure 3.4), and is present at lower frequencies in other NC speakers. Importantly, the closer affinity of this component to the other NC component (NC-West), suggests a more recent common origin for the two NC clusters than any of the others. The pairwise F_{ST} population tree confirms the clustering of all NC-speaking populations, with the south-eastern Bantu-speakers placed within the variation of other Bantu-speaking populations (red and green branches in Figure 3.6). Overall, is there any support for the presence of a predominantly non-Bantu component in south-eastern Africa? The results presented in this chapter suggest some degree of stratification among Bantu-speaking populations which matches the West vs. East NC linguistic split (Currie et al. 2013). The differential distribution of the two ADMIXTURE components in the NC populations, coupled with the gradual increase of the green component when moving from populations living in the west, to those living in the east and the southeast, hints to a degree of differentiation among Bantu-speaking populations shaped by founder events and drift associated with their geographic dispersal rather than massive gene introgression from other groups (Ramachandran et al. 2005). No other components, known ('Khoisan' or RHG) or unknown, are detected within the Mozambican sample at significant level. Despite the lack of a major non-Bantu component in Mozambique, other groups in southern Africa are characterised by such kind of contribution, in this case 'Khoisan' (Figure 3.4) in agreement with previous reports of 'Khoisan' admixture in southern African Bantu-speaking populations (Petersen et al. 2013; Schlebusch et al.

2012). This is confirmed by *f3* analyses that highlight several admixture events involving African populations. However, no signature of admixture between 'Khoisan' populations (or other hunter-gatherer groups) and Bantu speakers in Mozambique was identified. In fact, none of the tests in the form *f3(Target; Pop A, Pop B)* gave significant or negative values. Furthermore, the cross-validation analysis based on resampled data shows that the probability of having a false negative result is very low. Even though the absence of gene flow into the Mozambican sample here analysed cannot be completely excluded given the low number of analysed markers, I show that a high contribution from non-Bantu populations is highly unlikely in this population. I also noted that this might not be the case for all populations from this area as the analysed sample is possibly not representative of the region as a whole. Sporadic highly divergent mtDNA and Y chromosome lineages have been reported in Zambia and Mozambique despite the lack of robust population signatures of admixture (Barbieri et al. 2013b; Barbieri et al. 2013a; Batini et al. 2011; Marks et al. 2015). Similarly, I cannot refute a scenario where admixture occurred in the past but left no signature in modern day populations. Future analysis of ancient DNA will be crucial to further refine the model discussed in this work. Unfortunately, the climatic conditions in sub-Saharan Africa make the extraction of endogenous DNA still challenging but new methods developed in the last few years are promising as recently showed (Morris et al. 2014; Sirak et al. 2015). The observation of some degree of differentiation between Western and Eastern-South-Eastern Bantu speakers also suggests that higher SNP

densities might provide the power to identify and distinguish populations belonging to the two Bantu linguistic branches, which could be of help in elucidating the demic pattern of Bantu speakers dispersal as well as reconstruct the African ancestry of recently admixed populations (de Filippo et al. 2012; Montinaro et al. 2015).

In conclusion, this chapter's results underline the role played by both the dispersal of Bantu-speaking groups and the interaction with previous inhabitants in shaping the genetic and cultural variation of Bantu speaking populations of southern Africa. Such a combined effect has been suggested for other regions of Africa where Bantu-speaking groups are present (Patin et al. 2014). The integration of archaeological and linguistic data with a more complex demographic model is necessary to better understand the process through which languages, people and technology were spread following the so-called 'Bantu expansion' (Marks et al. 2015).

Chapter 4

CHARACTERIZATION OF THE VARIATION IN BANTU-SPEAKING POPULATIONS

4.1 Introduction

Almost a third of the 1.2 billion people in Africa speak a Bantu language, by far the largest language group in the continent (de Maret 2013). Part of the Niger-Congo phylum, the predominance of languages belonging to this language family across such a large area of the continent has sparked a large debate over its common origin and the processes through which it dispersed over time and space. Moreover, the wide distribution and close relatedness of these languages suggest that the distribution that we observe today is the result of a relatively recent and rapid dispersal – the so-called 'Bantu expansion'.

Academic interest in the Bantu expansion started, unsurprisingly, with linguistic scholars proposing in the 19th century that the similarities between the languages resulted from a large-scale migration from central-west Africa (de Maret 2013; Johnston 1886). However, linguistics alone cannot fully explain how the current distribution of Bantu languages across sub-Saharan Africa came to happen, as its spread could be the result of an actual movement of people or of a cultural diffusion – or any combination of the two –, with very different implications for the genetic diversity and relationships of the people speaking these languages. Genetic studies had here a key role by complementing the linguistic and archaeological findings. An increasing amount of genetic data on sub-Saharan African populations shows a shared ancestry among Bantu speakers, indicating an actual

migration of people across the African continent into southern Africa (Busby et al. 2016; de Filippo et al. 2012; S. Li et al. 2014; Tishkoff et al. 2009).

In Chapter 3 I tested the apparently unique genetic variation in a Bantu-speaking population in relation to admixture with local inhabitants and I suggested that demography more than gene-flow shaped such diversity. In order to expand the findings of the previous chapter and investigate the patterns of genetic variation in sub-Saharan African Niger-Congo-speaking populations, in this chapter I analyse genome-wide single nucleotide polymorphism (SNP) array data organized in two partially overlapping datasets – one consisting of several different African populations, most of which speaking Niger-Congo languages, and a second one, a subset of the former, focusing just on Bantu-speaking populations. Using these two datasets, I initially investigate the distribution of the genetic variation among Niger-Congo populations and analyse the patterns and signatures of admixture between the different populations in the datasets. Finally, I analyse the relationships between these samples especially focusing on the behaviour of the Bantu-speaking populations present. Altogether, this chapter highlights how the genetic and cultural diversity of sub-Saharan Africa was affected by its settlement by Bantu speakers. In particular, the results show how the interactions with communities local to the regions settled resulted in some levels of structure among Bantu-speaking people.

4.2 Results

4.2.1 Datasets

I have initially focused my attention on the 'Niger-Congo Genetic Dataset' (see Chapter 2, Methods) aiming to describe patterns of genetic variation found in sub-Saharan African populations speaking languages belonging to the Niger-Congo family (Figure 4.1 and Table 4.1). This dataset consists mainly of Niger-Congo populations (Bantu and non-Bantu

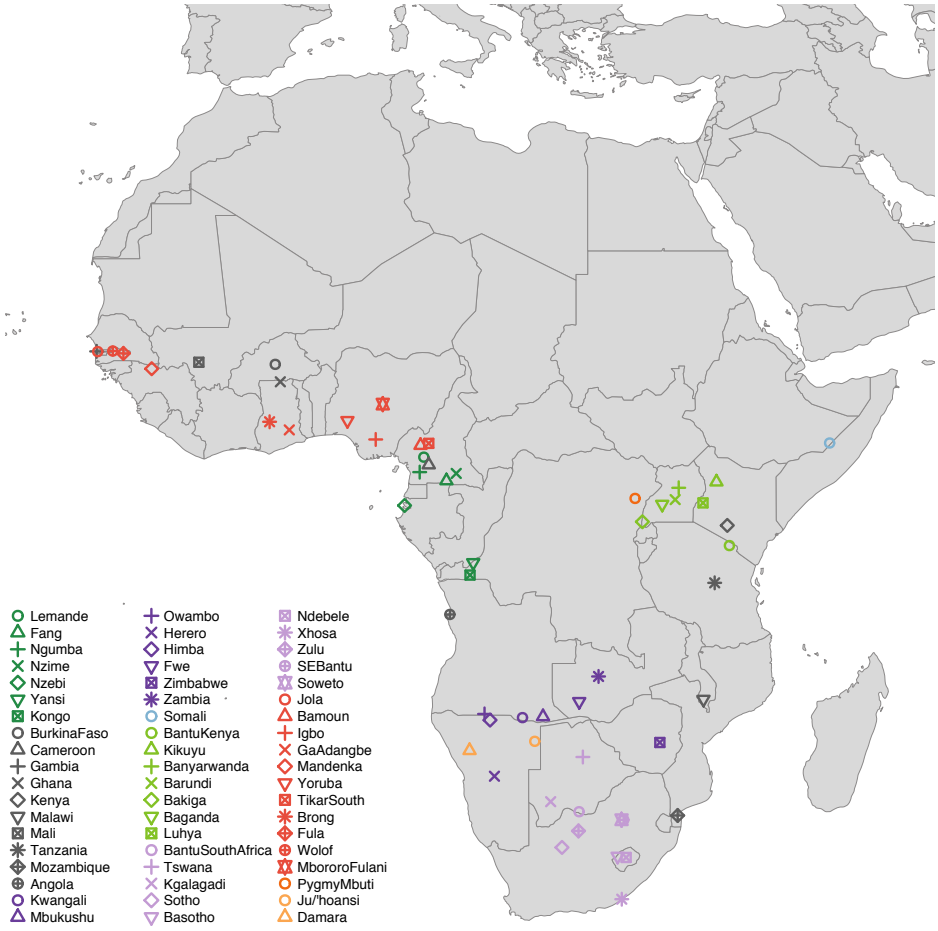


Figure 4.1 – Map of the samples included in the 'Niger-Congo Genetic Dataset'. Colours indicate different population groups: light orange, 'Khoisan'; dark orange, Pygmies; light blue, east African Somali; red, non-Bantu Niger-Congo; dark green, Bantu *cluster A*; light green, Bantu *cluster B*; dark purple, Bantu *cluster C*; light purple, Bantu *cluster D*; grey, populations with no linguistic information available.

Table 4.1 – Populations comprising each of the two datasets used on Chapter 4.

		Niger-Congo Genetic Dataset	*Bantu Genetic Dataset*
Non-Niger-Congo			
	CEU	x	
	Ju/'hoansi	x	
	Damara	x	
	PygmyMbuti	x	
	Somali	x	
Populations with no linguistic information available			
	Angola	x	
	Mozambique	x	
	Tanzania	x	
	Mali	x	
	Malawi	x	
	Kenya	x	
	Ghana	x	
	Gambia	x	
	Cameroon	x	
	BurkinaFaso	x	
Niger-Congo non-Narrow-Bantu			
	MbororoFulani	x	
	Wolof	x	
	Fula	x	
	Brong	x	
	TikarSouth	x	
	Yoruba	x	
	Mandenka	x	
	GaAdangbe	x	
	Igbo	x	
	Bamoun	x	
	Jola	x	
Niger-Congo Narrow Bantu			
Cluster A			
	Lemande	x	x
	Fang	x	x
	Ngumba	x	x
	Nzime	x	x
	Nzebi	x	x
	Yansi	x	x
	Kongo	x	x
Cluster B			
	Baganda	x	x
	Bakiga	x	x
	Banyarwanda	x	x
	Barundi	x	x
	Kikuyu	x	x
	Luhya	x	x
	BantuKenya	x	x
Cluster C			
	Herero	x	x
	Kwangali	x	x
	Owambo	x	x
	Fwe	x	x
	Himba	x	x
	Mbukushu	x	x
	Zambia	x	x
	Zimbabwe	x	x
Cluster D			
	Xhosa	x	x
	Basotho	x	x
	Kgalagadi	x	x
	Ndebele	x	x
	Tswana	x	x
	Zulu	x	x
	BantuSouthAfrica	x	x
	SEBantu	x	x
	Sotho	x	x
	Soweto	x	x

speakers) and populations for which no linguistic information is available but sampled in predominantly Niger-Congo-speaking countries. For clarity, the latter will be referenced as NLIA (No Linguistic Information Available) populations in this chapter. Additionally, I included a few reference populations to explore the relationships between potential sources of gene-flow from areas where Niger-Congo languages are spoken or outside Africa and Niger-Congo populations. In particular I included a European sample (CEU), an east African non-Niger-Congo population (a Cushitic-speaking population from Somalia), a rainforest hunter-gatherer (the Mbuti Pygmies), and two 'Khoisan'[9] populations: the Kx'a-speaking Ju/'hoansi, and the Khoe-speaking Damara. The Bantu populations in this dataset are then grouped in a new dataset, the 'Bantu Genetic Dataset' (Table 4.1). As explained before (see Chapter 2, Methods), the Bantu populations in the BG dataset here analysed are divided into different geographical clusters in Chapter 5 in order to implement different models of geographical dispersion. As this division helps to organize the samples in my dataset and allows a clearer description and interpretation of the results I decided to use them in this chapter as well. The Narrow Bantu populations present in the 'Niger-Congo Genetic Dataset' and the 'Bantu Genetic Dataset' here analysed are then divided in four geographical clusters: '*cluster A*' (west Africa), '*cluster B*'

[9] As mentioned before, the term 'Khoisan' is used in this thesis to refer to non-Bantu southern Africa populations sharing linguistic characteristics (click-rich non-Bantu languages) and does not necessarily imply genetic relationships (Güldemann and Fehn 2014).

(east Africa); '*cluster C*' (Namibia, Zambia, and Zimbabwe), and '*cluster D*' (rest of southern Africa – South Africa, Lesotho, and Botswana) (Table 4.1 and also highlighted with different colours in Figure 4.1).[10]

4.2.2 Principal Component Analysis

I initially analysed the 'Niger-Congo Genetic Dataset' using a genetic clustering method, the Principal Components Analysis (PCA). PCA revealed a 'F-shaped' distribution of the samples in the dataset (Figure 4.2). The first component (PC1, 42.53% of the total variation explained) separates Europe from all the African populations. Within Africa, the east African Somali population is located between the two groups (Figure 4.2A). Some of the Bantu populations from east Africa (shown in light green in Figure 4.2A, corresponding to *cluster B*) are spread on the first axis along the lower arm of the F shape in the direction of their geographical neighbours, the Cushitic Somali, and, further along the PC, Europe. Among these it is also possible to notice three of the Herero samples, indicating potential recent admixture in these individuals. Along the upper arm of the F shape, it is also noteworthy that both Fula (non-Bantu Niger-Congo) populations in the dataset (Fula and MbororoFulani) and a subset of the Gambia samples (as well as one individual from the Mandenka population) are also dispersed over the first component in the direction of Europe.

[10] See also Chapters 2 and 5 for more details on the geographical clustering of the Bantu-speaking populations in the BG dataset.

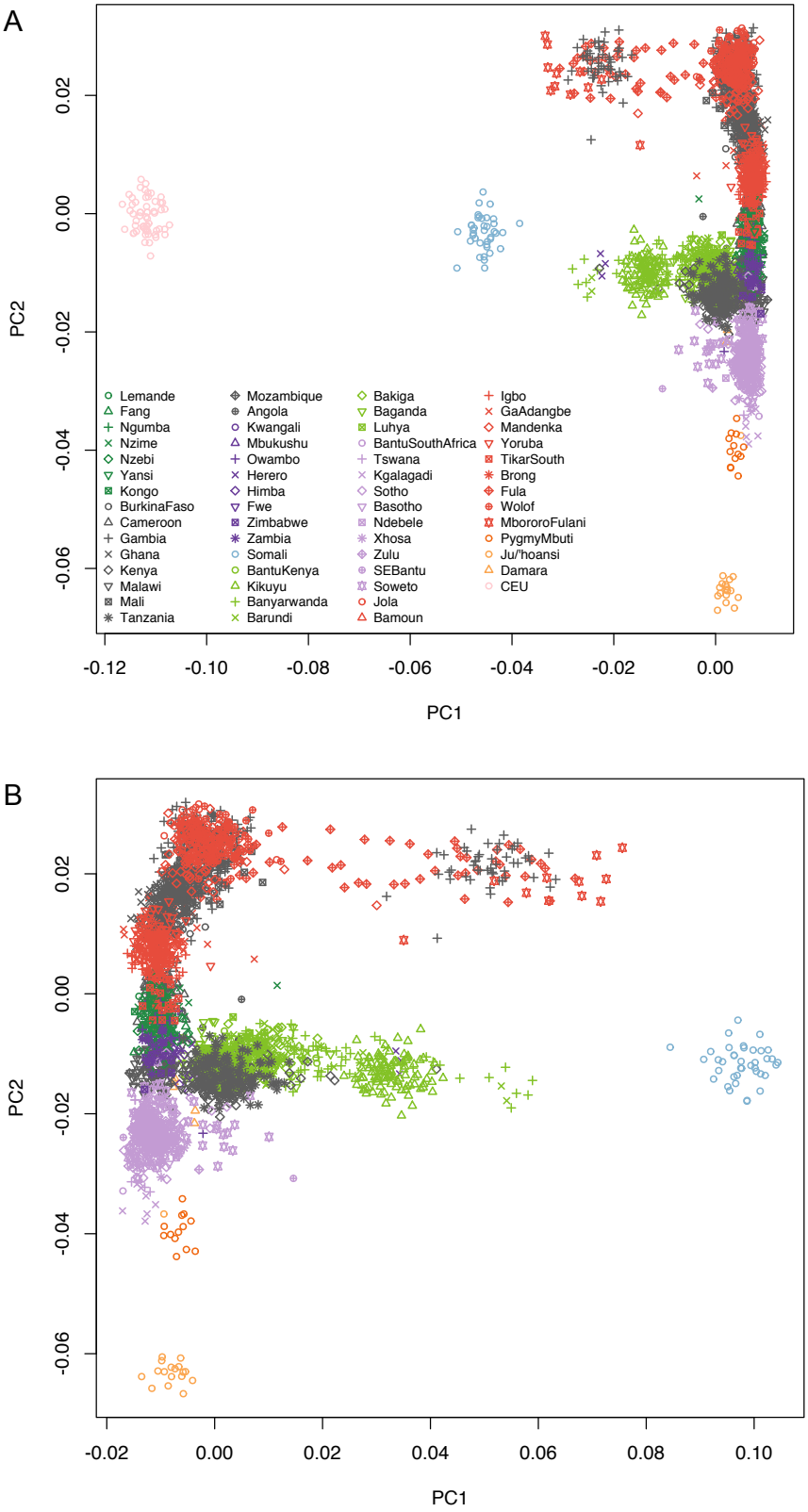


Figure 4.2 – Principal Component Analysis (PCA) plots of the 'Niger-Congo Genetic Dataset' for the first and second components. (A) Full dataset; (B) dataset excluding the European sample (CEU). Symbols and colours as in Figure 4.1: light orange, 'Khoisan'; dark orange, Pygmies; light blue, east African Somali; red, non-Bantu Niger-Congo; dark green, Bantu *cluster A*; light green, Bantu *cluster B*; dark purple, Bantu *cluster C*; light purple, Bantu *cluster D*; grey, populations with no linguistic information available; pink, Europe.

These two groups of Niger-Congo populations spreading along the two branches of the F shape are separated from each other on the second component.

I then removed the European population from the analysis for a better visualization of the African genetic variation and I reran the PC analysis. The F-shaped distribution of the samples continues to be present, in an even more pronounced way (Figure 4.2B). The first component (PC1, 20.78%) clearly shows the separation between non-Niger-Congo east Africa (Somali) and the other populations, with some of the non-Bantu Niger-Congo and the east-African Bantu populations spreading in the direction of the former, as seen before. The other non-Niger-Congo populations present in the dataset (with the exception of the Damara) – the Ju/'hoansi ('Khoisan') and the Mbuti Pygmies (rainforest hunter-gatherers) – separate from all the Niger-Congo populations on the second component (PC2, 19.21%). This separation is more pronounced for the Ju/'hoansi than for the Mbuti Pygmies, who are especially closer to Niger-Congo populations from southern Africa. The remaining non-Niger-Congo population in the dataset, the Damara, is shown to cluster within Niger-Congo populations in the analysis, and more specifically with Bantu populations, including their neighbour populations from Namibia, the Herero and the Himba (Supplementary Material 5N, 5U, and 5V). Niger-Congo populations also show a dispersion along the same second axis, with non-Bantu Niger-Congo on one side, south Bantu on the other, and east and west Bantu in between. All the NLIA populations used in this analysis (represented in grey

in Figure 4.2) clustered among other Niger-Congo populations in the dataset.

I then analysed the 'Bantu Genetic Dataset' (see Chapter 2, Methods), which focuses on Bantu-speaking populations (Figure 4.3 and Table 4.1). The Principal Component Analysis of this dataset shows the southeast African Bantu located on one side of the axis of the first principal component and the east African on the other, with the central-west and south-west Bantu populations between the two (Figure 4.4). Both the east and the South-East populations are also spread along the axis of the

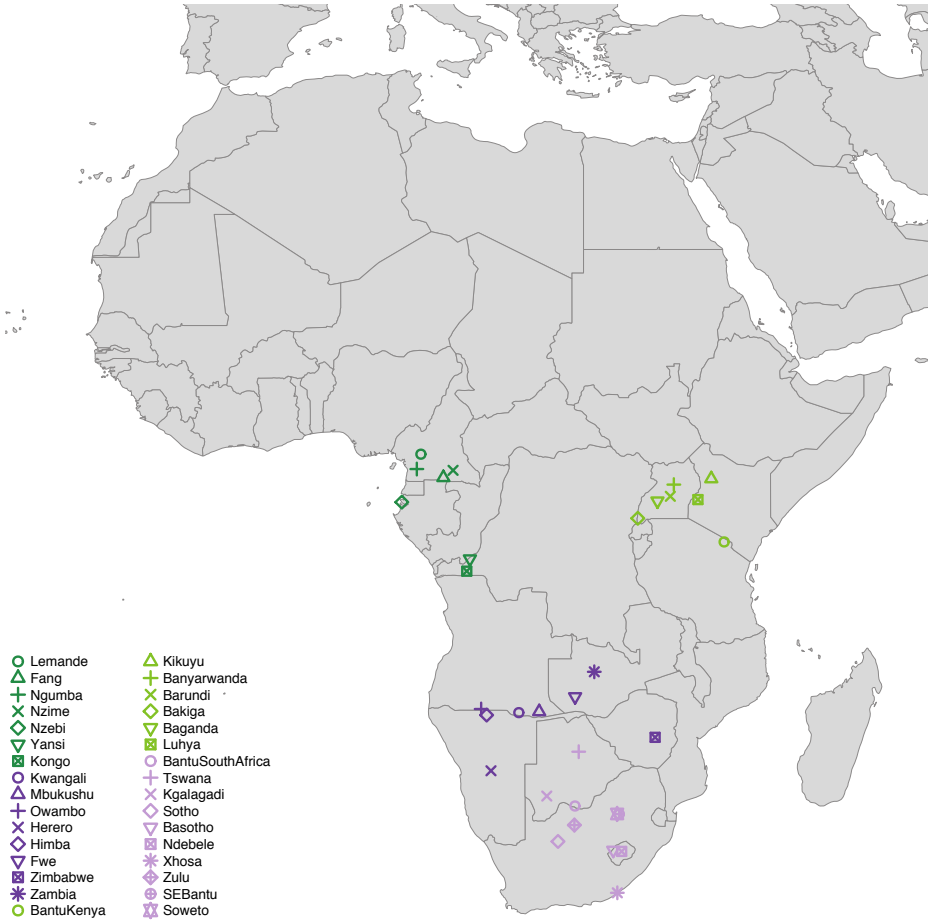


Figure 4.3 – Map of the samples included in the 'Bantu Genetic Dataset'. Colours indicate the four geographical clusters identified in Bantu populations (see Chapter 2): dark green, Bantu *cluster A*; light green, Bantu *cluster B*; dark purple, Bantu *cluster C*; light purple, Bantu *cluster D*.

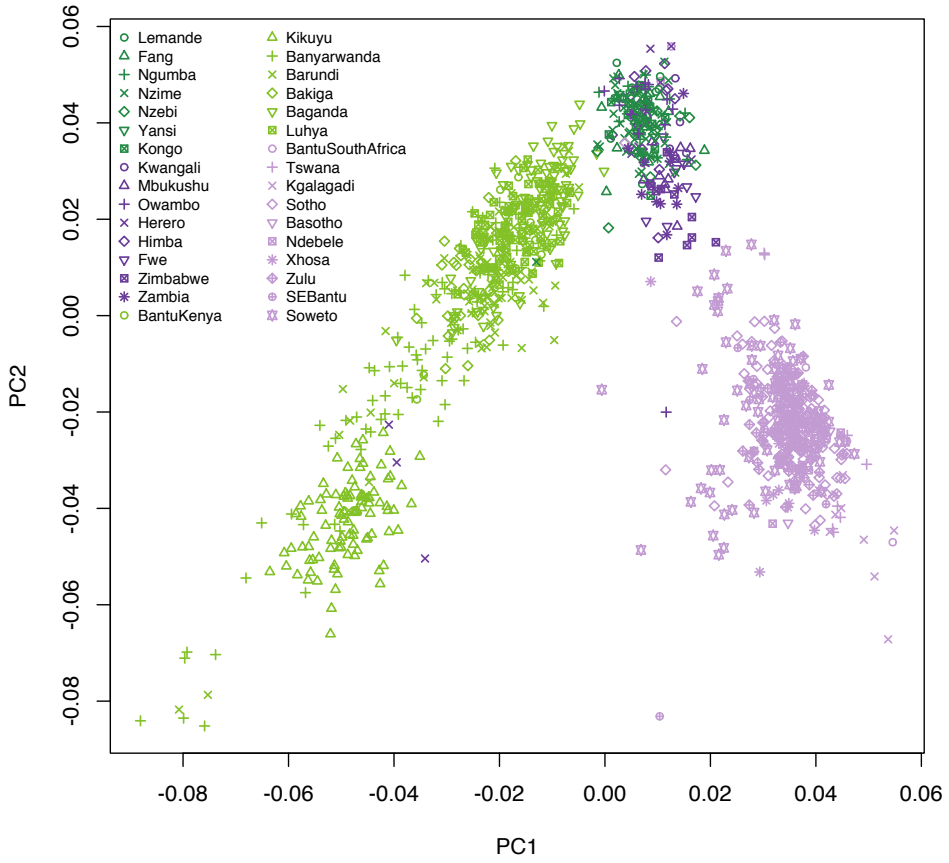


Figure 4.4 – Principal Component Analysis (PCA) plots of the 'Bantu Genetic Dataset' for the first and second components. Symbols and colours as in Figure 4.3: dark green, Bantu *cluster A*; light green, Bantu *cluster B*; dark purple, Bantu *cluster C*; light purple, Bantu *cluster D*.

second component (PC2), showing a further separation from the west and southwest populations along this axis. Even though populations follow this general pattern some exceptions can be seen at individual level (Supplementary Material 5 for the PCA plots). Individuals in the outermost range of the east African variation in the plot show a dispersal towards Europe/non-Bantu east Africa, which might be indication of a higher degree of admixture with these populations – this can be seen not only for some of the east African Bantu individuals (as is the case for some Barundi and Banyarwanda) but also for some southwestern Bantu individuals, as can be clearly seen for three Herero individuals and one Nzime that are all clearly

separated from the rest of the samples in their population. On the other hand, a symmetrical exception can also be seen for one Owambo individual that cluster instead closer to the southeast Bantu samples. This seems to be driving this individual closer to the Ju/'hoansi, which can indicate admixture with non-Bantu populations. Also noteworthy is one individual from the SEBantu population that seems to follow both these exceptions, being both drawn closer to the central-east populations (and, consequentially, Europe and Somali) and also closer to Ju/'hoansi and away from the edge of the southeast variation present in the PC analysis.

4.2.3 ADMIXTURE

An ADMIXTURE analysis was performed on the 'Niger-Congo Genetic Dataset'. The cross-validation (CV) method implemented in the software ADMIXTURE uses the consistency across different runs of the algorithm for the same number of ancestral populations (*K*), with the assumption that a well-supported value of *K* should have a lower CV error when compared to the others. In the case of the 'Niger-Congo Genetic Dataset', the lowest CV error was observed for *K*=8 (Figure 4.5). However, *K*=9 also showed a relatively low CV error. Analysing the results for both numbers of ancestral populations, the main difference between the results for these two values of *K* is that when considering nine ancestral populations (*K*=9) the Pygmy population is composed by a single RHG component where two different components (one 'Khoisan' and another observed in different proportions across multiple Bantu populations) were

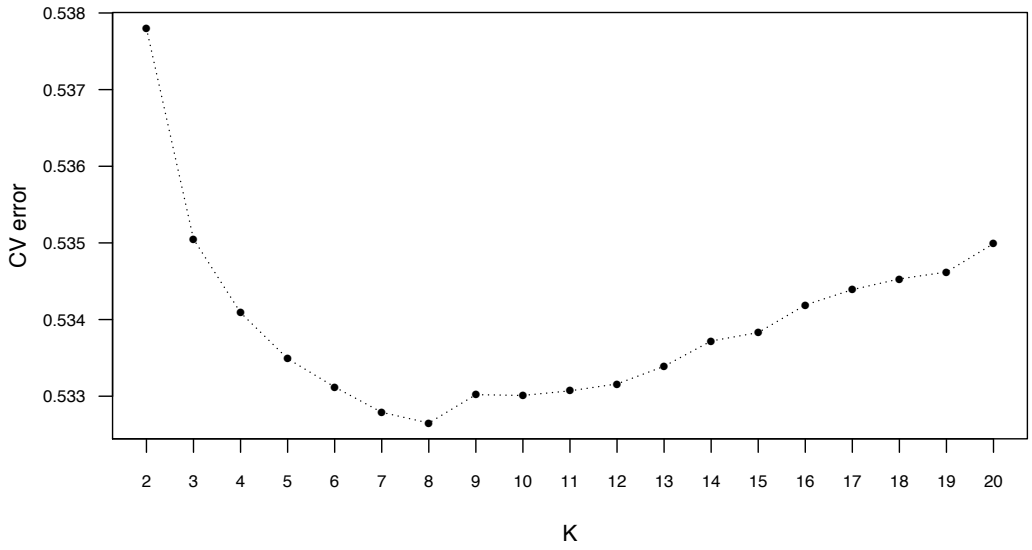


Figure 4.5 – Cross-validation (CV) plot for the ADMIXTURE analysis on the 'Niger-Congo Genetic Dataset'.

observed for *K*=8 (Figure 4.6). At *K*=9, components strongly associated with both language and geography can be identified: one each for Europe (pink), 'Khoisan' (light orange), RHG (dark orange), and east Africa (light blue); and five components mostly present in Niger-Congo populations (Figure 4.6 and Table 4.2). The five components found in Niger-Congo populations are non-randomly distributed and they too appear to be strongly associated with different groups within this family. Three of these five components seem more associated with west non-Bantu Niger-Congo populations: (*i*) the MbororoFulani sample from Nigeria is characterized almost exclusively (>80%) by a single component (represented in dark blue in Figure 4.6), which is also found in a significant proportion (~30%) only in the Fula sample from Gambia and in smaller quantity (12.5%) in a population sampled in the same country (coded here as Gambia); (*ii*) a component

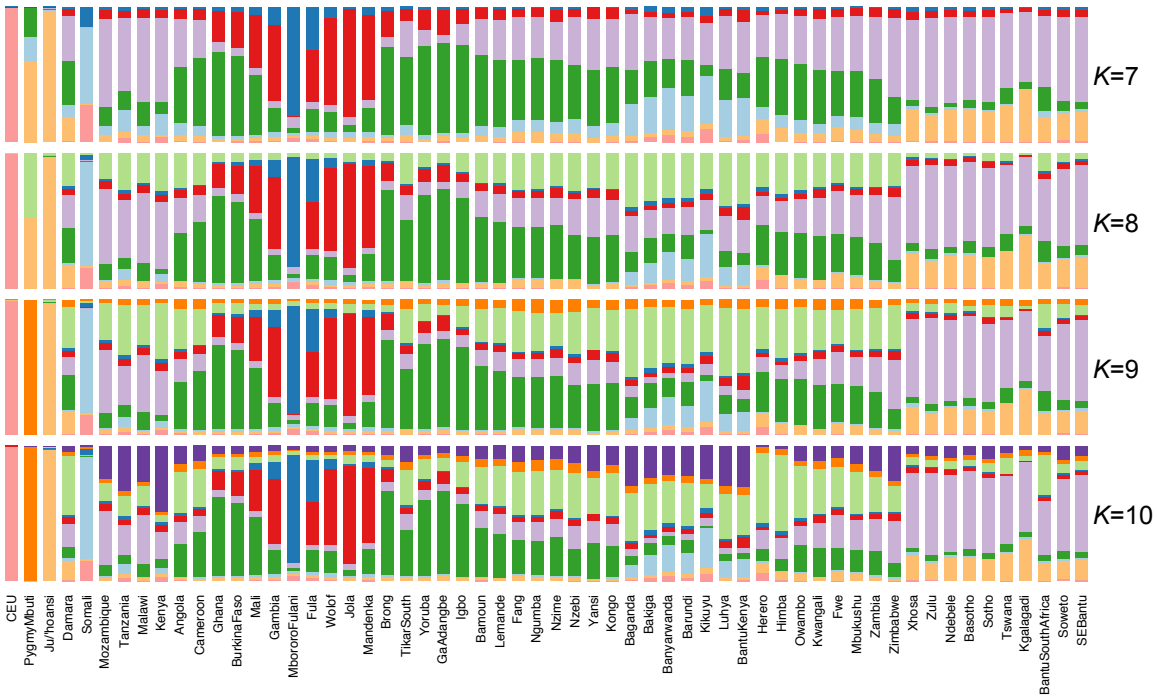


Figure 4.6 – 'Niger-Congo Genetic Dataset' Population ADMIXTURE plots from *K*=7 to *K*=10.

mainly found in western Niger-Congo populations (in red in Figure 4.6), which is not predominantly seen in any Bantu populations but is present in several non-Bantu Niger-Congo; and (*iii*) one other component found in populations closer to central-west Africa (higher amounts observed in Nigeria (Yoruba and Igbo), Burkina Faso, and Ghana (Ghana, Brong, and GaAdangbe) – dark green in Figure 4.6). This is also seen in different amounts across several Bantu populations, *cluster A* being the one showing the highest proportions (Figure 4.6 and Table 4.2). The Bantu populations on the dataset seem to be characterized by three different components observed in different amounts and combinations in different populations, mainly according to their geographic location (Figure 4.6 and 4.7). One of these components is also found in non-Bantu Niger-Congo populations from central-west Africa (represented in dark green in the plot); the other two

Table 4.2 – Median value per population of each ADMIXTURE component (at K=9) identified on the 'Niger-Congo Genetic Dataset'.

	Europe	Somali	Ju/'hoansi	Pygmy	NC Fulani	NC West	NC Yoruba	NC Bantu East	NC Bantu South	Non Niger-Congo	Niger-Congo
CEU	0.995	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.998	0.000
Somali	0.151	0.771	0.011	0.010	0.034	0.003	0.003	0.010	0.008	0.943	0.058
PygmyMbuti	0.000	0.000	0.002	0.997	0.000	0.000	0.000	0.000	0.001	0.999	0.001
MbororoFulani	0.045	0.022	0.014	0.012	0.805	0.003	0.029	0.032	0.039	0.093	0.908
Ju/'hoansi	0.000	0.000	0.973	0.003	0.001	0.001	0.007	0.010	0.005	0.976	0.024
Damara	0.008	0.016	0.162	0.051	0.018	0.050	0.256	0.306	0.132	0.237	0.762
BurkinaFaso	0.006	0.024	0.015	0.027	0.022	0.197	0.584	0.075	0.050	0.072	0.928
Cameroon	0.005	0.011	0.023	0.068	0.014	0.057	0.438	0.263	0.120	0.107	0.892
Gambia	0.013	0.024	0.020	0.030	0.125	0.513	0.180	0.048	0.047	0.087	0.913
Ghana	0.005	0.020	0.020	0.026	0.016	0.161	0.620	0.072	0.060	0.071	0.929
Kenya	0.037	0.046	0.020	0.026	0.027	0.062	0.044	0.411	0.326	0.129	0.870
Malawi	0.003	0.008	0.029	0.041	0.016	0.057	0.128	0.295	0.422	0.081	0.918
Mali	0.007	0.022	0.021	0.036	0.046	0.330	0.444	0.041	0.054	0.086	0.915
Tanzania	0.027	0.076	0.033	0.041	0.026	0.054	0.087	0.368	0.288	0.177	0.823
Mozambique	0.006	0.012	0.034	0.028	0.019	0.058	0.114	0.271	0.459	0.080	0.921
Angola	0.013	0.011	0.021	0.070	0.021	0.054	0.350	0.294	0.166	0.115	0.885
Jola	0.006	0.018	0.022	0.030	0.014	0.752	0.047	0.065	0.046	0.076	0.924
Bamoun	0.008	0.012	0.020	0.071	0.006	0.053	0.469	0.241	0.119	0.111	0.888
Igbo	0.003	0.013	0.023	0.049	0.012	0.050	0.614	0.153	0.082	0.088	0.911
GaAdangbe	0.006	0.020	0.034	0.039	0.014	0.108	0.653	0.069	0.056	0.099	0.900
Mandenka	0.009	0.025	0.024	0.029	0.048	0.580	0.183	0.049	0.052	0.087	0.912
Yoruba	0.004	0.015	0.022	0.033	0.014	0.081	0.635	0.120	0.076	0.074	0.926
TikarSouth	0.004	0.015	0.030	0.068	0.020	0.055	0.439	0.253	0.115	0.117	0.882
Brong	0.006	0.022	0.023	0.038	0.014	0.116	0.651	0.055	0.074	0.089	0.910
Fula	0.027	0.026	0.019	0.026	0.316	0.337	0.164	0.042	0.043	0.098	0.902
Wolof	0.011	0.036	0.017	0.025	0.073	0.592	0.159	0.038	0.049	0.089	0.911
Lemande	0.004	0.007	0.027	0.059	0.015	0.055	0.438	0.259	0.136	0.097	0.903
Fang	0.004	0.011	0.044	0.085	0.014	0.044	0.374	0.298	0.126	0.144	0.856
Ngumba	0.003	0.010	0.038	0.096	0.013	0.044	0.378	0.286	0.132	0.147	0.853
Nzime	0.003	0.013	0.032	0.095	0.020	0.062	0.392	0.249	0.134	0.143	0.857
Nzebi	0.001	0.013	0.036	0.090	0.016	0.048	0.317	0.312	0.167	0.140	0.860
Yansi	0.001	0.008	0.019	0.058	0.003	0.068	0.351	0.313	0.179	0.086	0.914
Kongo	0.002	0.010	0.024	0.074	0.010	0.065	0.345	0.308	0.162	0.110	0.890
BantuKenya	0.009	0.134	0.026	0.067	0.019	0.105	0.098	0.477	0.064	0.236	0.763
Kikuyu	0.060	0.304	0.039	0.037	0.032	0.055	0.038	0.348	0.085	0.440	0.558
Banyarwanda	0.039	0.202	0.038	0.051	0.031	0.041	0.140	0.415	0.044	0.330	0.671
Barundi	0.026	0.153	0.035	0.066	0.027	0.040	0.183	0.407	0.063	0.280	0.720
Bakiga	0.025	0.145	0.029	0.056	0.038	0.035	0.187	0.444	0.042	0.255	0.746
Baganda	0.013	0.089	0.024	0.067	0.024	0.047	0.157	0.500	0.078	0.193	0.806
Luhya	0.007	0.127	0.026	0.064	0.015	0.064	0.134	0.498	0.066	0.224	0.777
Kwangali	0.004	0.002	0.041	0.074	0.014	0.059	0.325	0.295	0.186	0.121	0.879
Mbukushu	0.004	0.012	0.067	0.069	0.014	0.039	0.281	0.298	0.217	0.152	0.849
Fwe	0.008	0.017	0.075	0.075	0.012	0.044	0.275	0.255	0.240	0.175	0.826
Owambo	0.002	0.015	0.061	0.055	0.024	0.037	0.336	0.336	0.135	0.133	0.868
Herero	0.062	0.012	0.098	0.054	0.035	0.026	0.298	0.308	0.108	0.226	0.775
Himba	0.006	0.004	0.084	0.070	0.035	0.018	0.321	0.356	0.107	0.164	0.837
Zambia	0.004	0.010	0.039	0.059	0.011	0.048	0.265	0.306	0.259	0.112	0.889
Zimbabwe	0.003	0.011	0.027	0.042	0.013	0.066	0.155	0.325	0.359	0.083	0.918
BantuSouthAfrica	0.002	0.008	0.165	0.030	0.020	0.039	0.154	0.181	0.402	0.205	0.796
Tswana	0.001	0.005	0.230	0.046	0.007	0.014	0.108	0.080	0.508	0.282	0.717
Kgalagadi	0.000	0.010	0.339	0.043	0.000	0.014	0.047	0.024	0.523	0.392	0.608
Basotho	0.005	0.015	0.191	0.044	0.012	0.028	0.064	0.040	0.601	0.255	0.745
Ndebele	0.005	0.021	0.190	0.028	0.012	0.038	0.037	0.073	0.595	0.244	0.755
Xhosa	0.010	0.024	0.201	0.027	0.018	0.044	0.048	0.056	0.573	0.262	0.739
Zulu	0.007	0.016	0.153	0.033	0.012	0.037	0.055	0.058	0.631	0.209	0.793
SEBantu	0.009	0.019	0.170	0.040	0.008	0.033	0.057	0.071	0.594	0.238	0.763
Soweto	0.015	0.014	0.159	0.034	0.012	0.034	0.073	0.097	0.563	0.222	0.779
Sotho	0.006	0.009	0.173	0.039	0.011	0.040	0.058	0.086	0.577	0.227	0.772

components are found in significant amounts almost exclusively in Bantu populations (represented in light green and purple in the ADMIXTURE plot – Figure 4.6 and Table 4.2). Interestingly, and in accordance with what was previously observed for the PC analysis, the only exception about the

Bantu-specificity of the components are the Damara[11]. This population shows an admixture profile typically observed in other Bantu-speaking populations, with almost 70% of this population profile belonging to one of the three 'Bantu components' and only 16% being from the 'Khoisan component'.

Non-Niger-Congo components, especially the 'Khoisan component', are also observed in considerable amounts in several of the Bantu populations, more noticeably on southern populations from *cluster D* (where all the populations show more than 10% of this component – Table 4.2). The marked spatial distribution of the different ADMIXTURE components in the Niger-Congo populations can easily be observed through their interpolation in maps of the African continent. I used these maps to name the components according to their location and frequency in the different ethnic groups included (Figure 4.7). As mentioned above, of the five components mainly found in Niger-Congo populations it is noticeable that two of the components are almost exclusively confined to non-Bantu populations from west Africa (components represented in red and dark blue in the maps, *NC West* and *NC Fulani* components, respectively). The remaining three components have a clear geographical distribution along the African continent. The *NC Yoruba* component shows a decrease in its proportion in populations from the northwest to the east (dark green in

[11] Not taking into account the NLIA populations in the dataset as these where sampled from majorly Bantu-speaking countries and, thus, might be expected to behave similarly to other Bantu-speaking populations.

Figure 4.7). The *NC Bantu East* component increases towards east Africa and in populations from the Bantu *cluster B* in the dataset (light green in Figure 4.7). Finally a component that has its highest proportion in populations from Bantu *cluster D*, in the southern regions of the continent (*NC Bantu South* component, represented in purple). The remaining four components identified in the analysis are modal to populations used as non-Niger-Congo references in my dataset: *Europe* component (pink), *Somali* component (non-Niger-Congo east Africa, represented in light blue), *Jul'hoansi* component (light orange), and *Pygmy* component (dark orange).

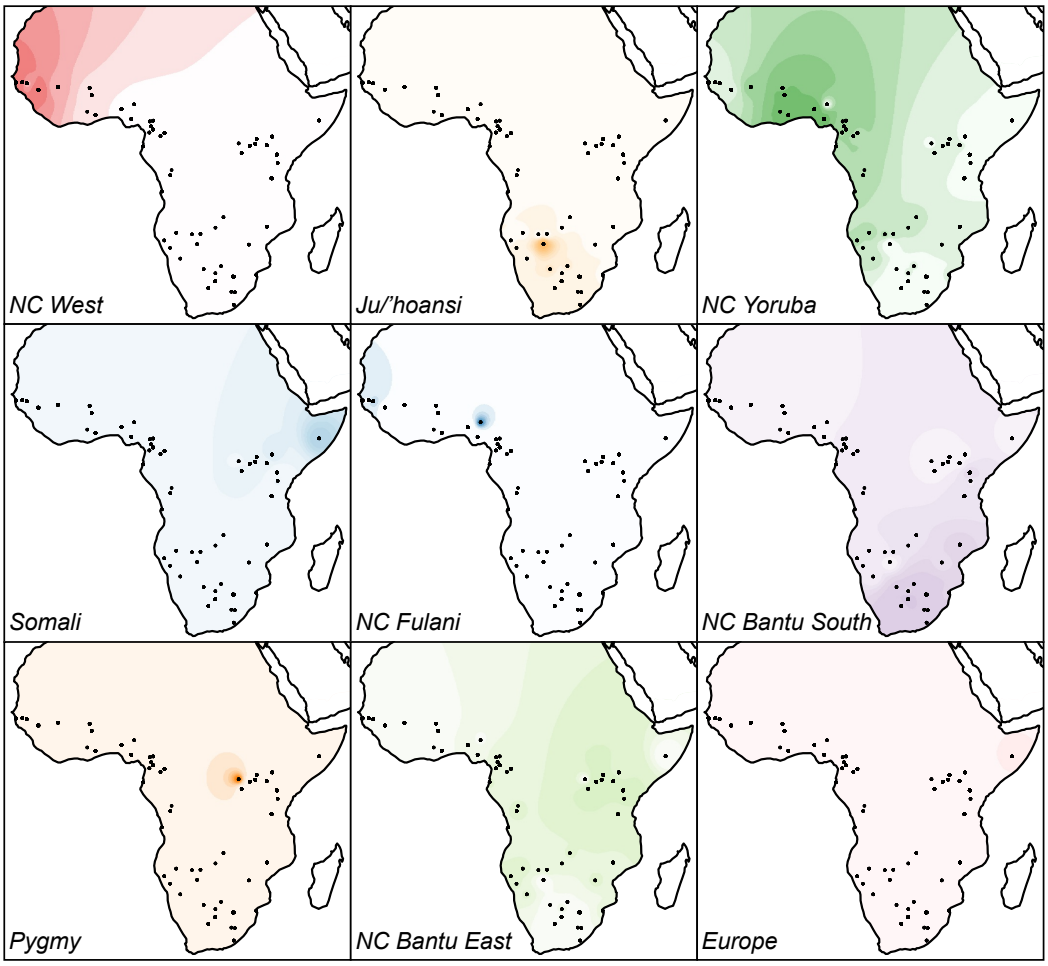


Figure 4.7 – Interpolation maps of the ADMIXTURE components identified on the 'Niger-Congo Genetic Dataset' (for *K*=9). Colours of the components on the maps are the same as in the K=9 ADMIXTURE plot in Figure 4.6.

4.2.4 Three-population tests (*f3*)

As described, non-Niger-Congo components were identified in different amounts in several of the Niger-Congo populations. In order to formally test for admixture with non-Niger-Congo groups in the history of Niger-Congo populations I performed the three-population tests with the format *f3(X; Yoruba,* '*Source*'*)* – with '*X*' being the population for which the admixture is going to be tested, and '*Source*' being either Europe (CEU), Somalia, Ju/'hoansi, or Mbuti RHG as examples of sources for non-Niger-Congo (Reich et al. 2009). Significant *f3* statistics (*Z-score* < −3) were observed for several of the trios of populations tested, which allowed the identification of several events of admixture (Figure 4.8, 4.9, 4.10, and 4.11).

4.2.4.1 Somalia

Statistically significant admixture events with non-Niger-Congo east Africa (Somali) were mainly found in neighbouring populations from central-east Africa (from Kenya, Uganda, and Tanzania) (Figure 4.8). On the other hand, two populations from west Africa (Gambia and Fula, both from the same country) also showed significant signals of admixture with the Somali sample. Notably, all of these populations (but not exclusively) also showed significant signatures of admixture with Europe.

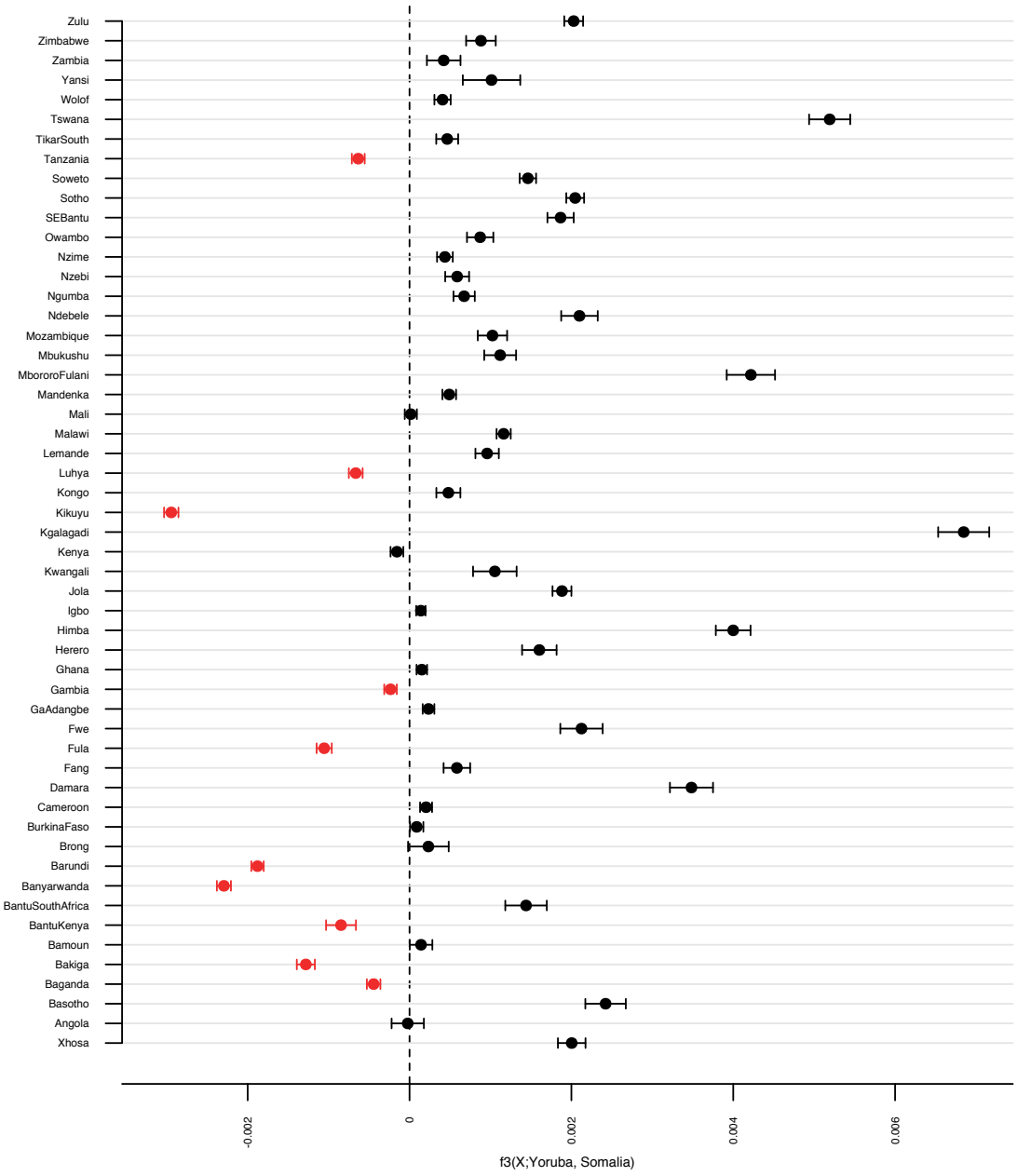


Figure 4.8 – Three-population (*f3*) test results with the format *f3*(X; Yoruba, Somalia) for the 'Niger-Congo Genetic Dataset'. Significant comparisons (Z-score < −3) are identified in red.

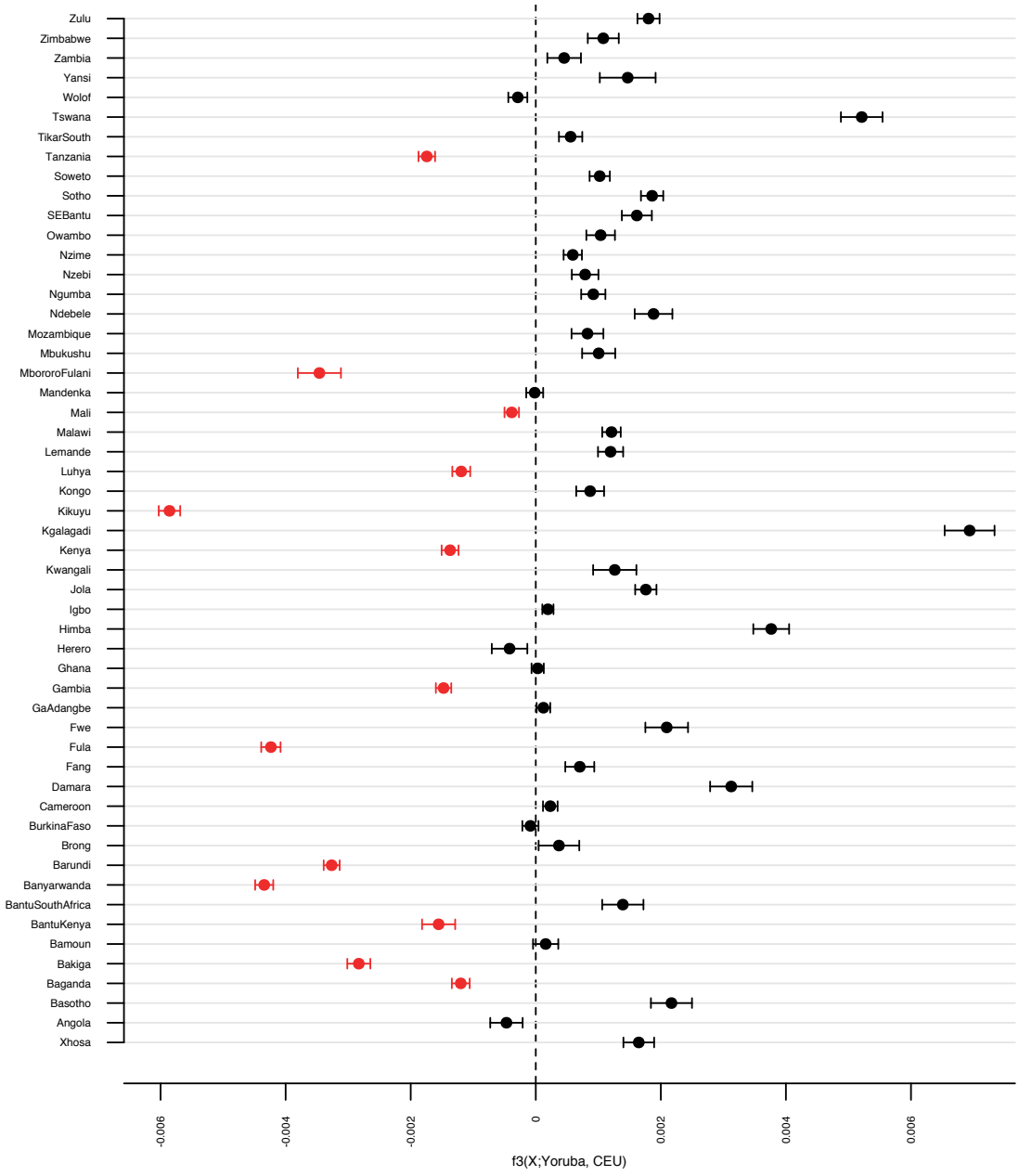


Figure 4.9 – Three-population (*f3*) test results with the format *f3*(X; Yoruba, CEU) for the 'Niger-Congo Genetic Dataset'. Significant comparisons (Z-score < −3) are identified in red.

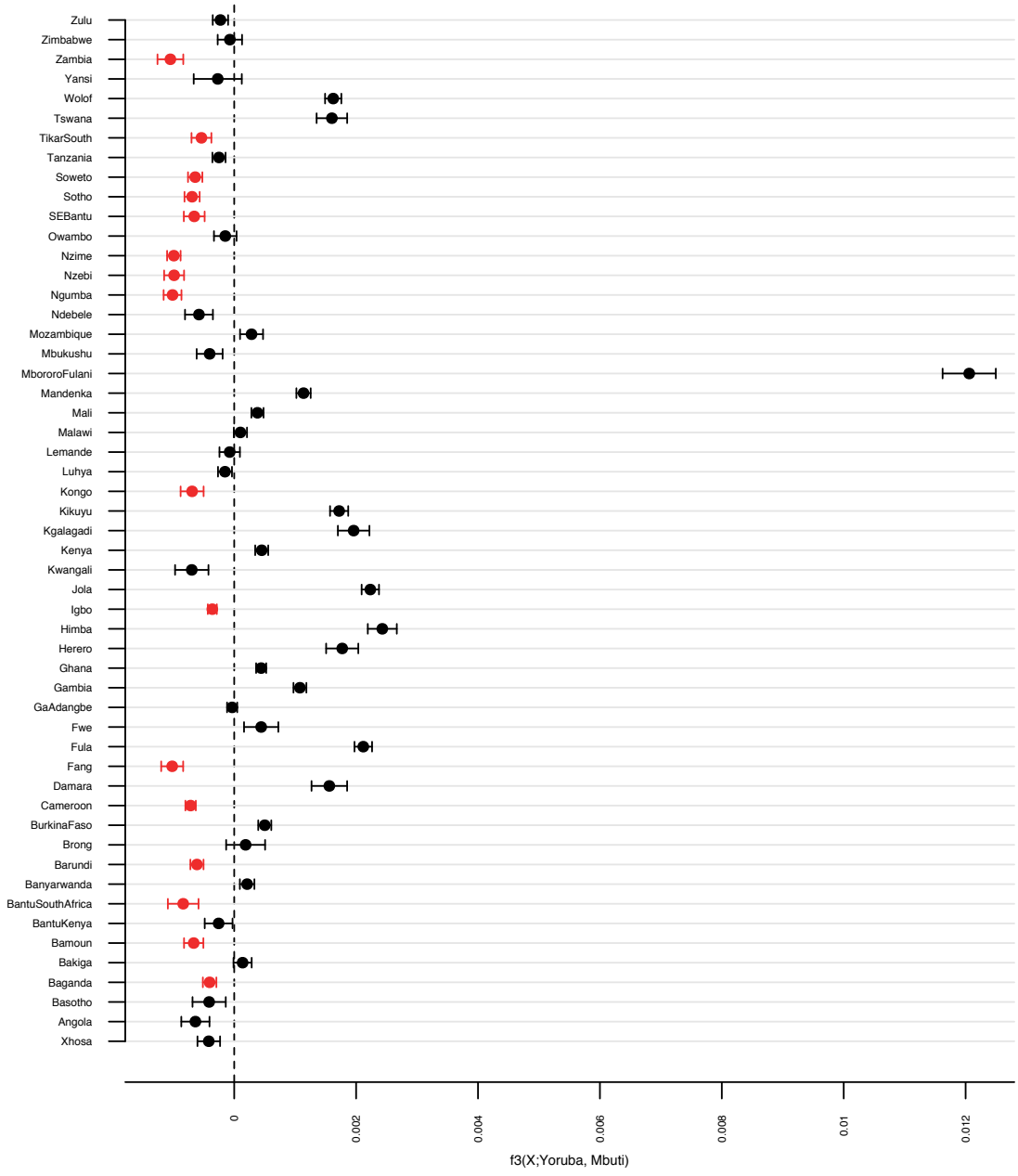


Figure 4.10 – Three-population (*f3*) test results with the format *f3*(X; Yoruba, Mbuti) for the 'Niger-Congo Genetic Dataset'. Significant comparisons (Z-score < −3) are identified in red.

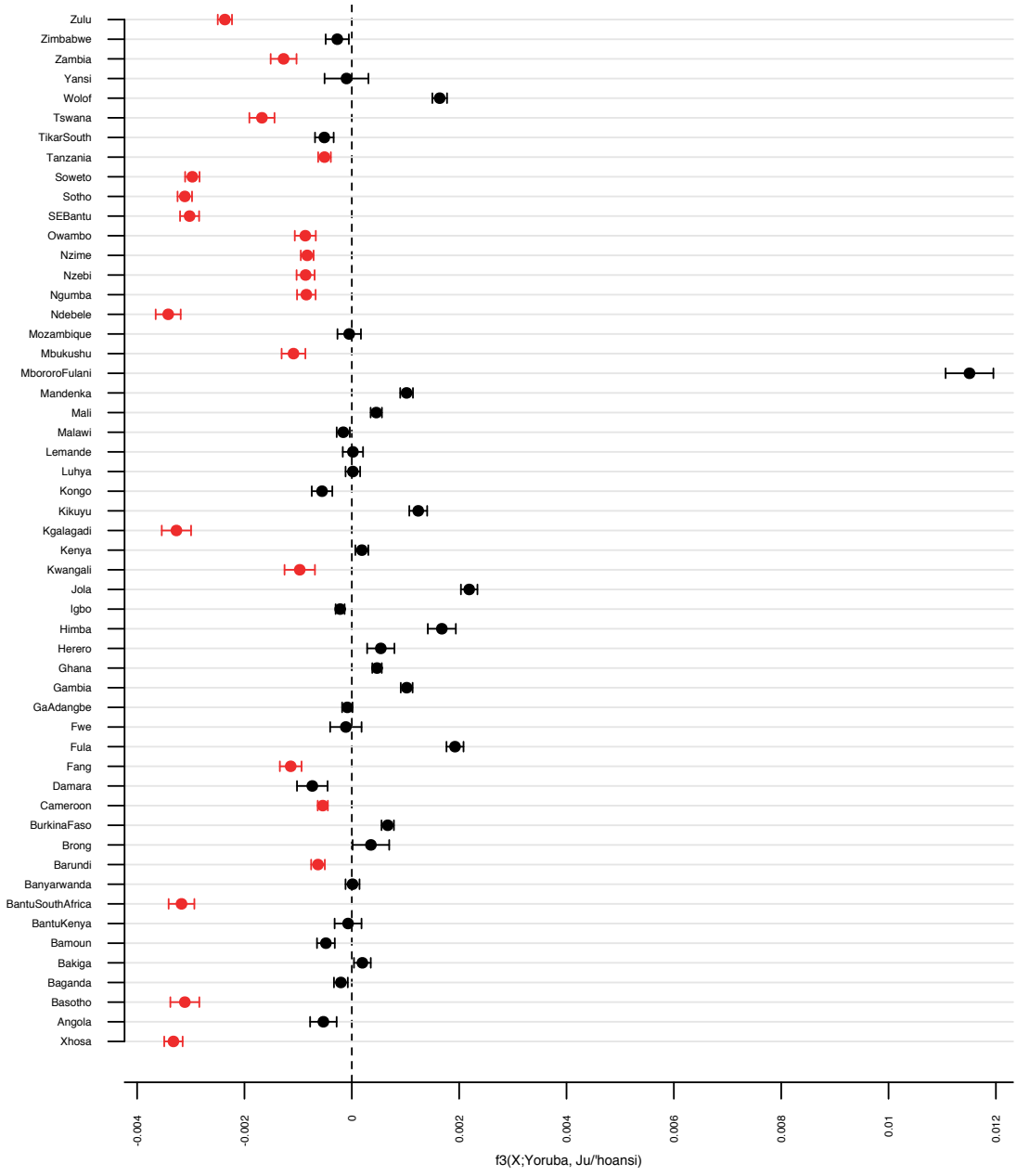


Figure 4.11 – Three-population (*f3*) test results with the format *f3*(X; Yoruba, Ju/'hoansi) for the 'Niger-Congo Genetic Dataset'. Significant comparisons (Z-score < −3) are identified in red.

4.2.4.2 Europe

As mentioned above, all the populations showing significant admixture signals with the Somali sample also showed them with the European sample (Figure 4.8 and 4.9). Additionally, three more populations from west Africa also showed significant signatures of admixture with Europe. However, it is important to note that the Somalia sample used as a source for possible admixture in this analysis also showed significant signatures of admixture with Europe itself (*f3* statistic = −0.00545; *Z-score* = −19.99; result not shown on Figure 4.9). As such, it is difficult to distinguish the specific impact of the two in some Niger-Congo populations.

4.2.4.3 Mbuti RHG

Signatures of admixture with the Mbuti Pygmies are mostly found in populations from central-west Africa (Figure 4.10). It is important to note here that most of the significant signatures of admixture with RHG are associated also with signatures of 'Khoisan' admixture, with only four populations with Pygmy signals of admixture also not showing significant signatures with 'Khoisan' (Figure 4.11).

4.2.4.4 'Khoisan'

Significant 'Khoisan' admixture in Niger-Congo populations was observed mostly in central-western and southern African populations (Figure 4.11). Signatures of 'Khoisan' admixture are also the most widespread in the

dataset, with 21 out of 56 populations showing significant signatures of admixture events with the Ju/'hoansi sample.

Overall, all Bantu populations from both *clusters B* and *D* (east and south Africa, respectively) show significant signatures of admixture with at least one of the populations used as sources. All Bantu populations from *cluster B* show signatures of admixture with both Europe and east Africa (Figure 4.8 and 4.9), and all populations from *cluster D* show significant admixture with 'Khoisan' (Ju/'hoansi) (Figure 4.11). Even if not all, several of the populations within *clusters A* and *C* also show significant signatures of admixture (5 out of 7 populations in *cluster A*, and 4 out of 8 in *cluster C*). The vast majority of the events identified in these two clusters are of admixture with Ju/'hoansi, with many of the populations presenting simultaneously signatures of admixture with Mbuti Pygmies (Figure 4.10 and 4.11).

4.2.5 F_{ST} population tree

To further explore the genetic relationships among populations in the two datasets I computed the genetic distance, F_{ST}, between all pairs of populations. I then used the matrix based on these estimates to generate a hierarchical clustering tree. The topology of the tree broadly reflects geography and ethno-linguistic clustering (Figure 4.12). The non-Niger-Congo populations are acting as outgroups to all the other populations in the

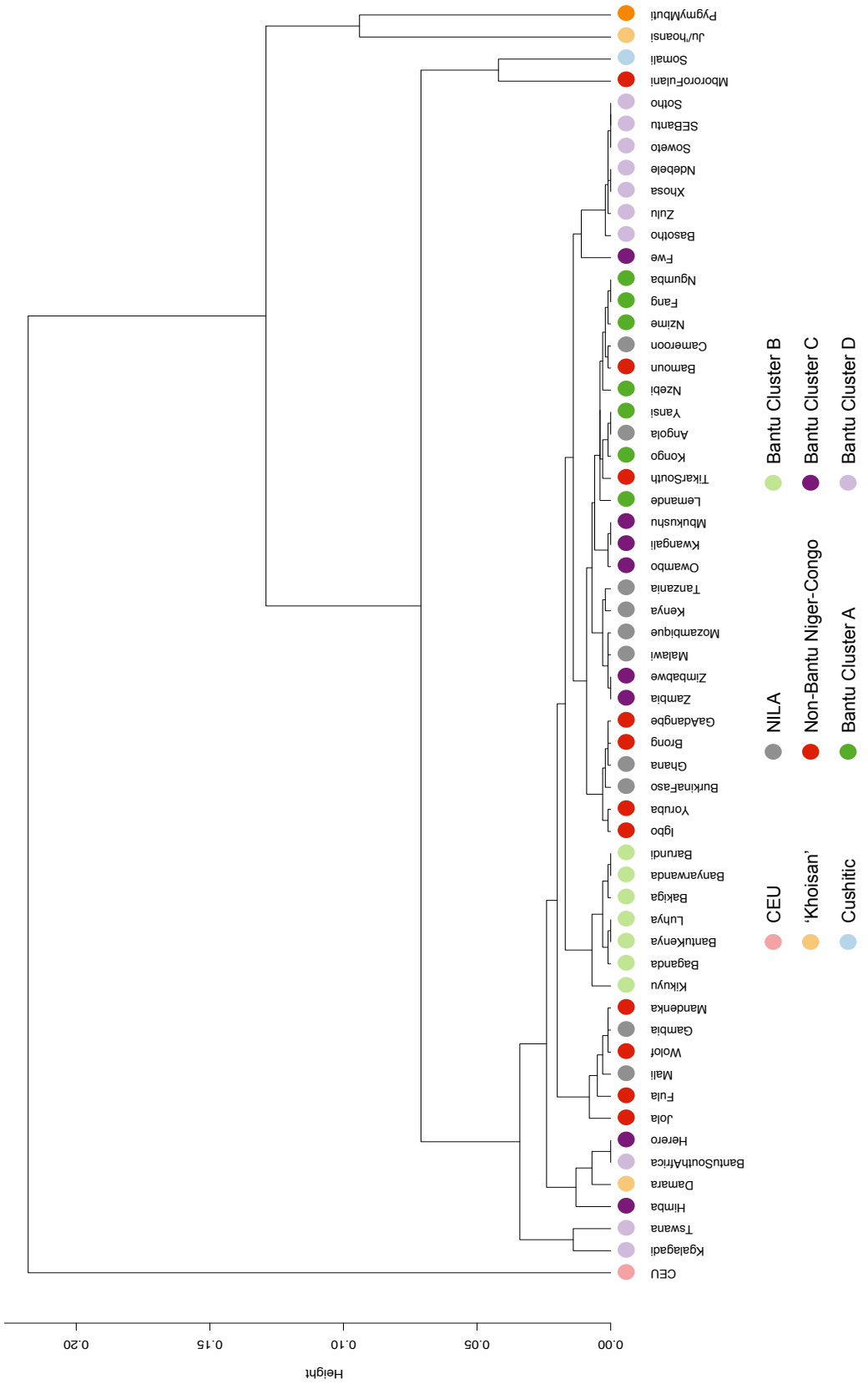


Figure 4.12 – Hierarchical population tree based on pairwise F_{ST} values for the 'Niger-Congo Genetic Dataset'. Coloured dots below branches correspond to the group colours in Figure 4.1.

dataset. The Damara are an exception to this, clustering with Bantu-speaking populations (namely their neighbour populations from Namibia, Himba and Herero) instead of clustering with the other 'Khoisan'-speaking population in the dataset, as also seen in the PCA and ADMIXTURE analyses. The NLIA populations included in this dataset (Table 4.1) cluster with the Niger-Congo populations, usually close to their neighbouring populations. However, not all populations cluster according to their geographic location and populations from the same region can be found in different clusters. This inconsistency in the geographical clustering is especially true for populations from southern areas of the continent (namely, Bantu *clusters C* and *D*). On the other hand, Bantu populations from *cluster B* all cluster in the same branch on the tree. Bantu populations from *cluster A* group together with other (but not all) populations from west Africa. Even though most the populations on this branch are from neighbouring regions to the location of *cluster A*, central-west Africa, a population from the south-west, Angola, is also present.

When the populations in the 'Bantu Genetic Dataset' are analysed, a similar pattern emerges. Populations do not always cluster according to their geographical origin (Figure 4.13). As before, all populations from *cluster B* (from north-east Africa) cluster together in the same branch of the tree. The most internal branch of the tree clusters all populations from *cluster A* and all but two from *cluster C*. The two absent populations from *cluster C* together with three others from the south-east (Kgalagadi, Tswana, and BantuSouthAfrica) act as outgroups to all the other

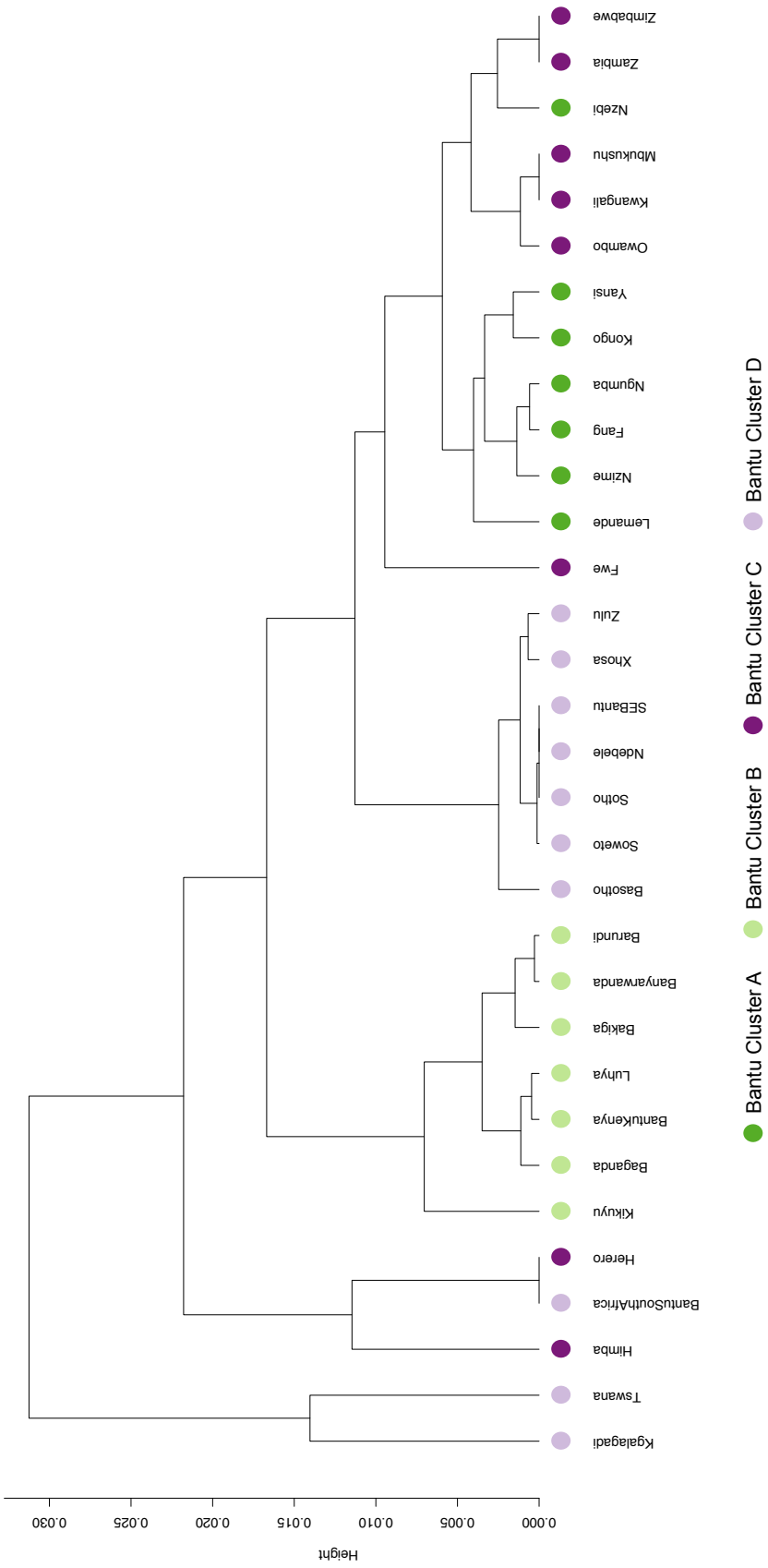


Figure 4.13 – Hierarchical population tree based on pairwise F_{ST} values for the 'Bantu Genetic Dataset'. Coloured dots below branches correspond to the group colours in Figure 4.3.

populations. The rest of the southeastern *cluster D* populations all cluster in the same branch of the tree.

4.2.6 Consolidation of the identified admixture events and the population tree

By analysing the results of the hierarchical F_{ST} tree jointly with the ADMIXTURE analyses and the formal tests of admixture, it is easy to observe that clustering, besides ethnicity and geography, is also affected by the admixture events experienced by the various populations (Figure 4.14).

The NLIA populations in the dataset cluster mostly with neighbouring populations from the same region. The main exception to this pattern seems to be samples from Kenya and Tanzania. They do not cluster with all the other northeastern populations in the dataset and instead cluster with populations from the south-east. However, the ADMIXTURE profile of these two populations is a clear indication of this clustering, as it resembles much more the pattern observed in southeastern populations (Figure 4.14). Namely, these populations show a high frequency of the *NC Bantu South* component (32.6% in Kenya, and 28.8% in Tanzania), a component almost absent in the other northeastern populations (< 9%; Table 4.2). A similar exception is also observed for the Angola sample. Even though this western population clusters with other populations from western regions, these are mostly from the north, namely Bantu populations belonging to *cluster A*.

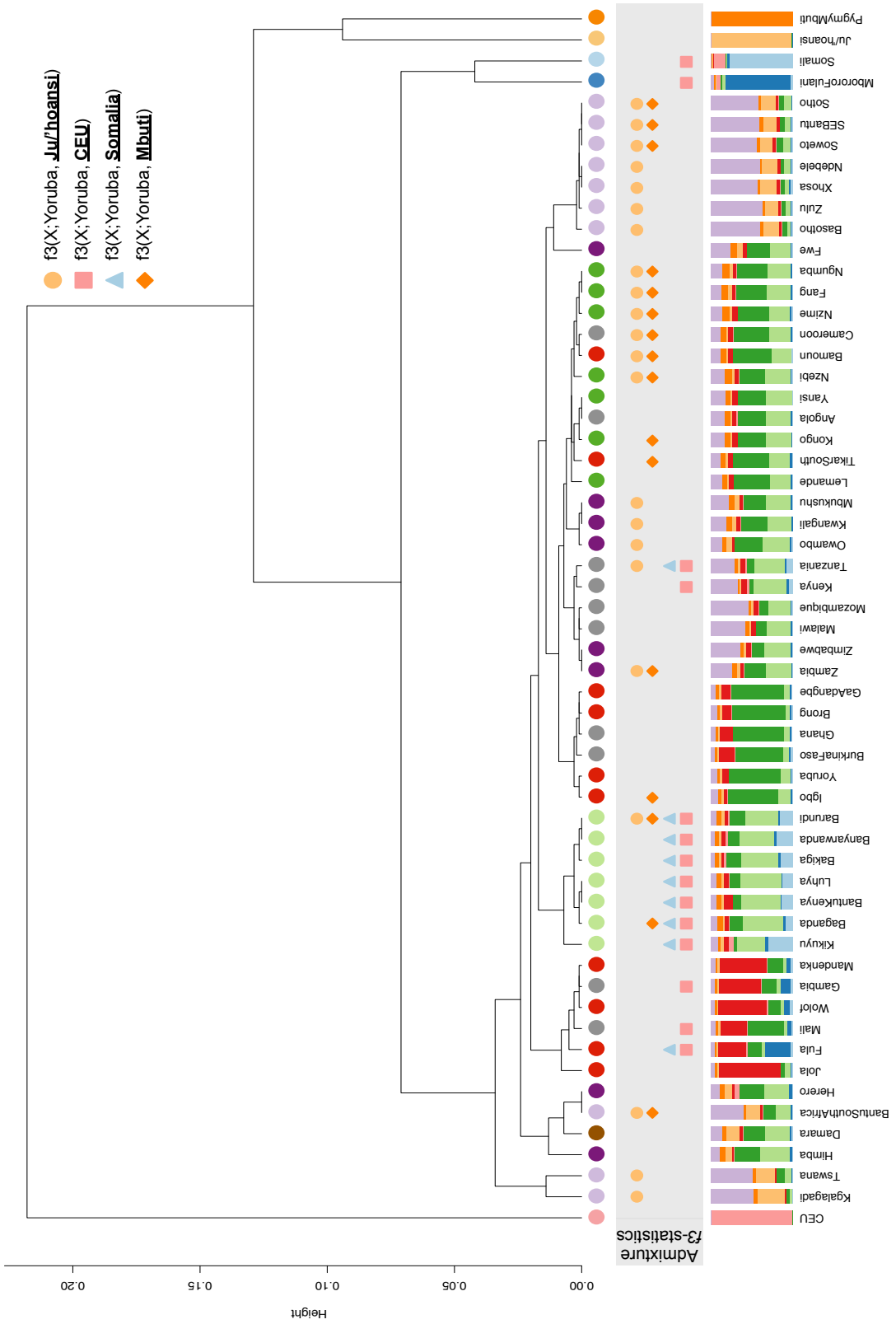


Figure 4.14 – Incorporation of the results of admixture events on the 'Niger-Congo Genetic Dataset' in the hierarchical population tree based on the F_{ST}. Coloured dots below branches correspond to the group colours in Figure 4.1; symbols correspond to significant comparisons (Z-score < −3) on the *f3 tests*; on the bottom of the figure, ADMIXTURE plot for *K*=9, ordered to mirror the topology of the tree.

The populations for which linguistic information is available also show deviations from what would be expected in a clustering uniquely driven by geography and ethnicity. The location of the Damara population in the tree – which, as stated above, is the only non-Niger-Congo population in the dataset to cluster among the Niger-Congo populations – is very easy to understand by interpreting these analyses together (Figure 4.14). The ADMIXTURE profile of the Damara is strikingly different from the other 'Khoisan' analysed (Ju/'hoansi) and much more similar to the pattern observed in its neighbours Himba and Herero. Namely, the frequency of the *Ju*/'*hoansi* component in the Damara is only 16.2% and more than a third of its components are Niger-Congo (a proportion similar to the one found in Bantu populations from South Africa) (Table 4.2).

4.3 Discussion

The dispersal of Bantu-speaking populations and their interaction with local inhabitants of the newly occupied regions have deeply shaped the genetic and cultural variation of sub-Saharan Africa. This phenomenon was briefly discussed in Chapter 3 in the context of the component found in southern Africa. Previously published works also address this aspect in relation to other regions (Patin et al. 2014; Patin et al. 2017; Tishkoff et al. 2009). Following the dispersion of Bantu speakers through Africa, the interactions between the settling agriculturalists and the local hunter-gatherers and pastoralists that previously inhabited those areas differed across the continent, shaping the genetic variation of the populations in different manners. These admixture dynamics can vary greatly through the African continent. For example, in the central African rainforest signals of a male-driven and non-symmetrical Bantu to RHG admixture have been identified (with some western RHG displaying more than 50% of Bantu ancestry (Patin et al. 2014)). On the other hand, admixture events identified in southern Africa seem to reflect an opposite sex-mediated admixture dynamic, with a female-driven 'Khoisan' to Bantu admixture (Barbieri et al. 2013b; Marks et al. 2015). Some southern African Bantu-speaking populations (as the South Africans Zulu and Xhosa) show a high frequency of 'Khoisan'-related mtDNA haplogroups L0d and L0k (Barbieri et al. 2013b; Marks et al. 2015; Schlebusch et al. 2013). The fact that the corresponding NRY haplogroups related to 'Khoisan' speakers are not observed in similar

frequencies is consistent with a 'Khoisan' female-driven admixture in Bantu speakers. Moreover, interactions between these two communities are also reflected by the fact that some southern African Bantu languages also have click consonants, an indication of possible cultural diffusion from 'Khoisan' (Barbieri et al. 2013a; Güldemann and Stoneking 2008). As a result of all these different admixture dynamics African populations – and Bantu-speaking populations in particular – show in the present-day very different genetic profiles throughout the continent (Barbieri et al. 2013b; Barbieri et al. 2014; Marks et al. 2015; Mitchell 2002; Patin et al. 2014; Patin et al. 2017; Pickrell et al. 2012; Pickrell et al. 2014; Tishkoff et al. 2009).

The pattern of variation across sub-Saharan Africa highlights the differences between populations located in different regions of the African continent, which appear to be influenced by admixture in modern populations. While early studies focusing on the Bantu speakers highlighted their relative genetic homogeneity when considered in the wider African context (Cavalli-Sforza et al. 1994) more recent studies have brought to light their heterogeneity (Choudhury et al. 2017; Patin et al. 2017). The investigation of the 'Niger-Congo Genetic Dataset', which I described in this chapter, allowed the identification of several different groups of Bantu-speaking populations. In the PC analysis, even though all the Bantu-speaking samples are close together and part of a continuous distribution, several subgroups can be identified, which mostly cluster according to their geographic location. The north-east samples are separated from the other

Bantus along the first component, in the direction of the east African Somali, with whom they all present significant signals of admixture.

However, it is important to note that geographic clustering is not always observed. The pairwise F_{ST} hierarchical tree combined with the results from the formal tests of admixture (*f3* tests) highlighted the role that admixture had in shaping the patterns of diversity observed in today Bantu populations, with several Bantu subgroups identified. The amounts and sources of admixture vary in different areas. European admixture was mainly observed in populations from east Africa, namely all populations from *cluster B* in the 'Bantu Genetic Dataset' showed significant signatures of admixture with the European sample, a possible outcome of a back migration into the continent after the first Out of Africa movements (Hodgson et al. 2014). On the other hand, 'Khoisan' signatures of admixture are more common in Bantu-speaking populations from central-west and southeast Africa. This is in agreement with results obtained in previous studies focusing on the admixture patterns between 'Khoisan' and Bantu populations in southern Africa, in which it was also demonstrated that mtDNA data mirrors the autosomal data (Barbieri et al. 2013b; Barbieri et al. 2013a; Marks et al. 2015; Rocha and Fehn 2016). This combination of the mtDNA and autosomal data – and the fact that it is not echoed by the Y chromosome – indicates that the admixture between these two groups was likely due to the assimilation of 'Khoisan' females into the Bantu settlements (Barbieri et al. 2013b; Barbieri et al. 2013a; Marks et al. 2015). However, it is important to note that the 'Khoisan' signatures of admixture identified are

many times also associated with signatures from Pygmy admixture, especially in central-west Africa (Figure 4.10, 4.11, and 4.14). This might be an indication that I do not have the right proxy for admixture in this region and, as a result, both populations are being identified as possible sources. Still, it is clear that the genetic structure of African Bantu populations is highly influenced by different dynamics of admixture between ethnic groups, mostly shaped by geographic proximity. In this context, the meeting of incoming farmers with some resident pastoral communities might have promoted these genetic exchanges due to the similar subsistence strategies.

Population differentiation among southern Bantu speakers has not been reported in many recent studies and they are often treated as a homogeneous group (Chimusa et al. 2015; Choudhury et al. 2017; Gurdasani et al. 2015; Schlebusch et al. 2012). A recent genome-wide study however – in accordance with what was reported earlier for uniparental and autosomal markers –, observed genetic divergence consistent with a division between the two main groups of southeastern Bantu speakers – the Nguni and the Sotho-Tswana speakers (Choudhury et al. 2017; Lane et al. 2002). In my datasets, I observe some differentiation among southeastern Bantu speakers as well. All of the Nguni speakers in my dataset (Xhosa, Ndebele, and Zulu) cluster together and with other southeastern populations. When considering the Sotho-Tswana speakers, the two samples of the Basotho population (coded as Sotho and Basotho) cluster together with the Nguni speakers above, while the other two Sotho-Tswana

speakers available (Kgalagadi and Tswana) do not cluster with these samples and act as outgroups to all the other Niger-Congo populations. It is important to note that Kgalagadi and Tswana have the highest frequency observed of the *Ju*/'*hoansi* component (33.9% and 23%, respectively) (Table 4.2). In addition, they show significant signals of admixture with Ju/'hoansi in my *f3* test (Figure 4.14). In this chapter, the Bantu populations analysed present some degree of differentiation, greatly influenced by admixture. Moreover, isolation by distance dynamics during the 'Bantu expansion' might have further impacted this differentiation, with founder events and drift associated with these movements possibly having a key role in these populations' genetic structure.

Also noteworthy, is the case of the Damara population. As stated before (see Chapter 2), the Damara are a Khoe-speaking population from Namibia. The Damara speak a dialect of the language spoken by the Nama, with whom they share not only the language but also much of the culture (Barnard 1992; Simons and Fennig 2018). However, the Damara have been shown to cluster with Niger-Congo populations instead of other 'Khoisan' populations (Pickrell et al. 2012). This same pattern is observed here. The Damara cluster close to other Niger-Congo populations from the south (namely from Namibia) in the PC analysis (Figure 4.2 and Supplementary Material 5N) and in the F_{ST} hierarchical tree (Figure 4.12). The ADMIXTURE analysis also shows an ADMIXTURE profile similar to those of the neighbouring Himba and Herero, Bantu-speaking populations from Namibia (Figure 4.6). Moreover, the *f3* tests including the Damara samples do not

show any significant signatures of admixture with any of the populations used as sources, including the ‘Khoisan’ (Figure 4.8, 4.9, 4.10, and 4.11). All of these signs support the hypothesis that the Damara are not a ‘Khoisan’ population but, in fact, a Niger-Congo population that experienced a language shift to Khoe by contact with neighbouring non-Bantu populations, possibly the Nama (Barnard 1992; G. T. Nurse et al. 1976; Pickrell et al. 2012). In fact, even the Damara folklore seems to support this claim, by claiming that their original language was Zulu, which they lost by contact with Khoe speakers (Barnard 1992).

The findings of this chapter highlight the differentiation observed in Bantu-speaking populations and the impact caused by the interaction with inhabitants of the regions into which the Bantu groups moved. Bantu populations from all the regions analysed in this chapter showed significant signals of admixture events with at least one possible source (26 out of the total 32 Bantu populations). Most of the times treated as a homogeneous group, it is important to consider the population structure of Bantu-speaking populations, namely in biomedical and ancestry research. The integration of archaeological and linguistic data with more complex demographic models might prove essential for a better understanding of the different processes through which languages and genetics spread across the African continent, in the so-called ‘Bantu expansion’.

Chapter 5

An Attempt to Reconcile the Genetics, Geography, and Languages of Bantu-speaking Populations

5.1 Introduction

The origin of the Bantu languages in present-day Cameroon near the border with Nigeria is well supported by linguistic studies (de Maret 2013; Grollemund et al. 2015). This language family is today present all across sub-Saharan Africa, having expanded east and south from its region of origin around 5000 years ago, possibly together with agriculture practices (Grollemund et al. 2015; Phillipson 2005; Vansina 1995). The field of population genetics has contributed to the investigation of the dispersal of this linguistic family and supported a model where this cultural expansion was attained through the diffusion of people, possibly also coupled with other cultural traits (eg iron smelting and agriculture) (Busby et al. 2016; de Filippo et al. 2012; Patin et al. 2017; Tishkoff et al. 2009). The suggestion of a demic process associated with the spread of cultural and linguistic traits called for additional investigations about the dynamics of this dispersal. Among the questions still contentious today, some address the actual routes taken by the populations during this expansion, as these have clear implications for the genetic diversity and relationships of present-day Bantu-speaking populations. Other questions focus as well on the degree of association between the cultural and genetic elements of this diffusion. However, poor resolution of the genetic markers investigated and limitations in representative sampling across the continent have restricted so far our ability to answer such questions.

The different routes suggested for the dispersal of these populations have been mostly based on the interpretation of linguistic data. Two main routes have been proposed so far, the major difference being related to when the Bantu-speaking populations passed the geographical barrier of the Equatorial forest. One of the models, the so-called 'early spilt model' (ES), states that Bantu populations would have split early in their evolutionary history, with one group moving directly south from their homeland near the present-day Cameroon-Nigeria border conquering most of west Africa, and other groups travelling east in the direction of the Great African Lakes. Only after reaching this region, would they travel south and settle across the eastern coast of the African continent. According to this model, the East and West branches of Bantu languages would be monophyletic clades (Currie et al. 2013; Rocha and Fehn 2016). The second most commonly supported model, the 'late split model' (LS), proposes that the separation of the East branch of Bantu languages would have only occurred after a first migration south across the Equatorial forest. South of the forest, this common population would have then split into two groups, one travelling further south along the coastline, and the other travelling east, both in the direction of the Great Lakes and further south along the east coast of the continent (Currie et al. 2013; Grollemund et al. 2015; Patin et al. 2017). As such, in this scenario, only the East branch of Bantu languages would form a monophyletic clade (Currie et al. 2013).

In order to investigate the patterns of dispersal during the 'Bantu expansion', in this chapter I integrate genetic and linguistic data from Bantu-

speaking populations with their geographical information. These populations, located in different regions of the African continent from their homeland to South Africa, allowed me to test different models of geographic dispersal. Moreover, I also attempt a method to mask the effects of recent admixture, which, as seen in the previous chapters, deeply shapes the genetic diversity observed today. My results highlight the different patterns of dispersal experienced by genes and words and the difficulties that arise from trying to reconcile them under a unifying model. Still, after trying to control for some potentially confounding variables in the results, genetics and linguistics show consistent results with the LS hypothesis of the division of East and West Bantu languages branches. I also show that populations from southern Africa in my dataset seem to experience strong levels of genetic drift, possibly as a result of their migration dynamics.

5.2 Results

5.2.1 Genetic datasets

In this data analyses chapter, I used the 'Bantu Genetic Dataset' and its subset, the 'Bantu Languages Genetic Dataset' (Figure 5.1 and Table 5.1; see also Chapter 2, Methods) to reconstruct and describe the relations between different aspects of Bantu populations' distribution. Namely, I investigate how the genetic, geographical, and linguistic distances of these populations are correlated with each other on the hypothesis of a

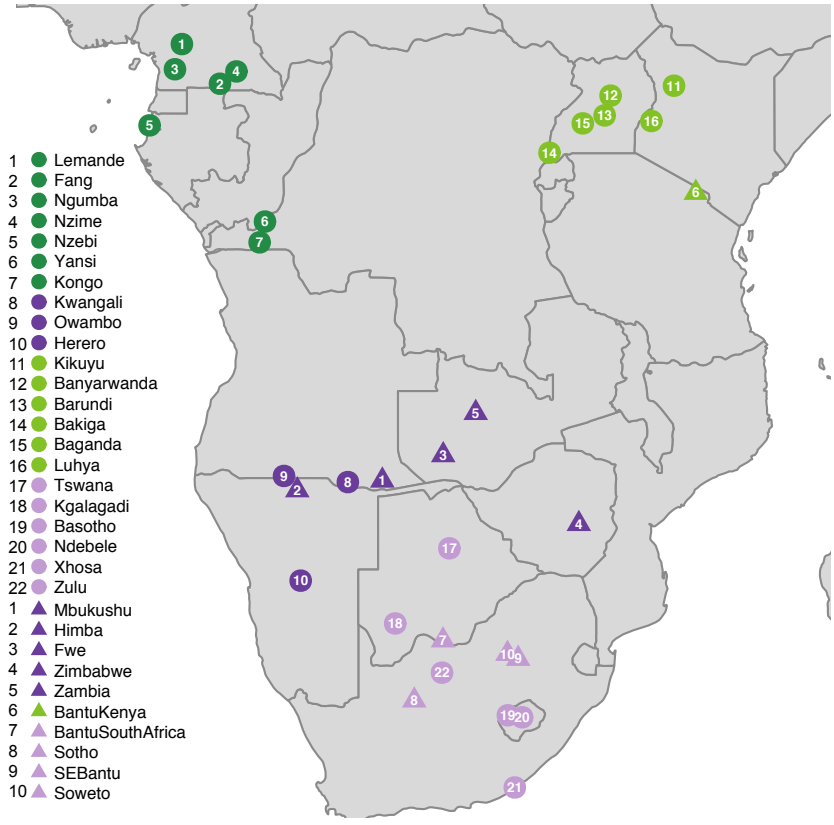


Figure 5.1 – Map with the location of samples included in the 'Bantu Genetic Dataset' (BG) and the 'Bantu Languages Genetic Dataset' (BLG).
Circles represent samples both in the BG and BLG datasets, while triangles represent samples only included in the BG dataset. Colours indicate the four geographical clusters identified in Bantu populations (see below): dark green, Bantu *cluster A*; light green, Bantu *cluster B*; dark purple, Bantu *cluster C*; light purple, Bantu *cluster D*.

Table 5.1 – Populations comprising each of the two datasets used on Chapter 5.

	Bantu Genetic Dataset	*Bantu Languages Genetic Dataset*
Cluster A		
Lemande	x	x
Fang	x	x
Ngumba	x	x
Nzime	x	x
Nzebi	x	x
Yansi	x	x
Kongo	x	x
Cluster B		
Baganda	x	x
Bakiga	x	x
Banyarwanda	x	x
Barundi	x	x
Kikuyu	x	x
Luhya	x	x
BantuKenya	x	
Cluster C		
West		
Herero	x	x
Kwangali	x	x
Owambo	x	x
Fwe	x	
Himba	x	
Mbukushu	x	
East		
Zambia	x	
Zimbabwe	x	
Cluster D		
Xhosa	x	x
Basotho	x	x
Kgalagadi	x	x
Ndebele	x	x
Tswana	x	x
Zulu	x	x
BantuSouthAfrica	x	
SEBantu	x	
Sotho	x	
Soweto	x	

co-distribution of genes and words spread by populations across the African continent. As stated before, even though all the populations present in the 'Bantu Genetic Dataset' (BG) speak languages belonging to the Narrow Bantu linguistic family, not all have detailed linguistic data to be able to be included in the linguistic distance matrix. As part of the analyses I perform in

this chapter involves the investigation of the correlation between different distance measures, I focused on a subset of the BG dataset by including only those populations for which the detailed linguistic information was available – the 'Bantu Languages Genetic Dataset' (BLG). In this context, I use the full BG dataset when analysing the correlation between genetic and geographical distances. However, in order to compare all the distances (genetic, geographic, and linguistic) for the same set of populations, correlations were also performed with the BLG dataset.

5.2.2 Geographical clustering of the samples

In order to organize the populations in my dataset and to interpret the results of the subsequent analyses in this chapter more easily, I started by investigating the possibility of clustering populations in my dataset in subgroups. To do this I initially performed a hierarchical clustering using the great-circle distance (GCD)[12] between pairs of populations in the BG dataset (Figure 5.2A). The tree obtained using this approach suggests the presence of four geographical clusters in my dataset – '*cluster A*', which includes populations from west Africa; '*cluster B*', grouping populations from the east; '*cluster C*', characterized by populations from Namibia, Zambia, and Zimbabwe; and '*cluster D*', with populations from the south (Figure 5.2B). In the BG dataset, *cluster C* can be further subdivided into two

[12] The great-circle line distance between two populations is the shortest distance between those two populations taking into account the curvature of the surface of the Earth.

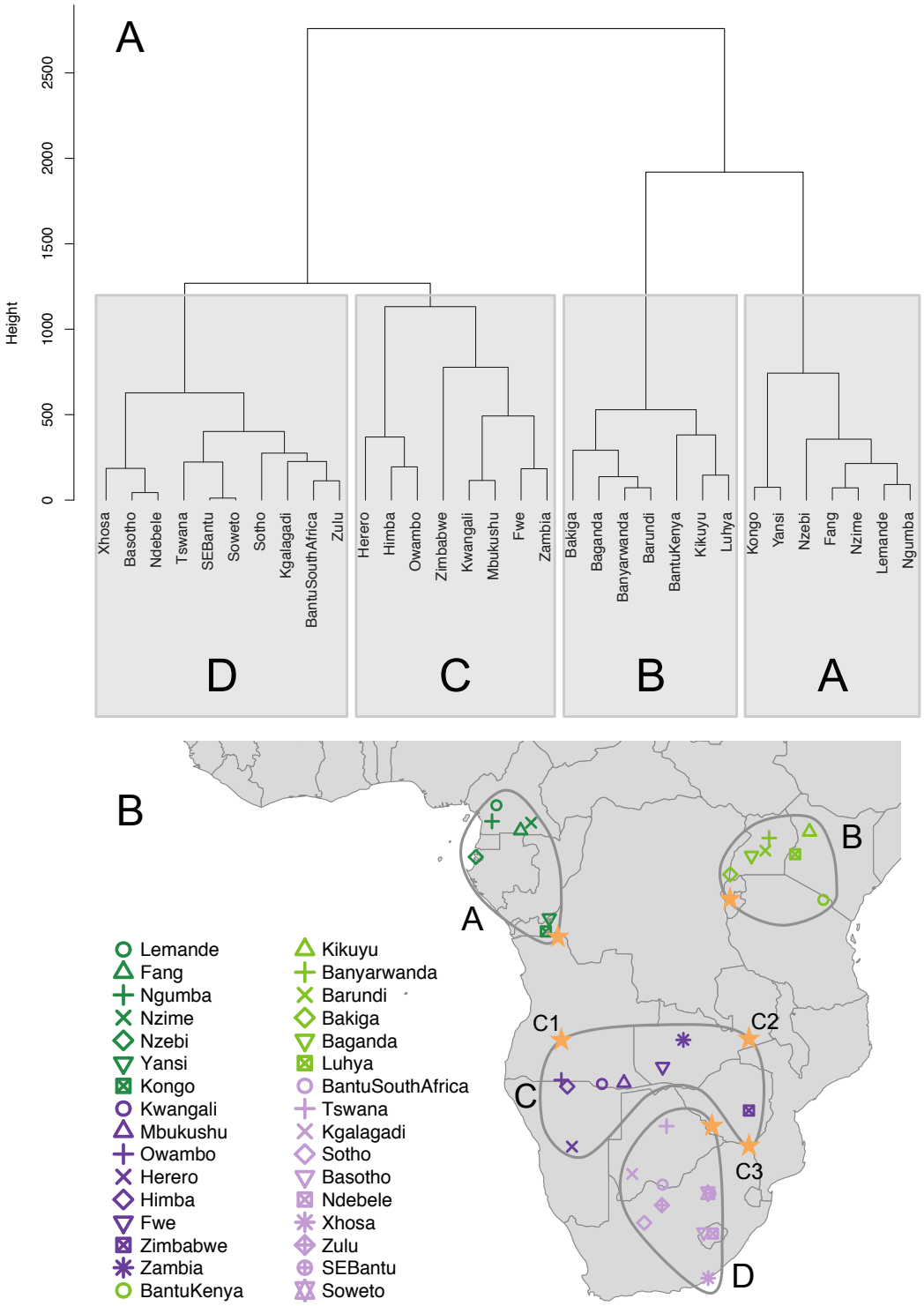


Figure 5.2 – Geographical clustering of the Bantu samples. (A) Hierarchical clustering of the great-circle distances (GCD) in the BG dataset; (B) location and composition of the four geographical clusters identified.

subgroups if linguistic affiliation is taken into account, as it is comprised of populations speaking languages from both East and West branches of the Narrow Bantu linguistic family (Table 5.1). In the BLG dataset this subdivision is not present as the East Bantu branch populations from *cluster C* are not present in this dataset as they could not be included in the linguistic distance matrix (Table 5.1).

The analysis of the distribution of the GCD within each of the four identified clusters further supports the identification of these clusters (Figure 5.3). The distribution of distances within each of the four clusters is very similar for the BLG dataset, showing no significant differences between clusters (Figure 5.3B). However, in the BG dataset *cluster C* presents higher levels of dispersion among the distances observed within the clusters, giving further support to the subdivision of this cluster into two subgroups (Figure 5.3A).

The differences between the two datasets as well as the populations included in each cluster are detailed in Table 5.1.

5.2.3 Choice of F_{ST} estimator

As described in the previous data analyses chapter (Chapter 4) and in the Methods chapter (Chapter 2), Bhatia and collaborators (Bhatia et al. 2013) recommend the use of the Hudson F_{ST} estimator when comparing population with different sample sizes as this is not sensitive to the ratio of sample sizes between pairs of populations and also as it does not tend to

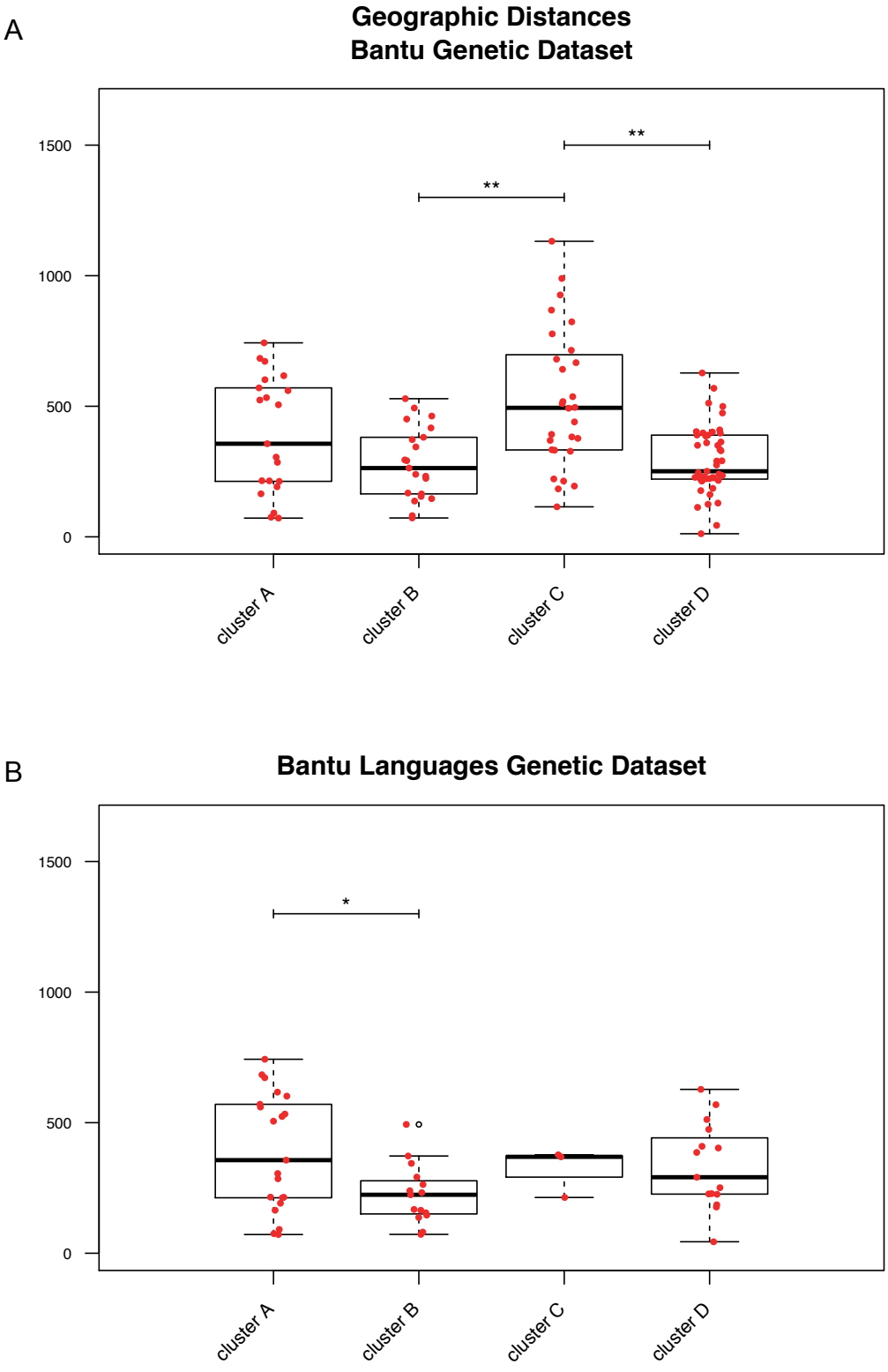


Figure 5.3 – Distribution of the geographic distances between populations within each cluster. Statistically significant differences after a Wilcoxon test are reported in the figure; *, *p-value*<0.05; **, adjusted *p-value*<0.0083, after Bonferroni correction.

overestimate the calculation of F_{ST}. However, as can be seen by the equation used by Bhatia and collaborators (Equation 1), for small sample sizes this F_{ST} estimator admits negative values – especially in the case of very similar allele frequencies. When these conditions are present, $(p_1 - p_2)^2$ might easily be smaller than $\frac{p_1(1-p_1)}{n_1-1} - \frac{p_2(1-p_2)}{n_2-1}$, which would yield a negative F_{ST} value as the denominator on the equation is always positive.

$$F_{ST}^{Hudson} = \frac{(p_1 - p_2)^2 - \frac{p_1(1-p_1)}{n_1-1} - \frac{p_2(1-p_2)}{n_2-1}}{p_1(1-p_2) + p_2(1-p_1)}, \qquad (1)$$

As some populations in my dataset have very small sample sizes, I decided to further investigate the impact of population size in F_{ST} calculations. I therefore initially investigated the pairwise F_{ST} distances in the 'Niger-Congo Genetic Dataset' analysed in the previous chapter, focusing on how the Hudson F_{ST} estimator would behave when using different sample sizes. While for a fixed sample size of 100 individuals only 4 out of 1653 pairwise distances are negative (0.2%), in the case of a fixed sample size of 10, 65.9% of the pairwise distances would be lower than zero (1089 out of 1653) (Figure 5.4).

Negative F_{ST} distance values do not have biological meaning and in addition I anticipated that negative values would be problematic to deal with further in the chapter. I decided to investigate if and how Hudson and the classic Wright's F_{ST} estimator are correlated (Figure 5.5). Both estimators showed to be strongly correlated (r^2=77.6%; p<0.05) and so I

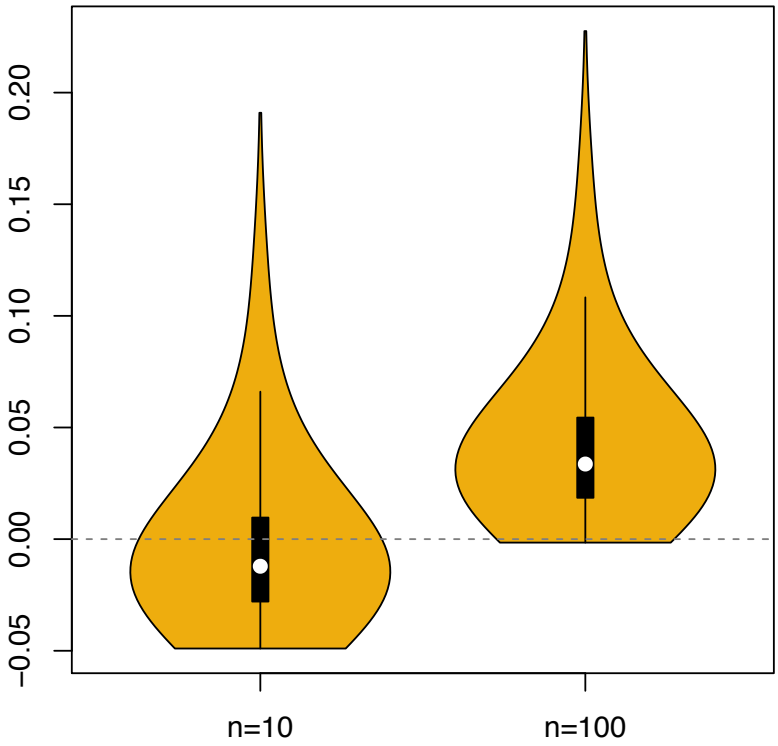


Figure 5.4 – Violin plot with the distribution of F_{ST} values for the Hudson F_{ST} estimator with fixed sample size, n, of 10 and 100.

decided to use the classic Wright's measure of F_{ST} as a ratio of variances (Equation 2) in all the subsequent analyses on this chapter (Wright 1949).

$$F_{ST}^{Wright} = \frac{var(p)}{p(1-p)}, \qquad (2)$$

5.2.4 Analysis of the correlation between pairwise distances

In order to test the correlation between the different distances for all pairs of populations – genetic distances (measured by the pairwise F_{ST} between populations), geographical distances (GCD), and linguistic distances (based on the number of differences between pairs of words) – I started by testing the normality of the distribution of my data through a

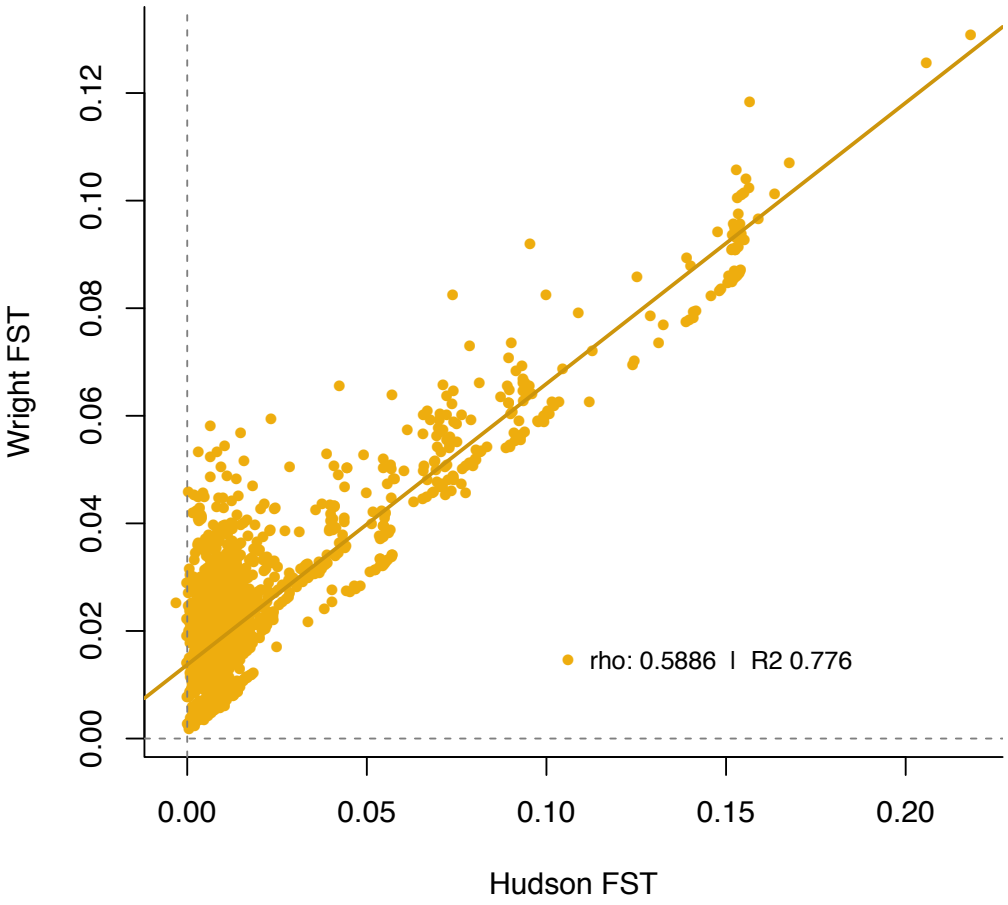


Figure 5.5 – Correlation between Hudson's and Wright's F_{ST} estimators.

Shapiro-Wilk test. For all the five distance matrices (genetic and geographical distances for both 'Bantu Genetic Dataset' and 'Bantu Languages Genetic Dataset' and linguistic distances only for the latter) I rejected the null hypothesis that the data were normally distributed. Given this, a Spearman's rank correlation coefficient – which allows for the data to be not normally distributed – was calculated between genetic distances (F_{ST}), geographical distances, and linguistic distances. Only the correlations between linguistic distances and both geographical and genetic distances showed a significant (p<0.05) positive correlation, with a Spearman coefficient of correlation, ρ, of 0.49 and 0.18, respectively (Figure 5.6).

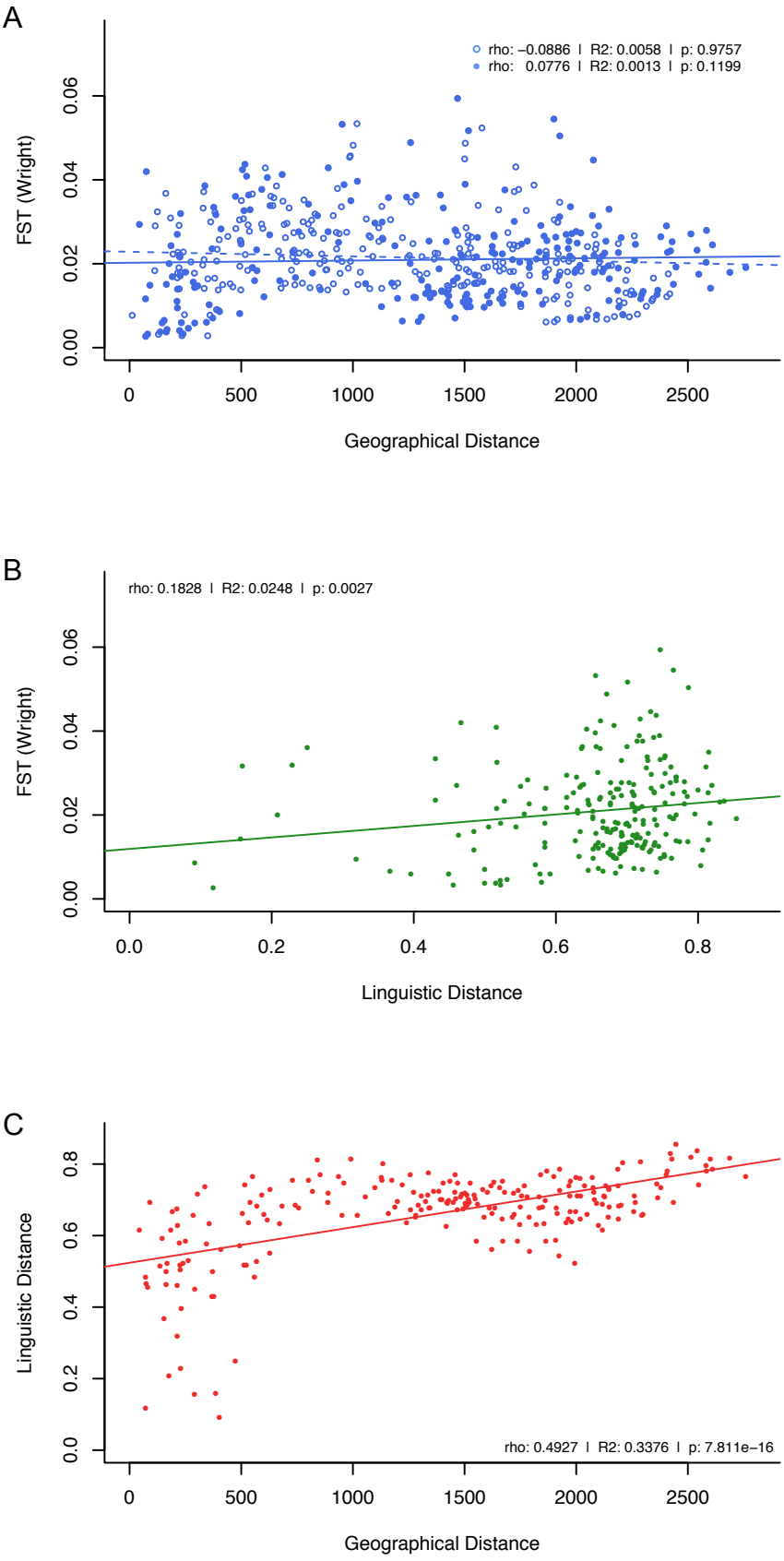


Figure 5.6 – Correlation between pairwise distances between Bantu-speaking populations. (A) Geographical and genetic (F_{ST}) distances; (B) linguistic and genetic (F_{ST}) distances; and (C) geographical and linguistic distances. Open circles (in panel A) correspond to the BG dataset and full circles to the BLG dataset (see main text).

However, even if significant, the amount of variation in linguistic distances explained by genetics (as measured by the coefficient of determination, r^2) is very small, with only 2% of the variation of the former being explained by the latter. Nonetheless, both correlations showed a positive coefficient of correlation, ρ, meaning that an increase in the linguistic distance leads to an increase in either the geographical or the genetic distances.

5.2.5 F_{ST} Population Tree

As seen in the results from the previous chapters, admixture played a major role in the dispersion of Bantu-speaking populations and in shaping the presently observed genetic diversity in these populations. A tree based on the newly calculated Wright F_{ST} values between pairs of populations on the 'Bantu Genetic Dataset' was constructed (Figure 5.7). As observed before (see Chapter 4) the topology of this tree does not always reflect the geographic location of the samples. None of the four geographical clusters (A–D) in the dataset have all their populations clustering together. In addition, most of the south-east samples from *cluster D* (with the exception from BantuSouthAfrica) are separated from all the other populations, as for Yansi (from the Democratic Republic of Congo), Fwe (from Zambia), and Kwangali (from Namibia).

In order to explore the impact of introgression in Bantu-speaking populations, I initially investigated the relationship between the components found in the BG dataset (and thus in its subset the BLG dataset). The

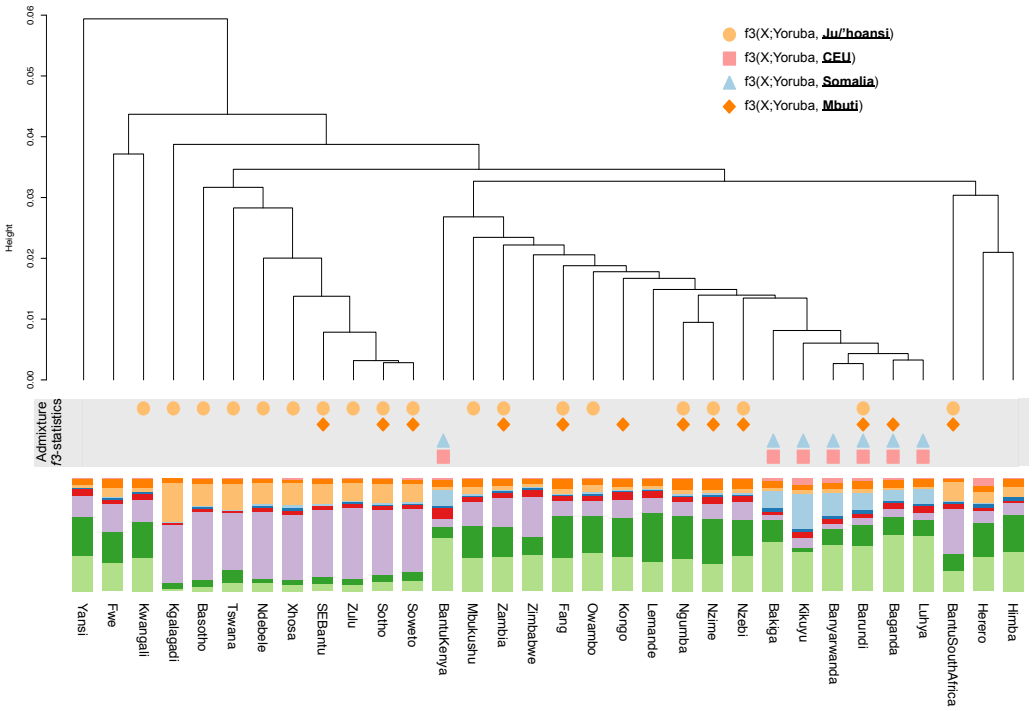


Figure 5.7 – Hierarchical population tree based on the F_{ST} for the 'Bantu Genetic Dataset'. The corresponding results on the analysis of admixture events (ADMIXTURE and *f3* tests) for the 'Niger-Congo Genetic Dataset' (see Chapter 4) are included below the tree: symbols correspond to significant comparisons (Z-score < −3) on the *f3 tests;* ADMIXTURE plot for *K*=9, ordered to mirror the topology of the tree above.

pairwise F_{ST} between the nine previously identified ADMIXTURE components (see Chapter 4) was calculated and used to generate a hierarchical tree (Table 5.2 and Figure 5.8). The three components primarily found in Bantu-speaking populations (*NC Yoruba*, *NC Bantu East* and *NC Bantu South* components, in dark and light green, and purple, respectively, in Figure 5.7) have all similar F_{ST} values between them (between 0.025 and 0.030; Table 5.2) and are similarly distant to a fourth component common in Niger-Congo speakers (*NC West*, values between 0.022 and 0.036). One of these Bantu components is the same found predominantly in Yoruba (*NC Yoruba*, dark green in Figure 5.7) and the other two show similar genetic distance values to it (0.025 and 0.028). Exploring the relationships with the non-Niger-Congo components identified (*Europe*, *Somali*, *Jul'hoansi*, and

Table 5.2 – F_{ST} between ADMIXTURE components at K=9. Identified on the 'Niger-Congo Genetic Dataset' (see Chapter 4).

	Europe	Somali	Ju/'hoansi	Pygmy	NC Fulani	NC West	NC Yoruba	NC Bantu East	NC Bantu South
Europe	–								
Somali	0.109	–							
Ju/'hoansi	0.209	0.143	–						
Pygmy	0.198	0.126	0.113	–					
NC Fulani	0.112	0.068	0.140	0.125	–				
NC West	0.154	0.074	0.121	0.103	0.053	–			
NC Yoruba	0.151	0.069	0.119	0.100	0.054	0.022	–		
NC Bantu East	0.153	0.078	0.113	0.099	0.065	0.036	0.028	–	
NC Bantu South	0.156	0.074	0.120	0.098	0.065	0.036	0.025	0.030	–

Pygmy components), the three Bantu components are also similarly distant to each of them. Of these, excluding the non-African *Europe* component, Bantu components all show the highest F_{ST} value with the *Ju*/'*hoansi* component (with values between 0.113 and 0.120). The hierarchical clustering approach based on the pairwise F_{ST} between the ADMIXTURE components identifies the close relationship between the Niger-Congo components, as they all cluster together (Figure 5.8).

5.2.6 Correction of genetic distances

In an attempt to control for admixture in Bantu-speaking populations I tested a method to mask the effects of admixture and generate a proxy reconstruction of the original 'unadmixed' populations (as previously

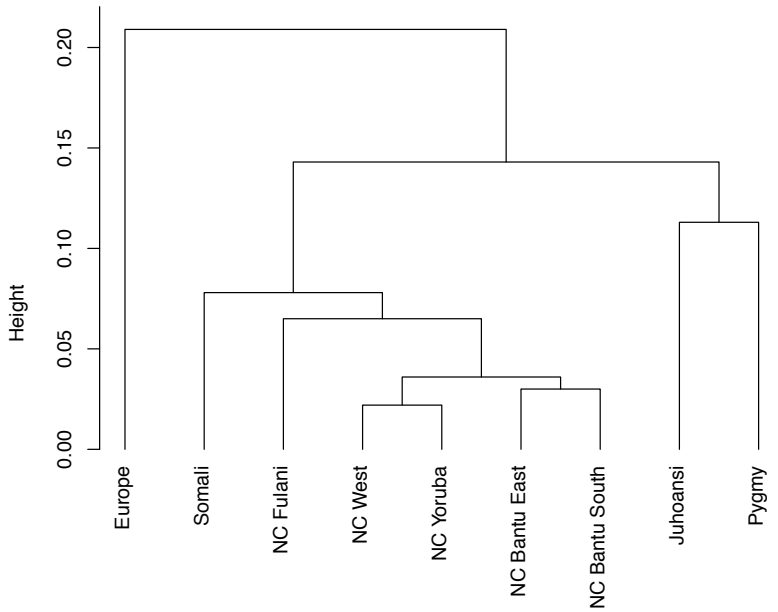


Figure 5.8 – Hierarchical tree of the F_{ST} between the ADMIXTURE components at K=9. (For the 'Niger-Congo Genetic Dataset', see Chapter 4)

explained in more detail in Chapter 2, Methods). Briefly, the method consisted in isolating the Niger-Congo-specific components[13] in all the admixed Bantu populations studied and using the allele frequencies of these ancestral components reconstructed by ADMIXTURE. Reassembling the Bantu populations as if they were composed only by the Niger-Congo modal ADMIXTURE components, I hope to remove the non-Niger-Congo signatures and reconstruct the ancestral populations. After this, I estimate the allele frequencies of these 'unadmixed' populations by combining the allele frequencies of each component according to their observed amount.

[13] With the exception of the *NC Fulani* component as this was not found in significant amounts in any of the Bantu populations (below 3.8%; Table 5.3) and was almost exclusively restricted to the MbororoFulani population.

Table 5.3 – Median value per Bantu-speaking population of each ADMIXTURE component. Identified for *K*=9 on the 'Niger-Congo Genetic Dataset' (see Chapter 4).

	Europe	Somali	Ju/'hoansi	Pygmy	NC Fulani	NC West	NC Yoruba	NC Bantu East	NC Bantu South	Niger-Congo	Non-Niger-Congo
Kgalagadi	0.000	0.010	0.339	0.043	0.000	0.014	0.047	0.024	0.523	0.608	0.392
Tswana	0.001	0.005	0.230	0.046	0.007	0.014	0.108	0.080	0.508	0.717	0.282
Himba	0.006	0.004	0.084	0.070	0.035	0.018	0.321	0.356	0.107	0.837	0.164
BantuSouthAfrica	0.002	0.008	0.165	0.030	0.020	0.039	0.154	0.181	0.402	0.796	0.205
Herero	0.062	0.012	0.098	0.054	0.035	0.026	0.298	0.308	0.108	0.775	0.226
Kikuyu	0.060	0.304	0.039	0.037	0.032	0.055	0.038	0.348	0.085	0.558	0.440
Baganda	0.013	0.089	0.024	0.067	0.024	0.047	0.157	0.500	0.078	0.806	0.193
BantuKenya	0.009	0.134	0.026	0.067	0.019	0.105	0.098	0.477	0.064	0.763	0.236
Luhya	0.007	0.127	0.026	0.064	0.015	0.064	0.134	0.498	0.066	0.777	0.224
Bakiga	0.025	0.145	0.029	0.056	0.038	0.035	0.187	0.444	0.042	0.746	0.255
Banyarwanda	0.039	0.202	0.038	0.051	0.031	0.041	0.140	0.415	0.044	0.671	0.330
Barundi	0.026	0.153	0.035	0.066	0.027	0.040	0.183	0.407	0.063	0.720	0.280
Zambia	0.004	0.010	0.039	0.059	0.011	0.048	0.265	0.306	0.259	0.889	0.112
Zimbabwe	0.003	0.011	0.027	0.042	0.013	0.066	0.155	0.325	0.359	0.918	0.083
Owambo	0.002	0.015	0.061	0.055	0.024	0.037	0.336	0.336	0.135	0.868	0.133
Kwangali	0.004	0.002	0.041	0.074	0.014	0.059	0.325	0.295	0.186	0.879	0.121
Mbukushu	0.004	0.012	0.067	0.069	0.014	0.039	0.281	0.298	0.217	0.849	0.152
Lemande	0.004	0.007	0.027	0.059	0.015	0.055	0.438	0.259	0.136	0.903	0.097
Kongo	0.002	0.010	0.024	0.074	0.010	0.065	0.345	0.308	0.162	0.890	0.110
Yansi	0.001	0.008	0.019	0.058	0.003	0.068	0.351	0.313	0.179	0.914	0.086
Nzebi	0.001	0.013	0.036	0.090	0.016	0.048	0.317	0.312	0.167	0.860	0.140
Nzime	0.003	0.013	0.032	0.095	0.020	0.062	0.392	0.249	0.134	0.857	0.143
Fang	0.004	0.011	0.044	0.085	0.014	0.044	0.374	0.298	0.126	0.856	0.144
Ngumba	0.003	0.010	0.038	0.096	0.013	0.044	0.378	0.286	0.132	0.853	0.147
Fwe	0.008	0.017	0.075	0.075	0.012	0.044	0.275	0.255	0.240	0.826	0.175
Basotho	0.005	0.015	0.191	0.044	0.012	0.028	0.064	0.040	0.601	0.745	0.255
Zulu	0.007	0.016	0.153	0.033	0.012	0.037	0.055	0.058	0.631	0.793	0.209
Xhosa	0.010	0.024	0.201	0.027	0.018	0.044	0.048	0.056	0.573	0.739	0.262
Ndebele	0.005	0.021	0.190	0.028	0.012	0.038	0.037	0.073	0.595	0.755	0.244
Soweto	0.015	0.014	0.159	0.034	0.012	0.034	0.073	0.097	0.563	0.779	0.222
SEBantu	0.009	0.019	0.170	0.040	0.008	0.033	0.057	0.071	0.594	0.763	0.238
Sotho	0.006	0.009	0.173	0.039	0.011	0.040	0.058	0.086	0.577	0.772	0.227

I then produced a tree after calculating the pairwise F_{ST} between the newly generated populations. The tree shows a clustering of populations more consistent with their geographic location. However, it is not entirely without some exceptions (Figure 5.9). Namely, the clade under which all populations from the west are clustered does not show consistency regarding latitude – with northwest and southwest populations clustering together under different clades – and also including three populations from the southeast.

I tested the impact of this correction by analysing the distribution of the F_{ST} values within each cluster and comparing it to the distribution before the correction. Before the correction, the geographic clusters presented high levels of intra-group genetic variation (Figure 5.10). After the correction, the pairwise F_{ST} values across the whole dataset became much lower (by several orders of magnitude), as expected if in fact the admixture in Bantu-

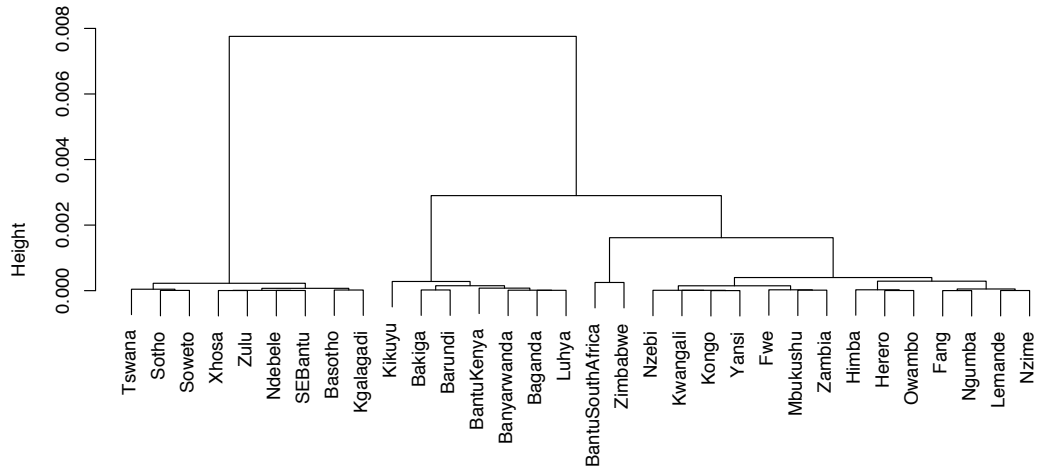


Figure 5.9 – Hierarchical F_{ST} population tree after the correction of the genetic distances.

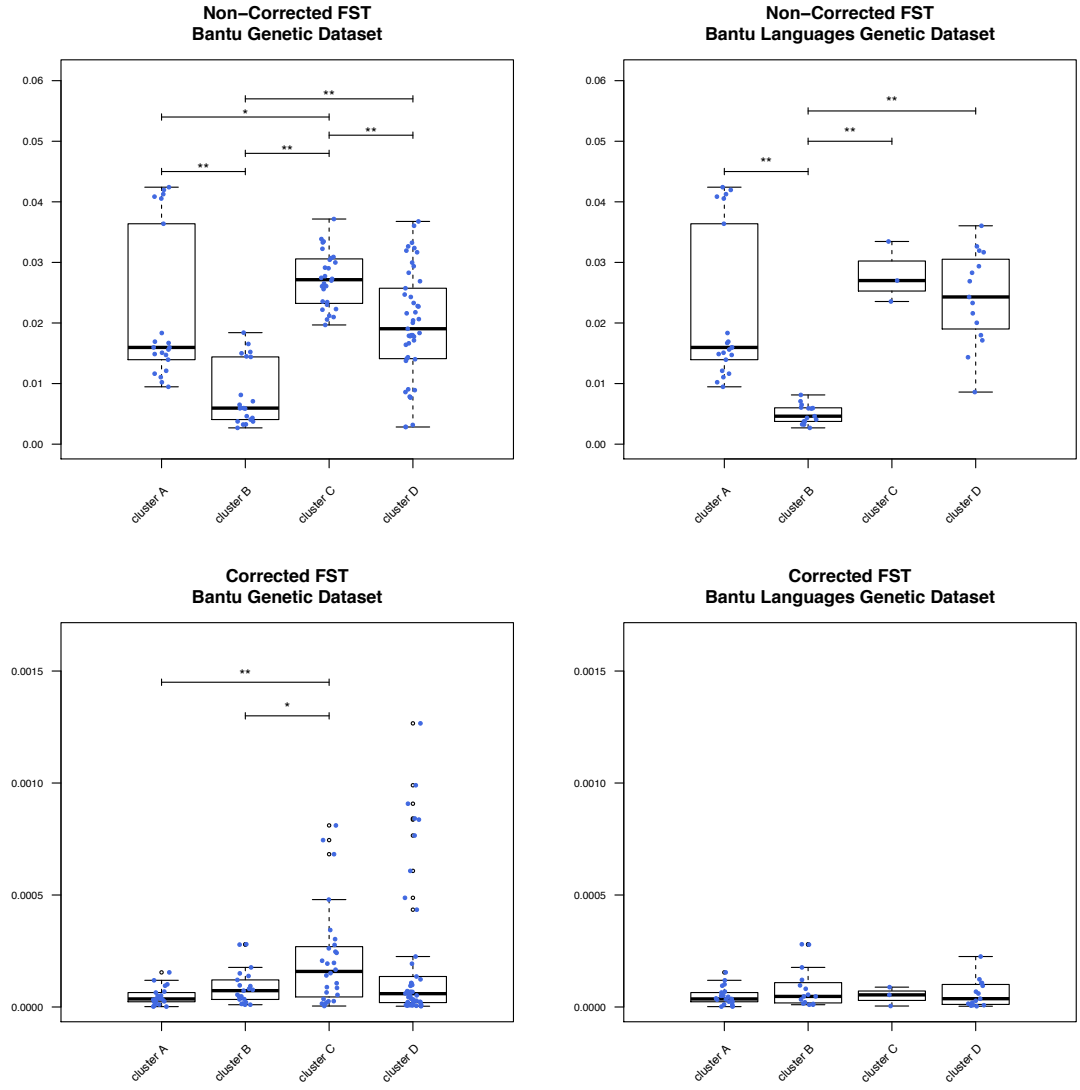


Figure 5.10 – Distribution of the pairwise genetic distances (F_{ST}) between populations within each cluster. Statistically significant differences after a Wilcoxon test are reported in the figure; *, *p-value*<0.05; **, adjusted *p-value*<0.0083, after Bonferroni correction.

speaking populations was being removed (Figure 5.10). Moreover, while the levels of genetic variation within each group also became much lower and more homogeneous, I still retain variability between groups and find a comparable larger variation when analysing these corrected genetic distances between each cluster and the other three (Figure 5.11).

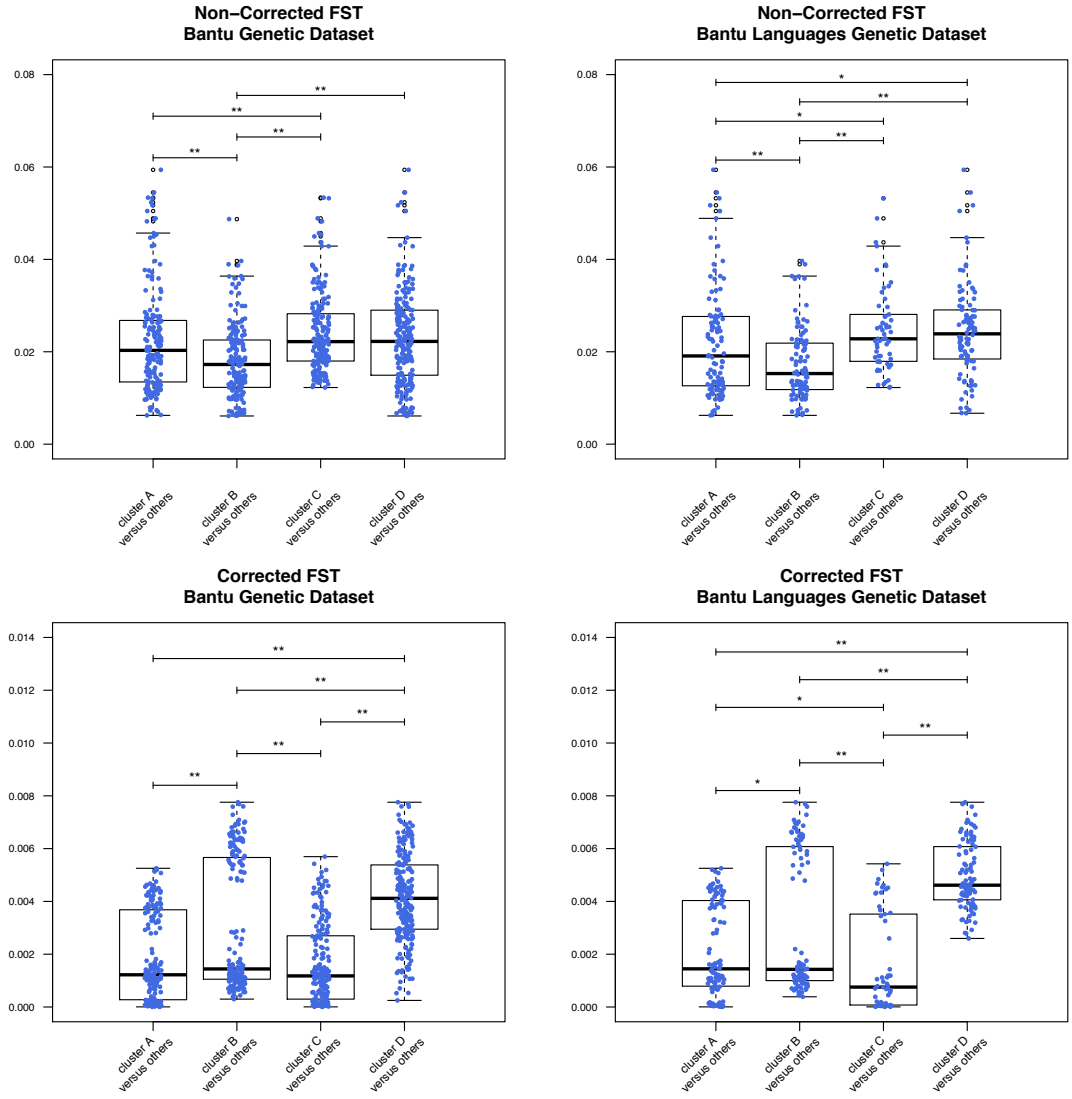


Figure 5.11 – Distribution of the pairwise genetic distances (F_{ST}) between each population from each cluster and all the populations from the other three clusters. Statistically significant differences after a Wilcoxon test are reported in the figure; *, *p-value*<0.05; **, adjusted *p-value*<0.0083, after Bonferroni correction.

As a way of further testing the validity of the correction method used here, I tested again the correlation between the new genetic distances[14] and both the geographic and the linguistic distances between

[14] For simplicity, the genetic distances after correction will be called 'corrected genetic distances' while the original distances (before the correction) will be called 'non-corrected genetic distances'.

pairs of populations. For the 'Bantu Languages Genetic Dataset' both these correlations were now significant (p<0.05) with a Spearman coefficient of correlation, ρ, of 0.72 (for the correlation between the corrected genetic distances and geography) and 0.52 (for the correlation of genetic and linguistic distances) (Figure 5.12). Fitting a linear model to all the correlations tested at this point the coefficient of determination, r^2, increases dramatically after the genetic correction. In the case of the correlation with the linguistic distances, the percentage of variation in these distances explained by the genetic distances increases from 2% to 17% after the correction (Figure 5.6, before, and Figure 5.12, after the correction). When analysing the correlation with geographic distances this is even more striking. Before the correction no statistically significant correlation between the two distances was observed, but the distribution of the corrected genetic distances and the geographic distances are now strongly correlated with more than 41% of the variation in the former being explained by the latter (Figure 5.6, before, and Figure 5.12, after the correction).

A similar pattern is also observed when considering the complete 'Bantu Genetic Dataset' for the correlation between genetic and geographical distances. As observed for the 'Bantu Languages Genetic Dataset', no statistically significant correlation between these two distances was observed before the correction to the genetic distances (Figure 5.6). However, after the correction 33% of the variation in the corrected genetic distances (measured by the coefficient of determination, r^2) is explained by geography (Figure 5.12).

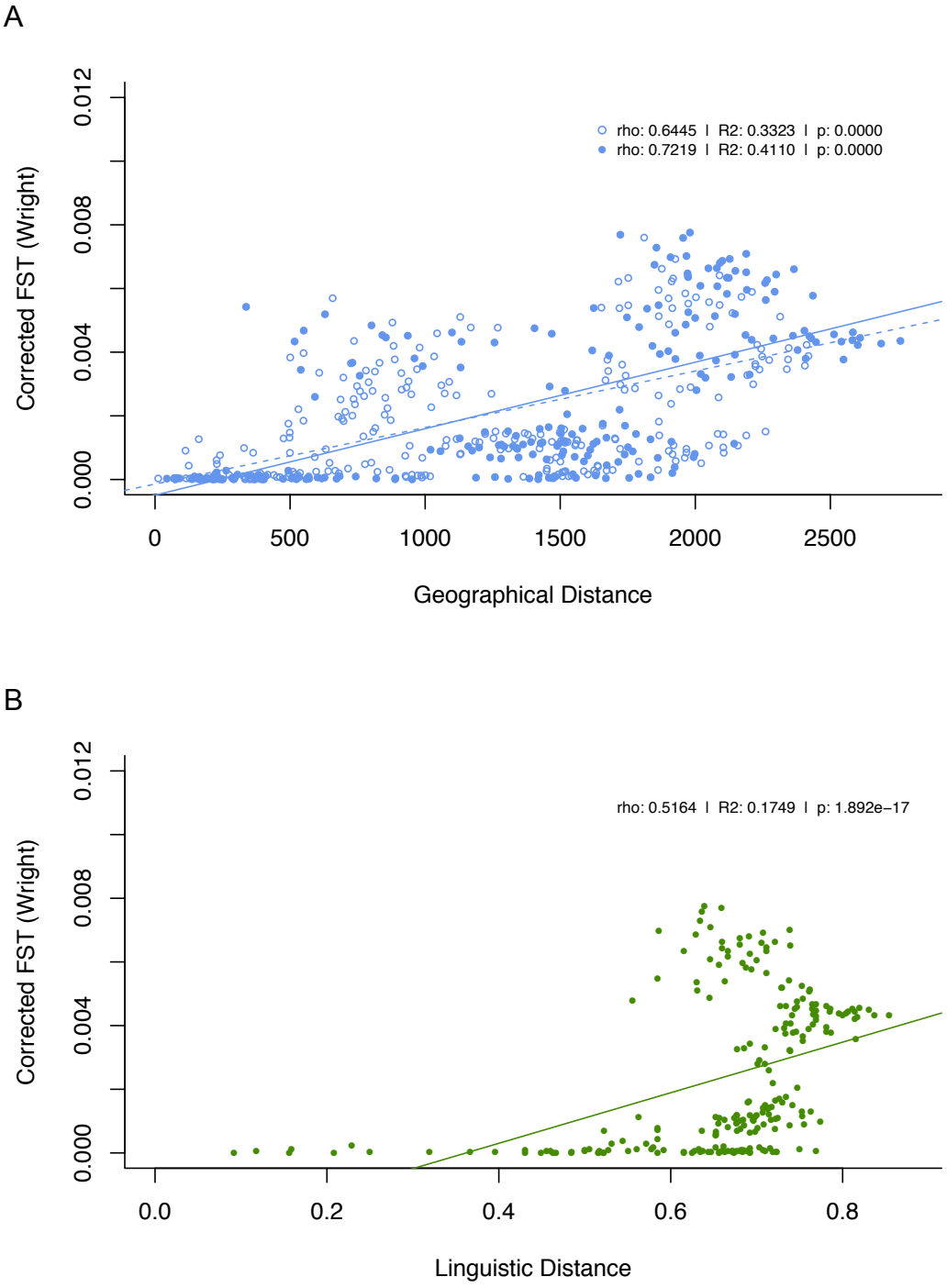


Figure 5.12 – Correlation between pairwise (A) geographical and (B) linguistic distances and the pairwise corrected genetic (F_{ST}) distances in Bantu-speaking populations. Open circles (in panel A) correspond to the complete BG dataset and full circles to the BLG dataset (see main text).

5.2.7 Model-based geographic distances

The geographical distances in all the aforementioned correlations tested were calculated as the GCD between all the possible pairs of sampling sites. However, this is not a true representation of the actual distance that needs to be travelled between any two sites, as these distances ignore the existence of deserts, forests, and other potential barriers to the movement of people across the African continent. As so, I decided to implement different models of geographical dispersion through the continent and thus get a better representation of the geographical distances between each pair of populations. As stated above, I performed a hierarchical clustering of the GCD between pairs of populations in the dataset in order to identify geographical clusters of populations (Figure 5.2A). I then divided the populations in my datasets into four geographical clusters – '*cluster A*' (west Africa); '*cluster B*' (east Africa); '*cluster C*' (Namibia, Zambia, and Zimbabwe); and '*cluster D*' (south Africa). Given the geographical location of these clusters in the African continent, I then identified one or more waypoints to represent the borders of each cluster, corresponding, as appropriate, to the combination of either the northernmost or the southernmost population coordinates with the easternmost or the westernmost coordinates in that cluster (Figure 5.2B). With this model-based geographic distances calculation I tried to get a more accurate measure of the real distance of different possible routes more likely taken by Bantu-speaking populations during their migratory path, avoiding, for

example, geographical barriers that might have hindered their movement across the continent.

I then selected different paths connecting the various clusters via the waypoints and use these combinations as a way to test for possible paths of dispersal. The different dispersal models are summarised in Figure 5.13. For each model the geographical distances for populations within each cluster were calculated as GCD while the distances between populations belonging to different clusters was calculated as the sum of the distances between each population to an appropriate waypoint in its cluster and the distance between different waypoints, connecting clusters according to the tested model. The way the different clusters are connected in each model is explained and illustrated in detail below (see also Figure 5.13).

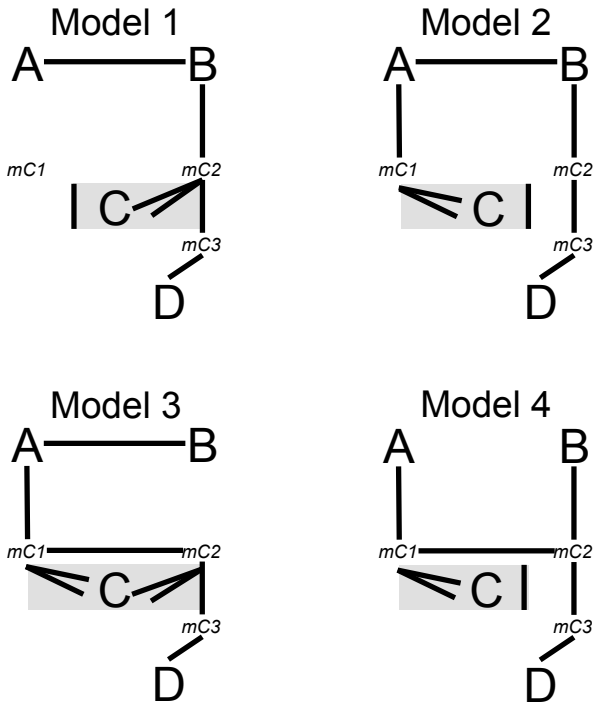


Figure 5.13 – Schematic representation of *models 1*, *2*, *3*, and *4* used for the calculation of the model-based pairwise geographical distances. Detailed maps for each model are included in Supplementary Material.

5.2.7.1 Model description

5.2.7.1.1 Model 1

The first model tested is described in Supplementary Material 6. In this model I assumed that Bantu populations would have had to travel from their place of origin in central-west Africa to the southern regions of the continent through the east region of the continent, where the Great Lakes are located. From there they would go south avoiding the Equatorial Forest and, after this barrier was left behind, back west.

5.2.7.1.2 Model 2

On the other hand, the second model tested assumes that Bantu-speaking populations could have travelled directly south from their homeland in central-west Africa through the coastline. It also assumes a path from the Bantu homeland in the direction of east Africa, which is then connected to populations in the southern regions of the continent. This model is represented in Supplementary Material 7.

5.2.7.1.3 Model 3

While the last two models include *cluster B*, in east Africa, as a midpoint in the migration of Bantu-speaking populations to the south, *model 3* (represented in Supplementary Material 8) assumes that east Africa was the endpoint of part of the migration of these populations. In this case,

populations would have expanded in the direction of east Africa directly from the Bantu homeland and no migration to the south would have happened from there. Instead, populations would migrate south directly from central west Africa. Here, as the Kalahari Desert could act like a difficult barrier to transpose by these population movements, populations from the south (*cluster D*) are connected to central-west Africa via *cluster C*.

5.2.7.1.4 Model 4

The final model tested, *model 4*, is described in Supplementary Material 9. As with *model 3* east Africa is also the end point of one of the migration paths in this case. However, in this model, no direct link between central west and east Africa is present and populations are assumed to have only travelled south from their homeland. According to this model, after the equatorial forest barrier was crossed there would have occurred a migration east and, then, back north to the region of the Great Lakes.

5.2.7.2 Model testing

I tested the correlation between the newly measured model-based geographical distances for the four models with both the linguistic (Table 5.4) and the genetic distances (Table 5.5).

As with the GCD, the correlations between the linguistic distances and the geographical distances were significant (p<0.05) for the four models

Table 5.4 – Coefficient of determination (r^2) for the correlation between the linguistic distances and the pairwise great-circle distances (GCD) and distances based on *models 1*, *2*, *3*, and *4*.

GCD	0.3376
model 1	**0.4542**
model 2	**0.4589**
model 3	0.2768
model 4	**0.4699**

Table 5.5 – Coefficient of determination (r^2) for the correlation between the non-corrected genetic distances and the pairwise great-circle distances (GCD) and distances based on *models 1*, *2*, *3*, and *4*.

GCD	*Bantu Languages Genetic Dataset*	*n.s.*
	Bantu Genetic Dataset	*n.s.*
model 1	*Bantu Languages Genetic Dataset*	0.0474
	Bantu Genetic Dataset	0.0122
model 2	*Bantu Languages Genetic Dataset*	0.0615
	Bantu Genetic Dataset	0.0241
model 3	*Bantu Languages Genetic Dataset*	0.0069
	Bantu Genetic Dataset	*n.s.*
model 4	*Bantu Languages Genetic Dataset*	*n.s.*
	Bantu Genetic Dataset	*n.s.*

tested. After fitting a linear model to all the correlations, the coefficient of determination, r^2, was compared between the model-based distances and the GCD to evaluate the improvement on the linguistic variation explained by the former over the latter. As can be seen in Table 5.4, the amount of variation explained through the model-based distances improves over the GCD for all models but *model 3*. Even though *model 4* is the one for which the amount of variation explained through geography is the largest overall, the values of the coefficient of determination for the three models showing

an improvement over CGD are all very similar (between 45.4% and 47.0% – Table 5.4).

As mentioned before, there was no significant correlation between the GCD and the non-corrected genetic distances. In this context, the model-based distances seem to show an improvement as for three out of the four models tested (*models 1* and *2* for both datasets, and *model 3* in the case of the 'Bantu Languages Genetic Dataset') this correlation is now significant ($p<0.05$). However, even in these cases the observed correlation is very weak – with only 0.7–6% of the genetic variation being explained by geography (Table 5.5).

However, as seen above, the correction to the genetic distances I applied before in this chapter led to a significant correlation with the GCD, explaining 33% and 41% of the variation for the 'Bantu Genetic Dataset' and the 'Bantu Languages Genetic Dataset', respectively (Figure 5.12). As such, the same corrected genetic distances were also tested against the model-based distances for the aforementioned models[15] (Table 5.6). In this case even though all the correlations with the model-based distances were again significant ($p<0.05$), they were not always an improvement when compared to the GCD – in both datasets a decrease in the percentage of variation is

[15] As stated before, all correlations between genetic and geography distances (including all corrections and geography models tested) were performed on both the complete 'Bantu Genetic Dataset' and its subset the 'Bantu Languages Genetic Dataset' for a more accurate comparison with the correlation with linguistic distances.

Table 5.6 – Coefficient of determination (r^2) for the correlation between the corrected genetic distances and the pairwise great-circle distances (GCD) and distances based on *models 1*, *2*, *3*, and *4*.

GCD	*Bantu Languages Genetic Dataset*	0.4110
	Bantu Genetic Dataset	0.3323
model 1	*Bantu Languages Genetic Dataset*	0.2685
	Bantu Genetic Dataset	0.2218
model 2	*Bantu Languages Genetic Dataset*	**0.4784**
	Bantu Genetic Dataset	0.3257
model 3	*Bantu Languages Genetic Dataset*	**0.7955**
	Bantu Genetic Dataset	**0.6794**
model 4	*Bantu Languages Genetic Dataset*	0.2614
	Bantu Genetic Dataset	0.2543

observed for *models 1* and *4*, and for the complete 'Bantu Genetic Dataset' this is also true for *model 2*. However, the amount of variation explained by model-based geographic distances almost doubled in the case of *model 3*, reaching values as high as 79.6% for the 'Bantu Languages Genetic Dataset', Table 5.6).

5.2.7.3 Model refinement

One of the geographical clusters, *cluster C*, contains populations belonging both to the East and West branches of the Narrow Bantu linguistic family (Table 5.1). As such, I was able to divide this cluster into two subgroups (*cluster C East* and *cluster C West*) by taking into account the linguistic information of the samples. In this context, I made some refinements to the models tested above by taking this into account (Figure 5.14). I focused these refinements on *models 2*, *3*, and *4*, as these are the ones explaining the higher levels of variation in both linguistic and corrected

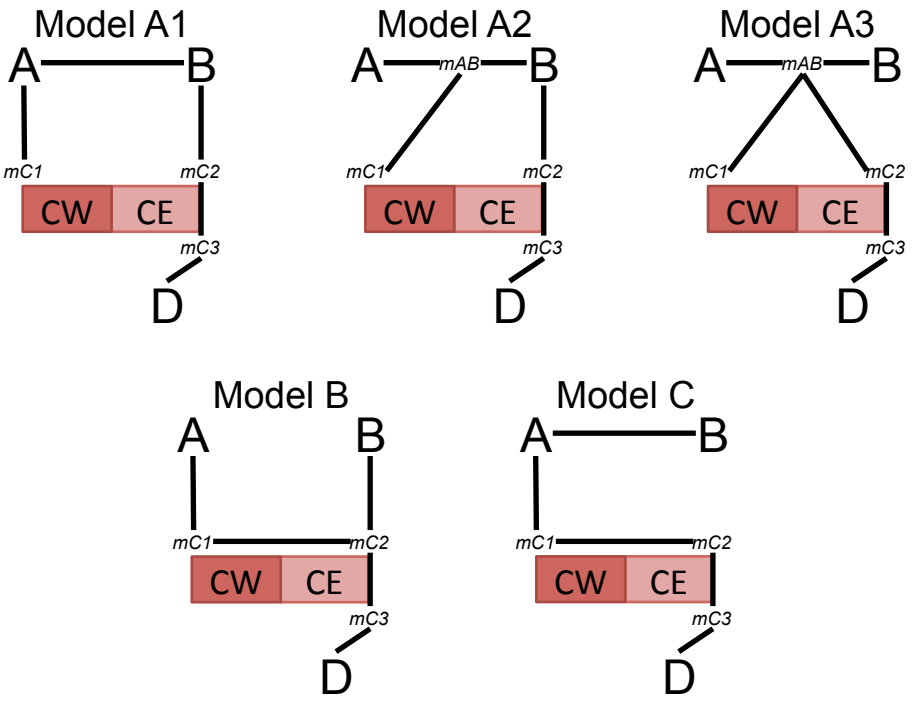


Figure 5.14 – Schematic representation of *models A1*, *A2*, *A3*, *B,* and *C* used for the calculation of the model-based pairwise geographical distances. Detailed maps for each model are included in Supplementary Material.

genetic distances (Table 5.4 and 5.6, respectively). However, it is important to stress that no populations from the East branch of the Narrow Bantu linguistic family are included in the 'Bantu Languages Genetic Dataset' and thus the refined *models A1*, *B*, and *C*, are, when using this dataset, identical to *models 2*, *4*, and *3*, respectively (Figure 5.13 and 5.14).

5.2.7.3.1 Model A1

Model A1 is designed as a refinement of the previously tested *model 2* and is illustrated in Supplementary Material 10. This refinement subdivides *cluster C* into two different clusters (*cluster C West* and *cluster C East*), according to their linguistic affiliation. While populations from the Western branch of the Bantu linguistic family in *cluster C* are connected, as

on *model 2*, directly to *cluster A* in central-west Africa, populations from the Eastern branch are now connected to the Bantu homeland via east Africa, through waypoint *mB*.

5.2.7.3.2 Model A2

As with *model A1*, *model A2* is also a refinement of *model 2*. While similar to the previous model, in this case a new waypoint (waypoint *mAB*) was included positioned halfway between waypoint *mA* and waypoint *mB* (Supplementary Material 11). Populations from the West branch of the Bantu linguistic family that form *cluster C* (*cluster C West*) are, in this model, connected via this new waypoint.

5.2.7.3.3 Model A3

Model A3 aims to be a further refinement of *model A2*. As with the previous model, a new waypoint was included (waypoint *mAB*), located halfway between waypoint *mA* and waypoint *mB*. However, contrary to what happened in the previous model, this new waypoint connects not only *cluster C West*, but *clusters A*, *B*, and *C East* are all connected to each other through this waypoint (Supplementary Material 12).

5.2.7.3.4 Model B

Model B is designed as a refinement of *model 4* by also subdividing *cluster C* in two different clusters according to the linguistic

affiliation regarding the West and East branches of the Bantu linguistic family. In this model, populations from *cluster C East* are now connected to the other clusters by the waypoint *mC2* while populations from *cluster C West* are, as in *model 4*, connected through the waypoint *mC1* (Supplementary Material 13).

5.2.7.3.5 Model C

Finally, *model C* is a refinement of the previous *model 3* by taking into account the same division of *cluster C* explained before. As with the previous model, the difference between this and the original *model 3* is that populations from *cluster C East* are now connected to the other clusters via the waypoint *mC2* (Supplementary Material 14).

5.2.7.4 Testing the refined models

These five new refined models were tested to assess the differences in the amount of variation of both linguistic and genetic distances explained by geography. As before, there is a generalized improvement in the correlation – as measured by the coefficient of determination, r^2 – when comparing the refined models with the GCD (Table 5.7 for linguistic distances, and Table 5.8 and 5.9 for genetic distances, before and after correction, respectively).

For the correlation with the linguistic distances, the only exception to this pattern is observed for one model (*model C*) for which a decrease in

Table 5.7 – Coefficient of determination (r^2) for the correlation between the linguistic distances and the pairwise great-circle distances (GCD) and distances based on *models 1*, *2*, *3*, *4*, *A1*, *A2*, *A3*, *B*, and *C*.

GCD	0.3376
model 1	**0.4542**
model 2	**0.4589**
model 3	0.2768
model 4	**0.4699**
model A1	**0.4589**
model A2	**0.5011**
model A3	**0.4602**
model B	**0.4699**
model C	0.2933

the coefficient of determination is observed (Table 5.7). Not surprisingly, the model for which this exception is observed is a refinement of the only model (*model 3*) for which a decrease in the coefficient was also observed as described above (Table 5.4).

As before, even though the non-corrected genetic distances (contrary to the GCD) show significant (p<0.05) correlations for some of the models tested, these correlations are always weak, with the coefficient of determination ranging between 1.1% and 7.3% (Table 5.8). The corrected genetic distances also show a similar pattern to the one observed for the four initial models with a dramatic improvement of the coefficient over the one observed for the non-corrected genetic distances. Even though this is not always the case when comparing the model-based distances with the GCD for the 'Bantu Genetic Dataset' (only two out of the five refined models – *model A3* and *model C* – show an improvement over GCD), for the 'Bantu Languages Genetic Dataset' this improvement is observed in four of the five models. Moreover, the correlation between the corrected genetic distances

Table 5.8 – Coefficient of determination (r^2) for the correlation between the non-corrected genetic distances and the pairwise great-circle distances (GCD) and distances based on *models 1*, *2*, *3*, *4*, *A1*, *A2*, *A3*, *B*, and *C*.

GCD		
	Bantu Languages Genetic Dataset	*n.s.*
	Bantu Genetic Dataset	*n.s.*
model 1	*Bantu Languages Genetic Dataset*	0.0474
	Bantu Genetic Dataset	0.0122
model 2	*Bantu Languages Genetic Dataset*	0.0615
	Bantu Genetic Dataset	0.0241
model 3	*Bantu Languages Genetic Dataset*	0.0069
	Bantu Genetic Dataset	*n.s.*
model 4	*Bantu Languages Genetic Dataset*	*n.s.*
	Bantu Genetic Dataset	*n.s.*
model A1	*Bantu Languages Genetic Dataset*	0.0615
	Bantu Genetic Dataset	0.0361
model A2	*Bantu Languages Genetic Dataset*	0.0732
	Bantu Genetic Dataset	0.0431
model A3	*Bantu Languages Genetic Dataset*	0.0455
	Bantu Genetic Dataset	0.0203
model B	*Bantu Languages Genetic Dataset*	*n.s.*
	Bantu Genetic Dataset	*n.s.*
model C	*Bantu Languages Genetic Dataset*	0.0110
	Bantu Genetic Dataset	*n.s.*

and the model-based geographic distances shows the strongest correlation of all the ones tested until this point, with 82% of the genetic variation being explained by geography for *model C* (for the 'Bantu Languages Genetic Dataset').

On the whole, it appears that genetics and linguistics seem to behave differently in relation to geography. Without taking into account the non-corrected genetic distances – due to a lack of or weak correlation with

Table 5.9 – Coefficient of determination (r^2) for the correlation between the corrected genetic distances and the pairwise great-circle distances (GCD) and distances based on *models 1*, *2*, *3*, *4*, *A1*, *A2*, *A3*, *B*, and *C*.

GCD	*Bantu Languages Genetic Dataset*	0.4110
	Bantu Genetic Dataset	0.3323
model 1	*Bantu Languages Genetic Dataset*	0.2685
	Bantu Genetic Dataset	0.2218
model 2	*Bantu Languages Genetic Dataset*	**0.4784**
	Bantu Genetic Dataset	0.3257
model 3	*Bantu Languages Genetic Dataset*	**0.7955**
	Bantu Genetic Dataset	**0.6794**
model 4	*Bantu Languages Genetic Dataset*	0.2614
	Bantu Genetic Dataset	0.2543
model A1	*Bantu Languages Genetic Dataset*	**0.4784**
	Bantu Genetic Dataset	0.3302
model A2	*Bantu Languages Genetic Dataset*	**0.4427**
	Bantu Genetic Dataset	0.3146
model A3	*Bantu Languages Genetic Dataset*	**0.6169**
	Bantu Genetic Dataset	**0.4813**
model B	*Bantu Languages Genetic Dataset*	0.2614
	Bantu Genetic Dataset	0.2615
model C	*Bantu Languages Genetic Dataset*	**0.8206**
	Bantu Genetic Dataset	**0.6646**

geography – the geographical model that seems to better explain the genetic distances variation is *model C*, with over 82% of the genetic variation (over 66% in the case of the 'Bantu Genetic Dataset'[16]) being explained by geography (Table 5.9). However, in the case of the variation in

[16] However, for the 'Bantu Genetic Dataset', geographical *model C* is not the one that explains the highest proportion of the variation in the corrected genetic distances, even if by a small margin. The best model is instead *model 3* – the one from which *model C* is a refinement – with geography explaining close to 68% of the variation in the genetic distances.

the linguistic distances, this model actually leads to a decrease of the coefficient of correlation when compared to the GCD. In the case of linguistic distances, the best model is instead *model A2*, with just over 50% of the variation explained by geography (Table 5.7).

5.2.7.5 Geographical distribution of the linguistic and genetic distances

The linguistic and genetic relationships between populations belonging to the same cluster can be very different across clusters. Geographical distances within clusters are similar across clusters, with clusters in the 'Bantu Languages Genetic Dataset' not showing any significant (after Bonferroni correction, adjusted *p-value*<0.0083) differences among themselves (while for the 'Bantu Genetic Dataset' *cluster C* does show significant (after Bonferroni correction, adjusted *p-value*<0.0083) differences with both *clusters B* and *D*) (Figure 5.3). This homogeneity in the geographical distances within the different clusters seems to indicate that I do not need to control for differences within clusters when interpreting the results for the different models. On the other hand, this does not seem to be the case for linguistic and genetic distances (Figure 5.15 and 5.10, respectively). Linguistic distances within clusters present very different ranges. C*luster A* shows the highest values among all the clusters, showing significant differences (after Bonferroni correction, adjusted *p-value*<0.0083) with all except *cluster C* (Figure 5.15). However, it is easily noticeable on the plot that *cluster C* does indeed seem to present much lower linguistic

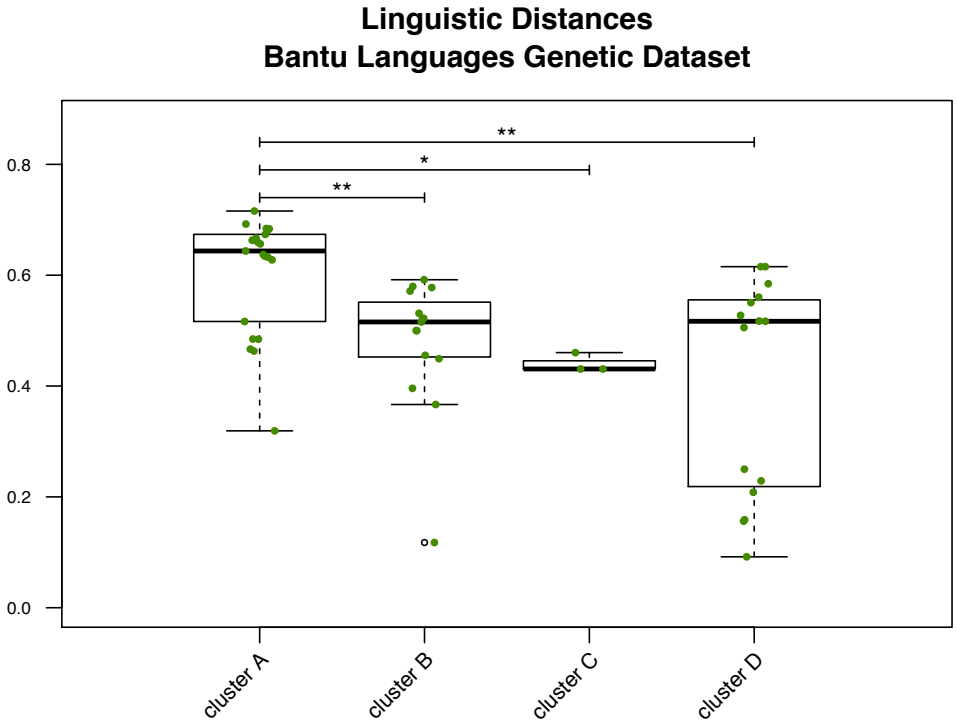


Figure 5.15 – Distribution of the pairwise linguistic distances between populations within each cluster. Statistically significant differences after a Wilcoxon test are reported in the figure; *, *p-value*<0.05; **, adjusted *p-value*<0.0083, after Bonferroni correction.

distances than *cluster A*. It is also important to note that for the BLG dataset, *cluster C* only includes three populations, which makes it very difficult to obtain statistically significant values with it. However, for the genetic distances the scenario observed is more complex than that for languages. Populations from each cluster show, before the correction method applied, higher degrees of variation of the distribution of pairwise F_{ST} distances amongst themselves, with the exception of *cluster B*. This more homogeneous cluster also presents the lowest pairwise F_{ST} among its populations of the four clusters (Figure 5.10).

When comparing the pairwise F_{ST} distances between populations from each cluster and populations from the other three, *cluster D* is the one for which the highest distances are observed, especially for the corrected genetic distances for which all comparisons between *cluster D* and the other clusters are statistically significant (after Bonferroni correction, adjusted *p-value*<0.0083) (Figure 5.11).

5.2.7.6 Modifications to the geographical models

I investigated if this particular behaviour by *cluster D* could explain the results testing a correlation between genetics and geography. The geographical model that explained the highest amount of genetic variation is, as stated before, *model C* (or the related unrefined geographical *model 3* in the case of the corrected genetic distances for the complete BG dataset). Interestingly, these two models are among those in which *cluster D* also shows larger geographical distances to all the other clusters (Figure 5.16).

Taking into consideration the different patterns of variation observed in the different distances I made a fixed set of modifications to all the models tested above as a way to assess if differences within and/or between clusters were in any way influencing the models being assessed. As mentioned, *clusters A* and *D* show the most divergent and interesting patterns. As mentioned above, while the geographic distances within both clusters do not seem to differ, populations from *cluster A* show the highest levels of linguistic distances within a cluster and populations from *cluster D*

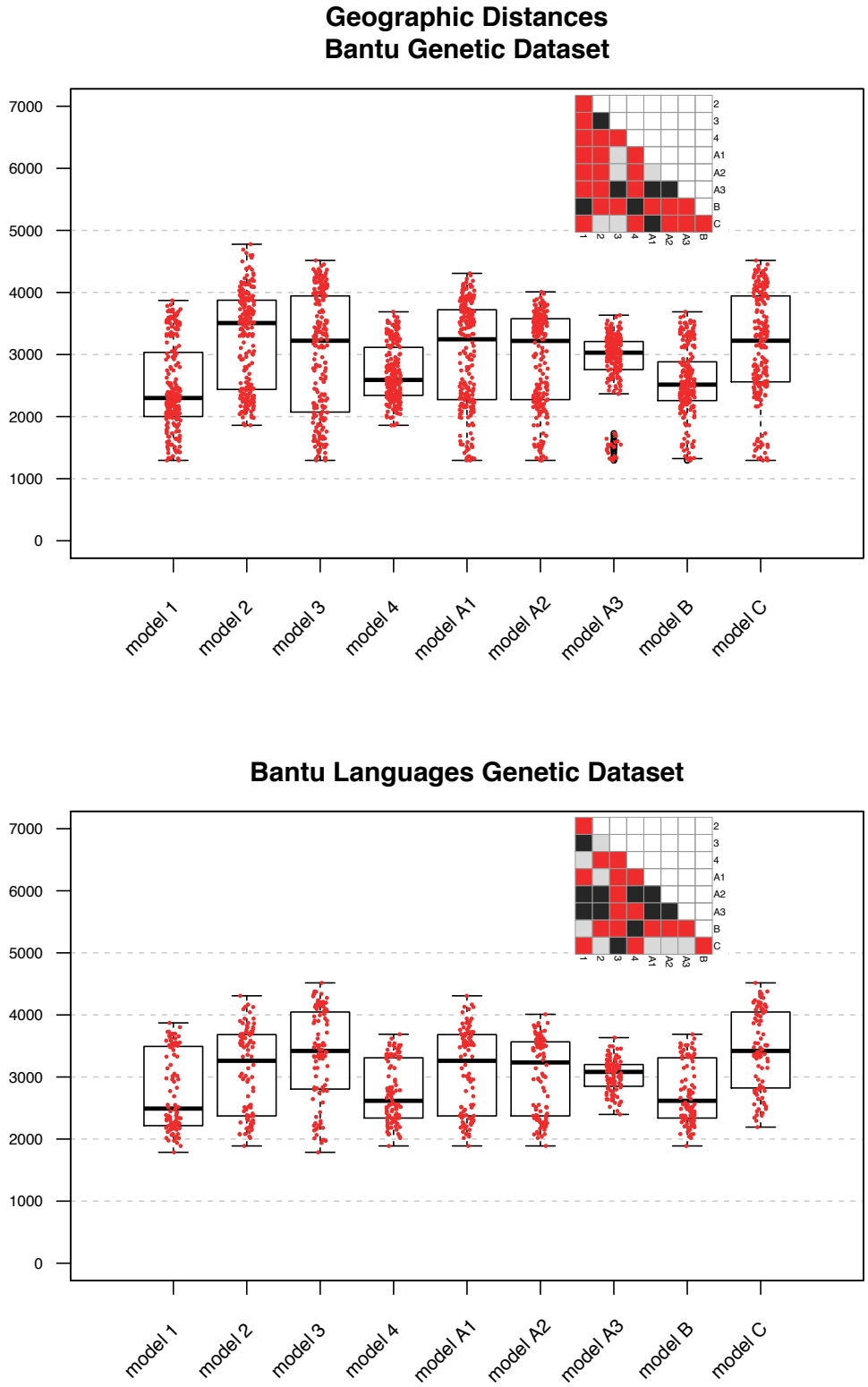


Figure 5.16 – Distribution of the pairwise geographic distances between each population from cluster D and all populations from the other 3 clusters for the nine models tested. Statistically significant differences after a Wilcoxon test are reported in a coloured matrix for each plot: grey, *p-value*>0.05; black, *p-value*<0.05; red, adjusted *p-value*<0.0083, after Bonferroni correction.

show the highest genetic distances to the other clusters (particularly after the genetic correction). As such, some of the tested modifications to the models expressly target these two clusters. For the nine models described and analysed above, the modifications performed were: (*i.-iv.*) remove, consecutively, the comparisons between populations within each cluster; (*v.*) remove all the comparisons between populations within a same cluster; (*vi.*) remove comparisons between populations within *cluster A* and within *cluster D*; (*vii.*) double the distances between *cluster D* and the three other clusters; (*viii.*) both remove comparisons within *cluster A* and within *cluster D*, and double the distances from all populations to populations in *cluster D* (tests *vi.* and *vii.* combined); (*ix.-xiii.*) increase the distance between *cluster D* and *cluster C* (through the waypoint *mC3*) by 100, 500, 1000, 5000, and 10,000 km; and (*xiv.*) completely remove *cluster D* from the analyses.

As the correlations with genetic distances before correction were either non-significant or very weak I decided to focus only on the corrected genetic distances for the analyses of these modifications to the models. The results from these modifications on the different models are summarized on Table 5.10 and 5.11 for the correlations of geographical distances with linguistic and corrected genetic distances, respectively. Languages and genetics seem to show a different response regarding these modifications. Linguistically, the model yielding the higher proportion of variation explained by geography was *model A2* (the same as pre-modifications) when simultaneously removing comparisons within *cluster A* and within *cluster D* (Table 5.10; modification *vi.*, when this model explains almost 57% of the

Table 5.10 – Coefficient of determination (r^2) for the correlation between the linguistic distances and the pairwise great-circle distances (GCD) and distances based on all models tested (including the modifications *i.-xiv.* tested for each model). Modifications: (*i.-iv.*) remove the comparisons between populations within (*i.*) cluster *A*, (*ii.*) cluster *B*, (*iii.*) cluster *C*, and (*iv.*) cluster *D*; (*v.*) remove all the comparisons between populations within a same cluster; (*vi.*) remove comparisons between populations within *cluster A* and within *cluster D*; (*vii.*) double the distances between *cluster D* and the three other clusters; (*viii.*) remove comparisons within both *cluster A* and *cluster D* and double the distances from all populations to populations in *cluster D*; (*ix.-xiii.*) increase the distance between *cluster D* and *cluster C* (through the waypoint *mC3*) by (*ix.*) 100, (*x.*) 500, (*xi.*) 1000, (*xii.*) 5000, and (*xiii.*) 10,000 km; (*xiv.*) completely remove *cluster D* from the analyses.

Modfications	GCD	model 1	model 2	model 3	model 4	model A1	model A2	model A3	model B	model C
Pre-modifications	0.3376	**0.4542**	**0.4589**	0.2768	**0.4699**	**0.4589**	**0.5011**	**0.4602**	**0.4699**	0.2933
i.	-	***0.5046***	***0.5000***	***0.2821***	***0.5307***	***0.5000***	***0.5596***	***0.5180***	***0.5307***	***0.3020***
ii.	-	0.3965	0.4081	0.2101	0.4127	0.4081	0.4527	0.4010	0.4127	0.2265
iii.	-	0.4388	0.4452	0.2614	0.4542	0.4452	0.4872	0.4444	0.4542	0.2777
iv.	-	0.4473	***0.4719***	0.2271	0.4618	***0.4719***	***0.5206***	0.4479	0.4618	0.2459
v.	-	0.1602	0.2756	*n.s.*	0.1617	0.2756	0.3384	0.1314	0.1617	*n.s.*
vi.	-	***0.4738***	***0.4958***	0.1926	***0.5011***	***0.4958***	***0.5691***	***0.4812***	***0.5011***	0.2143
vii.	-	0.3746	0.3452	0.2053	0.4248	0.3452	0.3657	0.3254	0.4248	0.2162
viii.	-	0.3777	0.3534	0.1434	0.4477	0.3534	0.3787	0.3011	0.4477	0.1563
ix.	-	***0.4566***	0.4539	***0.2771***	***0.4761***	0.4539	0.4951	0.4539	***0.4761***	0.2926
x.	-	***0.4558***	0.4324	0.2762	***0.4875***	0.4324	0.4690	0.4279	***0.4875***	0.2885
xi.	-	***0.4411***	0.4061	0.2723	***0.4810***	0.4061	0.4371	0.3981	***0.4810***	0.2816
xii.	-	0.3159	0.2921	0.2377	0.3435	0.2921	0.3030	0.2827	0.3435	0.2403
xiii.	-	0.2609	0.2483	0.2187	0.2761	0.2483	0.2537	0.2421	0.2761	0.2200
xiv.	-	0.3394	0.3407	***0.3407***	0.3681	0.3407	0.4068	0.4068	0.3681	***0.3407***

Table 5.11 – Coefficient of determination (r^2) for the correlation between the corrected genetic distances and the pairwise great-circle distances (GCD) and distances based on all models tested (including the modifications *i.-xiv.* tested for each model). Modifications: (*i.-iv.*) remove the comparisons between populations within (*i.*) cluster *A*, (*ii.*) cluster *B*, (*iii.*) cluster *C*, and (*iv.*) cluster *D*; (*v.*) remove all the comparisons between populations within a same cluster; (*vi.*) remove comparisons between populations within *cluster A* and within *cluster D*; (*vii.*) double the distances between *cluster D* and the three other clusters; (*viii.*) remove comparisons within both *cluster A* and *cluster D* and double the distances from all populations to populations in *cluster D*; (*ix.-xiii.*) increase the distance between *cluster D* and *cluster C* (through the waypoint *mC3*) by (*ix.*) 100, (*x.*) 500, (*xi.*) 1000, (*xii.*) 5000, and (*xiii.*) 10,000 km; (*xiv.*) completely remove *cluster D* from the analyses.

Modfications	GCD		model 1		model 2		model 3		model 4		model A1		model A2		model A3		model B		model C	
	BLG	*BG*	*BLG*	*BG*	*BLG*	*BG*	*BLG*	*BG*	*BLG*	*BG*	*BLG*	*BG*	*BLG*	*BG*	*BLG*	*BG*	*BLG*	*BG*	*BLG*	*BG*
Pre-modifications	0.4110	0.3323	0.2685	0.2218	**0.4784**	0.3257	**0.7955**	**0.6794**	0.2614	0.2543	**0.4784**	0.3302	**0.4427**	0.3146	**0.6169**	**0.4813**	0.2614	0.2615	**0.8206**	**0.6646**
i.	-	-	0.1989	0.1921	0.4250	0.2981	0.7813	0.6669	0.1901	0.2239	0.4250	0.3025	0.3855	0.2858	0.5852	0.4606	0.1901	0.2312	0.8113	0.6526
ii.	-	-	0.2199	0.1915	0.4432	0.2981	0.7909	0.6685	0.2120	0.2238	0.4432	0.3025	0.4053	0.2860	0.6017	0.4628	0.2120	0.2312	0.8200	0.6548
iii.	-	-	0.2597	0.1871	0.4718	0.2931	0.7937	0.6643	0.2525	0.2185	0.4718	***0.3450***	0.4357	***0.3196***	0.6130	***0.4886***	0.2525	0.2388	0.8195	0.6584
iv.	-	-	0.2207	0.1578	0.4423	0.2677	0.7869	0.6627	0.2126	0.1900	0.4423	0.2723	0.4042	0.2551	0.5967	0.4500	0.2126	0.1977	0.8155	0.6525
v.	-	-	0.0224	0.0125	0.2711	0.1066	0.7819	0.6093	*n.s.*	0.0220	0.2711	0.1655	0.2217	0.1314	0.5951	0.4199	*n.s.*	0.0432	***0.8400***	0.6222
vi.	-	-	0.1300	0.1169	0.3698	0.2279	0.7713	0.6469	0.1196	0.1464	0.3698	0.2321	0.3264	0.2133	0.5605	0.4253	0.1196	0.1541	0.8079	0.6398
vii.	-	-	***0.5957***	***0.5432***	***0.6361***	***0.4500***	***0.9197***	***0.8418***	***0.6505***	***0.5539***	***0.6361***	***0.5092***	***0.6357***	***0.5206***	***0.8190***	***0.7014***	***0.6505***	***0.5940***	***0.9343***	***0.8426***
viii.	-	-	***0.5107***	***0.4761***	***0.5621***	***0.3731***	***0.9042***	***0.8228***	***0.5794***	***0.4902***	***0.5621***	***0.4382***	***0.5601***	***0.4500***	***0.7844***	***0.6641***	***0.5794***	***0.5383***	***0.9227***	***0.8270***
ix.	-	-	***0.2972***	***0.2491***	***0.5001***	***0.3433***	***0.8089***	***0.6972***	***0.2915***	***0.2809***	***0.5001***	***0.3514***	***0.4664***	***0.3374***	***0.6376***	***0.5038***	***0.2915***	***0.2904***	***0.8320***	***0.6817***
x.	-	-	***0.4020***	***0.3503***	***0.5727***	***0.4042***	***0.8499***	***0.7519***	***0.4049***	***0.3773***	***0.5727***	***0.4245***	***0.5462***	***0.4162***	***0.7027***	***0.5751***	***0.4049***	***0.3955***	***0.8662***	***0.7341***
xi.	-	-	***0.5067***	***0.4506***	***0.6382***	***0.4625***	***0.8813***	***0.7925***	***0.5220***	***0.4712***	***0.6382***	***0.4930***	***0.6182***	***0.4893***	***0.7554***	***0.6324***	***0.5220***	***0.4967***	***0.8915***	***0.7734***
xii.	-	-	***0.7801***	***0.6813***	***0.7998***	***0.6277***	***0.9152***	***0.8148***	***0.8196***	***0.6858***	***0.7998***	***0.6631***	***0.7948***	***0.6639***	***0.8524***	***0.7248***	***0.8196***	***0.7082***	***0.9136***	***0.8009***
xiii.	-	-	***0.8269***	***0.7062***	***0.8306***	***0.6672***	***0.8992***	***0.7809***	***0.8556***	***0.7086***	***0.8306***	***0.6914***	***0.8281***	***0.6915***	***0.8598***	***0.7245***	***0.8556***	***0.7219***	***0.8971***	***0.7721***
xiv.	-	-	0.0738	0.0368	***0.5207***	***0.5492***	0.5207	0.5492	***0.7761***	***0.7609***	***0.5207***	0.2723	0.2998	0.1430	0.2998	0.2171	***0.7761***	***0.6121***	0.5207	0.4755

genetic variation). However, while for this model the removal of comparisons within *cluster A* only (modification *i.*) generated very similar results (approximately 56% of the linguistic variation explained), only removing comparisons within *cluster D* (modification *iv.*) gave a result closer to pre-modifications (52% of the linguistic variation explained). Focusing only on removing population comparisons within *cluster A* (near the Bantu languages homeland, the most linguistically diverse region occupied by Bantu speakers) leads to an overall increase of the proportion of variation explained by geography across the models (Table 5.10, *modification i.*). Apart from the already mentioned *model A2*, *model 4* (and *B* as both are identical for the BLG dataset; see section 5.2.7.3) shows the second highest amount of variation explained (53%).

The modification that yields the highest increase for the correlation of geography with the corrected genetic distances is doubling the distances between *cluster D* and the other 3 clusters – reaching values as high as 93.4% of the corrected genetic distances variation being explained by *model C* (Table 5.11). It should also be mentioned that when incrementally increasing the distance to *cluster D*, the coefficient of determination also increases accordingly. In the case of both *model 3* and *model C* this increase is only observed until I increase the distance by 5000 km starting to decrease again when I increase the distance by 10,000 km (the highest value tested). The full removal of comparisons with *cluster D* (modification *xiv.*) increased the amount of genetic variation explained in both BG dataset

and BLG dataset in *models 2*, *4*, and *B* (the latter explaining up to 77% of the genetic variation).

5.3 Discussion

Bantu-speaking populations share a striking genetic similarity across a vast area over most of sub-Saharan Africa. In fact, this shared genetic ancestry has been one of the arguments used to support the Bantu expansion being an actual movement of people across the continent and not just a cultural spread of the languages through neighbouring populations. Notwithstanding, it is not without interest that we can see that this genetic consistency does not mean uniformity, and that some degrees of differentiation are found among Bantu populations.

In the previous chapters, I analysed how demographic dynamics influenced the differentiation among Bantu populations (Chapter 3) and also the impact of admixture between different populations in the shaping of the genetic diversity currently observed in African Bantu-speaking populations (Chapter 4).

Taking all of that into account, in this last chapter my aim was to evaluate if and how genetics, geography, and linguistics were related considering the differentiation observed among these populations and how they combined to shape the current distribution that we can observe in the African continent in the present day.

The analysis of the distribution of the different distances measured between all the pairs of Bantu populations studied in this chapter, started by

showing a strong link between linguistic diversity and geography. The spread of agricultural techniques together with Bantu languages from central-west Africa towards east and south over the continental mass led to the present distribution of the variation. As movements of people accompanied this language dispersal, I expected the current genetic variation among Bantu speakers to be strongly defined by geographical proximity and linguistic similarities within the Bantu family, mirroring what is observed when analysing a broader ethno-linguistic diversity among African populations (Busby et al. 2016; Tishkoff et al. 2009). However, the initially analysed non-corrected genetic distances measured in my dataset did not show a significant link to either geography or linguistic distances.

When analysing the F_{ST}-based hierarchical tree, some results are especially noteworthy and not completely in accordance with what would be expected based on geography and linguistics alone. The existence of a cluster of Bantu-speaking southern Africa populations (most of the samples from *cluster D*, with the exception of BantuSouthAfrica) that diverge and are outgroups to other populations who speak similar languages is particularly noticeable (Figure 5.7). The composition of ADMIXTURE ancestral components of the populations from this group of populations seems to differ from the other populations in the dataset. The component that is mostly found in these populations (*NC Bantu South* component, represented in purple in Figure 5.7; Table 5.3) is present in other Niger-Congo populations but only in smaller amounts. However, in populations from this cluster this component is the most common and it is the only component

that represents more than 50% in any Bantu populations, which might be the cause for this cluster's behaviour. One possible explanation might be that this component was an unknown component present in now vanished inhabitants of the region that admixed with Bantu speakers arriving in this region.

Testing the pairwise F_{ST} values between the components found in the ADMIXTURE analysis I can see that both this component and the other two majorly seen in Bantu-speaking populations (*NC Yoruba* and *NC Bantu East* components, represented in dark and light green, respectively, in Figure 5.7) cluster together in the tree based on the components F_{ST} (Figure 5.8) and are all similarly distant to the *Jul'hoansi* component (values between 0.113 and 0.120; Table 5.2). Additionally, one of these three Bantu components (*NC Yoruba*) is the one predominantly found in Yoruba (used as a Niger-Congo reference) and the other two are similarly distant to it as well (0.025 and 0.028; Table 5.2). As such, there is no support for the hypothesis that this component was the result of a mixture between the incoming Bantu populations and other unknown south African inhabitants of the region. The data seem then to indicate that it is not the presence of the component itself driving these populations' pattern of variation but instead the composition and combination of different components in the various populations. One possible explanation for the current genetic diversity of southern African Bantu-speaking populations might be the experience of strong levels of genetic drift as the result of their migratory dynamics.

Even though the final number of markers analysed in this chapter did not allow me to proper analyse these dynamics in my dataset, a recent paper by Ananyo Choudhury and collaborators (Choudhury et al. 2017) analysed Runs Of Homozygosity (ROH) in some of the populations from *clusters B* and *D* (and other non-Bantu Niger-Congo) in my dataset (Figure 5.17). High levels of ROH provide an indication of the degree of within

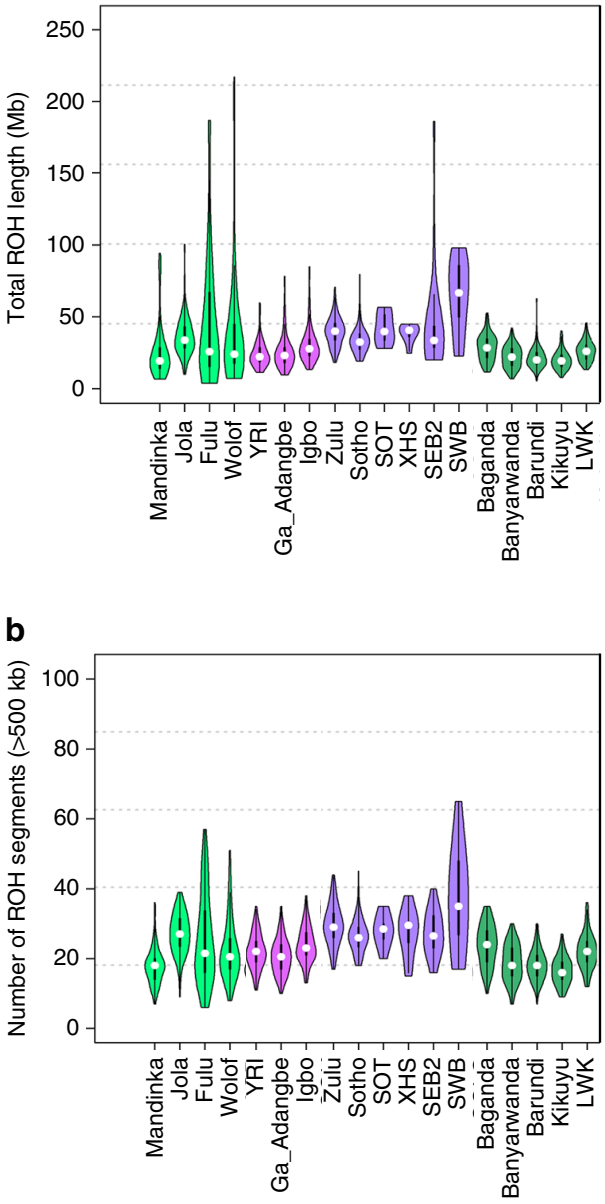


Figure 5.17 – Median per population of the (A) total Runs of Homozygosity (ROH) length in Mb; and the (B) number of ROH segments (<500 kb) for populations also present in my datasets. Adapted from (Choudhury et al. 2017).

population variation, which possibly suggests demographic processes accentuated by small population sizes and isolation. In Choudhury et al. (2017) analysis, south African Bantu populations show relatively higher lengths and higher number of ROH segments than other Niger-Congo in their dataset (Figure 5.17). This then also seems to support my previous hypothesis that southern African Bantu-speaking populations might have experienced strong drift effects during their migration into that region of the African continent shaping their currently observed genetic diversity.

In order to overcome some of the issues caused by the great influence of admixture in African populations and to uncover the demographic history of the Bantu speakers, I tried in this chapter a method to recover the signal of the ancestral Bantu populations, before the impact of admixture with native inhabitants from other non-Niger-Congo groups in their genomes. After this correction, there was a general increase in the homogeneity of the geographical clusters regarding the distribution of the pairwise F_{ST} distance between populations from a same cluster (Figure 5.10). Moreover, even though the genetic distances between pairs of populations decreased across my dataset – as expected after attempting to remove different degrees of admixture from related populations – the distances between groups are still significant and higher than the ones within groups (Figure 5.11). These two facts support the correction attempted, suggesting that I controlled and removed admixture in Bantu-speaking populations whilst still retaining the variability between groups.

The model-based measurement of geographic distances also led to a general increase in the amount of genetic and linguistic variation found on Bantu populations as explained by geography (Table 5.7 and 5.9). However, even if the corrected genetic distances are strongly correlated with geography this is not the case prior to the correction (Figure 5.6 and 5.12). This is yet another indication of the strong role played by admixture dynamics in shaping the genetic profile of present-day populations. Only after removing the admixture effects did the geographical pattern of the Bantu populations genetic diversity start to emerge. Nonetheless, even though both the genetic and linguistic diversity of Bantu-speaking populations are strongly linked with geography, it is important to note that both elements seem to be under different evolutionary pressures and are differently affected by geography as the best geographical model seems to differ for both.

Languages seem to be more dependent on the time factor, with older clusters of populations (near the Bantu homeland in central-west Africa) presenting higher levels of linguistic diversity than the clusters of populations that settled more recently in the southern areas of the continent (Figure 5.15). This seems to be affecting the results obtained for the geographical models as the removal of comparisons within the region (*cluster A* in the BLG dataset) leads to a generalized increase in the variation explained by geography (Table 5.9). Linguistically, the fit of most models also shows an improvement by simultaneously removing comparisons within *cluster D*, in south-east Africa (namely, explaining the

most linguistic variation across all analysis for *model A2*). The effect of *cluster D* in these results might be explained by the presence of linguistic structure observed between the two main groups of southeastern Bantu languages (Nguni and Sotho-Tswana). This structure can be identified in my dataset as plotted on Figure 5.15, with *cluster D* showing a Sotho/Tswana subgroup with lower linguistic diversity than the Nguni speakers.

Alternatively, genetics seem to be more influenced by demographic factors, with isolation and drift effects (combined with admixture as demonstrated on Chapter 4) likely playing a major role in the genetic differentiation among Bantu speakers. The geographical model that explained the highest amount of genetic variation (*model C*) was, additionally, among those in which higher geographical distances with populations from *cluster D* were observed (Figure 5.16). Moreover, the modifications to the models that produce the highest increases in the correlations are those involving increases in the geographical distances to *cluster D* (either by doubling the distances or increasing it by a fixed amount; Table 5.11). All this seems to indicate that the pairwise F_{ST} involving populations from *cluster D* might be higher than expected based solely on geographic distances. This can also be an indication that the most supported model might be only the one that is maximizing the geographical distances to populations from *cluster D* to accommodate this. As such, when effectively I removed populations from *cluster D* from the analysis the model

explaining most of the variation changes now to *model 4* or its refinement *model B*[17] (explaining 77.6% of the genetic variation for the 'Bantu Languages Genetic Dataset'; Table 5.11). These models of dispersal are in accordance with the most supported hypothesis regarding the split between the East and West branches of Bantu languages, the 'late-split' model, by both linguistic (Currie et al. 2013; Grollemund et al. 2015) and genetic data (Busby et al. 2016; Patin et al. 2017). Coincidently, for the linguistic dataset analysed here, when removing comparisons within the most linguistically diverse region (corresponding in the analysis to *cluster A*) the second most supported model is also *model 4* (and the identical, in that case, *model B*).

Altogether, it is noteworthy that after the correction the variation in the genetic diversity among Bantu-speakers is strongly linked with geography, more so than with the variation within the Bantu languages that they speak. As such, the current distribution of genetic variation in Bantu-speaking populations seems to be the result of an isolation by distance process, with travelling Bantu populations admixing with the inhabitants of the regions being settled. In order to fully understand the movements of these populations, ancient samples might be necessary. However, due to the geographic and climatic conditions of the African continent these are normally very difficult to obtain. Notwithstanding, an ancient sample from the

[17] As mentioned before, for the 'Bantu Languages Genetic Dataset' *model 4* and *model B* are identical since no populations from the East branch of the Narrow Bantu linguistic family are present in this dataset. For the 'Bantu Genetic Dataset' *model 4* explains a higher proportion of the variation than *model B* (76% against 61%; Table 5.11).

continent has been obtained and studied before (Gallego Llorente et al. 2015). Having access to unadmixed genetic references predating the migration of Bantu-speaking communities might be fundamental for a better insight and a deeper understanding into the movements and timings of the ‘Bantu expansion’ across the continent. However, this has been shown to be a difficult task. Therefore, decisive insights could also be gained by more modern samples to fill in the gaps in the geographic and linguistic coverage.

Chapter 6

FINAL REMARKS

The principal aim of this thesis is the characterization of the genetic variation in Bantu speakers and of the demographic dynamics that shaped their genetic makeup.

In Chapter 3 I investigated a genetic signal in Bantu-speaking populations from southeastern Africa, initially proposed as being the result of admixture with non-Bantu populations from the area (Sikora et al. 2011). I ultimately rejected that hypothesis and concluded instead that the observed component is, in fact, an outcome of the demographic dynamics associated with the Bantu speakers' expansion, resulting in the differentiation of Bantu-speaking groups.

The results of Chapter 3 steered me into a deeper investigation of the levels of differentiation in the genetic variation found among Niger-Congo populations. As such, Chapter 4 highlights the impact of the demographic events associated with the Bantu expansion in affecting the genetic variation of Bantu-speaking groups. I observed high levels of heterogeneity among Bantu-speaking populations, as a result of the interactions with prior inhabitants of the newly settled regions. Both aspects appear to have deeply shaped the genetic makeup observed in present-day Bantu speakers.

Chapter 5 analyses the relationships between different elements of Bantu-speaking populations. Namely, I investigated the behaviour of the genetic distances between Bantu speakers, their geographical dispersion

across the continent, and the relationship observed between their languages within the Narrow Bantu family. However, the attempted conciliation under the same model of dispersal between all these aspects proved to be far from straightforward.

As observed in Chapter 4, the genetic variation observed in present-day populations is strongly affected by admixture with other groups, introducing an additional source of variation. The work on Chapter 3 already focused on this aspect but the signal previously detected as admixture between the newly arrived Bantu agriculturalists and the earlier resident hunter-gatherers was instead unearthed as differentiation within Bantu speakers' genetic variation. This relates to some of the observations in Chapter 5. As I tried a method to control for admixture in Bantu-speaking populations it emerged that it was very important not only to correctly identify and control for the true signals of admixture but also to consider the differentiation among Bantu speakers independent of gene-flow from neighbouring populations. The genetic variation I observed in populations belonging to *cluster D* (South Africa) is a good example of this issue, as it forced me to understand the true nature of the components observed in these populations. Initially, the genetic pattern observed could be thought of as the result of admixture with pre-Bantu populations, but I propose instead that this is the result of strong genetic drift in the last stages of the Bantu expansion.

Altogether, it is noticeable that Bantu-speaking populations are far from a homogeneous group. However, despite the heterogeneity found in Bantu speakers, they are still treated as a single homogeneous group in several studies. This thesis shows that it is crucial that population structure within this group is taken into proper account. Genetic structure in human populations is an important confounding factor in genetic association studies. Thus, understanding the populations' genetic structure is decisive in biomedical research and medical genetics. Moreover, the genetic structure observed in these populations shows a strong link with geography, even more so than with the Bantu languages that they speak. It is important to note that even though I tried a method to control for genetic admixture with neighbouring populations from other ethnic groups and several models of Bantu speakers' dispersal, the linguistic data in all the analyses were fixed and not controlled/corrected for similar issues. Moreover, a fixed and sequential model of language dispersal might prove to be too rigid and not take into account dynamics happening during the evolution and dispersal of languages. An important direction of future work might be to try to address this exact issue by trying to control, for example, for the possibility of borrowing of words/language exchange between populations speaking different languages. However, the dynamics of dispersal between genetics and languages might be so different that a simple reconciliation over a geographical model might be impossible. While genetic borrowing implies movement of people, and, therefore, a slower process, language borrowing can easily happen between neighbouring populations within a generation

and words can travel great distances without any exchange of genetic material.

The levels of heterogeneity found and the confounding signals analysed also showed the vital importance of good sampling in this kind of work. In this thesis, several datasets had to be used in order to overcome ambiguity in the ethnicity of some samples, as well as a trade-off between the number of samples used and the number of markers. As such, it is of uttermost importance to have samples across all the area in which the Bantu-speaking populations settled during their dispersal. It is very important that these samples are well identified and located geographically and that the languages spoken by the various populations (as well as their ancestors) are well defined. Samples located near the starting point of the Bantu expansion in central-west Africa and the possible split areas in central Africa are particularly important to uncover the true relationships between these populations and to develop more sophisticated models of dispersal. Better sampling can also be reflected in the number of markers analysed, allowing the performing of more robust analyses, namely haplotype-based analyses that look into patterns of Linkage Disequilibrium (LD) in human genomes and thus allow for the dating of the events explored.

Ancient DNA samples might also be crucial for a better understanding of the dispersal of Bantu speakers across the African continent. Even though geographic and climatic conditions in Africa hinder the recovery of ancient genetic material this has been previously

demonstrated to work (Gallego Llorente et al. 2015; Schlebusch et al. 2017; Skoglund et al. 2017). Data from several ancient African samples from east and southern Africa have already been successfully retrieved and analysed and could bring new insights into the study of the dispersal of Bantu-speaking populations (Schlebusch et al. 2017; Skoglund et al. 2017). Complementing the currently available dataset of African populations with more modern samples from crucial areas along the route of Bantu expansion as well as ancient samples from other regions of the continent might prove invaluable for a full understanding of the processes and timings of the dispersal of Bantu-speaking populations.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009), 'Fast model-based estimation of ancestry in unrelated individuals', *Genome Research,* 19, 1655-64.

Alexander, D. H. and Lange, K. (2011), 'Enhancements to the ADMIXTURE algorithm for individual ancestry estimation', *BMC Bioinformatics,* 12.

Alexander, J. A. (1984), 'Early frontiers in southern Africa', in M. Hall, G. Avery, D. M. Avery, M. L. Wilson, and A. J. B. Humphreys (eds.), *Frontiers: Southern African archaeology today* (Oxford: B.A.R. International Series), 12-23.

Bahuchet, S. (2012), 'Changing language, remaining pygmy', *Human Biology,* 84, 11-43.

Barbieri, C., Butthof, A., Bostoen, K., and Pakendorf, B. (2013a), 'Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia', *European Journal of Human Genetics,* 21, 430-36.

Barbieri, C., Vicente, M., Rocha, J., Mpoloka, S. W., Stoneking, M., and Pakendorf, B. (2013b), 'Ancient substructure in early mtDNA lineages of southern Africa', *American Journal of Human Genetics,* 92, 285-92.

Barbieri, C., Güldemann, T., Naumann, C., Gerlach, L., Berthold, F., Nakagawa, H., Mpoloka, S. W., Stoneking, M., and Pakendorf, B. (2014), 'Unraveling the complex maternal history of southern African Khoisan populations', *American Journal of Physical Anthropology,* 153, 435-48.

Barnard, A. (1992), *Hunters and herders of southern Africa: a comparative ethnography of the Khoisan peoples* (Cambridge: Cambridge University Press).

Bastin, Y., Coupez, A., and Mann, M. (1999), *Continuity and divergence in the Bantu languages: perspectives from a lexicostatistic study* (Tervuren: Musée Royal de l'Afrique Centrale).

Batini, C., Ferri, G., Destro-Bisol, G., Brisighelli, F., Luiselli, D., Sanchez-Diz, P., Rocha, J., Simonson, T., Brehm, A., Montano, V., Elwali, N. E., Spedini, G., D'Amato, M. E., Myres, N., Ebbesen, P., Comas, D., and Capelli, C. (2011), 'Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages', *Molecular Biology and Evolution,* 28, 2603-13.

Behar, D. M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., Khusnutdinova, E. K., Balanovsky, O., Semino, O., Pereira, L., Comas, D., Gurwitz, D., Bonne-Tamir, B., Parfitt, T., Hammer, M.

F., Skorecki, K., and Villems, R. (2010), 'The genome-wide structure of the Jewish people', *Nature,* 466, 238-42.

Berger, L. R., Hawks, J., Dirks, P. H., Elliott, M., and Roberts, E. M. (2017), '*Homo naledi* and Pleistocene hominin evolution in subequatorial Africa', *Elife,* 6.

Berniell-Lee, G., Calafell, F., Bosch, E., Heyer, E., Sica, L., Mouguiama-Daouda, P., van der Veen, L., Hombert, J. M., Quintana-Murci, L., and Comas, D. (2009), 'Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages', *Molecular Biology and Evolution,* 26, 1581-9.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013), 'Estimating and interpreting FST: the impact of rare variants', *Genome Research,* 23, 1514-21.

Blench, R. (2013), 'Language, linguistics, and archaeology', in Peter Mitchell and Paul J. Lane (eds.), *The Oxford Handbook of African Archaeology* (Oxford: Oxford University Press), Chapter 4.

Bohannan, P. and Plog, F. (1967), *Beyond the frontier: social process and cultural change* (American Museum sourcebooks in anthropology; Garden City: Natural History Press).

Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., Amenga-Etego, L. N., Enimil, A., Apinjoh, T., Ndila, C. M., Manjurano, A., Nyirongo, V., Doumba, O., Rockett, K. A.,

Kwiatkowski, D. P., Spencer, C. C., and Malaria Genomic Epidemiology, N. (2016), 'Admixture into and within sub-Saharan Africa', *Elife,* 5.

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994), *The history and geography of human genes* (Princeton: Princeton University Press).

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015), 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *Gigascience,* 4.

Chimusa, E. R., Meintjies, A., Tchanga, M., Mulder, N., Seoighe, C., Soodyall, H., and Ramesar, R. (2015), 'A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations', *PLoS Genetics,* 11.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M. J., Joubert, F., Meintjes, A., Mulder, N., Ramesar, R., Rees, J., Scholtz, K., Sengupta, D., Soodyall, H., Venter, P., Warnich, L., and Pepper, M. S. (2017), 'Whole-genome sequencing for an enhanced understanding of genetic variation among south Africans', *Nature Communications,* 8.

Coelho, M., Sequeira, F., Luiselli, D., Beleza, S., and Rocha, J. (2009), 'On the edge of Bantu expansions: mtDNA, Y chromosome and

lactase persistence genetic variation in southwestern Angola', *BMC Evolutionary Biology,* 9, 80.

Currie, T. E., Meade, A., Guillon, M., and Mace, R. (2013), 'Cultural phylogeography of the Bantu Languages of sub-Saharan Africa', *Proceedings of the Royal Society B: Biological Sciences,* 280.

Darwin, C. (1871), *The descent of man, and selection in relation to sex*, 2 vols. (London,: J. Murray).

de Filippo, C., Barbieri, C., Whitten, M., Mpoloka, S. W., Gunnarsdottir, E. D., Bostoen, K., Nyambe, T., Beyer, K., Schreiber, H., de Knijff, P., Luiselli, D., Stoneking, M., and Pakendorf, B. (2011), 'Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups', *Molecular Biology and Evolution,* 28, 1255-69.

de Filippo, C., Bostoen, K., Stoneking, M., and Pakendorf, B. (2012), 'Bringing together linguistic and genetic evidence to test the Bantu expansion', *Proceedings of the Royal Society B: Biological Sciences,* 279, 3256-63.

de Maret, P. (2013), 'Archaeologies of the Bantu expansion', in Peter Mitchell and Paul J. Lane (eds.), *The Oxford Handbook of African Archaeology* (Oxford: Oxford University Press), Chapter 43.

Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglia, A., Tofanelli, S., Spedini, G., and Capelli, C. (2004), 'Variation of

female and male lineages in sub-Saharan populations: the importance of sociocultural factors', *Molecular Biology and Evolution,* 21, 1673-82.

Diamond, J. and Bellwood, P. (2003), 'Farmers and their languages: the first expansions', *Science,* 300, 597-603.

Eggert, M. K. H. (2005), 'The Bantu problem and African archaeology', in Ann Brower Stahl (ed.), *African Archaeology: A Critical Introduction* (Blackwell Studies in Global Archaeology; Oxford: Blackwell Publishing), 301-26.

Ehret, C. (2001), 'Bantu expansions: re-envisioning a central problem of early African history', *International Journal of African Historical Studies,* 34, 5-41.

Gallego Llorente, M., Jones, E. R., Eriksson, A., Siska, V., Arthur, K. W., Arthur, J. W., Curtis, M. C., Stock, J. T., Coltorti, M., Pieruccini, P., Stretton, S., Brock, F., Higham, T., Park, Y., Hofreiter, M., Bradley, D. G., Bhak, J., Pinhasi, R., and Manica, A. (2015), 'Ancient Ethiopian genome reveals extensive Eurasian admixture in eastern Africa', *Science,* 350, 132-32.

Greenberg, J. H. (1955), *Studies in African linguistic classification* (New Haven: Compass Publishing Company).

--- (1963), 'The languages of Africa', *International Journal of American Linguistics,* 29, 5-177.

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., and Pagel, M. (2015), 'Bantu expansion shows that habitat alters the route and pace of human dispersals', *Proceedings of the National Academy of Sciences of the United States of America,* 112, 13296-301.

Güldemann, T. and Stoneking, M. (2008), 'A historical appraisal of clicks: a linguistic and genetic population perspective', *Annual Review of Anthropology,* 37, 93-109.

Güldemann, T. and Fehn, A. M. (2014), *Beyond "Khoisan": historical relations in the Kalahari basin* (Current issues in linguistic theory; Amsterdam: John Benjamins Publishing Company).

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., Doumatey, A. P., Asiki, G., Seeley, J., Sisay-Joof, F., Jallow, M., Tollman, S., Mekonnen, E., Ekong, R., Oljira, T., Bradman, N., Bojang, K., Ramsay, M., Adeyemo, A., Bekele, E., Motala, A., Norris, S. A., Pirie, F., Kaleebu, P., Kwiatkowski, D., Tyler-Smith, C., Rotimi, C., Zeggini, E., and Sandhu, M. S. (2015), 'The African Genome Variation Project shapes medical genetics in Africa', *Nature,* 517, 327-32.

Guthrie, M. (1967), *Comparative Bantu; an introduction to the comparative linguistics and prehistory of the Bantu languages* (Farnborough: Gregg).

Heine, B. and Nurse, D. (2000), *African languages: an introduction* (Cambridge: Cambridge University Press).

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., and Myers, S. (2014), 'A genetic atlas of human admixture history', *Science,* 343, 747-51.

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J. M., Kidd, J. M., Rodriguez-Botigue, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A. A., Underhill, P. A., Comas, D., Kidd, K. K., Norman, P. J., Parham, P., Bustamante, C. D., Mountain, J. L., and Feldman, M. W. (2011), 'Hunter-gatherer genomic diversity suggests a southern African origin for modern humans', *Proceedings of the National Academy of Sciences of the United States of America,* 108, 5154-62.

Henn, B. M., Botigue, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlaoui-Zid, K., Zalloua, P. A., Moreno-Estrada, A., Bertranpetit, J., Bustamante, C. D., and Comas, D. (2012a), 'Genomic ancestry of north Africans supports back-to-Africa migrations', *PLoS Genetics,* 8.

Henn, B. M., Cavalli-Sforza, L. L., and Feldman, M. W. (2012b), 'The great human expansion', *Proceedings of the National Academy of Sciences of the United States of America,* 109, 17758-64.

Hodgson, J. A., Mulligan, C. J., Al-Meeri, A., and Raaum, R. L. (2014), 'Early back-to-Africa migration into the Horn of Africa', *PLoS Genetics,* 10.

Holden, C. J. (2002), 'Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis', *Proceedings of the Royal Society B: Biological Sciences,* 269, 793-9.

Hublin, J. J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., and Gunz, P. (2017), 'New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*', *Nature,* 546, 289-92.

Hudson, R. R., Slatkin, M., and Maddison, W. P. (1992), 'Estimation of levels of gene flow from DNA sequence data', *Genetics,* 132, 583-9.

Huxley, T. H. (1863), *Evidence as to Man's Place in Nature* (London: Williams & Norgate).

International HapMap Consortium (2003), 'The international HapMap project', *Nature,* 426, 789-96.

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., and Tyler-Smith, C. (2013), *Human evolutionary genetics* (2nd edn.; New York: Garland Science).

Johnston, H. (1886), *The Kilima-njaro expedition: a record of scientific exploration in eastern Equatorial Africa, and a general description of the natural history, languages, and commerce of the Kilima-njaro district* (London: Kegan Paul & Trench).

Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E., Seielstad, M. T., and Batzer, M. A. (2000), 'The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data', *American Journal of Human Genetics,* 66, 979-88.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., Thorsteinsdottir, U., and Stefansson, K. (2012), 'Rate of de novo mutations and the importance of father's age to disease risk', *Nature,* 488, 471-5.

Lane, A. B., Soodyall, H., Arndt, S., Ratshikhopha, M. E., Jonker, E., Freeman, C., Young, L., Morar, B., and Toffie, L. (2002), 'Genetic substructure in south African Bantu-speakers: evidence from

autosomal DNA and Y-chromosome studies', *American Journal of Physical Anthropology,* 119, 175-85.

Legendre, P. and Legendre, L. (2012), *Numerical ecology* (Developments in environmental modelling; Amsterdam: Elsevier).

Li, H. and Durbin, R. (2011), 'Inference of human population history from individual whole-genome sequences', *Nature,* 475, 493-6.

Li, S., Schlebusch, C., and Jakobsson, M. (2014), 'Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples', *Proceedings of the Royal Society B: Biological Sciences,* 281.

Lohse, K. and Frantz, L. A. F. (2014), 'Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes', *Genetics,* 196.

Manica, A., Prugnolle, F., and Balloux, F. (2005), 'Geography is a better determinant of human genetic differentiation than ethnicity', *Human Genetics,* 118, 366-71.

Manica, A., Amos, W., Balloux, F., and Hanihara, T. (2007), 'The effect of ancient population bottlenecks on human phenotypic variation', *Nature,* 448, 346-8.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W. M. (2010), 'Robust relationship inference in genome-wide association studies', *Bioinformatics,* 26, 2867-73.

Mardis, E. R. (2008), 'The impact of next-generation sequencing technology on genetics', *Trends in Genetics,* 24, 133-41.

Marks, S. J., Montinaro, F., Levy, H., Brisighelli, F., Ferri, G., Bertoncini, S., Batini, C., Busby, G. B., Arthur, C., Mitchell, P., Stewart, B. A., Oosthuizen, O., Oosthuizen, E., D'Amato, M. E., Davison, S., Pascali, V., and Capelli, C. (2015), 'Static and moving frontiers: the genetic landscape of southern African Bantu-speaking populations', *Molecular Biology and Evolution,* 32, 29-43.

Meyer, M., Arsuaga, J. L., de Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martinez, I., Gracia, A., Bermudez de Castro, J. M., Carbonell, E., Viola, B., Kelso, J., Prufer, K., and Paabo, S. (2016), 'Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins', *Nature,* 531, 504-7.

Mitchell, P. (2002), *The archaeology of southern Africa* (Cambridge world archaeology; Cambridge: Cambridge University Press).

Mitchell, P., Plug, I., Bailey, G., and Woodborne, S. (2008), 'Bringing the Kalahari debate to the mountains: late first millennium AD hunter-gatherer/farmer interaction in highland Lesotho', *Before Farming,* 2008, article 4.

Mitchell, P. (2009), 'Hunter-gatherers and farmers: some implications of 1,800 years of interaction in the Maloti-Drakensberg region of southern Africa', *Senri Ethnological Studies,* 73, 15-46.

--- (2010), 'Genetics and southern African prehistory: an archaeological view', *Journal of Anthropological Sciences,* 88, 73-92.

--- (2013), 'Early farming communities of southern and south-central Africa', in Peter Mitchell and Paul J. Lane (eds.), *The Oxford Handbook of African Archaeology* (Oxford: Oxford University Press), Chapter 45.

Montinaro, F., Busby, G. B., Pascali, V. L., Myers, S., Hellenthal, G., and Capelli, C. (2015), 'Unravelling the hidden ancestry of American admixed populations', *Nature Communications,* 6, article 6596.

Montinaro, F., Busby, G. B., Gonzalez-Santos, M., Oosthuitzen, O., Oosthuitzen, E., Anagnostou, P., Destro-Bisol, G., Pascali, V. L., and Capelli, C. (2017), 'Complex ancient genetic structure and cultural transitions in southern African populations', *Genetics,* 205, 303-16.

Morris, A. G., Heinze, A., Chan, E. K., Smith, A. B., and Hayes, V. M. (2014), 'First ancient mitochondrial human genome from a prepastoralist southern African', *Genome Biology and Evolution,* 6, 2647-53.

Mullis, K. B. and Faloona, F. A. (1987), 'Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction', *Methods Enzymol,* 155, 335-50.

Nachman, M. W. and Crowell, S. L. (2000), 'Estimate of the mutation rate per nucleotide in humans', *Genetics,* 156, 297-304.

Newman, J. L. (1995), *The peopling of Africa: a geographic interpretation* (New Haven: Yale University Press).

Nurse, D. and Tucker, I. (2002), 'A survey report for the Bantu languages', *SIL Electronic Survey Reports*.

Nurse, G. T., Lane, A. B., and Jenkins, T. (1976), 'Sero-genetic studies on the Dama of south west Africa', *Annals of Human Biology,* 3, 33-50.

Nychka, D., Furrer, R., Paige, J., and Sain, S. (2015), 'fields: tools for spatial data', R package version 9.0 <http://www.image.ucar.edu/fields>

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2017), 'vegan: community ecology package', R package version 2.4-3 <https://CRAN.R-project.org/package=vegan>

Pakendorf, B., Bostoen, K., and de Filippo, C. (2011), 'Molecular perspectives on the Bantu expansion: a synthesis', *Language Dynamics and Change,* 1, 50-88.

Paradis, E., Claude, J., and Strimmer, K. (2004), 'APE: analyses of phylogenetics and evolution in R language', *Bioinformatics,* 20, 289-90.

Patin, E., Siddle, K. J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Regnault, B., Lemee, L., Gravel, S., Hombert, J. M., Van der Veen, L., Dominy, N. J., Perry, G. H., Barreiro, L. B., Verdu, P., Heyer, E., and Quintana-Murci, L. (2014), 'The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists', *Nature Communications,* 5.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J. M., Pereira, J. B., Fernandes, V., Pereira, L., Van der Veen, L., Mouguiama-Daouda, P., Bustamante, C. D., Hombert, J. M., and Quintana-Murci, L. (2017), 'Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America', *Science,* 356, 543-46.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012), 'Ancient admixture in human history', *Genetics,* 192, 1065-93.

Pembleton, L. W., Cogan, N. O., and Forster, J. W. (2013), 'StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations', *Molecular Ecology Resources,* 13, 946-52.

Petersen, D. C., Libiger, O., Tindall, E. A., Hardie, R. A., Hannick, L. I., Glashoff, R. H., Mukerji, M., Indian Genome Variation, C., Fernandez, P., Haacke, W., Schork, N. J., and Hayes, V. M. (2013), 'Complex patterns of genomic admixture within southern Africa', *PLoS Genetics,* 9.

Phillipson, D. W. (2005), *African archaeology* (3rd edn.; Cambridge: Cambridge University Press).

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S. W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P. R., Lachance, J., Mountain, J., Bustamante, C. D., Berger, B., Tishkoff, S. A., Henn, B. M., Stoneking, M., Reich, D., and Pakendorf, B. (2012), 'The genetic prehistory of southern Africa', *Nature Communications,* 3, 1143.

Pickrell, J. K. and Pritchard, J. K. (2012), 'Inference of population splits and mixtures from genome-wide allele frequency data', *PLoS Genetics,* 8.

Pickrell, J. K., Patterson, N., Loh, P. R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014), 'Ancient west Eurasian ancestry in southern and eastern Africa', *Proceedings of the National Academy of Sciences of the United States of America,* 111, 2632-7.

Posnansky, M. (1961), 'Pottery types from archaeological sites in east-Africa', *Journal of African History,* 2, 177-98.

Prugnolle, F., Manica, A., and Balloux, F. (2005), 'Geography predicts neutral genetic diversity of human populations', *Current Biology,* 15, 159-60.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007), 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *American Journal of Human Genetics,* 81, 559-75.

QGIS Development Team (2014), 'QGIS geographic information system', Open Source Geospatial Foundation <http://qgis.osgeo.org>

R Core Team (2016), 'R: a language and environment for statistical computing', R Foundation for Statistical Computing, Vienna, Austria <https://www.R-project.org/>

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. (2005), 'Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa', *Proceedings of the National Academy of Sciences of the United States of America,* 102, 15942-7.

Ray, N., Currat, M., Berthier, P., and Excoffier, L. (2005), 'Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations', *Genome Research,* 15, 1161-7.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009), 'Reconstructing Indian population history', *Nature,* 461, 489-94.

Rexová, K., Bastin, Y., and Frynta, D. (2006), 'Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data', *Naturwissenschaften,* 93, 189-94.

Richter, D., Grun, R., Joannes-Boyau, R., Steele, T. E., Amani, F., Rue, M., Fernandes, P., Raynal, J. P., Geraads, D., Ben-Ncer, A., Hublin, J. J., and McPherron, S. P. (2017), 'The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age', *Nature,* 546, 293-96.

Rocha, J. and Fehn, A. M. (2016), 'Genetics and demographic history of the Bantu', *eLS*.

Salas, A., Richards, M., De la Fe, T., Lareu, M. V., Sobrino, B., Sanchez-Diz, P., Macaulay, V., and Carracedo, A. (2002), 'The making of the African mtDNA landscape', *American Journal of Human Genetics,* 71, 1082-111.

Schlebusch, C. M., Skoglund, P., Sjodin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G., Soodyall, H., and Jakobsson, M. (2012), 'Genomic variation in seven Khoe-San groups reveals adaptation and complex African history', *Science,* 338, 374-9.

Schlebusch, C. M., Lombard, M., and Soodyall, H. (2013), 'MtDNA control region variation affirms diversity and deep sub-structure in populations from southern Africa', *BMC Evolutionary Biology,* 13.

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., and Jakobsson, M. (2017), 'Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago', *Science,* 358, 652-55.

Sikora, M., Laayouni, H., Calafell, F., Comas, D., and Bertranpetit, J. (2011), 'A genomic analysis identifies a novel component in the genetic

structure of sub-Saharan African populations', *European Journal of Human Genetics,* 19, 84-8.

Simons, G. F. and Fennig, C. D. (2018), *Ethnologue: languages of the world, twenty-first edition* (Dallas: SIL International. Online version: http://www.ethnologue.com).

Sirak, K. A., Fernandes, D. M., Connell, S., and Pinhasi, R. (2015), 'No longer the 1%: Optimizing ancient DNA yields from Saharan African samples', Poster session presented at: The 84th Annual Meeting of the American Association of Physical Anthropologists (St. Louis).

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney, E., Michel, M., Stewardson, K., Cerezo-Román, J. I., Chiumia, C., Crowther, A., Gomani-Chindebvu, E., Gidna, A. O., Grillo, K. M., Helenius, I. T., Hellenthal, G., Helm, R., Horton, M., López, S., Mabulla, A. Z. P., Parkington, J., Shipton, C., Thomas, M. G., Tibesasa, R., Welling, M., Hayes, V. M., Kennett, D. J., Ramesar, R., Meyer, M., Pääbo, S., Patterson, N., Morris, A. G., Boivin, N., Pinhasi, R., Krause, J., and Reich, D. (2017), 'Reconstructing prehistoric African population structure', *Cell,* 171, 59-71.

Sobrino, B., Brion, M., and Carracedo, A. (2005), 'SNPs in forensic genetics: a review on SNP typing methodologies', *Forensic Science International,* 154, 181-94.

Stringer, C. (2002), 'Modern human origins: progress and prospects', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,* 357, 563-79.

Stringer, C. and Galway-Witham, J. (2017), 'Palaeoanthropology: on the origin of our species', *Nature,* 546, 212-14.

Tattersall, I. (2009), 'Human origins: out of Africa', *Proceedings of the National Academy of Sciences of the United States of America,* 106, 16018-21.

Tishkoff, S. A. and Williams, S. M. (2002), 'Genetic analysis of African populations: human evolution and complex disease', *Nature Reviews Genetics,* 3, 611-21.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., and Williams, S. M. (2009), 'The genetic structure and history of Africans and African Americans', *Science,* 324, 1035-44.

Vansina, J. (1979), 'Bantu in the crystal ball, I', *History in Africa,* 6, 287-333.

--- (1995), 'New linguistic evidence and the Bantu expansion', *Journal of African History,* 36, 173-95.

Weir, B. S. and Cockerham, C. C. (1984), 'Estimating F-statistics for the analysis of population structure', *Evolution,* 38, 1358-70.

Willoughby, P. R. (2007), *The evolution of modern humans in Africa: a comprehensive guide* (Lanham: Altamira Press).

Wood, B. and Collard, M. (1999), 'The human genus', *Science,* 284, 65-71.

Wood, B. (2005), *Human evolution: a very short introduction* (Very short introductions; Oxford: Oxford University Press).

Wright, S. (1949), 'The genetical structure of populations', *Annals of Human Genetics,* 15, 323-54.

CONTRIBUTORS

Regarding the thesis of Miguel González dos Santos, submitted for the degree of Doctor of Philosophy, Hilary Term 2018, with the title 'Human Genetic Variation and the Dispersal of Bantu-Speaking Populations.'

All chapters have been planned, conceived, and edited with the help of supervisor Cristian Capelli.

Chapter 1: General Introduction. My own work.

Chapter 2: Material and Methods. My own work.

Chapter 3: Genome-Wide SNP Analysis of Southern African Populations Provides New Insights into the Dispersal of Bantu-Speaking Groups. This chapter is based on a published manuscript. I performed the analysis and wrote the chapter, and the corresponding manuscript, with the support of Francesco Montinaro. The samples had been previously collected by Sarah J. Marks and Cristian Capelli (with the assistance of George B.J. Busby, Ockie Oosthuizen, Erica Oosthuizen, and Simon Marks) and Sarah J. Marks genotyped them. The other authors of the manuscript (George B.J. Busby, Paolo Anagnostou, Giovanni Destro-Bisol, Vincenzo Pascali) contributed to the discussion of the results.

Chapter 4: Characterization of the Variation in Bantu-Speaking Populations. I wrote the chapter and performed all the analysis. I collected the Angolan samples and performed the DNA extraction for genotyping of these and other new samples provided by collaborators as indicated in Table 2.1. Other samples included in the analyses have been previously genotyped by members of Prof. Capelli research group.

Chapter 5: An Attempt to Reconcile Genetics, Geography, and Languages of the Bantu Expansion. I wrote the chapter and performed all the analysis. Rebecca Grollemund assembled the linguistic matrix of words used.

Chapter 6: Final Remarks. My own work.

SUPPLEMENTARY MATERIAL

Supplementary Material 1 – Cross-validation (CV) plot for the ADMIXTURE analysis with the 'Mozambique Genetic Dataset'. 189

Supplementary Material 2 – ADMIXTURE plots from *K*=2 to *K*=8 for the 'Mozambique Genetic Dataset'. 190

Supplementary Material 3 – *f3 tests* for the 'Mozambique Genetic Dataset'. 191

Supplementary Material 4 – Published version of Chapter 3. 195

Supplementary Material 5 A—EEE – Principal Component Analysis (PCA) plots for the first and second components for the 'Niger-Congo Genetic Dataset' (after removing the CEU sample) highlighted for each individual population. 205

Supplementary Material 6 – Detailed representation of *model 1* used for the calculation of the model-based pairwise geographical distances. 216

Supplementary Material 7 – Detailed representation of *model 2* used for the calculation of the model-based pairwise geographical distances. 217

Supplementary Material 8 – Detailed representation of *model 3* used for the calculation of the model-based pairwise geographical distances. 218

Supplementary Material 9 – Detailed representation of *model 4* used for the calculation of the model-based pairwise geographical distances. 219

Supplementary Material 10 – Detailed representation of *model A1* used for the calculation of the model-based pairwise geographical distances. 220

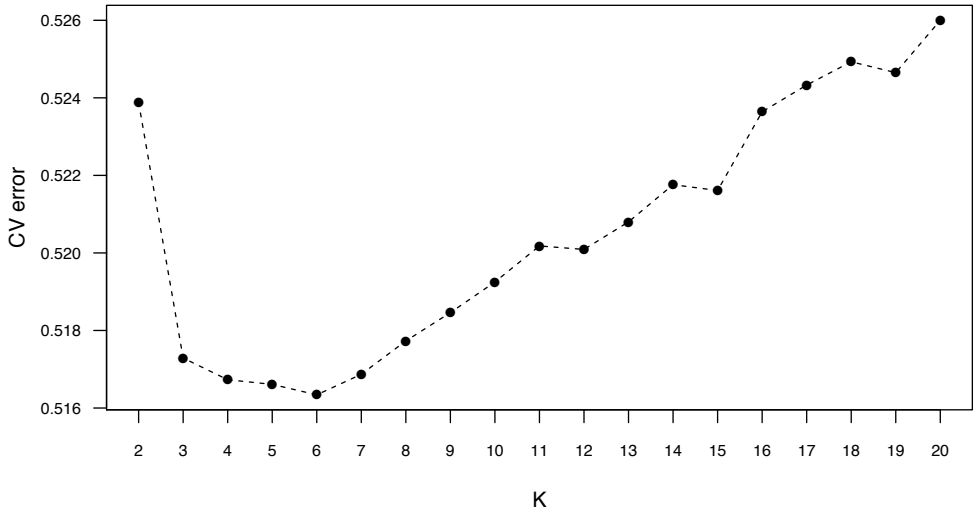
Supplementary Material 11 – Detailed representation of *model A2* used for the calculation of the model-based pairwise geographical distances. 221

Supplementary Material 12 – Detailed representation of *model A3* used for the calculation of the model-based pairwise geographical distances. 222

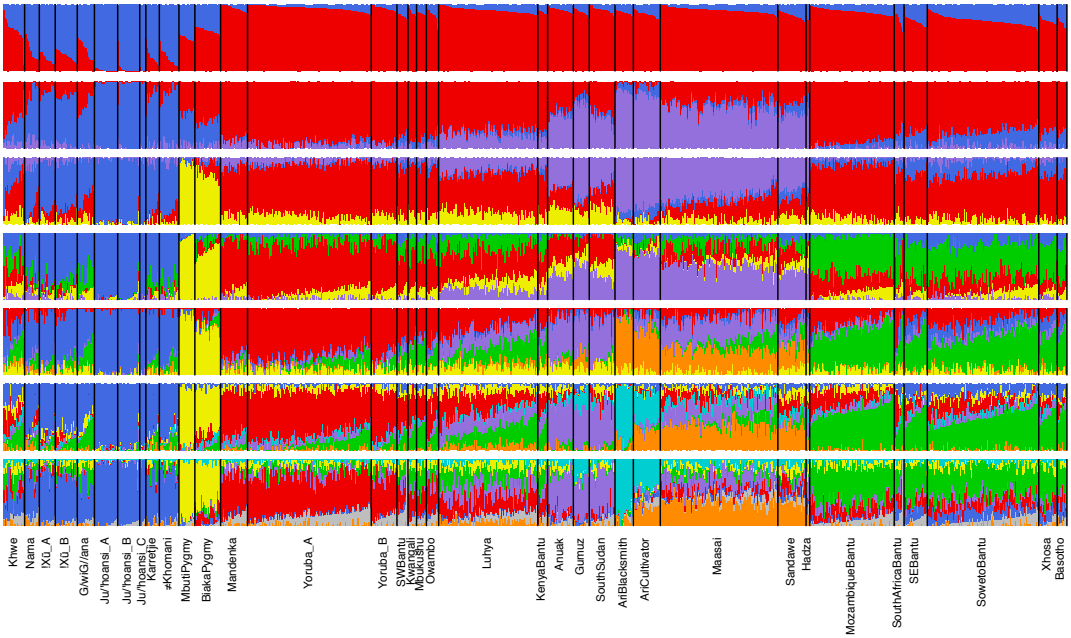
Supplementary Material 13 – Detailed representation of *model B* used for the calculation of the model-based pairwise geographical distances. 223

Supplementary Material 14 – Detailed representation of *model C* used for the calculation of the model-based pairwise geographical distances. 224

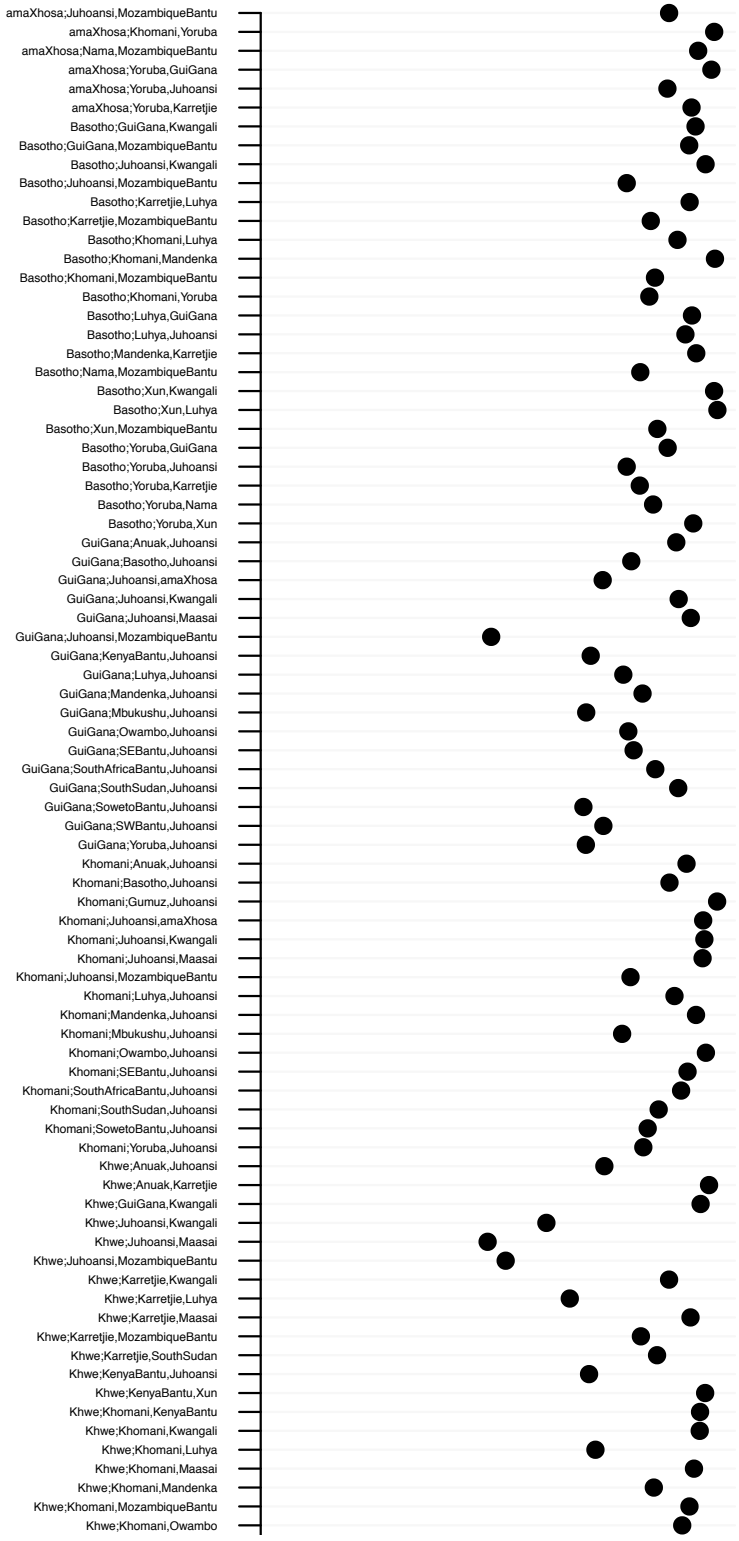
Supplementary Material 1 – Cross-validation (CV) plot for the ADMIXTURE analysis with the 'Mozambique Genetic Dataset'. Lowest value was observed for K=6 (0.5163).



Supplementary Material 2 – ADMIXTURE plots from *K*=2 to *K*=8 for the 'Mozambique Genetic Dataset'.



Supplementary Material 3 – *f3 tests* for the 'Mozambique Genetic Dataset'.Significant comparisons (Z-score < −3) for the f3 tests between all possible trios of populations.



Khwe;Khomani,SouthSudan
Khwe;Khomani,Yoruba
Khwe;Luhya,Juhoansi
Khwe;Mandenka,Juhoansi
Khwe;Mandenka,Karretjie
Khwe;Mandenka,Xun
Khwe;Mbukushu,Juhoansi
Khwe;Mbukushu,Karretjie
Khwe;Mbukushu,Nama
Khwe;Mbukushu,Xun
Khwe;Nama,Luhya
Khwe;Nama,MozambiqueBantu
Khwe;Owambo,Juhoansi
Khwe;Owambo,Karretjie
Khwe;Owambo,Xun
Khwe;Sandawe,Juhoansi
Khwe;SouthSudan,Juhoansi
Khwe;SowetoBantu,Juhoansi
Khwe;SWBantu,Juhoansi
Khwe;Xun,Kwangali
Khwe;Xun,Luhya
Khwe;Xun,Maasai
Khwe;Xun,MozambiqueBantu
Khwe;Xun,SouthSudan
Khwe;Yoruba,GuiGana
Khwe;Yoruba,Juhoansi
Khwe;Yoruba,Karretjie
Khwe;Yoruba,Nama
Khwe;Yoruba,Xun
Luhya;Anuak,Basotho
Luhya;Anuak,MozambiqueBantu
Luhya;Anuak,SowetoBantu
Luhya;AriBlacksmith,MozambiqueBantu
Luhya;AriBlacksmith,Owambo
Luhya;AriCultivator,MozambiqueBantu
Luhya;Basotho,SouthSudan
Luhya;Gumuz,MozambiqueBantu
Luhya;Maasai,MozambiqueBantu
Luhya;Mbukushu,SouthSudan
Luhya;Owambo,Maasai
Luhya;Sandawe,MozambiqueBantu
Luhya;SEBantu,SouthSudan
Luhya;SouthSudan,MozambiqueBantu
Luhya;SowetoBantu,SouthSudan
Maasai;AriBlacksmith,MozambiqueBantu
Maasai;AriBlacksmith,Owambo
Maasai;AriBlacksmith,Yoruba
Maasai;AriCultivator,KenyaBantu
Maasai;AriCultivator,Kwangali
Maasai;AriCultivator,Luhya
Maasai;AriCultivator,Mandenka
Maasai;AriCultivator,MozambiqueBantu
Maasai;AriCultivator,Owambo
Maasai;AriCultivator,SWBantu
Maasai;AriCultivator,Yoruba
Maasai;SowetoBantu,AriCultivator
Mbukushu;GuiGana,Kwangali
Nama;Anuak,Juhoansi
Nama;Anuak,Karretjie
Nama;AriBlacksmith,Juhoansi
Nama;AriBlacksmith,Karretjie
Nama;AriCultivator,Juhoansi
Nama;Basotho,Juhoansi
Nama;Gumuz,Juhoansi
Nama;Gumuz,Karretjie
Nama;Juhoansi,amaXhosa
Nama;Juhoansi,Kwangali
Nama;Juhoansi,Maasai
Nama;Juhoansi,MozambiqueBantu
Nama;Karretjie,Maasai
Nama;Karretjie,Sandawe
Nama;Karretjie,SouthSudan
Nama;KenyaBantu,Juhoansi
Nama;Khomani,Anuak
Nama;Khomani,Gumuz
Nama;Khomani,KenyaBantu
Nama;Khomani,Maasai
Nama;Khomani,Sandawe
Nama;Luhya,Juhoansi
Nama;Mandenka,Juhoansi
Nama;Mbukushu,Juhoansi

Nama;Owambo,Juhoansi
Nama;Sandawe,Juhoansi
Nama;SEBantu,Juhoansi
Nama;SouthAfricaBantu,Juhoansi
Nama;SouthSudan,Juhoansi
Nama;SowetoBantu,Juhoansi
Nama;SWBantu,Juhoansi
Nama;Yoruba,Juhoansi
SEBantu;GuiGana,Kwangali
SEBantu;GuiGana,MozambiqueBantu
SEBantu;Juhoansi,Kwangali
SEBantu;Juhoansi,MozambiqueBantu
SEBantu;Karretjie,Kwangali
SEBantu;Karretjie,Luhya
SEBantu;Karretjie,MozambiqueBantu
SEBantu;Khomani,KenyaBantu
SEBantu;Khomani,Kwangali
SEBantu;Khomani,Luhya
SEBantu;Khomani,Mandenka
SEBantu;Khomani,MozambiqueBantu
SEBantu;Khomani,Yoruba
SEBantu;Luhya,GuiGana
SEBantu;Luhya,Juhoansi
SEBantu;Mandenka,GuiGana
SEBantu;Mandenka,Juhoansi
SEBantu;Mandenka,Karretjie
SEBantu;Mandenka,Nama
SEBantu;Mandenka,Xun
SEBantu;Mbukushu,Juhoansi
SEBantu;Mbukushu,Karretjie
SEBantu;Mbukushu,Nama
SEBantu;Nama,Kwangali
SEBantu;Nama,Luhya
SEBantu;Nama,MozambiqueBantu
SEBantu;Xun,Kwangali
SEBantu;Xun,Luhya
SEBantu;Xun,MozambiqueBantu
SEBantu;Yoruba,GuiGana
SEBantu;Yoruba,Juhoansi
SEBantu;Yoruba,Karretjie
SEBantu;Yoruba,Nama
SEBantu;Yoruba,Xun
SouthAfricaBantu;AriCultivator,SWBantu
SouthAfricaBantu;GuiGana,MozambiqueBantu
SouthAfricaBantu;Juhoansi,MozambiqueBantu
SouthAfricaBantu;Karretjie,MozambiqueBantu
SouthAfricaBantu;Karretjie,SWBantu
SouthAfricaBantu;Khomani,Luhya
SouthAfricaBantu;Khomani,Mandenka
SouthAfricaBantu;Khomani,MozambiqueBantu
SouthAfricaBantu;Khomani,SWBantu
SouthAfricaBantu;Khomani,Yoruba
SouthAfricaBantu;Khwe,SWBantu
SouthAfricaBantu;Luhya,GuiGana
SouthAfricaBantu;Luhya,Juhoansi
SouthAfricaBantu;Mandenka,GuiGana
SouthAfricaBantu;Mandenka,Juhoansi
SouthAfricaBantu;Mandenka,Karretjie
SouthAfricaBantu;Mandenka,Nama
SouthAfricaBantu;Mandenka,Xun
SouthAfricaBantu;MbutiPygmy,SWBantu
SouthAfricaBantu;Nama,Luhya
SouthAfricaBantu;Nama,MozambiqueBantu
SouthAfricaBantu;Nama,SWBantu
SouthAfricaBantu;Owambo,GuiGana
SouthAfricaBantu;Owambo,Juhoansi
SouthAfricaBantu;SEBantu,SWBantu
SouthAfricaBantu;SowetoBantu,SWBantu
SouthAfricaBantu;SWBantu,amaXhosa
SouthAfricaBantu;SWBantu,Basotho
SouthAfricaBantu;SWBantu,BiakaPygmy
SouthAfricaBantu;SWBantu,GuiGana
SouthAfricaBantu;SWBantu,Juhoansi
SouthAfricaBantu;SWBantu,Luhya
SouthAfricaBantu;SWBantu,MozambiqueBantu
SouthAfricaBantu;SWBantu,Sandawe
SouthAfricaBantu;SWBantu,Xun
SouthAfricaBantu;Xun,Luhya
SouthAfricaBantu;Xun,MozambiqueBantu
SouthAfricaBantu;Yoruba,GuiGana
SouthAfricaBantu;Yoruba,Juhoansi

SouthAfricaBantu;Yoruba,Juhoansi
SouthAfricaBantu;Yoruba,Karretjie
SouthAfricaBantu;Yoruba,Nama
SouthAfricaBantu;Yoruba,Xun
SouthSudan;Gumuz,SWBantu
SowetoBantu;Anuak,Karretjie
SowetoBantu;GuiGana,Kwangali
SowetoBantu;GuiGana,MozambiqueBantu
SowetoBantu;Juhoansi,Kwangali
SowetoBantu;Juhoansi,MozambiqueBantu
SowetoBantu;Karretjie,Kwangali
SowetoBantu;Karretjie,Luhya
SowetoBantu;Karretjie,MozambiqueBantu
SowetoBantu;Karretjie,SouthSudan
SowetoBantu;KenyaBantu,GuiGana
SowetoBantu;KenyaBantu,Xun
SowetoBantu;Khomani,KenyaBantu
SowetoBantu;Khomani,Kwangali
SowetoBantu;Khomani,Luhya
SowetoBantu;Khomani,Mandenka
SowetoBantu;Khomani,MozambiqueBantu
SowetoBantu;Khomani,Yoruba
SowetoBantu;Khwe,MozambiqueBantu
SowetoBantu;Luhya,GuiGana
SowetoBantu;Luhya,Juhoansi
SowetoBantu;Maasai,GuiGana
SowetoBantu;Mandenka,GuiGana
SowetoBantu;Mandenka,Juhoansi
SowetoBantu;Mandenka,Karretjie
SowetoBantu;Mandenka,Nama
SowetoBantu;Mandenka,Xun
SowetoBantu;Mbukushu,GuiGana
SowetoBantu;Mbukushu,Juhoansi
SowetoBantu;Mbukushu,Karretjie
SowetoBantu;Mbukushu,Khomani
SowetoBantu;Mbukushu,Nama
SowetoBantu;Mbukushu,Xun
SowetoBantu;Nama,Kwangali
SowetoBantu;Nama,Luhya
SowetoBantu;Nama,MozambiqueBantu
SowetoBantu;Owambo,GuiGana
SowetoBantu;Owambo,Juhoansi
SowetoBantu;Owambo,Karretjie
SowetoBantu;SouthSudan,GuiGana
SowetoBantu;Xun,Kwangali
SowetoBantu;Xun,Luhya
SowetoBantu;Xun,MozambiqueBantu
SowetoBantu;Yoruba,GuiGana
SowetoBantu;Yoruba,Juhoansi
SowetoBantu;Yoruba,Karretjie
SowetoBantu;Yoruba,Nama
SowetoBantu;Yoruba,Xun
Xun;Anuak,Juhoansi
Xun;AriBlacksmith,Juhoansi
Xun;AriCultivator,Juhoansi
Xun;Basotho,Juhoansi
Xun;BiakaPygmy,Juhoansi
Xun;Gumuz,Juhoansi
Xun;Hadza,Juhoansi
Xun;Juhoansi,amaXhosa
Xun;Juhoansi,Kwangali
Xun;Juhoansi,Maasai
Xun;Juhoansi,MozambiqueBantu
Xun;KenyaBantu,Juhoansi
Xun;Khwe,Juhoansi
Xun;Luhya,Juhoansi
Xun;Mandenka,Juhoansi
Xun;Mbukushu,Juhoansi
Xun;Owambo,Juhoansi
Xun;Sandawe,Juhoansi
Xun;SEBantu,Juhoansi
Xun;SouthAfricaBantu,Juhoansi
Xun;SouthSudan,Juhoansi
Xun;SowetoBantu,Juhoansi
Xun;SWBantu,Juhoansi
Xun;Yoruba,Juhoansi
−14
−12
−10
−8
−6
−4
Z−score

Supplementary Material 4 – Published version of Chapter 3.

González-Santos M, Montinaro F, Oosthuizen O, Oosthuizen E, Busby GB, Anagnostou P, Destro-Bisol G, Pascali V, Capelli C 2015. “Genome-Wide SNP Analysis of Southern African Populations Provides New Insights into the Dispersal of Bantu-Speaking Groups.” *Genome Biology and Evolution*. 7:2560–2568. doi: 10.1093/gbe/evv164

GBE

Genome-Wide SNP Analysis of Southern African Populations Provides New Insights into the Dispersal of Bantu-Speaking Groups

Miguel González-Santos[1,†], Francesco Montinaro[1,2,†], Ockie Oosthuizen[3], Erica Oosthuizen[3], George B.J. Busby[1,4], Paolo Anagnostou[5,6], Giovanni Destro-Bisol[5,6], Vincenzo Pascali[2], and Cristian Capelli[1,*]

[1]Department of Zoology, University of Oxford, United Kingdom

[2]Institute of Legal Medicine, Catholic University, Rome, Italy

[3]School of Medicine, University of Namibia, Windhoek, Namibia

[4]Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom

[5]Dipartimento di Biologia Ambientale, Università "La Sapienza", Rome, Italy

[6]Istituto Italiano di Antropologia, Rome, Italy

*Corresponding author: E-mail: cristian.capelli@zoo.ox.ac.uk.

[†]These authors contributed equally to this work.

Accepted: August 14, 2015

Abstract

The expansion of Bantu-speaking agropastoralist populations had a great impact on the genetic, linguistic, and cultural variation of sub-Saharan Africa. It is generally accepted that Bantu languages originated in an area around the present border between Cameroon and Nigeria approximately 5,000 years ago, from where they spread South and East becoming the largest African linguistic branch. The demic consequences of this event are reflected in the relatively high genetic homogeneity observed across most of sub-Saharan Africa populations. In this work, we explored genome-wide single nucleotide polymorphism data from 28 populations to characterize the genetic components present in sub-Saharan African populations. Combining novel data from four Southern African populations with previously published results, we reject the hypothesis that the "non-Bantu" genetic component reported in South-Eastern Africa (Mozambique) reflects extensive gene flow between incoming agriculturalist and resident hunter-gatherer communities. We alternatively suggest that this novel component is the result of demographic dynamics associated with the Bantu dispersal.

Key words: Southern Africa, Bantu speakers, admixture, genome-wide SNPs.

Introduction

The genetic structure of African populations is the result of both ancient and more recent episodes of migration and admixture (Tishkoff et al. 2009; Hellenthal et al. 2014; Pickrell et al. 2012, 2014). Among these historical events, the expansion of Bantu-speaking agropastoralist populations had a substantial impact on the continental distribution of genetic diversity. Bantu languages (part of the Niger-Congo linguistic group, NC) are thought to have originated near the present border between Cameroon and Nigeria approximately 5,000 years before present (Newman 1995; Diamond and Bellwood 2003). From its place of origin, this branch of the NC linguistic phylum spread East and South across sub-Saharan Africa, together with agricultural techniques and the use of iron (Newman 1995; Diamond and Bellwood 2003). Today, Bantu is the largest African linguistic family in both geographical extension and number of speakers, indicative of the impact this migration had on the continent (de Filippo et al. 2012). However, although there is a general consensus on the place and time of the origin of this movement, the actual routes used are still under debate, with two main hypotheses having recently been tested. These hypotheses differ mainly on the geographical and chronological dimensions of the split between East and West Bantu-speaking groups (de Filippo et al. 2012; Currie et al. 2013). Together with the route followed, another important aspect to be considered is the

© The Author(s) 2015. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

degree of interaction that populations involved in this expansion had with the groups inhabiting the regions they were moving into. For example, the arrival of Bantu-speaking people in the southern regions of the continent could have led either to isolation or to admixture with the inhabitant pastoralists and hunter-gatherers (Destro-Bisol et al. 2004; Mitchell et al. 2008; Mitchell 2009; de Filippo et al. 2012; Patin et al. 2014).

Previous investigations have focused on the demographic dynamics of the Bantu dispersal (Barbieri, Vicente, et al. 2013; Schlebusch et al. 2013; Barbieri et al. 2014; Patin et al. 2014; Pickrell et al. 2012, 2014). Within this context, the differentiation of a population sample from Mozambique from other African populations suggested the presence of a specific South-Eastern component within the continent. This component, found at a proportion higher than 50%, has been interpreted as derived from an ancestral population (presumably related to hunter-gatherer populations) inhabiting the area before the arrival of expanding Bantu speakers (Sikora et al. 2011). The predominance of such a component in extant populations could be explained by the cultural shift of a foraging community (fig. 1*a*) or substantial gene flow into the arriving Bantu-speaking groups (fig. 1*b*). Alternatively, the reported Mozambique uniqueness could simply be the result of the dispersal process whose genetic signature would be expected to be shared with populations related to those living in Mozambique. This scenario might include or not some degree of gene flow from hunter-gatherer populations (fig. 1*c*).

To test these hypotheses and further characterize the reported Mozambican component, we analyzed new genome-wide single nucleotide polymorphism (SNP) array data from 35 sub-Saharan African populations from 868 individuals (including 33 novel genotypes from Namibia and Lesotho) (fig. 2). Our results place the so-called "South-Eastern component" within the broader genetic variation of Southern Africa. Additionally, we find no evidence in the Mozambican population for substantial gene flow between Bantu and hunter-gatherer populations from this region.

Results

Relationships among individuals were explored using principal component analysis (PCA) (fig. 3). The first component (PC1) shows a separation of Khoisan-speaking groups from the other populations, with East non-NC populations separating along the second component (PC2) (fig. 3*a*). To further analyze these results, a second PCA was performed removing the East non-NC (Afro-Asiatic, Nilo-Saharan, and linguistic isolates) populations. In this, the Khoisan-speaking groups and the NC-speaking populations are separated along the first component (PC1), the rainforest hunter-gatherers (RHG, also known as Pygmies) located among the two (fig. 3*b*). RHG are separated from the other populations along the second component (PC2). The NC speakers are spread along the same axis, the Mandenka and Yoruba individuals on one side and various South-Eastern Bantu speakers on the other (fig. 3*b*). We then performed ADMIXTURE analysis, and obtained for $K = 6$ the most suitable number of ancestral components (supplementary fig. S1, Supplementary Material online). Five groups emerge that we associate for simplicity to the linguistic group spoken by the majority of the populations characterized by the indicated component: Khoisan (here considered as Southern African populations speaking click-rich languages; Güldemann and Fehn 2014) (blue), Nilo-Saharan (East-1; purple) and Afro-Asiatic (East-2; orange) in East Africa, RHG (yellow), and NC speakers, the latter being characterized by two major components (red and green in fig. 4*a*). These two components are present in all NC populations but their amount varies among groups. One of these components (red) dominates populations speaking West Bantu and other NC languages, whereas the other one (green) is more common among South-Eastern Bantu speakers (fig. 4*a*). The main component of the Mozambican sample (green) is similarly present in populations from Lesotho and South Africa (even for higher values of *K*, supplementary fig. S5, Supplementary Material online). Notably, the Basotho from Lesotho and the amaXhosa/SE Bantu/SowetoBantu from South Africa also show a non-Bantu component (Khoisan) but additional components (Khoisan or others) are virtually absent in samples from Mozambique. The spatial distribution of these components was visualized by interpolation maps (fig. 4*b*–*g*). The two main components present in NC populations mapped on to different geographic regions, so that these two can be broadly identified as NC-West and NC-Southeast (fig. 4*f* and *g*, respectively). The Afro-Asiatic and Nilo-Saharan components are mostly restricted to populations from Central-East Africa (fig. 4*c* and *d*) and the RHG component was geographically restricted to Pygmy populations (fig. 4*e*).

The genetic distances among the six components were estimated through F_{ST}. The largest values are observed when the Khoisan component is compared with the East Africa ones, followed by the comparisons of the first with both components from NC populations. The two NC components (West and East) are the most genetically close, their F_{ST} value (0.042) at least two times smaller than any other comparison (table 1).

Although STRUCTURE-like analyses are useful to identify patterns of gene flow between groups, they are not a formal test of admixture as similar profiles can be the result of several different population histories. For this reason, we carried out *f3* tests (Reich et al. 2009) for windows of 100 SNPs between all the possible trios of populations in our data set and reported all the comparisons in supplementary figure S6 and tables S2 and S3, Supplementary Material online. As expected, this analysis identified most of the known admixture events that characterized African populations, involving Khoisan and Bantu populations in Southern Africa.

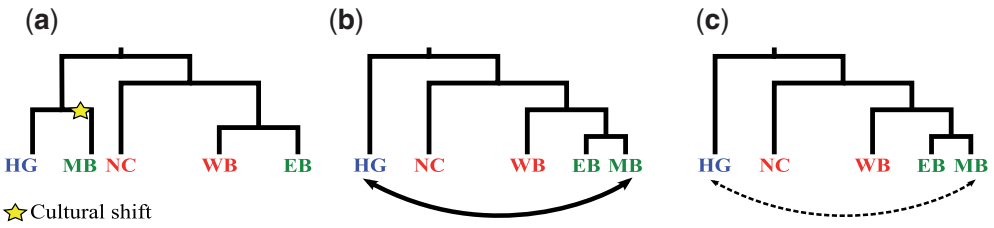


Fig. 1.—Alternative scenarios for the Mozambique Bantu (MB) differentiation: (*a*) Cultural shift after a split from a non-Bantu, hunter-gatherer population (HG), (*b*) substantial gene flow into a Bantu-speaking group, (*c*) differentiation of an "Eastern" Bantu component (EB, including the MB population), coupled or not with limited gene flow. NC, Niger-Congo; WB, Western-Bantu.

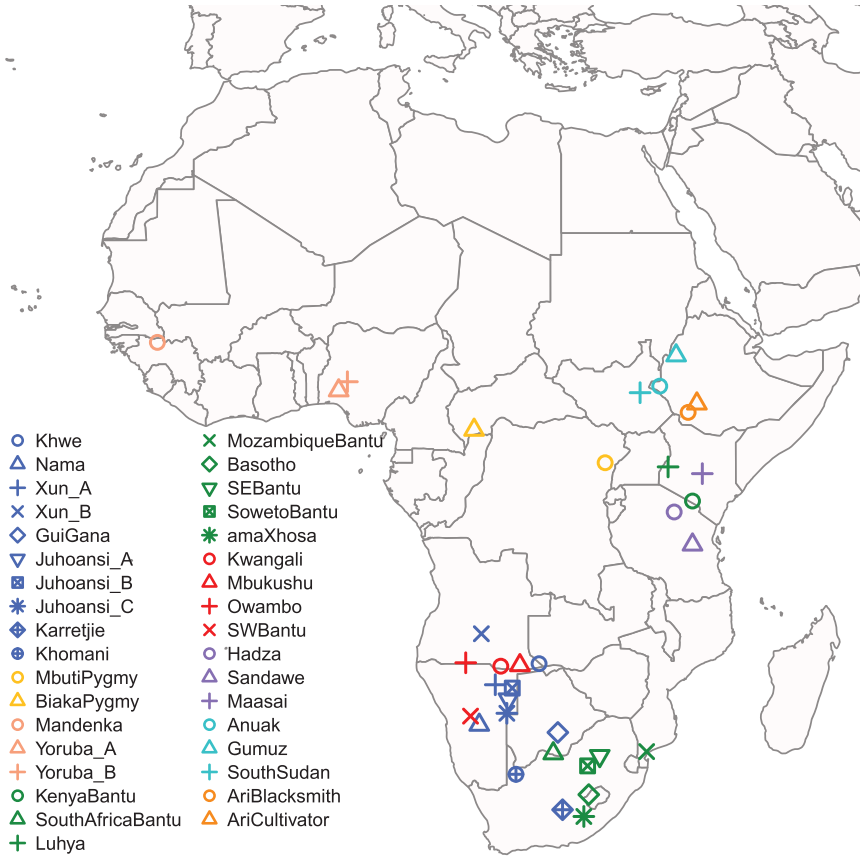


Fig. 2.—Map of the populations analyzed. The colors indicate different population groups: Blue, Khoisan speakers; yellow, Pygmies; salmon, non-Bantu Niger-Congo speakers; green, East-Bantu speakers; red, West-Bantu speakers; purple, language isolates; turquoise, Nilo-Saharan; and orange, Afro-Asiatic. Information on the populations included can be found in the supplementary material, Supplementary Material online.

Interestingly, despite the low number of markers used, the analysis identified East African ancestry in six (out of seven) Khoisan populations, as previously reported (Pickrell et al. 2012). On the other side, the analysis did not find signature of admixture in Karretjie, Owambo, and Kwangali since significant values of the statistics were not observed. However, even though not significant, for these three populations we observed negative values of *f3* statistics 20, 16, and 38 times,

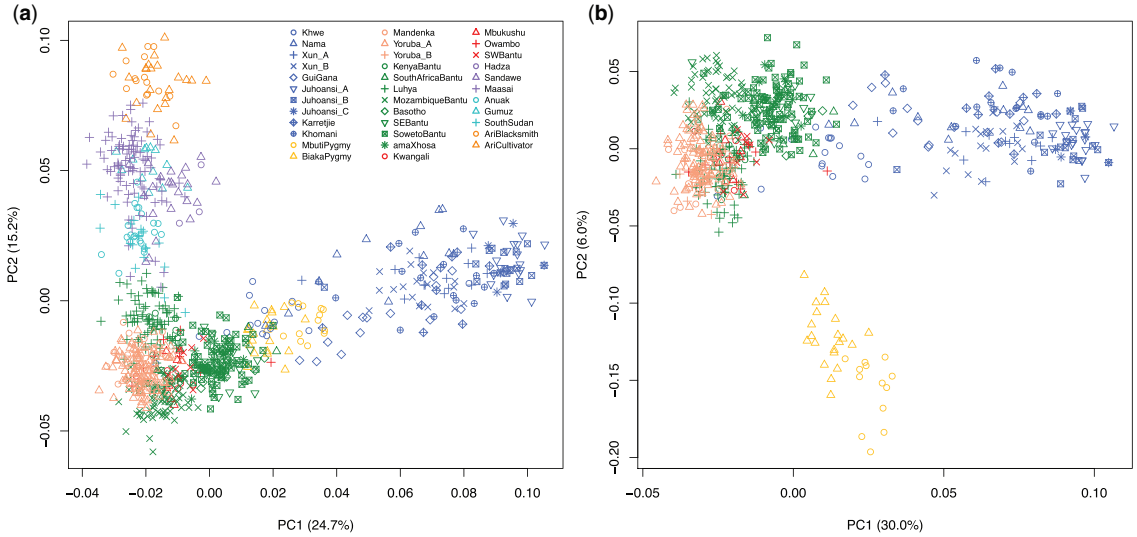


Fig. 3.—Principal Components (PC) plots for the first and second components: (*a*) Full data set; (*b*) Data set excluding Afro-Asiatic, Nilo-Saharan, and language isolates. Symbols and colors as in figure 2: Blue, Khoisan speakers; yellow, Pygmies; salmon, non-Bantu Niger-Congo speakers; green, East-Bantu speakers; red, West-Bantu speakers; purple, language isolates; turquoise, Nilo-Saharan; and orange, Afro-Asiatic.

respectively. No signature of admixture was found in Mozambique since no statistically significant values of *f3* statistics were observed. In addition, contrary to what reported for Karretjie, Owambo, and Kwangali, none of the *f3* involving Mozambique as target has nonsignificant negative values.

Given the low number of markers analyzed, we performed a cross-validation (CV) *f3* analysis in 1,900 resampled samples composed by different ancestry of Yoruba and of Juhoansi, Sandawe, and Mbuti, as described in the Materials and Methods section. The *f3* statistics for all the simulated samples were always significant, with the exception of Yoruba95%Juhoansi5% and Yoruba95%Mbuti5% where the tests characterized by a *Z*-score lower than −3 were, respectively, 99% and 98%. However, it must be stressed that even for the nonsignificant tests the *Z*-scores were always below −2 (Juhoansi: −2.19; Mbuti: −2.98, −2.90).

The pairwise F_{ST} population tree shows a separation of the main groups present in our data set, first with a group of Afro-Asiatic, Nilo-Saharan and linguistic isolates, then the Khoisan speakers (except the Khwe, a highly admixed population as can be seen in fig. 4*a*) and the NC-speaking populations (fig. 5), with the RHG located between the last two. In line with what is observed for the ADMIXTURE components, the South-Eastern populations grouped with the other NC populations, the internal branching order mirroring the West–East–Southeast linguistic subdivisions (Currie et al. 2013).

Discussion

The spread of Bantu-speaking groups across Southern Africa has significantly shaped the distribution of genetic variation of Sub-Sahara Africa (Tishkoff et al. 2009). However, such dispersal did not occur in a void, as hunter-gatherers and pastoralists were already present in the areas that Bantu-speaking farmers were entering. The interactions between residents and newcomers were different in different areas, resulting in different degrees of integration and admixture between these groups (Mitchell 2002; Barbieri, Butthof, et al. 2013; Barbieri et al. 2014; Marks et al. 2014; Patin et al. 2014; Pickrell et al. 2012, 2014). Archaeological and genetic data suggest that foraging groups were once present all across Southern Africa, with a much larger distribution than the previously observed (Mitchell 2010; Barbieri, Vicente, et al. 2013). Their former presence over a wide geographic area has raised the possibility of the assimilation of these non-Bantu groups into farming communities through complex integration dynamics (Mitchell 2002), with the legacy of these events being recovered in contemporary Bantu-speaking populations (Sikora et al. 2011; Barbieri, Butthof, et al. 2013; Marks et al. 2014). It is within this context that the previously reported "South-Eastern" African component had been interpreted: The genetic signature of a pre-Bantu community assimilated into the Bantu speakers of Mozambique (Sikora et al. 2011). Archaeological evidence supports the idea that the dispersal of Bantu speakers was relatively fast along the Eastern part of Africa, which suggested reduced interactions between the

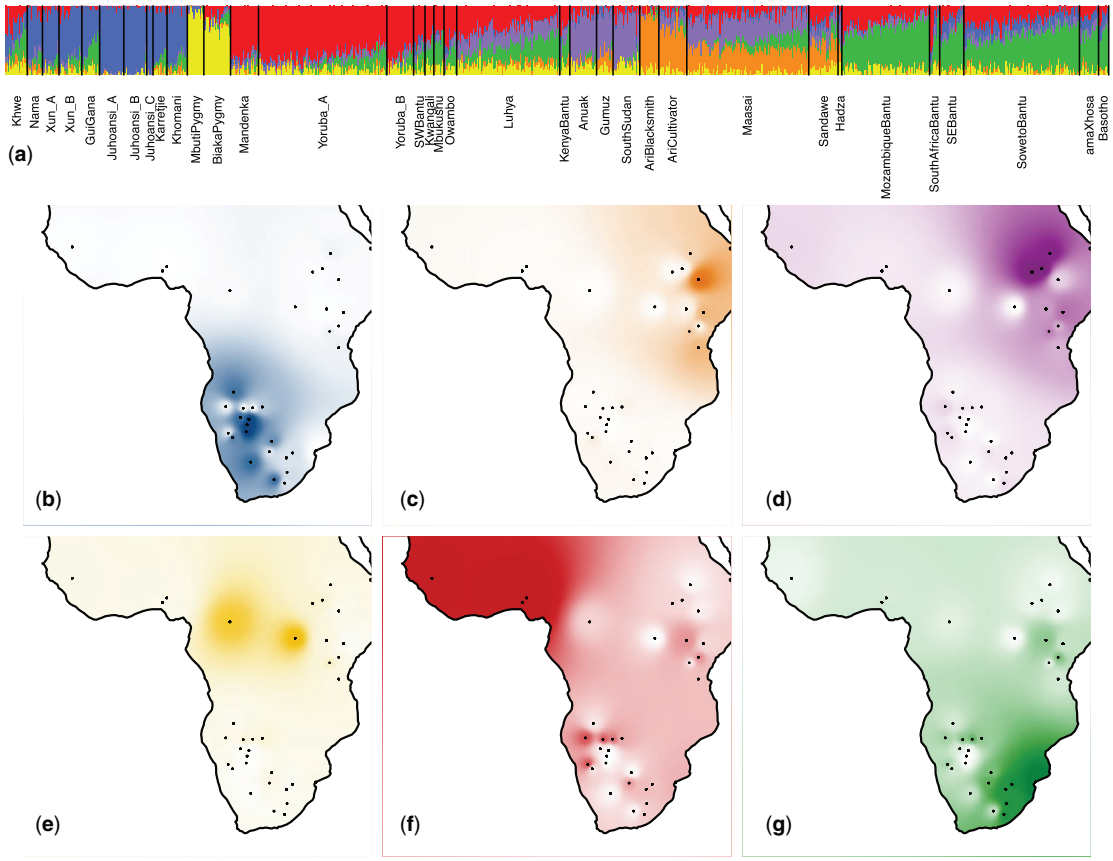


Fig. 4.—(*a*) ADMIXTURE plot for $K = 6$; (*b–g*) interpolation maps of the ADMIXTURE components for $K = 6$. Colors of the components on the maps are the same as in the ADMIXTURE plot.

Table 1

F_{ST} between ADMIXTURE Components at $K = 6$

	NC-East	NC-West	East-1	East-2	Pygmies	Khoisan
NC-East	—					
NC-West	0.042	—				
East-1	0.099	0.081	—			
East-2	0.123	0.109	0.110	—		
Pygmies	0.097	0.095	0.132	0.137	—	
Khoisan	0.129	0.132	0.165	0.158	0.102	—

Bantu-speaking populations and the ones previously present in the occupied areas (Alexander 1984). The change in the pace of migration occurred much further south, around Lesotho and South Africa, where the ecological conditions encountered were inappropriate for the crops used by these farming communities. The slowing of the process favoured more extended interaction between communities and possibly facilitated gene-flow (Bohannan and Plog 1967; Alexander 1984). The different admixture dynamics experienced by Bantu-speaking communities in different areas of South-Eastern Africa resulted in different degree of assimilation of hunter-gatherers and pastoralist communities (Marks et al. 2014).

The previously reported differentiation of the Mozambique sample from other Bantu-speaking populations was originally interpreted as evidence for a pre-Bantu genetic component in South-Eastern Africa (Sikora et al. 2011). When we reanalyzed the Mozambican data within a data set comprising several Southern African Bantu-speaking populations, the reported uniqueness of this population disappeared. In the second component of the PCA, Mozambique appears as part of a cluster of several sub-Saharan African populations speaking NC languages and distributed along a Northeast–Southwest continuum (fig. 3*b*). The ADMIXTURE components present in

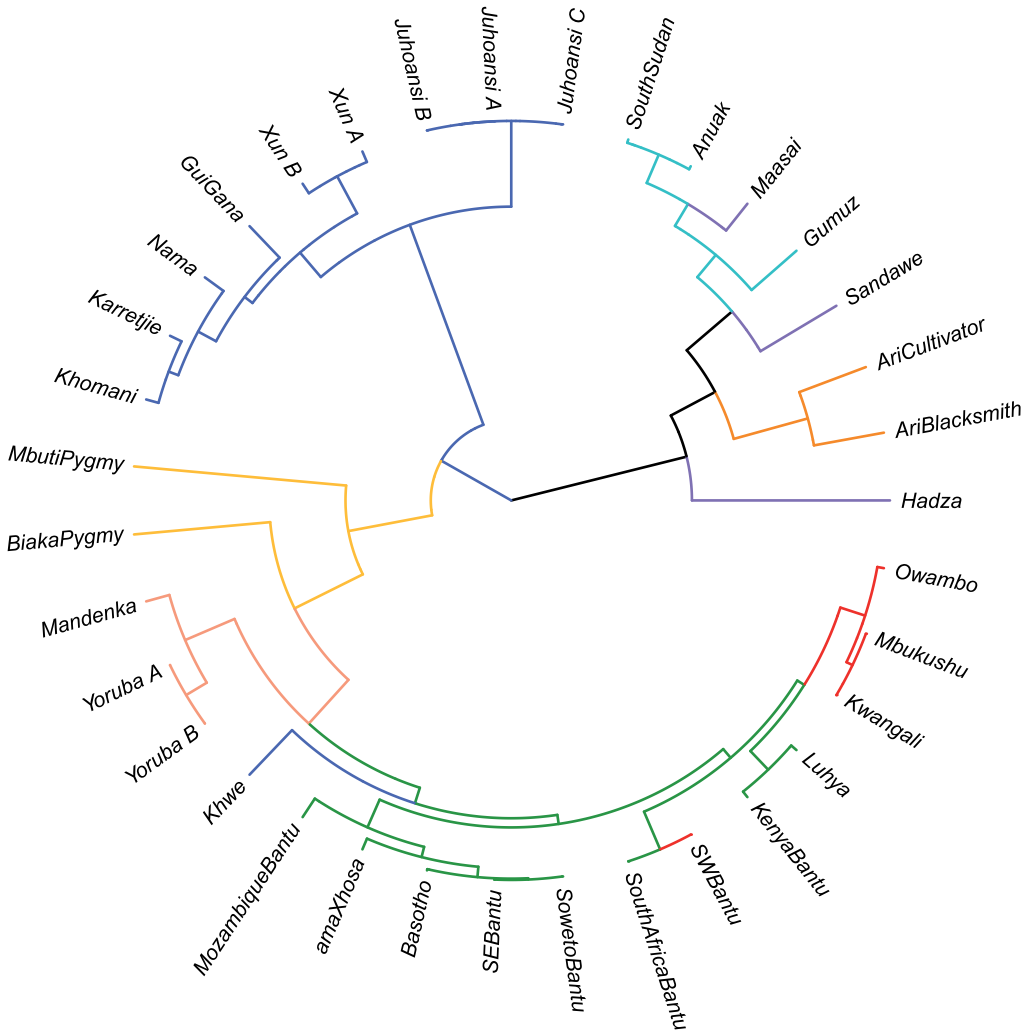


Fig. 5.—Hierarchical population tree based on pairwise F_{ST} values. Color of the branches corresponds to the color of the populations in figure 2.

the Mozambican samples are the same as those of other East and South-Eastern African populations from Kenya, South Africa, and Lesotho (fig. 4*a*). The major component in Mozambique is also the major component in other populations from the South-Eastern region (green in fig. 4*a*), and is present at lower frequencies in other NC speakers. Importantly, the closer affinity of this component to the other NC component (NC-West) suggests a more recent common origin for the two NC clusters than any of the others. The pairwise F_{ST} population tree confirms the clustering of all NC-speaking populations, with the South-Eastern Bantu-speakers placed within the variation of other Bantu-speaking populations (red and green branches in fig. 5). Overall, is there any support for the presence of a predominantly non-Bantu component in South-Eastern Africa? Our results suggest some degree of stratification among Bantu-speaking populations which matches the West versus East NC linguistic split (Currie et al. 2013). The differential distribution of the two ADMIXTURE components in the NC populations, coupled with the gradual increase of the green component when moving from populations living in the west, to those living in the East and the Southeast, hints to a degree of differentiation among Bantu-speaking populations shaped by founder events and drift associated with their geographic dispersal rather than massive gene introgression from other groups (Ramachandran et al. 2005). No other components, known (Khoisan or RHG) or unknown, are detected within the Mozambican sample at significant level.

Despite the lack of a major non-Bantu component in Mozambique, we note that other groups in Southern Africa are characterized by such kind of contribution, in this case Khoisan (fig. 4*a*) in agreement with previous reports of Khoisan admixture in Southern African Bantu-speaking populations (Schlebusch et al. 2012; Petersen et al. 2013). This is confirmed by *f3* analyses that highlight several admixture events involving African populations. However, we failed to identify any signature of admixture between Khoisan populations (or other hunter-gatherer groups) and Bantu speakers in Mozambique. In fact, none of the tests in the form *f3*(*Target; Pop A, Pop B*) gave significant or negative values. Furthermore, our CV analysis based on resampled data shows that the probability of having a false negative result is very low. Even though the absence of gene flow into the Mozambican sample here analyzed cannot be completely excluded given the low number of analyzed markers, we show that a high contribution from non-Bantu populations is highly unlikely in this population. We also noted that this might not be the case for all populations from this area as the analyzed sample is possibly not representative of the region as a whole. Sporadic highly divergent mtDNA and Y chromosome lineages have been reported in Zambia and Mozambique despite the lack of robust population signatures of admixture (Batini et al. 2011; Barbieri, Butthof, et al. 2013; Barbieri, Vicente, et al. 2013; Marks et al. 2014). Similarly, we cannot refute a scenario where admixture occurred in the past but left no signature in modern day populations. Future analysis of ancient DNA will be crucial to further refine the model discussed in this work. Unfortunately, the climatic conditions in Sub-Saharan Africa make the extraction of endogenous DNA still challenging but new methods developed in the last few years are promising as recently showed (Morris et al. 2014; Sirak et al. 2015). The observation of some degree of differentiation between Western and Eastern–South–Eastern Bantu speakers also suggests that higher SNP densities might provide the power to identify and distinguish populations belonging to the two Bantu linguistic branches, which could be of help in elucidating the demic pattern of Bantu speakers dispersal as well as reconstruct the African ancestry of recently admixed populations (de Filippo et al. 2012; Montinaro et al. 2015).

In conclusion, our results underline the role played by both the dispersal of Bantu-speaking groups and the interaction with previous inhabitants in shaping the genetic and cultural variation of Bantu-speaking populations of Southern Africa. Such combined effect has been suggested for other regions of Africa where Bantu-speaking groups are present (Patin et al. 2014). The integration of archaeological and linguistic data with a more complex demographic model is necessary to better understand the process through which languages, people, and technology were spread following the so-called "Bantu expansion" (Marks et al. 2014).

Materials and Methods

Samples

Saliva samples from unrelated individuals were collected in Namibia and Lesotho using the Oragene DNA collection kits (DNA Genotek, Inc., Ottawa, ON, Canada) and DNA was extracted according to manufacturer's protocols. The samples presented here were collected during three field trips. The Basotho speakers were collected in Lesotho in 2009 (Marks et al. 2012, 2014); the Owambo, Mbukushu, and Kwangali were collected in Namibia in 2010 and 2012. All participants were healthy adults from whom appropriate informed consent was obtained. These investigations received ethical approval by the Oxford Tropical Research Ethics Committee, the Lesotho Ministry of Health and Social Welfare, the Lesotho Ministry of Local Government, the Lesotho Ministry of Tourism, Environment and Culture, and the Namibian Ministry of Health and Social Services. Ethnic and linguistic information about the donors, as well as their parents and grandparents if known, was also collected.

A total of 33 individuals (8 Basotho, 10 Owambo, 8 Mbukushu, and 7 Kwangali) were genotyped using the Illumina Human 610-Quad BeadChip (Illumina, San Diego, CA) (Basotho and Owambo) and the Human Omni5-Quad BeadChip (Illumina) (Mbukushu and Kwangali). The analyzed data described in this article are available on CC's group website (https://capelligroup.wordpress.com/data/).

Data Set

We combined our data with available genome-wide SNP genotypes from different African populations, using the software PLINK version 1.9 (Chang et al. 2015; Purcell and Chang 2015) (fig. 2).

The assembled data set was pruned through quality control (QC) filtering both SNPs and individuals. Each single population data set was initially filtered to remove SNPs and individuals with a call rate below 0.9. After the merge, an additional QC step was performed, removing SNPs and individuals with a call rate below 0.98. Related individuals up to the second degree were removed from our data set using the software KING (Manichaikul et al. 2010). To overcome the effects of markers in strong linkage disequilibrium, we removed SNPs with an $r^2 > 0.4$ using a sliding window of 200 SNPs, shifted at 25 SNP intervals (Behar et al. 2010).

To detect and remove samples with strong European admixture from our data set, the CEU data from HapMap Phase 3 (http://hapmap.ncbi.nlm.nih.gov/, last accessed July 30, 2015) (International HapMap Consortium 2003) were initially included in the analysis. We then performed an ADMIXTURE run (Alexander et al. 2009) with the default options and individuals characterized with more than 10% of the European

component (for $K = 4$) were removed. The final data set was composed of a total of 868 individuals from 35 populations (International HapMap Consortium 2003, 2007; Li et al. 2008; Henn et al. 2011; Sikora et al. 2011; Schlebusch et al. 2012; May et al. 2013; Petersen et al. 2013) typed on 1,747 SNPs (fig. 2 and supplementary table S1, Supplementary Material online). Similar number of SNPs has been successfully used to recover evident and robust patterns of genetic diversity in African populations (Sikora et al. 2011).

Statistical Analysis

Population genetic structure was initially explored through PCA with PLINK software version 1.9 (Chang et al. 2015; Purcell and Chang), and we used ADMIXTURE (version 1.23) (Alexander et al. 2009) to further explore genetic variation, using a range of putative source clusters (K) from 2 to 20. The most supported value of K was estimated using the CV procedure, as implemented in ADMIXTURE (Alexander et al. 2009; Alexander and Lange 2011). We computed F_{ST} between the components to assess genetic differentiation (Alexander et al. 2009). The spatial distribution of the ADMIXTURE components ($K = 6$, see Results) was visualized by plotting the fraction of each component in the analyzed populations on a map with the interpolation plugin of the QGIS software, using the Inverse Distance Weighting method with a distance coefficient of 3 (QGIS Development Team 2014). Weir and Cockerham's pairwise F_{ST} between populations was calculated using the R package StAMPP (Weir and Cockerham 1984; Pembleton et al. 2013). The R packages stats and ape were used to build and visualize, respectively, a hierarchical tree with these values, using the complete linkage method (Paradis et al. 2004; Legendre P and Legendre L 2012; R Core Team 2014).

Given the fact that ADMIXTURE analysis is not a formal text of gene-flow, we carried out a *f3* test (Reich et al. 2009) for windows of 100 markers using the threepop companion software in the TreeMix suite (Pickrell and Pritchard 2012), and reported *f3* statistics characterized by a value lower than −3. The *f3* statistics has been demonstrated to be robust, even with ascertainment bias (Patterson et al. 2012). Briefly, in a *f3* test with the form *f3*(*Target; Pop A, Pop B*) a significantly negative value of the statistic highlights a complex phylogeny of the Target population, that has a certain amount of ancestry from populations related to A and B. However, a positive *f3* value does not necessarily imply the absence of admixture. For this reason, and given the low number of markers used in this analysis, we performed a *f3*(*Target; Yoruba, Pop 1*) test on 100 simulated samples composed by Yoruba and Pop1, with proportion alpha and 1-alpha respectively, where Pop1 is represented by Juhoansi, Sandawe or Mbuti, and alpha $\in\{0.05, 0.10, \ldots, 0.95\}$.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Wenner-Gren Foundation, the University of Oxford Boise Fund, the John Fell Oxford University Press (OUP) Research Fund, and Fundação para a Ciência e Tecnologia (grant number SFRH/BD/90648/2012 to M.G.-S.). The authors thank all the people who donated their DNA samples making this work possible and all the various people and institutions that helped with the organization of the fieldwork and the collection of the samples. They also thank Sarah J. Marks and Simon Marks for the help with the collection of some of the Namibian samples, and Dr Martin Sikora and Professor Jaume Bertranpetit for sharing the Mozambique data with them.

Literature Cited

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Alexander JA. 1984. Early frontiers in southern Africa. In: Hall M, Avery G, Avery DM, Wilson ML, Humphreys AJB, editors. Frontiers: southern African archaeology today. Oxford: B.A.R. International Series. p. 12–23.

Barbieri C, Butthof A, Bostoen K, Pakendorf B. 2013. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur J Hum Genet. 21:430–436.

Barbieri C, et al. 2014. Unraveling the complex maternal history of Southern African Khoisan populations. Am J Phys Anthropol. 153:435–448.

Barbieri C, Vicente M, et al. 2013. Ancient substructure in early mtDNA lineages of southern Africa. Am J Hum Genet. 92:285–292.

Batini C, et al. 2011. Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. Mol Biol Evol. 28:2603–2613.

Behar DM, et al. 2010. The genome-wide structure of the Jewish people. Nature 466:238–242.

Bohannan P, Plog F. 1967. Beyond the frontier; social process and cultural change. Garden City (NY): Published for the American Museum of Natural History by Natural History Press.

Chang CC, et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7.

Currie TE, Meade A, Guillon M, Mace R. 2013. Cultural phylogeography of the Bantu Languages of sub-Saharan Africa. Proc R Soc Lond B Biol Sci. 280:20130695.

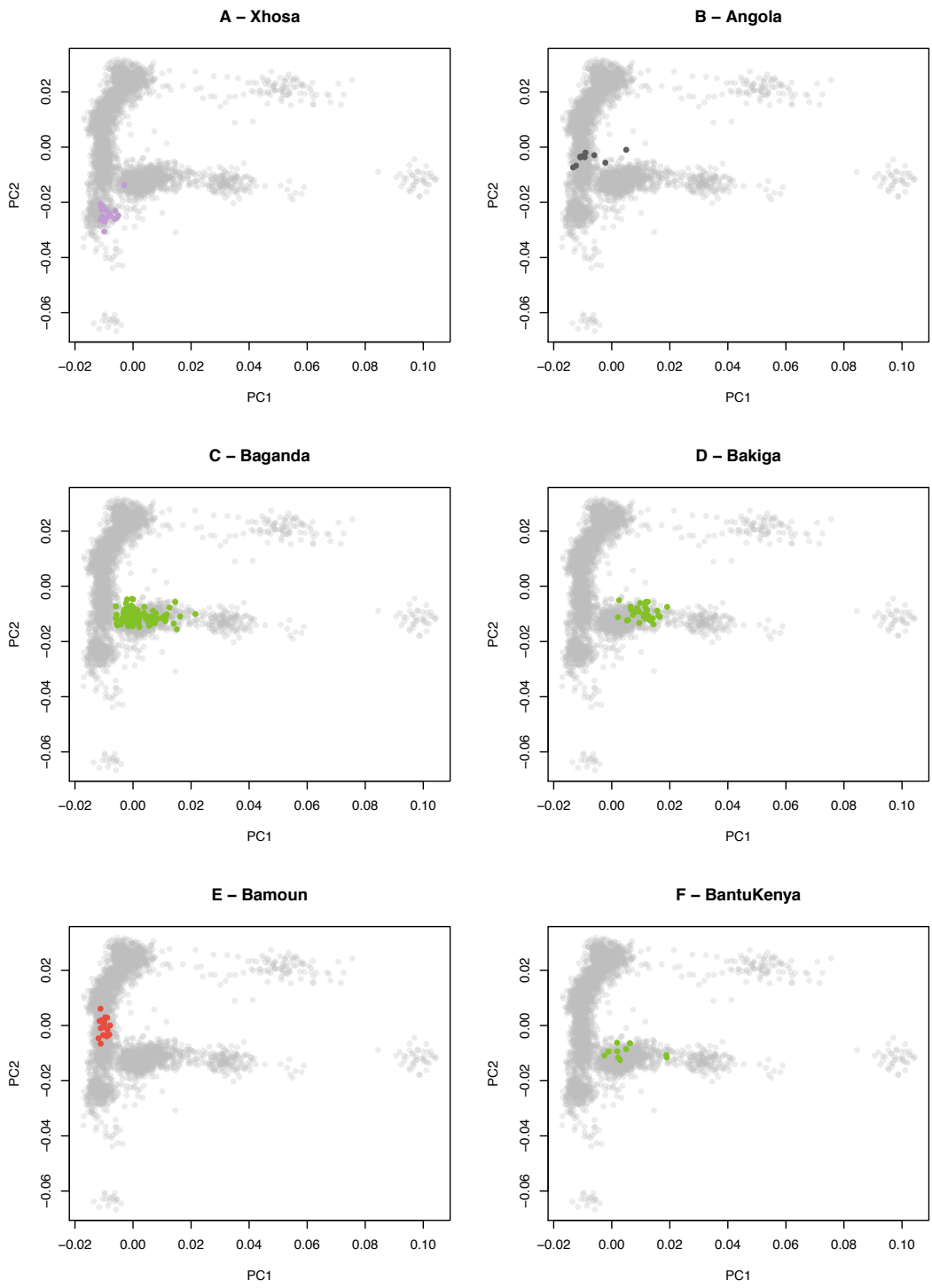
de Filippo C, Bostoen K, Stoneking M, Pakendorf B. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc R Soc Lond B Biol Sci. 279:3256–3263.

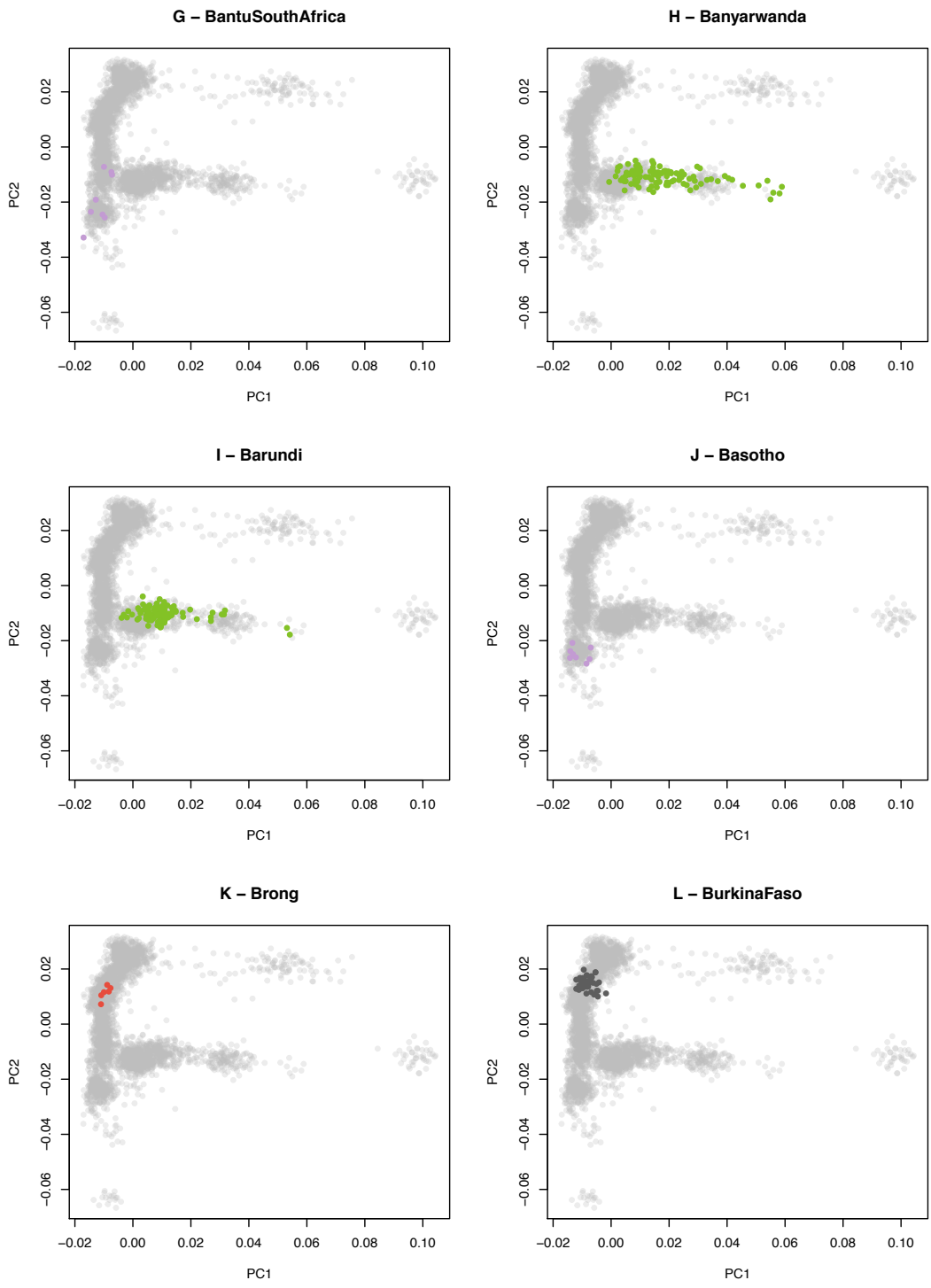
Destro-Bisol G, et al. 2004. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol. 21:1673–1682.

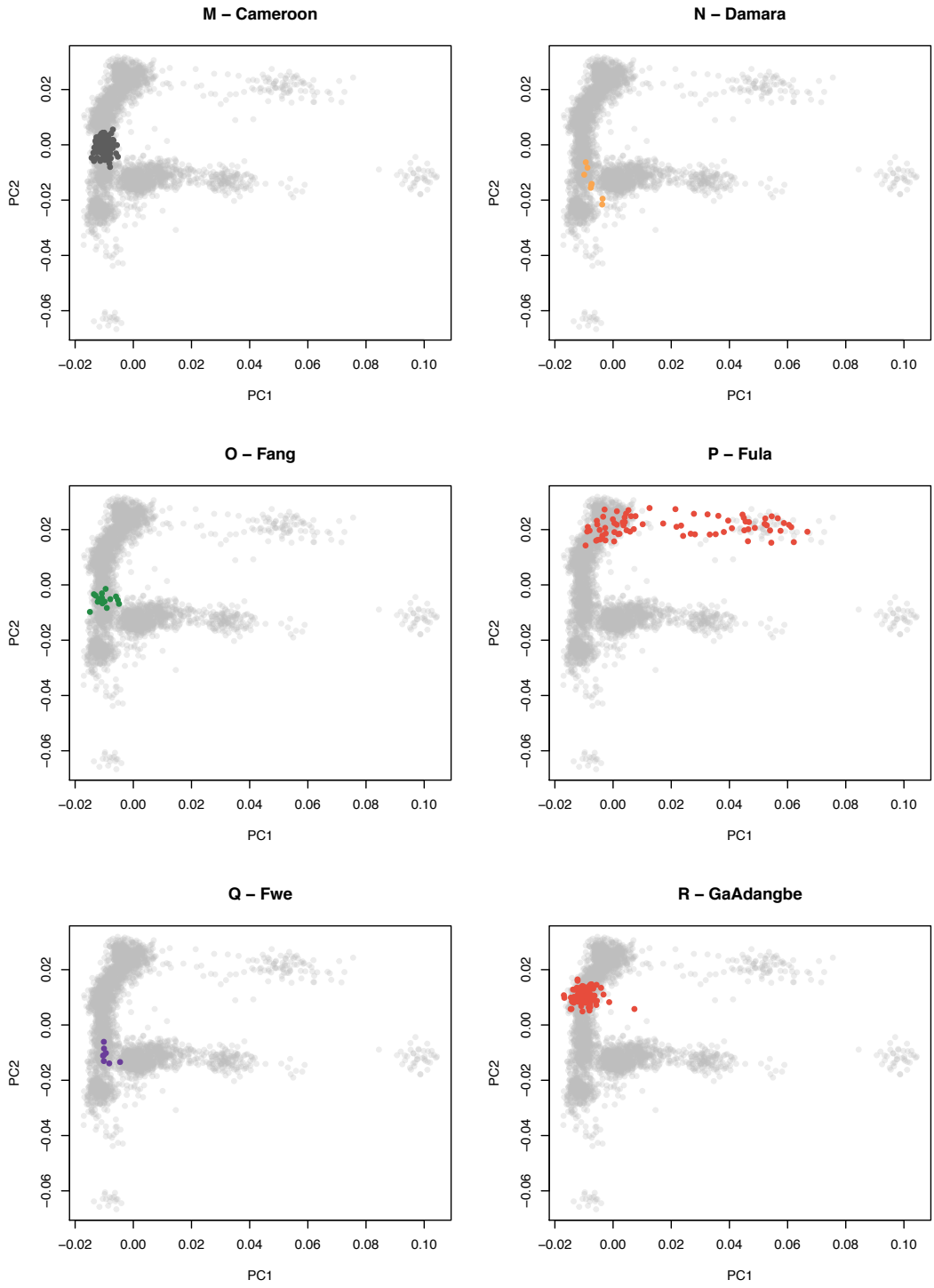
Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. Science 300:597–603.
Güldemann T, Fehn A-M. 2014. Beyond "Khoisan": historical relations in the Kalahari basin. Amsterdam/Philadelphia (PA): John Benjamins Publishing Company.
Hellenthal G, et al. 2014. A genetic atlas of human admixture history. Science 343:747–751.
Henn BM, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A. 108:5154–5162.
International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.
International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.
Legendre P, Legendre L. 2012. Numerical ecology. Amsterdam: Elsevier Science.
Li JZ, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.
Manichaikul A, et al. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867–2873.
Marks SJ, et al. 2014. Static and moving frontiers: the genetic landscape of Southern African Bantu-speaking populations. Mol Biol Evol. 32:29–43.
Marks SJ, Levy H, Martinez-Cadenas C, Montinaro F, Capelli C. 2012. Migration distance rather than migration rate explains genetic diversity in human patrilocal groups. Mol Ecol. 21:4958–4969.
May A, et al. 2013. Genetic diversity in black South Africans from Soweto. BMC Genomics 14:644.
Mitchell P. 2002. The archaeology of southern Africa. Cambridge/New York: Cambridge University Press.
Mitchell P. 2009. Hunter-gatherers and farmers: some implications of 2000 years of interaction in the Maloti-Drakensberg region of southern Africa. In: Ikeya K, Ogawa H, Mitchell PJ, editors. Interactions between hunter-gatherers and farmers: from prehistory to present. Osaka (Japan): National Museum of Ethnology. p. 15–46.
Mitchell P. 2010. Genetics and southern African prehistory: an archaeological view. J Anthropol Sci. 88:73–92.
Mitchell P, Plug I, Bailey G. 2008. Bringing the Kalahari debate to the mountains: domestic livestock and other evidence of late first millennium AD hunter-gatherer/farmer interaction from highland Lesotho. Before Farming [online version] 2008:1–22.
Montinaro F, et al. 2015. Unravelling the hidden ancestry of American admixed populations. Nat Commun. 6:6596.
Morris AG, Heinze A, Chan EK, Smith AB, Hayes VM. 2014. First ancient mitochondrial human genome from a prepastoralist southern African. Genome Biol Evol. 6:2647–2653.
Newman JL. 1995. The peopling of Africa: a geographic interpretation. New Haven (CT): Yale University Press.
Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.
Patin E, et al. 2014. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat Commun. 5:3163.
Patterson N, et al. 2012. Ancient admixture in human history. Genetics 192:1065–1093.
Pembleton LW, Cogan NO, Forster JW. 2013. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Mol Ecol Resour. 13:946–952.
Petersen DC, et al. 2013. Complex patterns of genomic admixture within southern Africa. PLoS Genet. 9:e1003309.
Pickrell JK, et al. 2012. The genetic prehistory of southern Africa. Nat Commun. 3:1143.
Pickrell JK, et al. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A. 111:2632–2637.
Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8:e1002967.
Purcell S, Chang C. 2015. PLINK version 1.9.
QGIS Development Team. 2014. QGIS Geographic Information System: Open Source Geospatial Foundation.
R Core Team. 2014. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Ramachandran S, et al. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 102:15942–15947.
Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461:489–494.
Schlebusch CM, et al. 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338:374–379.
Schlebusch CM, Lombard M, Soodyall H. 2013. MtDNA control region variation affirms diversity and deep sub-structure in populations from southern Africa. BMC Evol Biol. 13:56.
Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J. 2011. A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. Eur J Hum Genet. 19:84–88.
Sirak KA, Fernandes DM, Connell S, Pinhasi R. 2015. No longer the 1%: optimizing ancient DNA yield from Saharan African samples. Poster session presented at: The 84th Annual Meeting of the American Association of Physical Anthropologists; March 25–28, 2015; St. Louis, MO.
Tishkoff SA, et al. 2009. The genetic structure and history of Africans and African Americans. Science 324:1035–1044.
Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population-structure. Evolution 38:1358–1370.

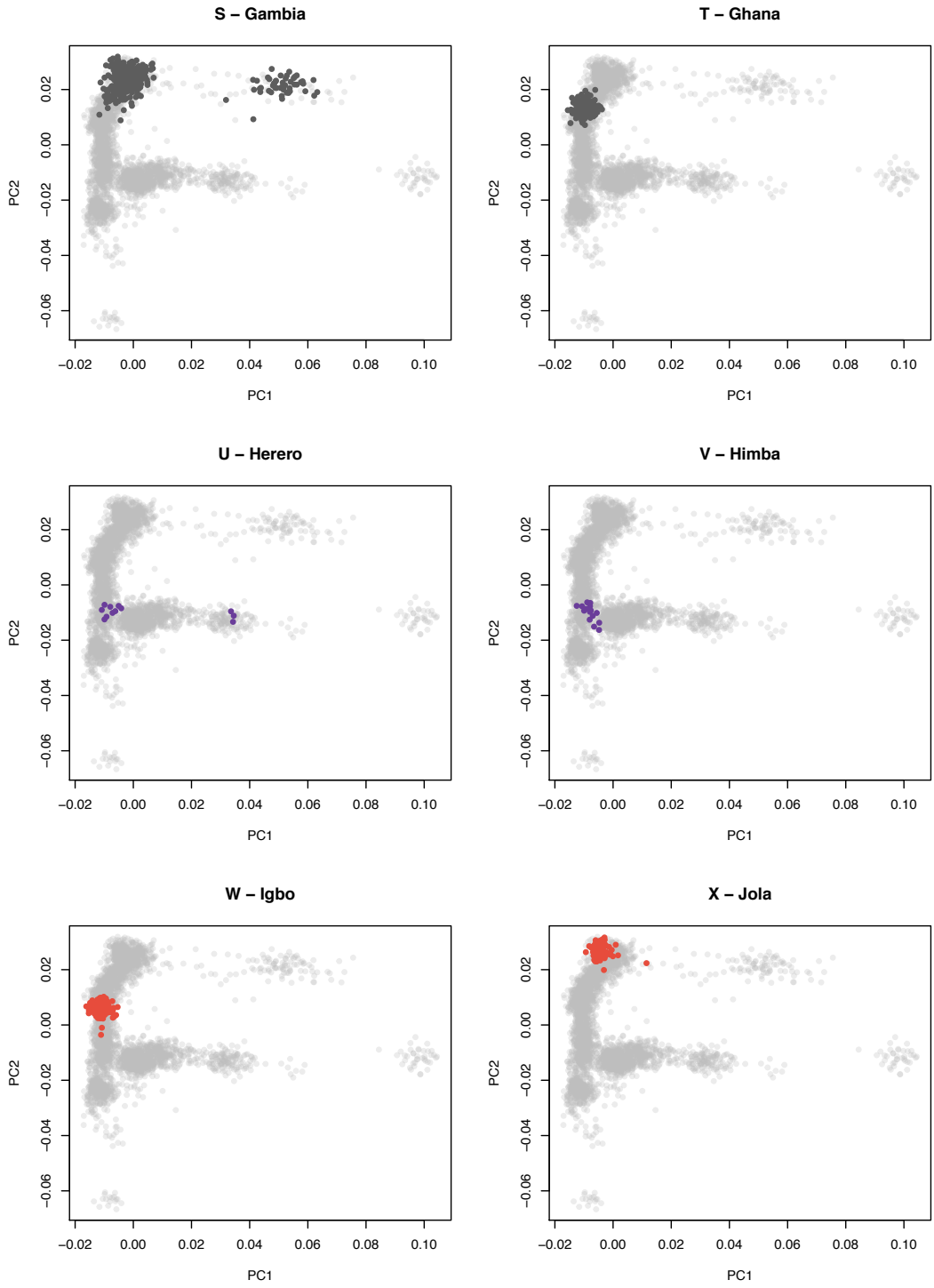
Associate editor: Naruya Saitou

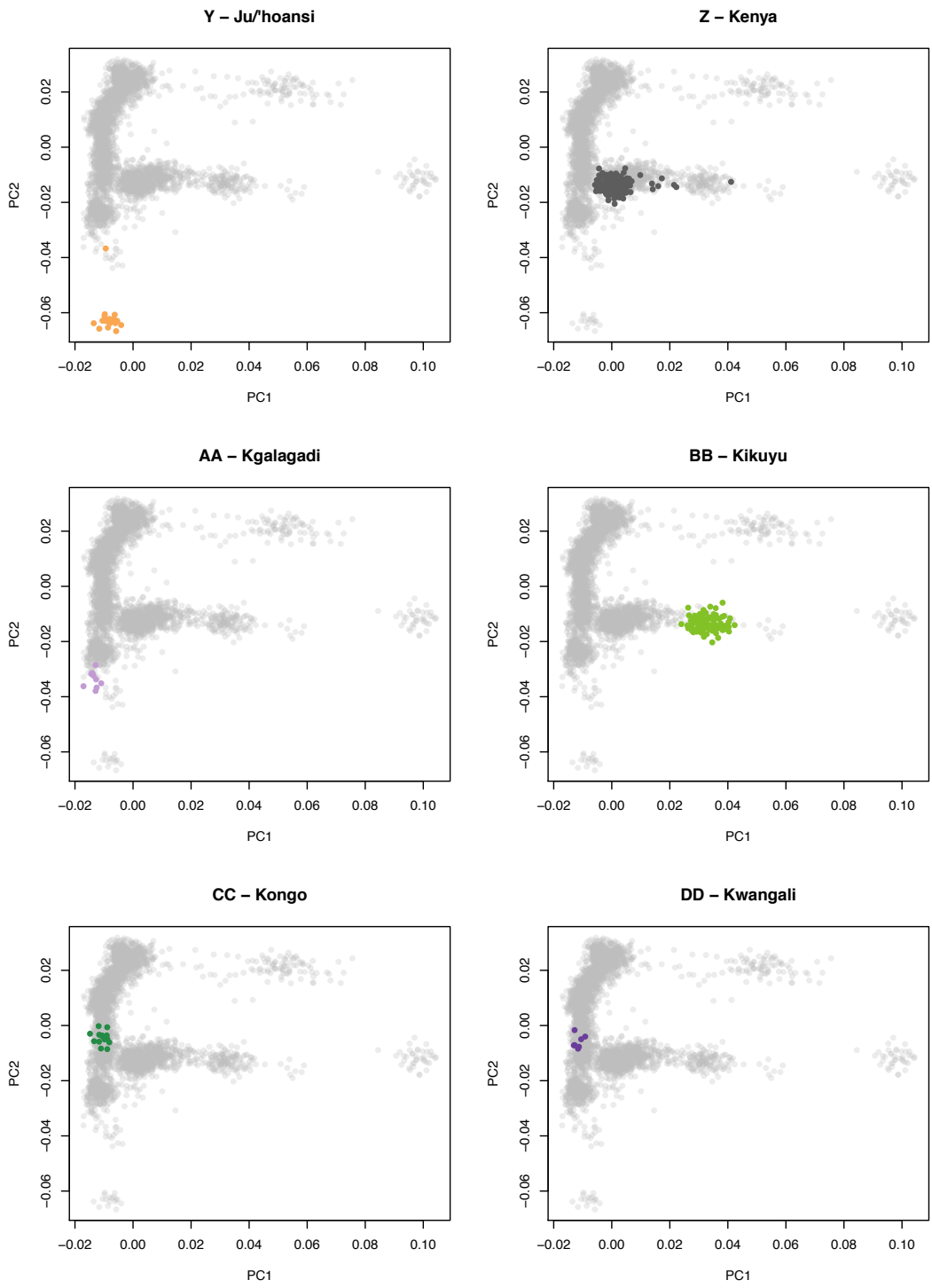
Supplementary Material 5 A—EEE – Principal Component Analysis (PCA) plots for the first and second components for the 'Niger-Congo Genetic Dataset' (after removing the CEU sample) highlighted for each individual population. In each plot all populations are represented in light grey except for one for which the colour correspond to the one on Figure 4.2: light orange, 'Khoisan' speakers; dark orange, Pygmies; light blue, East Africa Somali; dark grey, populations with no linguistic information available; red, non-Bantu Niger-Congo; dark green, Bantu *cluster A*; light green, Bantu *cluster B*; dark purple, Bantu *cluster C*; light purple, Bantu *cluster D*.

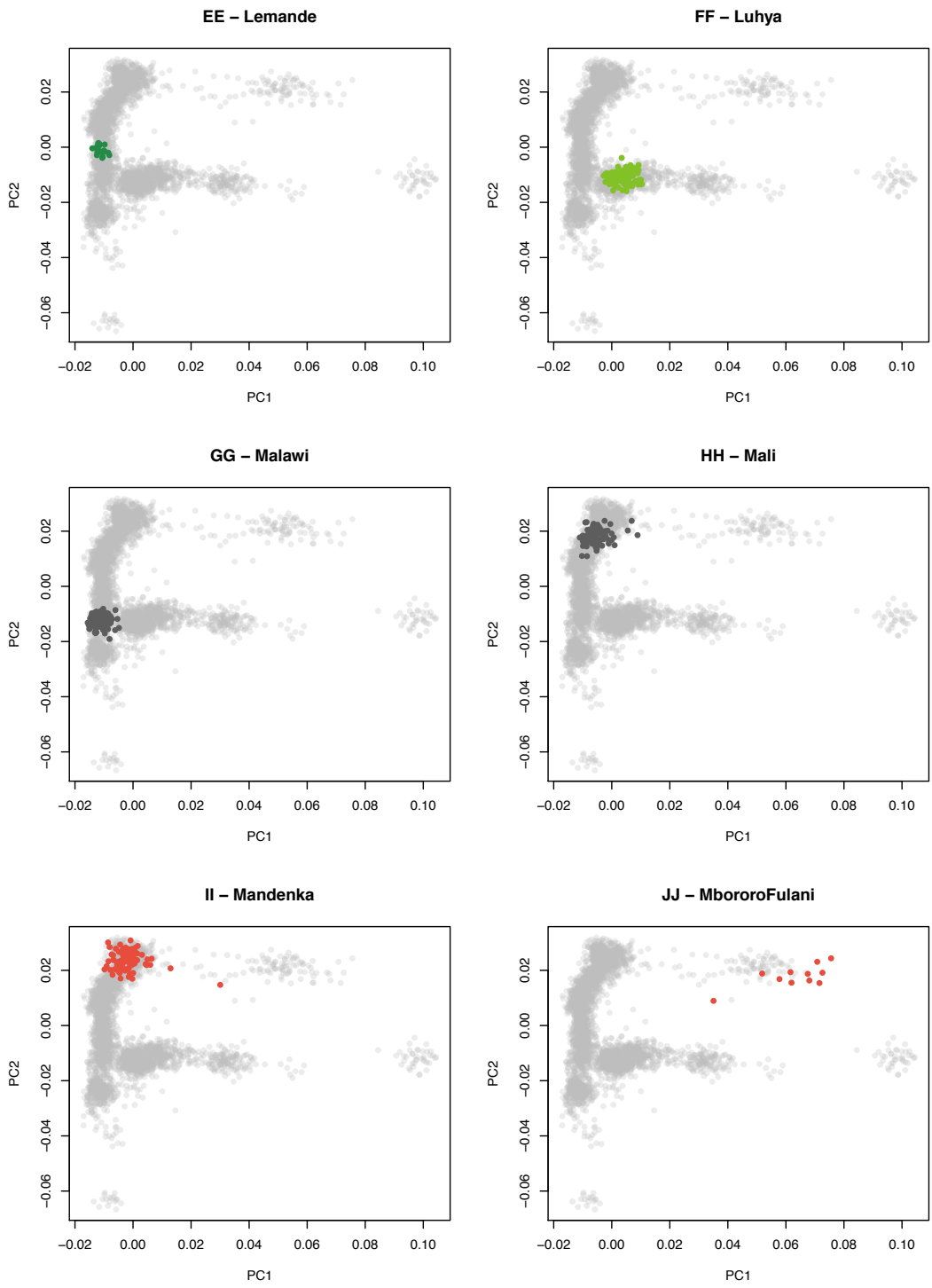

A – Xhosa
B – Angola
C – Baganda
D – Bakiga
E – Bamoun
F – BantuKenya
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
−0.02
0.00
0.02

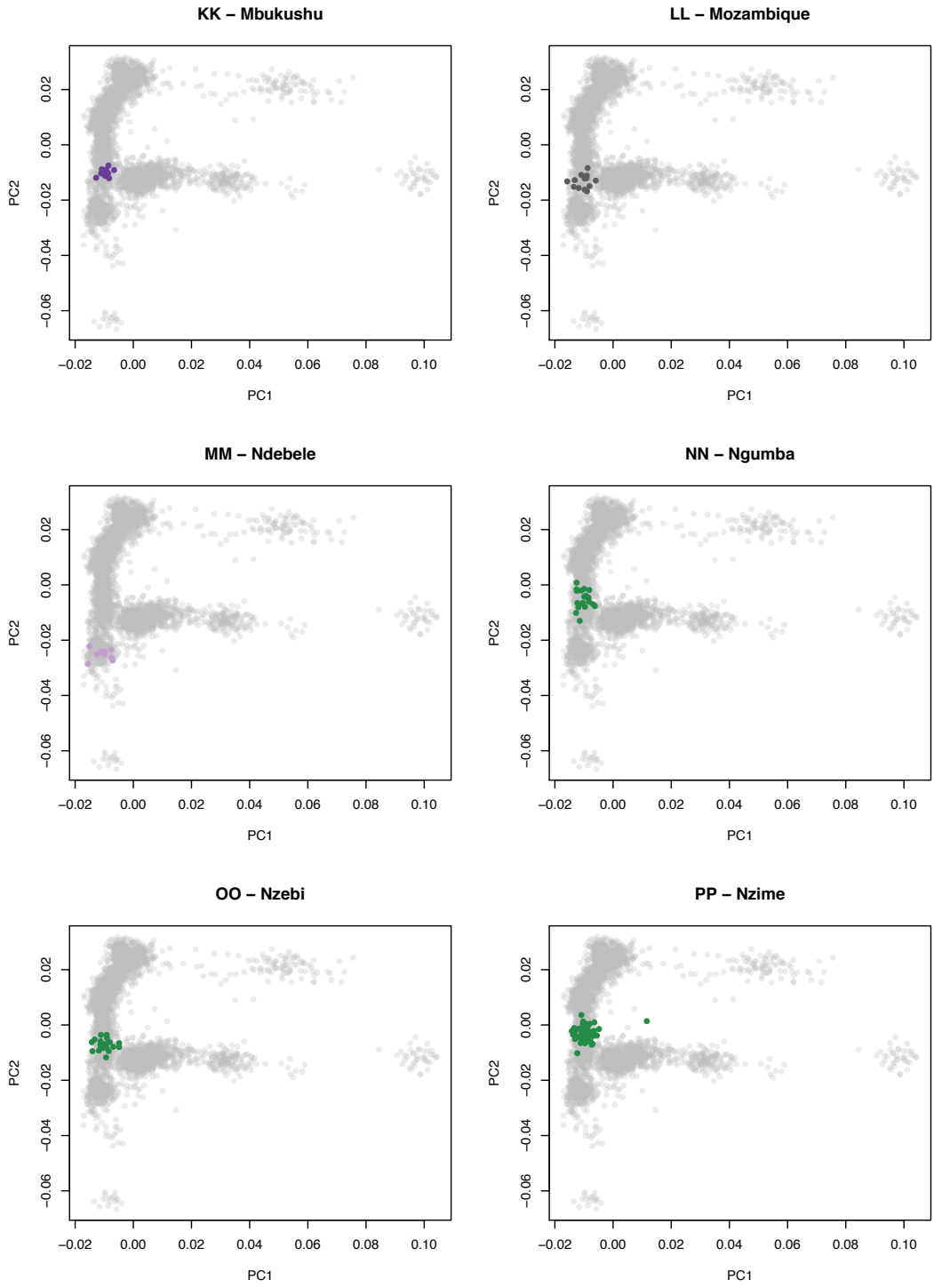

G – BantuSouthAfrica
H – Banyarwanda
I – Barundi
J – Basotho
K – Brong
L – BurkinaFaso
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
−0.02
0.00
0.02

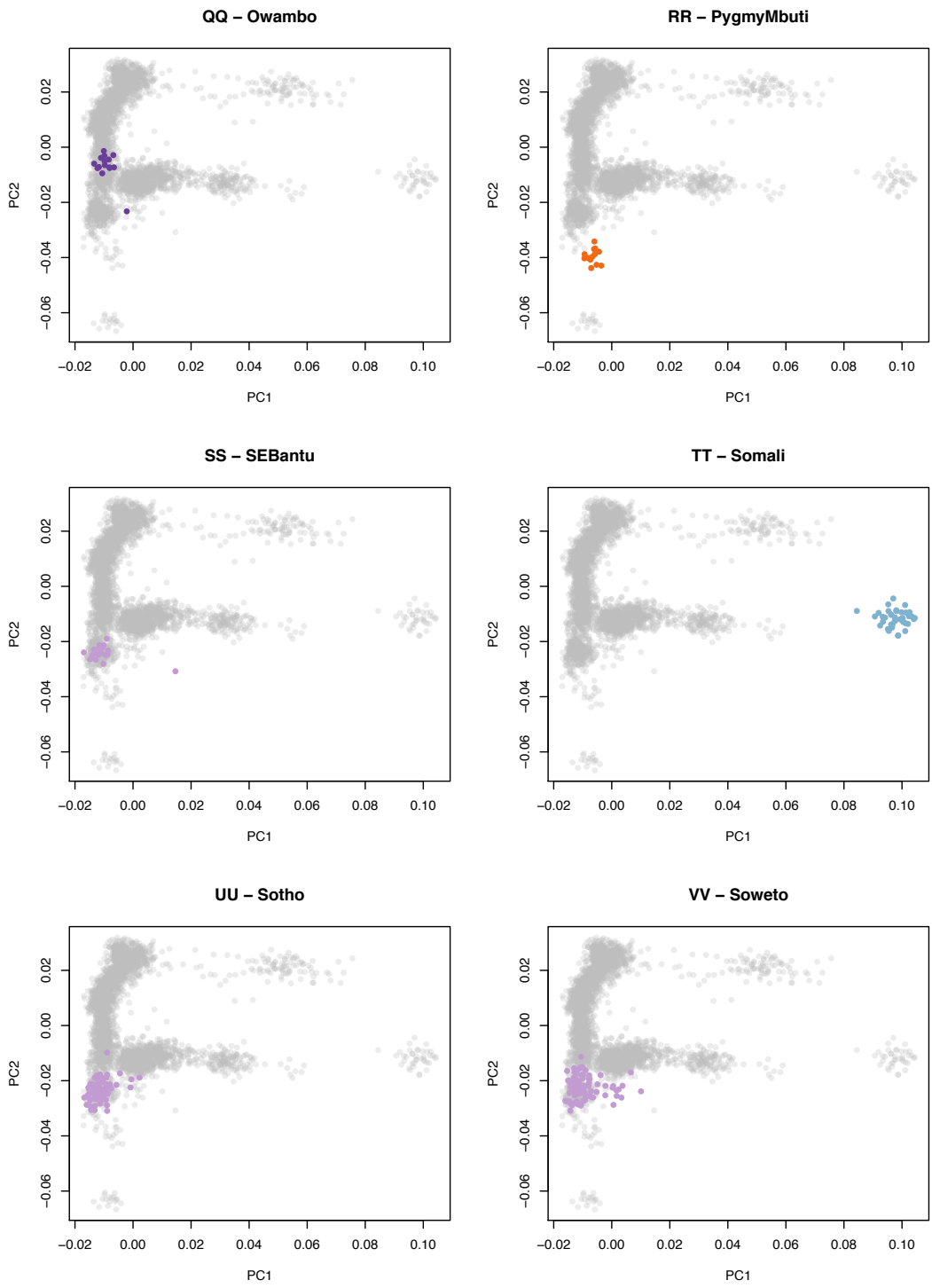

M – Cameroon
N – Damara
O – Fang
P – Fula
Q – Fwe
R – GaAdangbe
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
0.02
0.00
−0.02

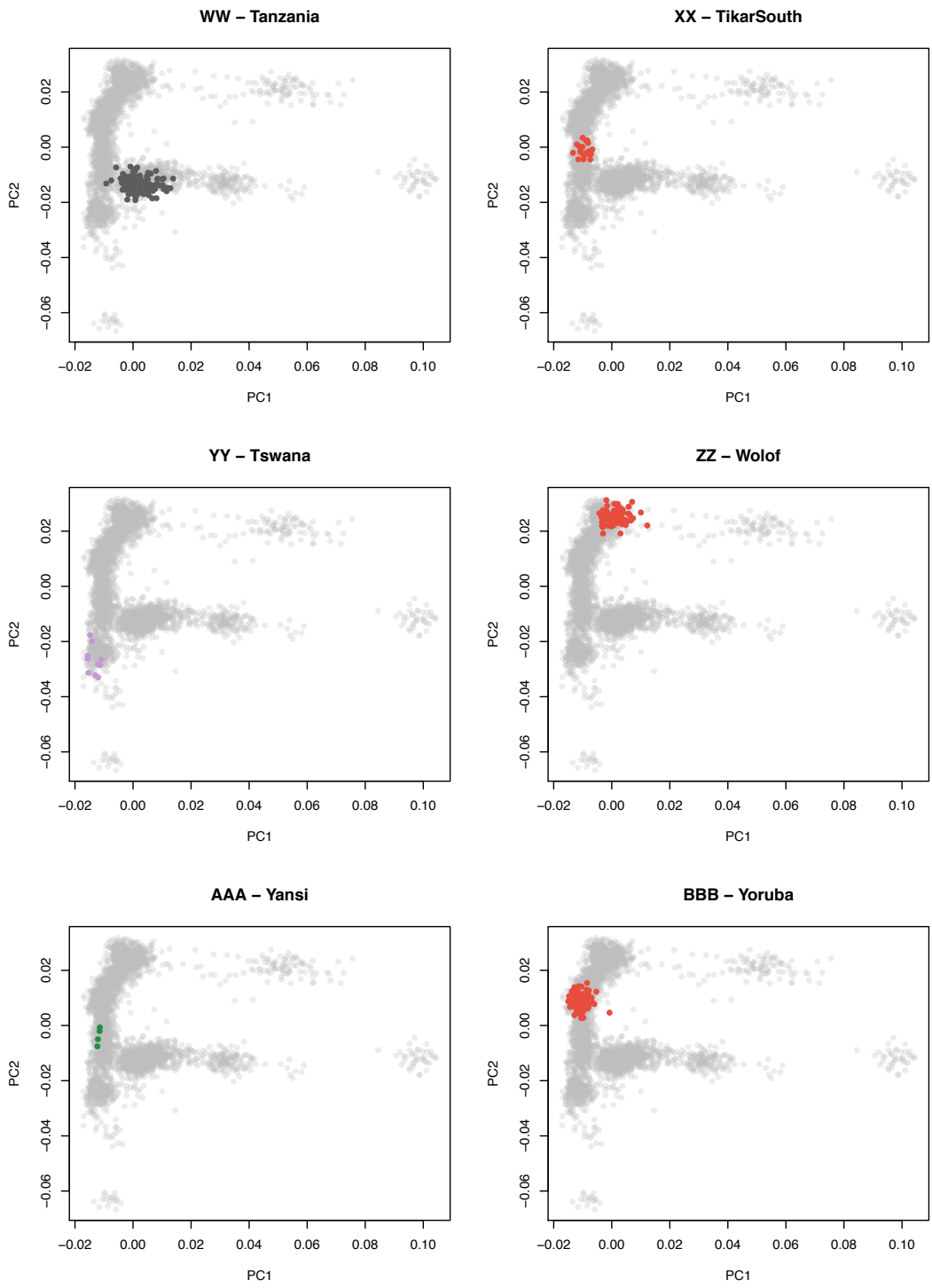

S – Gambia
T – Ghana
U – Herero
V – Himba
W – Igbo
X – Jola
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
−0.02
0.00
0.02

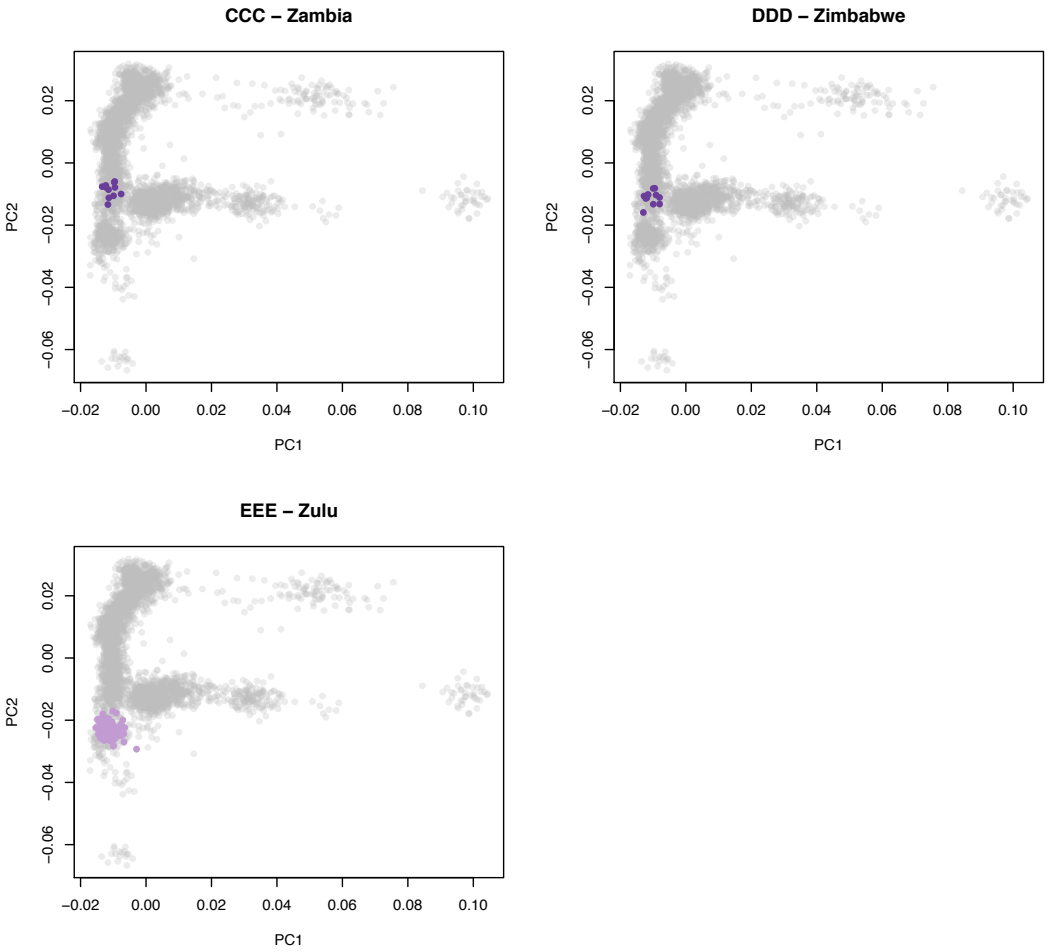

Y – Ju/'hoansi
Z – Kenya
AA – Kgalagadi
BB – Kikuyu
CC – Kongo
DD – Kwangali
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
0.02
0.00
−0.02


EE – Lemande
FF – Luhya
GG – Malawi
HH – Mali
II – Mandenka
JJ – MbororoFulani
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
−0.02
0.00
0.02

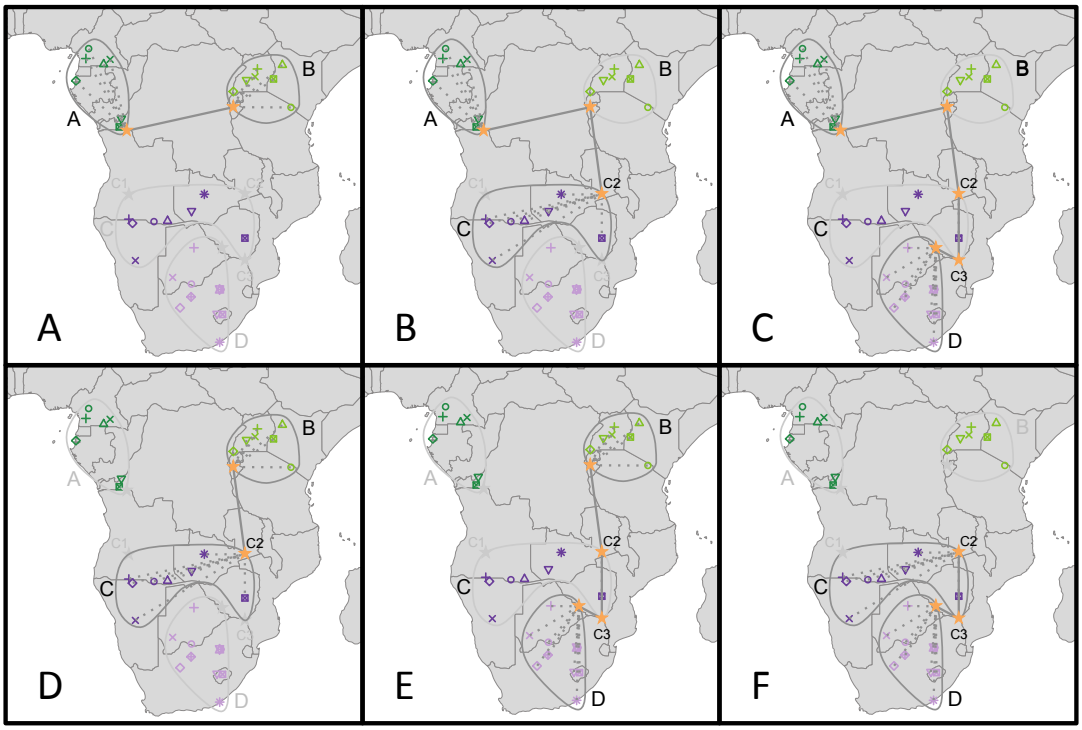

KK – Mbukushu
LL – Mozambique
MM – Ndebele
NN – Ngumba
OO – Nzebi
PP – Nzime
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
−0.02
0.00
0.02


QQ – Owambo
RR – PygmyMbuti
SS – SEBantu
TT – Somali
UU – Sotho
VV – Soweto
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
−0.02
0.00
0.02

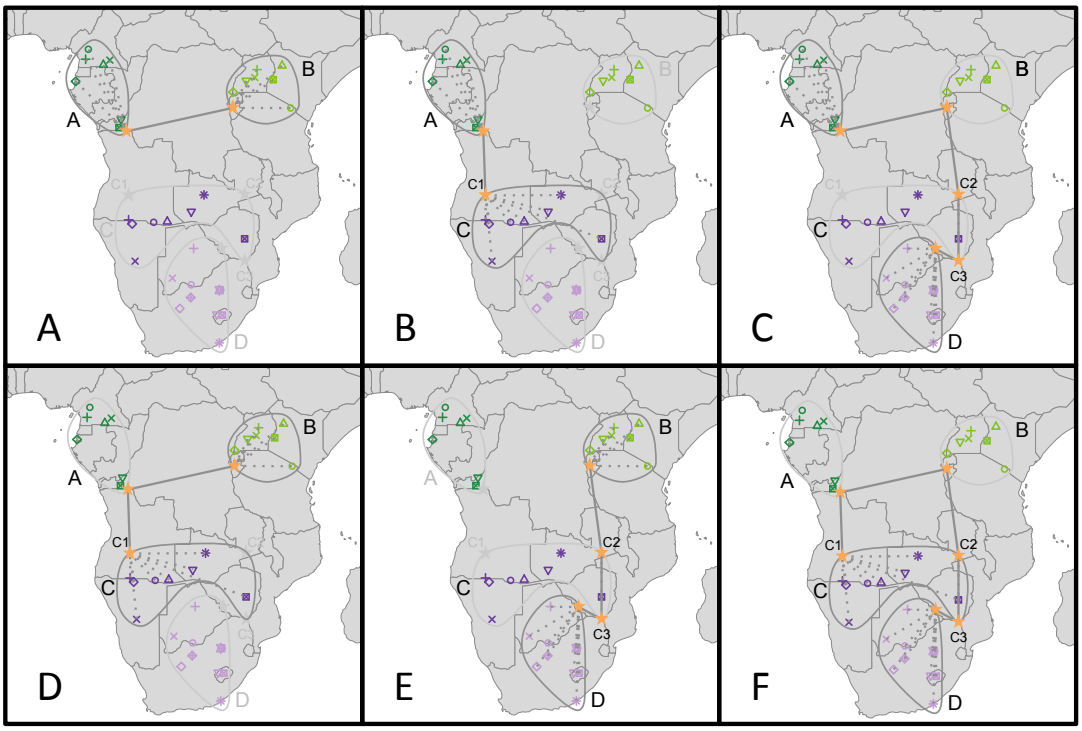

WW – Tanzania
XX – TikarSouth
YY – Tswana
ZZ – Wolof
AAA – Yansi
BBB – Yoruba
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04
−0.02
0.00
0.02


CCC – Zambia
DDD – Zimbabwe
EEE – Zulu
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.06
−0.04

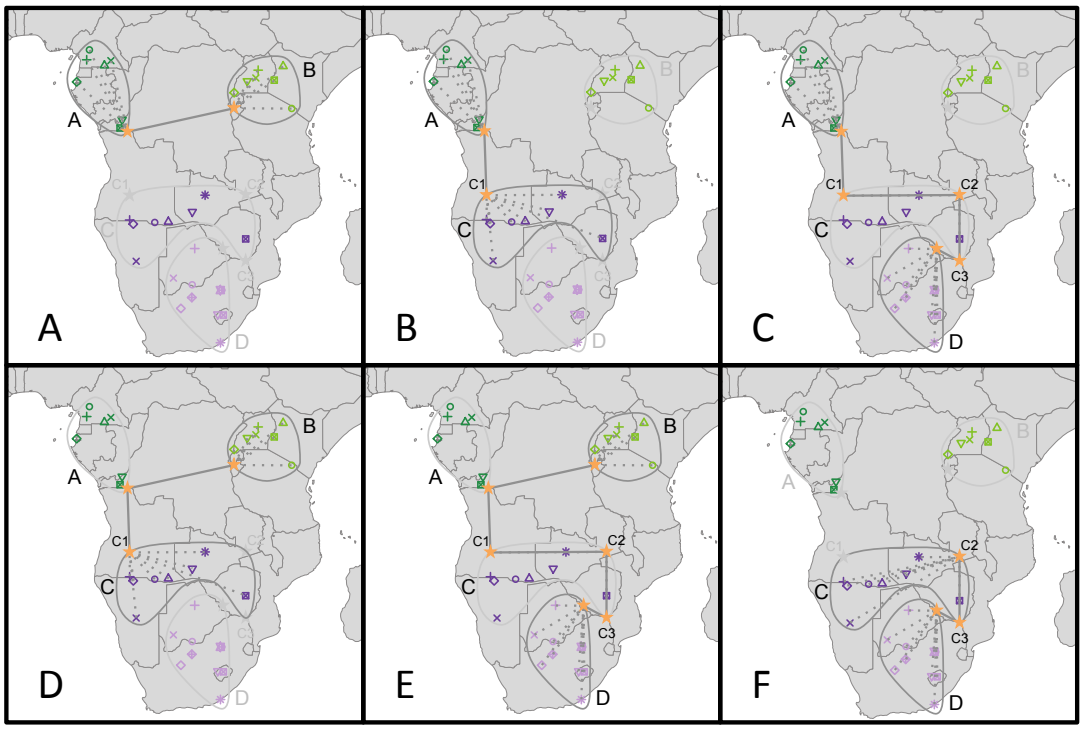
Supplementary Material 6 – Detailed representation of *model 1* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster C*; (C) *cluster A* and *cluster D*; (D) *cluster B* and *cluster C*; (E) *cluster B* and *cluster D*; (F) *cluster C* and *cluster D*.



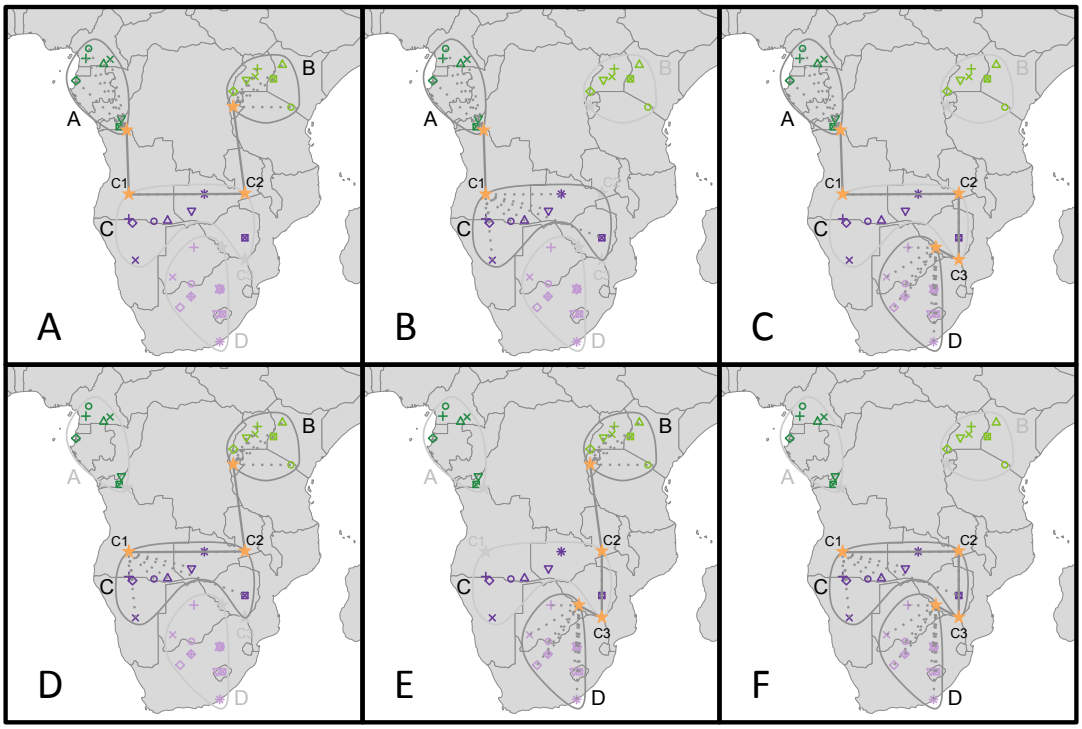
Supplementary Material 7 – Detailed representation of *model 2* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster C*; (C) *cluster A* and *cluster D*; (D) *cluster B* and *cluster C*; (E) *cluster B* and *cluster D*; (F) *cluster C* and *cluster D*.



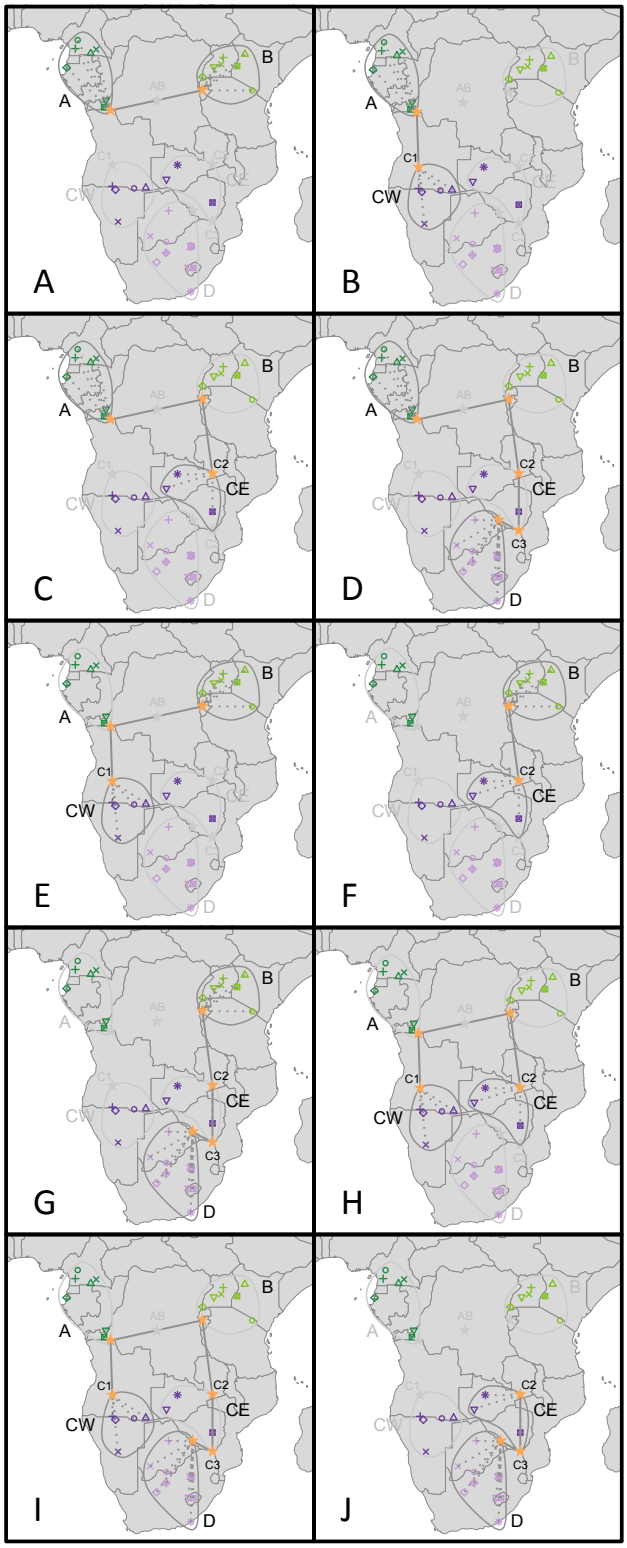
Supplementary Material 8 – Detailed representation of *model 3* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster C*; (C) *cluster A* and *cluster D*; (D) *cluster B* and *cluster C*; (E) *cluster B* and *cluster D*; (F) *cluster C* and *cluster D*.



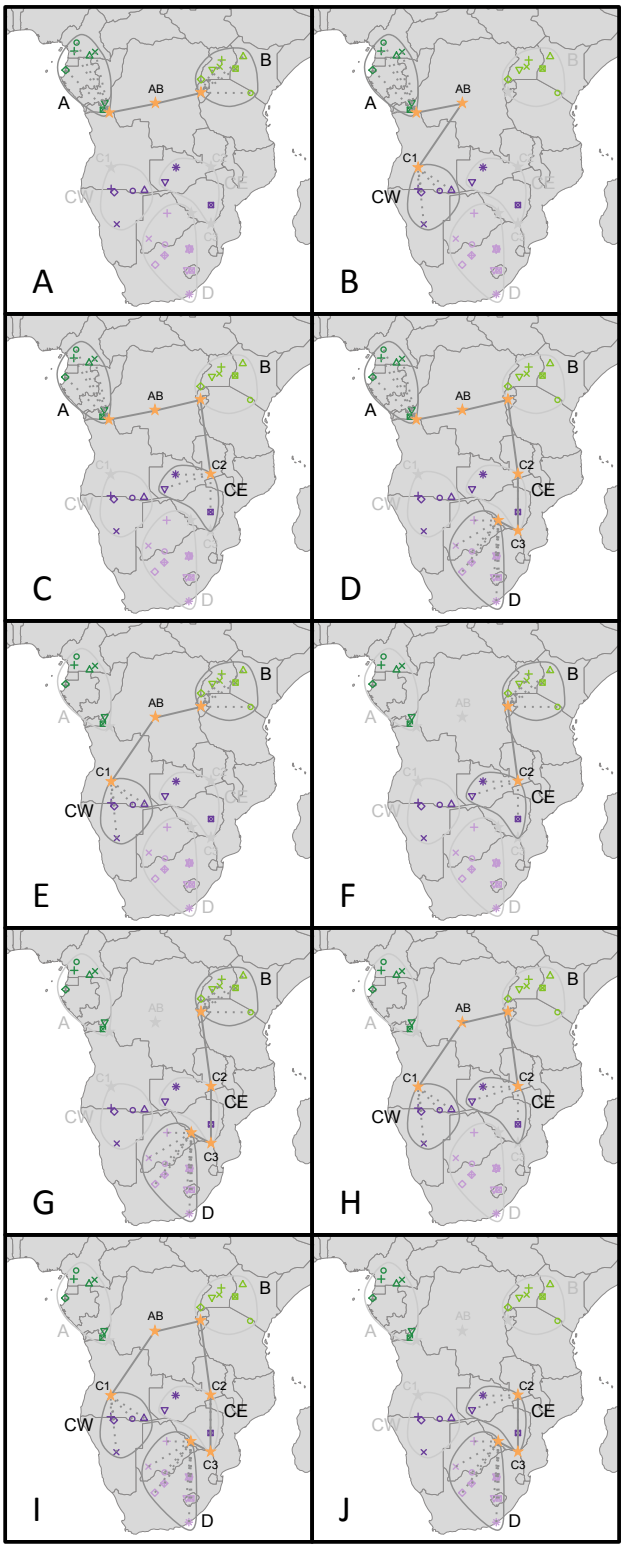
Supplementary Material 9 – Detailed representation of *model 4* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster C*; (C) *cluster A* and *cluster D*; (D) *cluster B* and *cluster C*; (E) *cluster B* and *cluster D*; (F) *cluster C* and *cluster D*.



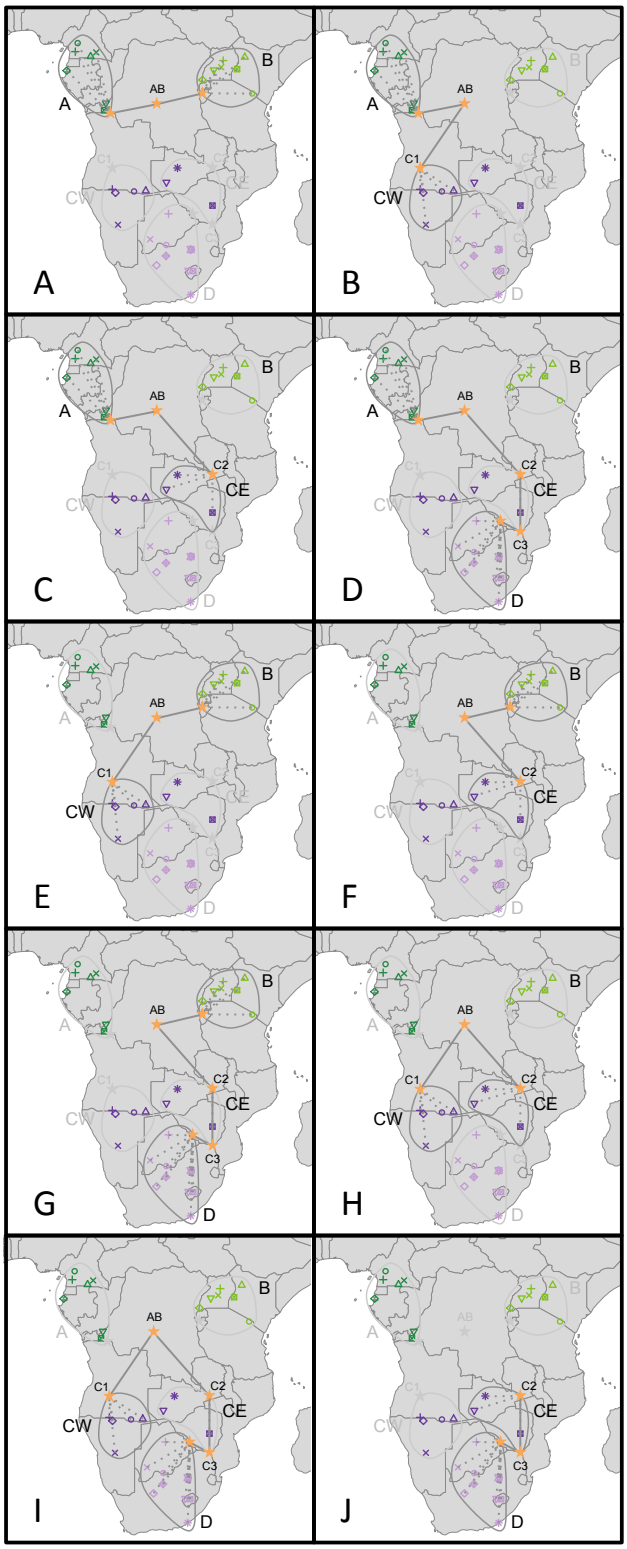
Supplementary Material 10 – Detailed representation of *model A1* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster CW*; (C) *cluster A* and *cluster CE*; (D) *cluster A* and *cluster D*; (E) *cluster B* and *cluster CW*; (F) *cluster B* and *cluster CE*; (G) *cluster B* and *cluster D*; (H) *cluster CW* and *cluster CE*; (I) *cluster CW* and *cluster D*; (J) *cluster CE* and *cluster D*.



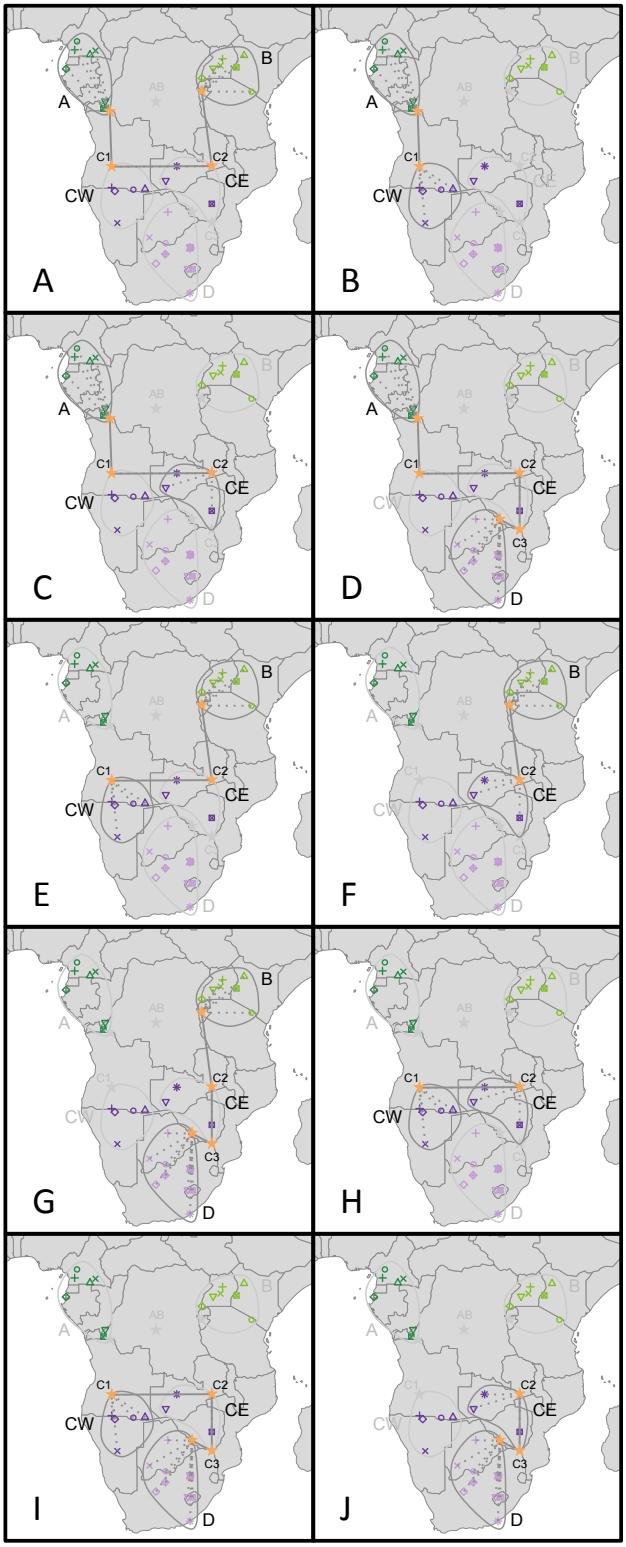
Supplementary Material 11 – Detailed representation of *model A2* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster CW*; (C) *cluster A* and *cluster CE*; (D) *cluster A* and *cluster D*; (E) *cluster B* and *cluster CW*; (F) *cluster B* and *cluster CE*; (G) *cluster B* and *cluster D*; (H) *cluster CW* and *cluster CE*; (I) *cluster CW* and *cluster D*; (J) *cluster CE* and *cluster D*.



Supplementary Material 12 – Detailed representation of *model A3* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster CW*; (C) *cluster A* and *cluster CE*; (D) *cluster A* and *cluster D*; (E) *cluster B* and *cluster CW*; (F) *cluster B* and *cluster CE*; (G) *cluster B* and *cluster D*; (H) *cluster CW* and *cluster CE*; (I) *cluster CW* and *cluster D*; (J) *cluster CE* and *cluster D*.



Supplementary Material 13 – Detailed representation of *model B* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster CW*; (C) *cluster A* and *cluster CE*; (D) *cluster A* and *cluster D*; (E) *cluster B* and *cluster CW*; (F) *cluster B* and *cluster CE*; (G) *cluster B* and *cluster D*; (H) *cluster CW* and *cluster CE*; (I) *cluster CW* and *cluster D*; (J) *cluster CE* and *cluster D*.



Supplementary Material 14 – Detailed representation of *model C* used for the calculation of the model-based pairwise geographical distances.
Each map represents how one cluster is connected to one other cluster, showing all possibilities; (A) *cluster A* and *cluster B*; (B) *cluster A* and *cluster CW*; (C) *cluster A* and *cluster CE*; (D) *cluster A* and *cluster D*; (E) *cluster B* and *cluster CW*; (F) *cluster B* and *cluster CE*; (G) *cluster B* and *cluster D*; (H) *cluster CW* and *cluster CE*; (I) *cluster CW* and *cluster D*; (J) *cluster CE* and *cluster D*.