Università
degli Studi
di Ferrara

PH.D. Course
in
Evolutionary Biology and Ecology

In cooperation with:
Università degli Studi di Parma
Università degli Studi di Firenze

CYCLE XXXII

COORDINATOR Prof. Guido Barbujani

Inference of human migration from genomic and linguistic data

Scientific/Disciplinary Sector (SDS) BIO/18

Candidate
Dott. Silva Santos Patrícia Alexandra

Supervisor
Prof. Barbujani Guido

Co-Supervisor
Prof. Ghirotto Silvia

Co-Supervisor
Dott. Gonzalez-Fortes Gloria

Year 2016/2019

Contents

Note: This thesis follows the citation style of Heredity journal.

List of tables

List of figures

List of abbreviations

SNP	Single-Nucleotide Polymorphisms
NRY	Non-Recombinant portion of the Y chromosome
NGS	Next-Generation Sequencing
PCM	Parametric Comparison Method
IE	Indo-European
UL	Uralic
AL	Altaic
yBP	years Before Present
HPD	Highest Probability Density
AMH	Anatomical Modern Humans
LBK	Linearbandkeramik
ABC	Approximate Bayesian Computation
ENA	European Nucleotide Archive
VCF	Variant Call Format
PCA	Principal Component Analysis
Kb	Kilobase
AP	Ancient Population
MP	Modern Population
IBD	Isolation By Distance
MRCA	Most Recent Common Ancestor
SuSt	Summary Statistics
ABC-RF	Approximate Bayesian Computation-Random Forest
SFS	Site Frequency Spectrum
FDSS	Frequency Distribution of Segregating Sites
bp	Base Pair
CE	Classification Error
TP	True Positives
LDA	Linear Discriminant Analysis
FP	False Positives

Summary

Demographic events in human history are expected to leave traces in languages and genes, hence Darwin's intuition that the best possible description of linguistic relationships among populations would be their phylogenetic tree. In the 90's, Sokal & Cavalli-Sforza tested empirically Darwin's hypothesis for the first time, concluding that linguistic and genetic distances are strongly correlated. However, many questions remained open, due to the lack at the time of suitable genetic data, and to the limitations of linguistic comparisons based on lexicon. Indeed, lexical comparisons have proved to be reliable for closely related languages, but are useless for large-scale comparisons, across language families. Recent methodological developments focusing on syntax,–the abstract rules to combine words into sentences, which in principle are universally comparable–, rather than lexicon, are now enabling more sophisticated quantitative studies across language families. Moreover, the advent of Next-Generation Sequencing (NGS) technologies have revolutionized the field of genetics, as it allows to sequence rapidly and cheaply entire genomes, dramatically increasing the quantity of genetic information available from worldwide populations.

In this thesis, we took advantage of the state-of-art methods for the comparison of linguistic and genomic data. The thesis is divided into two chapters:

(1) In the first chapter, we combined linguistic and genomic data to shed light on the origin and spread dynamics of the Indo-European (IE), Uralic (UR) and Altaic (AL) linguistic families in Eurasia. We showed that languages of these families form three well-distinct clusters, but UR linguistic outliers share evident similarities with their geographical neighbours. Remarkably matching patterns of resemblance are observed comparing genomes in contemporary populations, i.e., with Western UR speakers appearing genetically closer, in parallel shades, to their IE-speaking neighbours, and the Eastern Khanty showing similarities with AL speakers. Finally, we tried to interpret some of the observed historical patterns through a comparison between ancient and modern DNA variation, suggesting a South-to-North spread of UR languages in current Finland. Therefore, this study points out –and is able to quantify– plausible secondary convergence in the syntax of languages of different families, providing evidence that such interference effects were accompanied, and possibly caused, by equally measurable demographic exchanges.

(2) In the second chapter, we proposed a new Approximate Bayesian Computation (ABC) framework in which genomic and linguistic data would be simultaneously considered in the analysis of demographic models, and would also allow inference about biological and

cultural evolution. We first assessed the power of the linguistic framework, to understand to what extent the proposed method is actually able to correctly identify the demographic history. After the validation, we applied this new ABC approach to the analysis and comparison of Bantu-expansion models.

One of the most significant moments in African history is the expansion of Bantu-speaking populations starting 5.000 yBP. Bantu languages represent the largest African language family occupying a vast territory and spoken by millions of people. Multidisciplinary studies associated this expansion with the transition from hunter-gatherer societies to food producers, resulting in population growth and dispersal. However, the dynamics of this process are all but established. Two main hypotheses have been proposed: an Early split of Bantu farmers into Western and Eastern, at the north of the rainforest; or a Late split, in which the Eastern group branches off the Western group at south of the rainforest.

We applied our new ABC framework to compare the Early and Late-Split models using for the first time whole-genome data, from Bantu-speaking individuals, together with linguistic data from the dataset of Grollemund *et al.* (2015). We designed six demographic models that represent alternatives scenarios for the Bantu expansion dynamics and we simulated linguistic and genetic data based on these models. Analyzing the linguistic data, our results seem to better support the Late Split against the other competing hypothesis, although further analyses seem necessary to reach a solid conclusion. For the genetic data, on the contrary, our results were not satisfactory. We tested the demographic models used for languages with our genetic data, however human DNA carries the consequences of the accumulation of diversity over long periods of time, whereas languages do not preserve well old signals. Simulate genetic data using the linguistic demographic models, in which the first population emerges at recent times (6.000 yBP), originates no genetic variation between individuals. Consequently, we need to design new demographic models introducing an ancestral Bantu population, to take into account the ancestral genetic diversity. Ultimately, with these two extended datasets, combined with the power produced by the present ABC method, we expect to reveal details of the past history of Bantu population with an unprecedented definition.

Riassunto

Gli eventi demografici nella storia umana lasciano tracce nelle lingue e nei geni, secondo l'intuizione di Darwin per cui la migliore descrizione possibile delle relazioni linguistiche tra le popolazioni sarebbe il loro albero filogenetico. Negli anni 90, Sokal e Cavalli-Sforza hanno testato empiricamente l'ipotesi di Darwin per la prima volta, concludendo che distanze linguistiche e genetiche sono fortemente correlate. Tuttavia, molte domande sono rimaste aperte, a causa della mancanza al momento di dati genetici adeguati e delle limitazioni dei confronti linguistici basati sul lessico. Il lessico permette confronti fra lingue strettamente correlate, ma è invece inutile per i confronti su larga scala, fra lingue di famiglie diverse che, per definizione, non hanno etimologie in comune. Recenti sviluppi metodologici hanno permesso analisi a livello di sintassi (le regole astratte attraverso cui le parole sono combinate in frasi, in linea teorica confrontabili fra qualunque coppia di lingue) che oggi permettono studi comparativi fra diverse famiglie linguistiche. Inoltre, l'avvento delle tecnologie di *Next-Generation Sequencing* (NGS) ha rivoluzionato gli studi genetici, in quanto è ora possibile sequenziare rapidamente e a bassi costi l'intero genoma, aumentando in maniera spettacolare la quantità di informazioni genetiche disponibili sulle popolazioni di tutto il mondo.

In questa tesi abbiamo sfruttato questi metodi all'avanguardia per il confronto di dati linguistici e genomici. L'elaborato è diviso in due capitoli:

(1) Nel primo capitolo, abbiamo combinato i dati linguistici e genomici per spiegare l'origine e le dinamiche di diffusione nell'Eurasia delle famiglie linguistiche indoeuropea (IE), uralica (UR) e altaica (AL). Le lingue di queste famiglie formano tre gruppi ben distinti, ma gli outlier linguistici UR condividono in modo sostanziale la somiglianza con i loro vicini geografici. Confrontando i genomi nelle popolazioni contemporanee si osservano livelli di somiglianza notevolmente simili: le popolazioni occidentali di lingua UR appaiono geneticamente più simili ai loro vicini di lingua IE e i Khanty (UR orientali) mostrano somiglianze con le popolazioni che parlano lingue AL. Infine, abbiamo cercato di interpretare le distribuzioni di variabilità genetica e linguistica confrontando DNA antico e moderno. Questi test suggeriscono una diffusione da sud a nord delle lingue UR nell'attuale Finlandia. Pertanto, questo studio sottolinea - ed è in grado di quantificare - una plausibile convergenza secondaria nella sintassi di lingue di famiglie diverse, fornendo prove che tali effetti di interferenza sono stati accompagnati, e forse causati, da scambi demografici ugualmente misurabili.

(2) Nel secondo capitolo, abbiamo proposto un nuovo framework di Calcoli Bayesiani Approssimati (ABC) in cui abbiamo testato diversi modelli demografici alla luce dei dati genomici e linguistici, allo scopo di ricostruire aspetti dell'evoluzione biologica e culturale. Abbiamo prima valutato il potere dell'analisi linguistica, per capire fino a che punto il metodo proposto è effettivamente in grado di identificare correttamente la vera storia demografica. Dopo la convalida, abbiamo applicato questo nuovo framework ABC all'analisi e al confronto dei modelli di espansione delle popolazioni di lingua Bantu.

Uno dei momenti più importanti della storia africana è l'espansione delle popolazioni di lingua Bantu iniziato 5.000 anni fa. Le lingue bantu rappresentano la più grande famiglia di lingue africane, occupano un vasto territorio e sono parlate da milioni di persone. Studi multidisciplinari hanno associato questa espansione con il passaggio dalle società di cacciatori-raccoglitori a quelle di coltivatori, quando l'accumulo di cibo ha permesso l'aumento delle dimensioni delle popolazioni e la loro conseguente espansione. Tuttavia, le dinamiche di questa espansione sono oggetto di dibattito. Sono state proposte due ipotesi principali: una divisione precoce (*Early split*) degli agricoltori Bantu a ovest ed est, avvenuto separandosi a nord della foresta pluviale; o una divisione tardiva (*Late split*), in cui il gruppo est si distacca dal gruppo ovest a sud della foresta pluviale.

Abbiamo applicato il nostro nuovo framework ABC per testare i modelli *Early* e *Late split* usando per la prima volta i dati di interi genomi, da individui di lingua bantu, combinati ai dati linguistici dal set di Grollemund *et al.* (2015). Abbiamo ipotizzato sei modelli demografici che descrivono le dinamiche di espansione dei Bantu e simulato dati linguistici e genetici basati su questi modelli. Analizzando i dati linguistici, i nostri risultati sembrano supportare l'ipotesi di *Late split* rispetto alle altre ipotesi concorrenti, anche se è necessario per poter prendere una posizione chiara saranno necessari approfondimenti. Per quanto riguarda i dati genetici, al contrario, i nostri risultati non sono stati del tutto soddisfacenti. Abbiamo testato il modello demografico utilizzato per le lingue con i nostri dati genetici, ma abbiamo constatato che tuttavia, nel DNA si ritrova una diversità genetica generata attraverso lunghi periodi di tempo, mentre le lingue non conservano segnali più antichi. Simulare dati genetici usando i modelli demografici linguistici, in cui la prima popolazione emerge in tempi recenti (6.000 anni fa), non origina alcuna variazione genetica tra individui. Di conseguenza, dobbiamo progettare nuovi modelli demografici introducendo una popolazione ancestrale di Bantu per tenere conto della diversità genetica che questa popolazione portava con sé prima del differenziarsi delle lingue. Attraverso questi due set di

dati estesi, combinati alla potenza prodotta dall'attuale metodo ABC, ci aspettiamo di rivelare dettagli della storia passata della popolazione Bantu con una definizione senza precedenti.

Keywords

Whole-genome data

Languages

Culture

Human Migrations

Introduction

Genes and Languages

In *The Origin of Species* (1859), Darwin proposed that linguistic change along human history tends to be correlated with the biological differentiation of populations. In fact, factors isolating populations from each other, such as geographical distance and barriers to migration, are likely to promote both biological and cultural divergence, whereas factors favouring contacts should have the opposite effect at both levels (Sokal, 1988; Barbujani and Pilastro, 1993; Longobardi, 2003; Creanza *et al.*, 2015). Exceptions do exist, e.g. when a small group imposes its language upon a larger population, causing a cultural change not matched at the genetic level, a phenomenon called élite dominance (Renfrew, 1992). However, despite élite dominance and other processes of horizontal language transmission creating local mismatches, parallel genetic and linguistic change are more the rule than the exception (Cavalli-Sforza *et al.*, 1988; Sokal, 1988; Poloni *et al.*, 1997; Belle and Barbujani, 2007; Gray *et al.*, 2009; Henn *et al.*, 2012; Longobardi *et al.*, 2015). This implies that linguistic diversity offers a set of testable hypotheses about the demographic processes shaping genetic diversity, and vice versa.

In the 80's, Sokal & Cavalli-Sforza turned this idea into a vigorous research programme. They set the basis for an innovative and interdisciplinary approach based on the comparison of gene frequencies between population groups previously defined by linguistic criteria. Common demographic processes affecting different populations would result in a parallel evolution of genetic and linguistic variants. On the other hand, when differentiated genetic groups (either because of the consequences of isolation by distance and/or barriers to gene flow) share similar languages, cultural processes should be called to explain the linguistic diffusion (Barbujani and Pilastro, 1993). Sokal's & Cavalli-Sforza's groups were the firsts to test empirically Darwin's hypothesis concluding that linguistic and genetic distances are strongly correlated (Cavalli-Sforza *et al.*, 1988; Sokal, 1988). Their effects, and successive ones, clarified several aspects of human demographic history, especially in Europe (Barbujani *et al.*, 1995).

Many questions, however, remained open, due to the paucity at the time of suitable genetic data, and to the limitations of linguistic comparisons based on lexicon. Indeed, the first studies aimed at inferring the relationships between genetics and language only had available the allelic frequencies of a very limited number of markers, from which genetic distances between pairs of populations were estimated (Cavalli-Sforza *et al.*, 1988). These first studies suffered from the restrictions on the amount of differences that can accumulate in a little set of variants, and from the large variance of the effects of genetic drift upon allele frequencies. If the split between two populations is ancient, the genetic distances computed may not adequately reflect the amount of genetic and linguistic differentiation. Lastly, the population size may affect the rates of genetic and linguistic change, but not necessarily equally (Chen *et al.*, 1995). During the last years, the development of new DNA markers and sequencing technologies have experienced enormous advances, which now allow one to overcome some of the old limits and to move forward into the comparison of genomes and languages.

New era of Next-Generation Sequencing (NGS) technologies

First attempts to quantify the genetic variability in humans were done using gel electrophoresis of soluble proteins (Hubby and Lewontin, 1966). Subsequently, the development of DNA markers like microsatellites and single-nucleotide polymorphisms (SNPs) allowed a better screening of the human DNA. Additionally, analysis based on monoparental markers, namely human mitochondrial DNA and the non-recombinant portion of the Y chromosome (NRY), helped to understand the evolutionary history on the paternal and maternal lines, and contributed greatly to biogeographical studies (Johnson *et al.*, 1983; Cann *et al.*, 1987; Underhill *et al.*, 2000; Rosenberg *et al.*, 2002; Tishkoff *et al.*, 2009). However, the amount of genetic information that can be retrieved from these techniques is still limited, since only a few thousands of variants can be simultaneously typed in an individual (Levy *et al.*, 2007; Wheeler *et al.*, 2008).

The advent of Next-Generation Sequencing (NGS) technologies has revolutionized the field of genetics, turning it into something quantitatively different, which deserved the new name of genomics. The crucial technological progress was represented by the methods for sequencing the entire genome of an individual without previous information on specific sequences, or Next-Generation Sequencing (hereafter: NGS), which work at affordable cost in relatively short times. Since its implementation we have seen a tremendous increase of the quantity of genetic information available from worldwide populations. Having access to full genomic information provides a more comprehensive and a less biased view of the genetic diversity (Schraiber and Akey, 2015), helping to infer otherwise elusive aspects of human evolution and demographic history, like ancient admixture events related to past migrations (Tennessen *et al.*, 2012; Raghavan *et al.*, 2015; Skoglund *et al.*, 2015; Consortium *et al.*, 2015). Also, the NGS technology has ignited the era of palaeogenomics, allowing the sequencing of complete genomes from prehistoric remains (albeit at low coverage; see below). This wealth of ancient DNA data has opened direct windows into the past, allowing one to study the changes in genetic diversity over time (Rasmussen *et al.*, 2010; Raghavan *et al.*, 2014; Lazaridis *et al.*, 2014; Haak *et al.*, 2015).

In parallel to the development of high throughput sequencing techniques, several biostatistical methods have been designed to make the most of the whole genome data for the inference of human population demography (Li and Durbin, 2012; Schiffels and Durbin, 2014; Terhorst *et al.*, 2017). This way, the field is slowly moving from the description of genetic patterns (in time and place) to hypothesis testing for inference about the underlying evolutionary and demographic processes. A common caveat of all these methods is a limited time depth. By using a few number of individuals we can only infer the evolutionary history from ancient past until ~ 2.000 years Before Present (yBP) (Schiffels and Durbin, 2014).

Linguistic features: lexical and syntactic

There are approximately 7.000 different languages spoken in the world today, culturally variable at every level of their structure, from the sound system, to the grammar and semantics (Greenhill *et al.*, 2010; Levinson and Gray, 2012). This diversity may reflects the legacy of thousands of years of cultural evolution. But, when language is used to study the population structure, how is this diversity to be measured? Phonemic, syntactic, and grammatical properties can be quantified for each language. Languages can be analysed in an analogous way to DNA sequences, containing grammatical and phonological structures and vocabularies (lexicons). Similarity between words can arise by horizontal transmission, language borrowing, or vertical transmission, from parents to offspring (like genes). Moreover, processes of mutation and random drift can act on languages, in the same way it happens on genes (Atkinson and Gray, 2006; Shijulal *et al.*, 2011). Thus, there is no doubt that using language similarities/dissimilarities as a clue of evolutionary relatedness can help to inference the population history (Colonna *et al.*, 2010).

The simplest and most straightforward approach is to use the lexical information (words and morphemes) to measure lexical divergence between groups of languages. The closer languages or language families are related, the greater the number of words (cognates: David 2011 (Crystal, 2011)) sharing a meaning and an etymology, and hence showing descent from a common linguistic ancestor. These lexical comparisons have proved to be reliable for closely related languages, belonging to the same linguistic family. However, languages are classified in different linguistic families when they share no recognizable cognates, and hence no identifiable common ancestor (Rowe and Levine, 2015). Therefore, when one is to compare populations on the large geographical scale, i.e. across language families, the lexical tool proves inadequate (Colonna *et al.*, 2010; Longobardi and Guardiano, 2013; Longobardi *et al.*, 2015). Although attempts have been made to reconstruct linguistic phylogenies by comparing lexical items of different language families (see e.g. (Ruhlen, 1991, 1994)) there is a serious, perhaps unsurmountable, problem of distinguishing real cognates from random coincidences. The inference of linguistic relationships comparing vocabulary items (words/morphemes) and their sound structures, which dissolve with time (plausible estimates place this limit around 8000 ± 2000 years ago), can create serious problems due to the resemblance among words simply due to chance: e.g. English *much*, *day*, *have* and Spanish *mucho*, *dia*, *haber* are false cognates. By contrast, real cognates (e.g. English *full* and Italian *pieno*) may not look alike at all (Atkinson and Gray, 2006; Colonna *et al.*, 2010; Longobardi and Guardiano, 2013; Longobardi *et al.*, 2015; Greenhill *et al.*, 2017). Another example is the concept *tooth*, which has a cognate set that unites English *tooth*, German *Zahn*, Italian *dente* and French *dent* as etymologically related (figure 1) (Shijulal *et al.*, 2011).

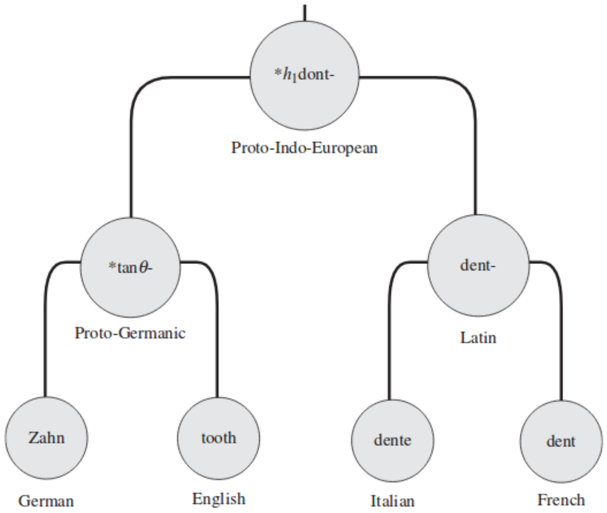


Figure 1. Etymological reconstruction of the concept tooth. The English and German word forms have descended from the Proto-Germanic ancestor. The Italian and French words are descendants of Latin, and the Proto-Germanic and Latin forms stem from Proto-Indo-European.

Syntax (abstract rules to combine words into sentences) appears more measurable, universally comparable and stable than the lexicon. The abstract nature of the grammatical features makes them comparable between distant language families and less susceptible to subjective interpretations than the lexicon, which relies on previous linguistic work to identify sound correspondences and cognate items. Additionally, grammatical structure is more resistant to change and borrowing than the lexicon (Greenhill *et al.*, 2017). Thus, the use of languages structures may extend the time depths at which language data can be compared, retaining a phylogenetic signal beyond the current temporal ceiling on the reconstruction of language history (Dunn, 2005; Longobardi and Guardiano, 2009; Colonna *et al.*, 2010).

In 2009, Longobardi *et al.* developed a new taxonomic technique, namely the Parametric Comparison Method (PCM), which is based exclusively on syntax comparison (Longobardi and Guardiano, 2009, 2013). This method is built on the principle that the core grammar of any natural language be represented by a string of binary symbols, each symbol coding the value of a linguistic parameter. Such strings of symbols can be unambiguously collated and language distances precisely measured (Colonna *et al.*, 2010; Longobardi and Guardiano, 2013). Following this method, it was created a dataset (Ceolin, 2019) composed by 97 binary parameters for 65 Eurasian languages belonging to different linguistic families (see an example in table 1). So far, this method has only seldom been used. The very simple reason is, the starting point for a lexical comparison is a dictionary of the two languages, which is generally already available, whereas a reconstruction of grammatical and syntactic rules of a language requires a deep, and time-consuming, analysis with a native speaker and a linguistic expert who has acquired the language being studied.

Table 1. An example; First five parameters of the dataset for seven different languages in the format defined in Longobardi & Guardiano (2009). In each row we have a grammatical parameter that is encoded as '+' and '-'. The symbol '0' encodes the neutralizing effect of implicational dependencies across parameters. Uncertain states are indicated by '?'. From the fifth column we have the languages analysed, in this example we showed the Sicilian, Neapolitan, Italian, Spanish, French, Portuguese and Romanian languages.

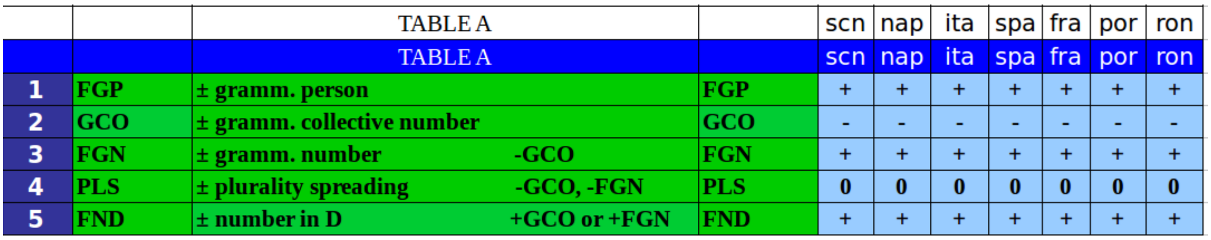


		TABLE A		scn	nap	ita	spa	fra	por	ron
		TABLE A		scn	nap	ita	spa	fra	por	ron
1	FGP	± gramm. person	FGP	+	+	+	+	+	+	+
2	GCO	± gramm. collective number	GCO	-	-	-	-	-	-	-
3	FGN	± gramm. number -GCO	FGN	+	+	+	+	+	+	+
4	PLS	± plurality spreading -GCO, -FGN	PLS	0	0	0	0	0	0	0
5	FND	± number in D +GCO or +FGN	FND	+	+	+	+	+	+	+

Abbreviations: FGP – Grammaticalized person; GCO – Grammaticalized number; FGN – Grammaticalized number; PLS – Plurality spreading; FND – Number in D (elements normally associated with the D-area, such as 'articles' (i.e. Person/Number/definiteness/count markers)); scn – Sicilian; nap – Neapolitan; ita – Italian; spa – Spanish; fra – French; por – Portuguese; ron – Romanian.

In this thesis, we took advantage of the state-of-art methods described above for the comparison of linguistic and genomic data. The thesis is divided in two chapters: in the first chapter, we applied the PCM method for the comparison of different linguistic families. In specific, we were interested in understanding the genomic and linguistic diversity across different language families from Eurasian populations: Indo-European (IE), Uralic (UR) and Altaic (AL). In the second chapter, we developed a new Approximate Bayesian Computation (ABC) framework in which genomic and linguistic data (lexicon) are simultaneously considered in the analysis of demographic models. We applied this new ABC framework to the analysis of Bantu-expansion models.

CHAPTER I:

Parallel signatures of past human migration in linguistic and genomic diversity of Western/Central Eurasia

1.1 Introduction

1.1.1 Some general concepts on genetic relationships and linguistic barriers

Populations tend to differentiate genetically from their neighbours because of divergence due to random genetic drift and/or natural selection pressure for adaptation to different environmental conditions (Cavalli-Sforza *et al.*, 1993; Newberry *et al.*, 2017). Genetic drift, i.e., the change in the frequency of an existing allele in a population, depends on the effective size of each population. In the Paleolithic when population densities were low, drift was an important cause of local genetic differentiation, which led to a patchy geographical distribution of the genetic diversity. On the contrary, population expansions tend to add detectable patterns of genetic gradients around their areas of origin and can extend to large regions in a few thousand years. Their genetic effects are relatively stable and overlapping expansions that took place at different times can often be distinguished from each other by statistical methods (Cavalli-Sforza *et al.*, 1993).

Differences between populations can be reduced by gene flow, i.e., the exchange of individuals or gametes. The rate of gene flow depends on the relationships between populations. In general, geographic distance is a major factor limiting gene flow, but there are also physical barriers as mountains, seas and deserts that have the same consequences (Barbujani, 1997). Besides the physical barriers, other obstacles that may affect the patterns of genetic variation are cultural barriers. Those can be social, religious, or the more stable and easier to locate in space, language barriers (Barbujani, 1997). Linguistic differences are themselves barriers to gene flow, which reinforce genetic differentiation to some degree (Chen *et al.*, 1995). Usually, when choosing a partner, humans do not tend to cross linguistic boundaries, so they preferentially choose partners that speak the same language and consequently live nearby. In the long run, this may contribute to creating reproductive barriers between populations speaking different languages, resulting in genetic divergence (Barbujani, 1997).

In Europe, despite the low average levels of genetic differentiation between populations, there is a close correspondence between genetic and geographic distance (figure 1.1) (Novembre *et al.*, 2008). The existence of a linguistic component in addition to geographic differentiation and the near ubiquity of linguistic boundaries along the zones of rapid genetic change suggest historical, as well as geographic factors to be responsible for the genetic differentiation of the European speakers of different language families.

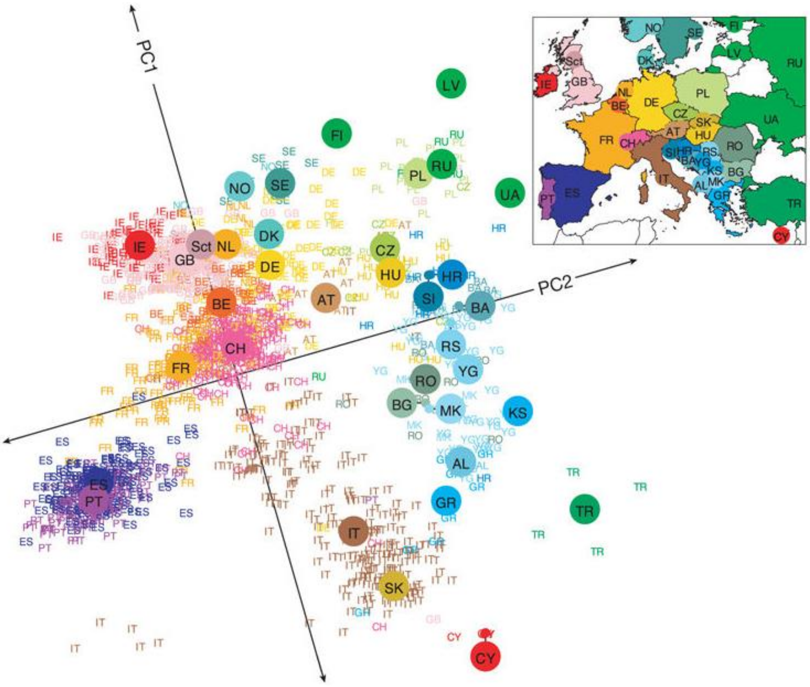


Figure 1.1. Two-dimensional summary of genetic data from 1.387 Europeans. Small coloured labels represent individuals and large coloured points represent median principal component axis one (PC1) and principal component axis one (PC2) values for each country. The inset map provides a key to the labels. A geographic map of Europe arise from this analysis, which is emphasized by the rotation of the PC axes. Abbreviations: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia, Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia (image from Novembre *et al.* (2008)).

1.1.2 Genetic structure in Eurasia

Europe has been investigated at the genetic level more, and for longer, than any other part of the world. Large genetic gradients were discovered in studies based on protein polymorphisms, already 40 years ago (Menozzi *et al.*, 1978; Sokal and Menozzi, 1982), and later confirmed at the DNA level (Chikhi *et al.*, 1998, 2002; Novembre *et al.*, 2008). The causes of such a strong geographical structuring have been debated for a long time. Currently, the consensus is that contemporary Europeans appear to mainly trace their ancestry to three founding groups, namely Western hunter-gatherers who entered Europe in Paleolithic times, early Neolithic farmers from the Near East, and a third group from the East (Lazaridis *et al.*, 2014), later identified with Bronze-Age people who dispersed from the Pontic-Caspian steppes (Haak *et al.*, 2015). The proportions of these ancestral components in modern genomes vary geographically; the Paleolithic component generally accounts for a minor fraction of the European genome variation, whereas the Neolithic and Bronze-age components prevail, respectively, in Southern and Northern Europe (Haak *et al.*, 2015). As for Asia, both its geographic extension and the relative paucity of ancient samples yielding amplifiable DNA, make it difficult to synthetically describe its genetic structure. Studies in the pre-genomic era showed a general correspondence between linguistic and genetic clusters in Central Asia (Barbujani and Pilastro, 1993; Heyer and Mennecier, 2009), with some exceptions, though (Kraaijenbrink *et al.*, 2014) and need be confirmed by state-of-the-art data and analyses. In a recent analysis of whole genomes, De Barros et al. (2018) identified six main genetic components in modern and ancient samples, differently distributed in Western, Central and Southern Asia (de Barros Damgaard *et al.*, 2018). Taking all together, modern and ancient genomes have revealed extensive population migrations, replacements and admixture events since the Paleolithic.

1.1.3 Main linguistic families in Eurasia and their genetic connections

Research in historical linguistics suggests that groups or families of languages can be further classed together into larger units, generally termed macrofamilies, for which some common origin can be postulated. The most important of such macrofamilies, already proposed by Holger Pedersen at the beginning of the 20th century, is Nostratic ((Kaiser and Shevoroshkin, 1988); see also (Renfrew, 1991)). This (controversial) macrofamily partly overlaps with

Greenberg's (Greenberg, 2000, 2002) Eurasiatic, composed by the languages presented in figure 1.2.

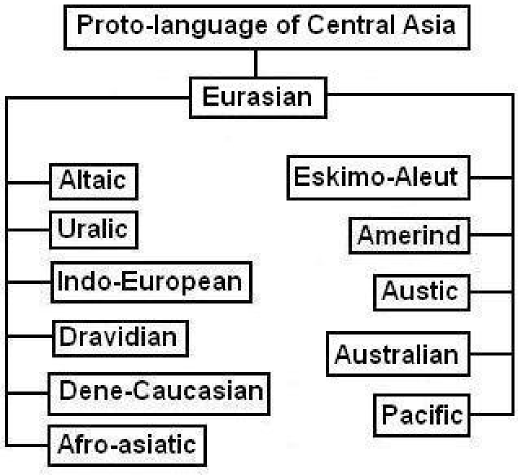


Figure 1.2. Eurasian linguistic families.

Here, we shall focus on the Indo-European (IE), Uralic (UL) and Altaic (AL) linguistic families (figure 1.3).

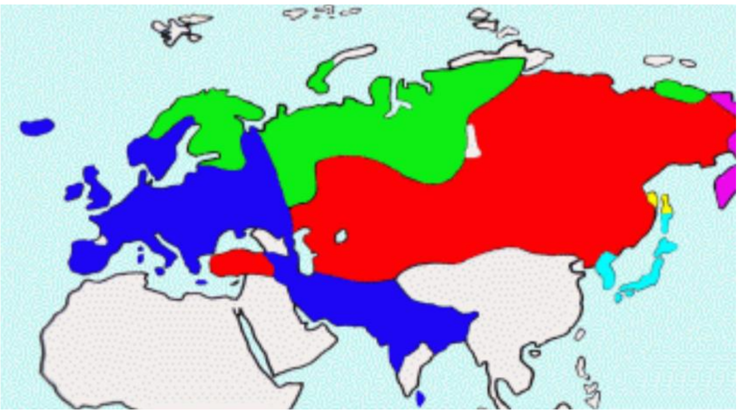


Figure 1.3. Map showing the geographic distribution of the three Eurasian linguistic families under study: Indo-European in blue, Uralic in green and Altaic in red. Korean and Japanese languages form also part of the Altaic macrofamily but are not represented in this map. In the figure is represented the most inclusive view about Altaic and Uralic language families, since there are uncertainties about the status of singles languages.

Europe, from the linguistic standpoint, is by any standards rather uniform. With just five exceptions (Basque, Finn, Estonian, Hungarian and Turkish) all European languages are classified within the IE family. IE is the main language family with 445 languages currently spoken. Despite a long tradition of genetic and linguistic studies (Menozzi *et al.*, 1978; Cavalli-Sforza *et al.*, 1988; Sokal *et al.*, 1990; Barbujani and Pilastro, 1993; Gray and Atkinson, 2003; Novembre *et al.*, 2008; Bouckaert *et al.*, 2012; Longobardi and Guardiano, 2013) it is still unclear whether: (1) early IE languages come from the Pontic-Caspian steppes approximately 5.000 years ago and spread in Europe in the Bronze Age (Gimbutas, 1979); or (2) from Anatolia and spread with the dispersal of early Neolithic farmers around 8.000 – 9.500 years ago (figure 1.4) (Renfrew, 1987).

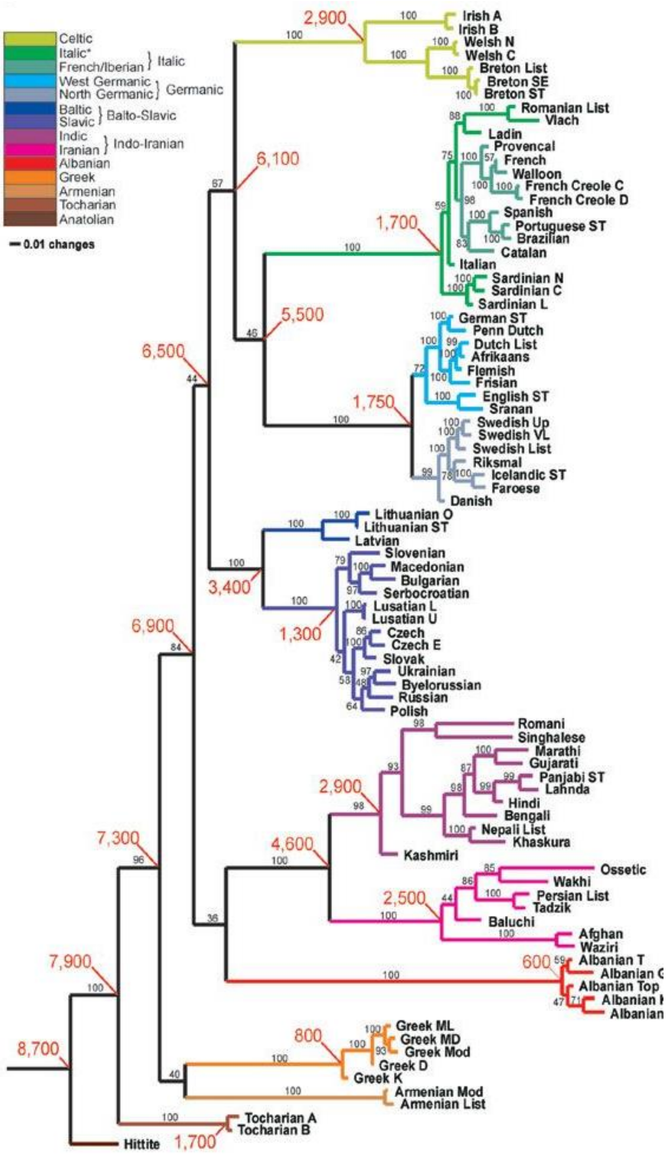


Figure 1.4. Linguistic tree of the Indo-European languages based on lexical comparisons. The main language groupings are colour coded. Branch lengths are proportional to the inferred maximum-likelihood estimates of evolutionary change per cognate. Values above each branch (in black) express the bayesian posterior probabilities as a percentage. Values in red show the inferred ages of nodes in years before present. *Italic also includes the French/Iberian subgroup (image from (Gray and Atkinson, 2003)).

As far for UR linguistic family, it includes 38 languages spoken by populations that are found around the northern boreal zone, from Western Siberia to the Baltic Sea in Europe. The origin of the Uralic family has been hypothesized in the Volga-Oka area around 6.000 years ago (Janhunen, 2009a; Honkola *et al.*, 2013), in the form of a protolanguage that split into two main branches, the Samoyed and the Finno-Ugric (figure 1.5). After that, the Finno-Ugric would have spread and diverged toward north and northwest of the Uralic mountains, evolving into the modern Finnic, Saami and Permian languages, while the Udmurt and Mari persisted around the Volga (Abondolo, 1998; Tambets *et al.*, 2018). The time of the spread and the cultural material associated to the proto-Uralic expansion is highly debated and heavily depends on the estimates of the split times in the Uralic language tree, which however vary widely (Kallio, 2006; Janhunen, 2009a; Honkola *et al.*, 2013).

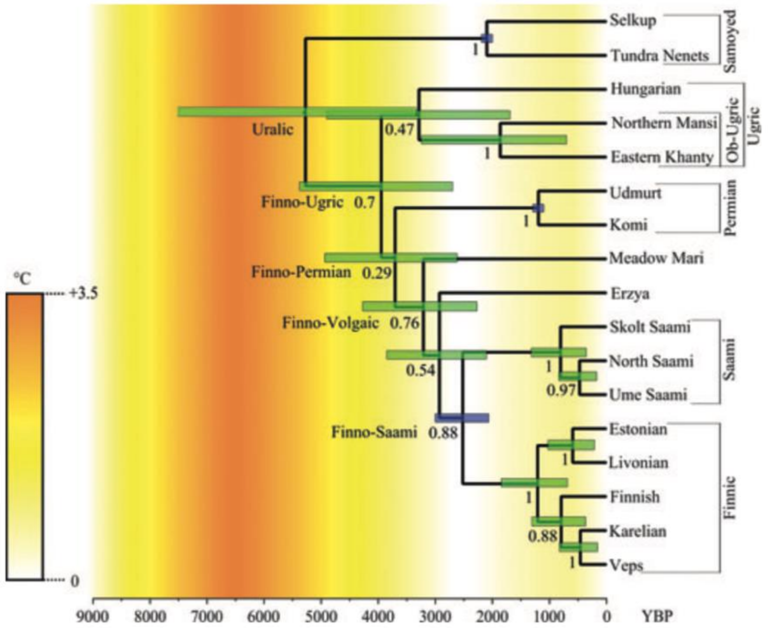


Figure 1.5. Linguistic tree of the Uralic languages. In the x-axis is presented the time in yBP. Green bars represent the 95% highest probability density (HPD) for the divergence times. Values of the posterior probabilities are presented outside of the nodes. Calibration points (Samoyed, Permian and Finno-Saami) are labelled with blue bars indicating the uniform prior of the calibration points. Names of different protolanguages are marked on the nodes of the tree, and the names of different subclades are on the right margins. The colour scale of the picture describes the temperature changes with relation to current temperature (+3.5 – 0 ºC red-white) of the Northeastern Europe/East Europe tundra (west side of the Ural Mountains) (image from (Honkola *et al.*, 2013)).

Genetic studies based on modern and ancient samples have reported a Siberian genome component in the individuals from Uralic speaking areas in Europe (mainly Finland and Estonia) and suggested a Siberian origin for the spread of the Uralic languages into Northeast Europe (Tambets *et al.*, 2018; Lamnidis *et al.*, 2018; Saag *et al.*, 2019). However, the relationship between this Siberian component and the expansion of the Uralic languages is not straightforward, as its presence in the ancient populations predates most linguistic estimates of the spread of the extant Finnic in the area (Honkola *et al.*, 2013; Lamnidis *et al.*, 2018). Our special focus on UR speakers as carriers of information about human demographic history is prompted by previous results, showing that the Westernmost UR-speaking populations in Europe display peculiar properties in their gene-syntax-geography relations (Longobardi *et al.*, 2015).

Finally, Altaic is a controversial family (Georg *et al.*, 1999; Vovin, 2005; Robbeets, 2005; Ceolin, 2019) often argued to include Turkic, Mongolian, and Tungusic languages (as well as, for some scholars, Japanese, Korean and Ainu) (figure 1.6) (Georg *et al.*, 1999; Vovin, 2005). Linguists supporting the hypothesis that the AL family only includes Turkic, Mongolian, and Tungusic claim that the common linguistic features with Japanese, Korean and Ainu is a results of language borrowing rather than descent from a common ancestral language. Speakers of both AL and UR are currently scattered in the Northern Eurasia region spanning from Eastern Europe and Anatolia to Siberia.

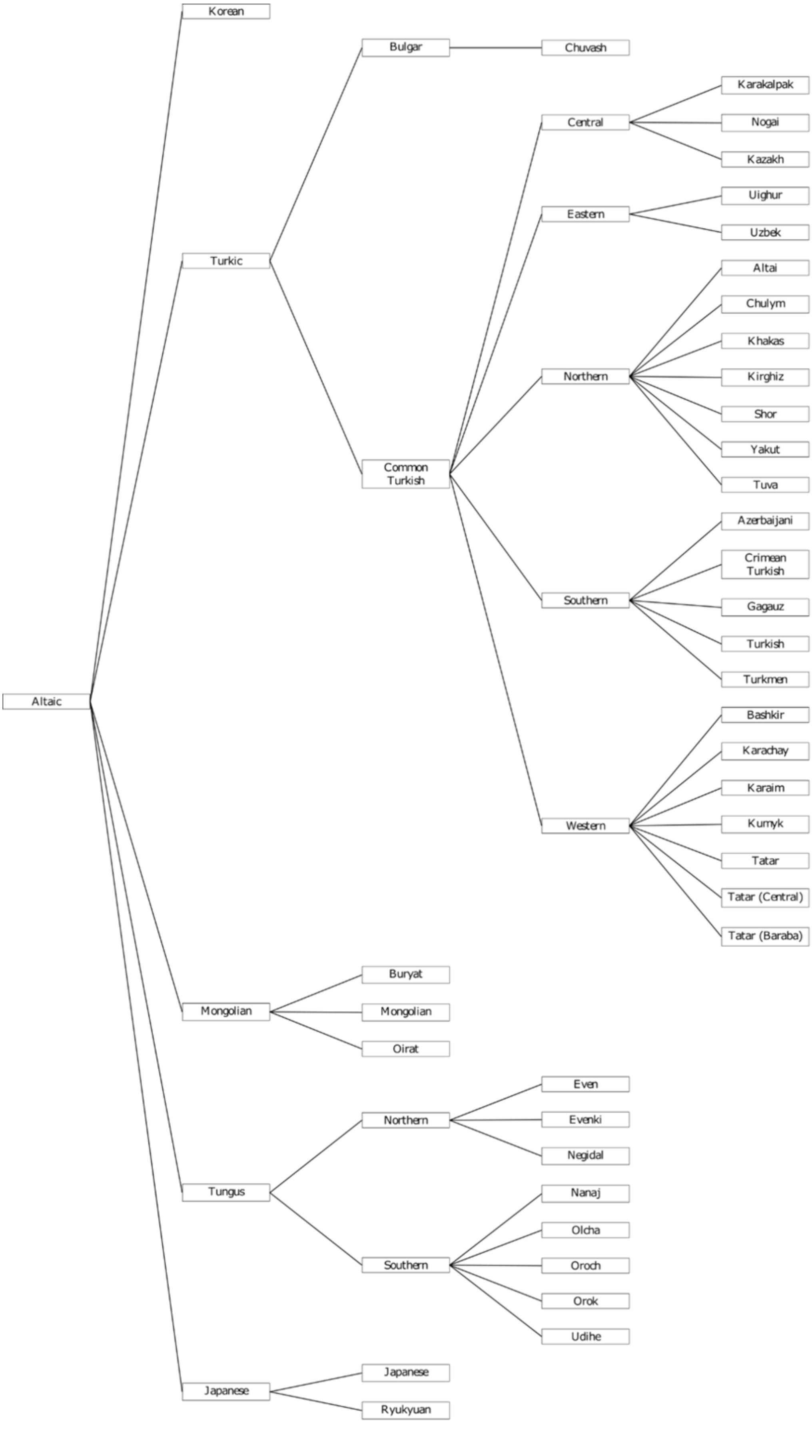


Figure 1.6. Linguistic tree of the Altaic languages (tree from Ruhlen 1991).

1.1.4. Main cultural changes associated with demographic events in Europe

After the arrival of the Anatomical Modern Humans (AMH) into Europe, the first main demographic event affecting the continent as a whole seems the introduction of farming, approximately 8.000 years ago (Menozzi *et al.*, 1978; Cavalli-Sforza *et al.*, 1993; Allentoft *et al.*, 2015; Haak *et al.*, 2015; Olalde *et al.*, 2015; Mathieson *et al.*, 2015; González-Fortes *et al.*, 2017). This transition, called the Neolithic revolution, was characterised by the shift from hunting-gathering to farming as the main subsistence technology. Its demographic impact was doubtless large (Menozzi et al. 1978) and there is reason to believe it was a cause, and possibly the main cause, for the spread of IE languages in Europe (Renfrew hypothesis - (Sokal and Menozzi, 1982; Renfrew, 1987; Sokal, 1988; Chikhi *et al.*, 1998, 2002)). Archaeological data suggests that these early farmers were settled initially in the Near East and/or Anatolia, and subsequently spread farming and associated technologies into Europe. The spread of farming out of the Near East followed two main routes (figure 1.7): a first expansion through the Northern Mediterranean coastline represented by the Impressa and Cardial cultures, and a second expansion represented by the Linearbandkeramik (LBK) culture that followed the Danube River into Central Europe (Ammerman and Cavalli-Sforza, 1984; Chikhi *et al.*, 2002; Olalde *et al.*, 2015).

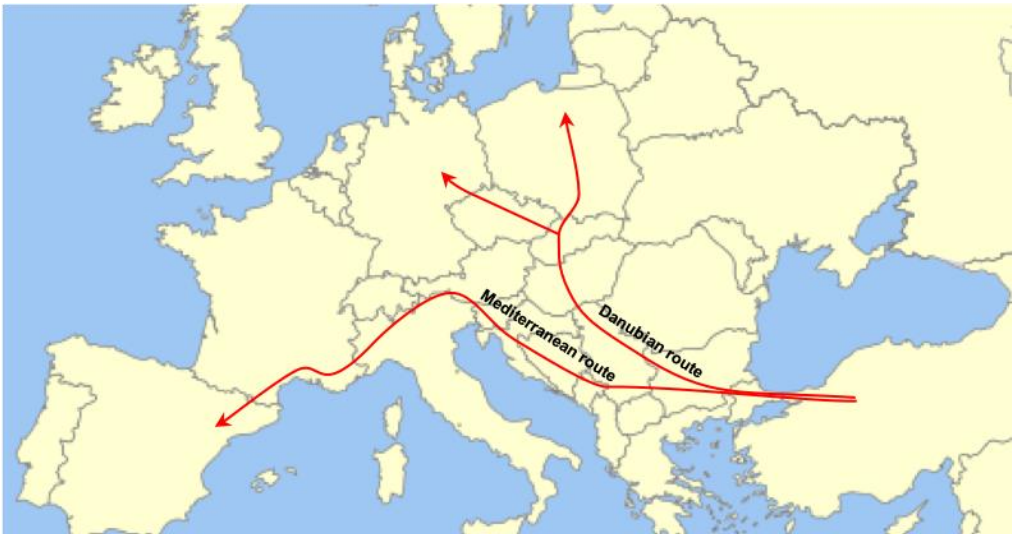


Figure 1.7. Two main routes of the Neolithic expansion. The Mediterranean route, which cover parts of the Adriatic coast, Italy, southern France and part of the Iberian Peninsula; and the Danubian route that passes through Austria, Czech Republic, Germany, northern France, Benelux, Poland, and Ukraine.

The process leading to the adoption of farming in Europe was not a simple migration of people from one original place to a destination. Theoretical (Ammerman and Cavalli-Sforza, 1973) and empirical (Barbujani and Pilastro, 1993; Barbujani *et al.*, 1995; Chikhi *et al.*, 2002) work show that the gradient observed across the whole Europe (figure 1.8) could only be generated under four conditions, namely: (a) an initial difference in the gene pool between expanding (i.e. near Eastern/Anatolian farmers) and recipient (i.e. European hunter-gatherers) populations; (b) demographic growth in the farming populations; (c) range expansion of the farming population; and (d) not immediate admixture between the two groups at their encounter, so that the farmers will keep growing in numbers, whereas the hunter gatherers will not (see (Barbujani, 2013)). In short, this process, a demic diffusion, is characterised by the expansion into additional territories of a population whose size is increasing.

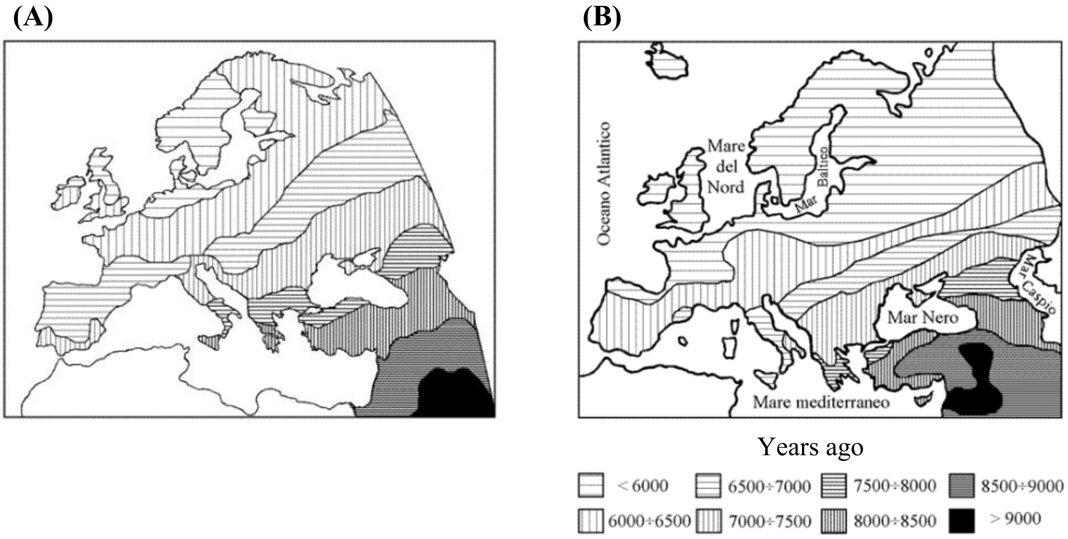


Figure 1.8. (A) A summary of genetic; **(B)** and agriculture diffusion in Europe. Genetic data analyzed in (A) was obtained from 95 classical polymorphisms (for more detail see Cavalli-Sforza 1997 (PNAS)). In (A) and (B) it is clear the genetic and cultural gradient South to North across the whole Europe, starting approximately 9.000 years ago (black region in the map) until 6.000 years ago.

Studies using modern and ancient genetic data support the spread of agriculture in Europe by population dispersal rather than by cultural transmission (Cavalli-Sforza *et al.*, 1993; Chikhi *et al.*, 1998, 2002; Bramanti *et al.*, 2009; Haak *et al.*, 2015; Olalde *et al.*, 2015; Mathieson *et al.*, 2015; González-Fortes *et al.*, 2017). Farming and the IE languages, seem to have spread together, strongly suggesting that both came from the Near East, as proposed by Cavalli-Sforza and Renfrew (Renfrew, 1987; Barbujani and Pilastro, 1993; Cavalli-Sforza, 1997). In addition, studies using linguistic data observed that the basal position occupied by Anatolian languages like Hittite in the IE family tree could be explained by an Anatolian homeland (figure 1.4) (Gray and Atkinson, 2003; Bouckaert *et al.*, 2012).

Until recently, no historic process documented after the Neolithic (10.000 years ago) seemed to have been associated with population growth to the degree sufficient to cause such a strong patterning of genetic variation. Genetic and linguistic variation in most of Eurasia might, by and large, reflect the same generating process: languages of the Nostratic family spread as people moved, and hence probably owe their diffusion to processes of Neolithic dispersal from a common homeland somewhere in Southwest Asia (Barbujani & Pilastro 1993). However, ancient DNA studies of prehistoric samples recently found genomic evidence of a possible, second major migration, this time from the Pontic steppes into central Europe, at the end of the late Neolithic (Gamba *et al.*, 2014; Allentoft *et al.*, 2015; Haak *et al.*, 2015; Mathieson *et al.*, 2015; González-Fortes *et al.*, 2017). Nomadic herders known as the Yamnaya, an early Bronze Age culture that came from the grasslands, or steppes, of modern-day Russia and Ukraine, expanded Westwards, bringing with them metallurgy and animal herding skills. These technological advancements might have represented the factor providing the expanding people with some advantage in survival and reproductive ability over the previously settled populations, resulting in a mode of spread roughly resembling the one described for the Neolithic demic diffusion. The Yamnaya probably interbred with local Europeans, who were descendants of both the farmers and hunter-gatherers. Within a few hundred years, the Yamnaya contributed to at least half of central Europeans' genetic ancestry. By contrast, a recent analysis of Asian genomes suggested that the spread of IE languages in South Asia and Anatolia may have little, if anything, to do, with migration from the Pontic-Caspian steppes (de Barros Damgaard *et al.*, 2018). Thus, the main argument in favour of the Anatolian hypothesis (that major language change requires major migration) can now also be applied to the Steppe hypothesis (Gimbutas' hypothesis - (Gimbutas, 1979)) (figure 1.9). Indeed, Haak *et al.* (2015) and later ancient DNA studies have claimed a steppe origin of the modern IE languages in Europe (Allentoft *et al.*, 2015; Haak *et al.*, 2015).

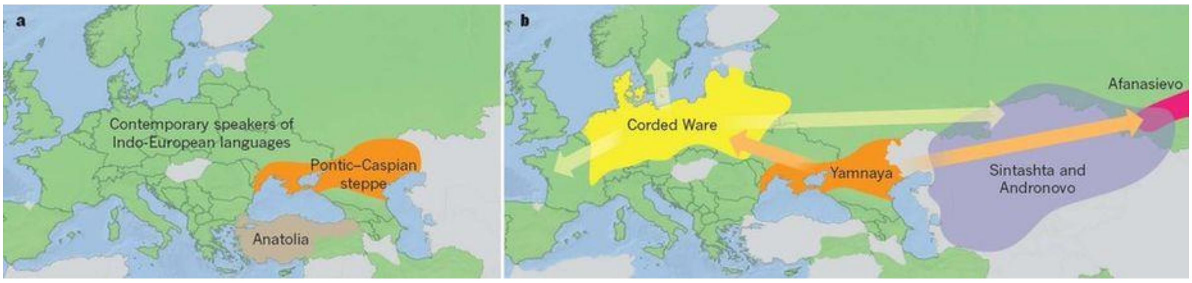


Figure 1.9. The spread of Indo-European languages. Indo-European languages have been spoken across a broad area of Eurasia throughout recorded history (territories in which these languages are spoken today are marked in green). Two geographical origins for these languages have been proposed: **a)** Anatolia and the **b)** Pontic–Caspian steppe.

1.2 Aim

This study is part of a larger project called LanGeLin (Language and Gene Lineages) that aims to build up comparable phylogenetic trees of strategically chosen languages and populations, and therefore to test in the strongest possible way Darwin's expectation about their eventual congruity, both on the local and global scales. Differences at the biological and linguistic level are likely to retain historical information that generate robust phylogenies.

To solve the problem represented by the scarce, or nonexisting, detectable relationships between lexical items in languages of different linguistic families, in this study, we propose to describe linguistic relationships at the syntactical, rather than lexical, level through the PCM. Focussing on Eurasian populations speaking languages of the IE, UR and AL families, we compared the syntax and the genomes of several AL- UR- and IE-speaking populations with the available modern and ancient genomes in the area of interest. Our multidisciplinary approach comparing grammars and genomes will ultimately help us better understand the evolution of this cultural and biological diversity in Western/Central Eurasia. In particular, we shall try to understand whether, under particular conditions, secondary contacts have played a non-negligible role in determining syntactic convergence (Thomason and Kaufman, 1988), and whether the existing evidence tends to support Anatolia or the Pontic Steppes as original homeland of the IE languages.

1.3 Methods

1.3.1 Genomic dataset

The dataset analysed in this study comprises the high-coverage sequenced genomes of 45 individuals from 17 populations from Eurasia (figure 1.10 and table 1.1). The samples were collected from Pagani *et al.* (2016) (Pagani *et al.*, 2016) and downloaded from the public database ENA (European Nucleotide Archive). For the sake of equal representation, a random subset of three individuals per population was chosen for populations with a larger sample size, to perform all the analyses.

Ancient and modern Genome-Wide SNP array data from Patterson *et al.* (2012), Lazaridis *et al.* (2014), Mathieson *et al.* (2015), Allentoft *et al.* (2015) and Haak *et al.* (2015) was used to perform Outgroup *f3*-statistics and *qpAdm* analysis (Supplementary Tables S1.1 and S1.2, respectively) (Patterson *et al.*, 2012; Lazaridis *et al.*, 2014; Allentoft *et al.*, 2015; Haak *et al.*, 2015; Mathieson *et al.*, 2015).

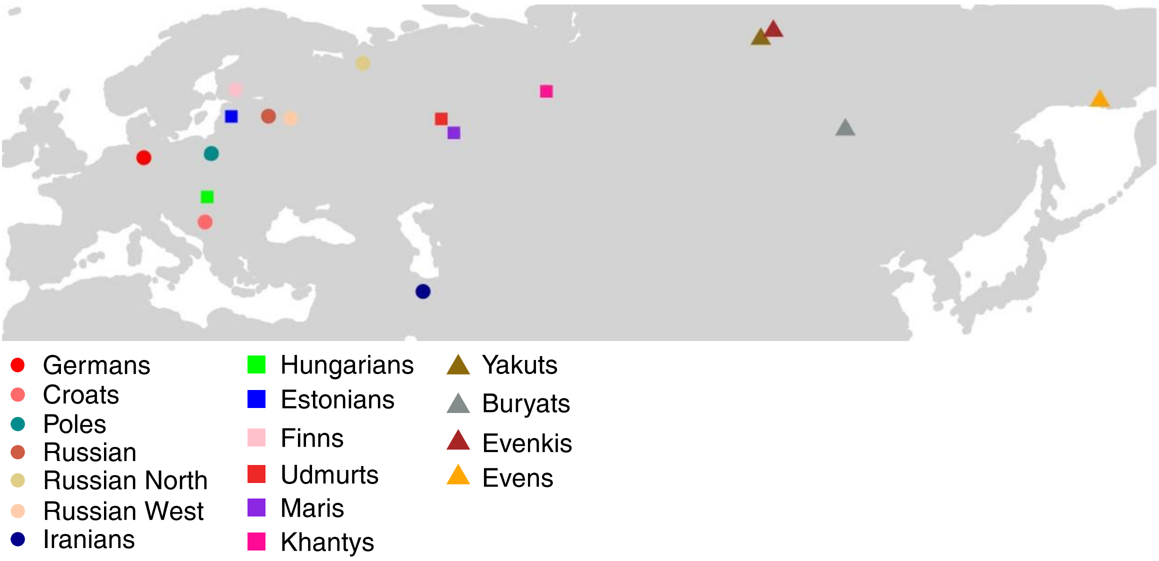


Figure 1.10. Map with the location of the individuals used in the analysis. Populations speaking an IE, UR and AL language are represented by circles, squares and triangles, respectively.

1.3.2 Dataset preparation

Samples from Pagani *et al.* (2016) (Pagani *et al.*, 2016) were in Complete Genomics MasterVar format files (reads mapped against the human genome reference hg19/GRCh37). To convert the MasterVar file into a Variant Call Format (VCF) the cgatool mkvf (version 1.8.0.1) from Complete Genomics was used. The VCF file created only contains SNP variants with a quality above 40 dB, which means variants called with a high confidence. All the VCF files from the different individuals were merged using BCFtools (version v1.6-36) merge with the option '-m none' to output the multiallelic sites in different lines. All the duplicated variants were excluded from the data. The VCF files were phased using SHAPEIT2 (Delaneau *et al.*, 2012) (version v2.r837) using the 1000 Genomes phase 3 haplotypes as a reference panel, as recommended (Abecasis *et al.*, 2010; Consortium *et al.*, 2015). Heterozygous sites not present in the 1000 Genomes data were left unphased. In the end, 11.931.455 autosomal SNPs were obtained.

Table 1.1. Whole-genome samples collected for the populations under study.

Sample Size	Sample ID	Populations	Country	Region	Coverage	Language Family	Reference
3	Est1, Est2, Est3	**Estonian**	Estonia	Europe	>40	**Uralic**	Pagani *et al.* 2016
3	Fin1, Fin2, Fin3	**Finnish**	Finland	Europe	>40	**Uralic**	Pagani *et al.* 2016
3	Rus1, RusPi1, RusPs2	**Russian (North and West)**	Russia	Europe	>40	**Indo-European**	Pagani *et al.* 2016
3	Pole1, Pole2, Pole3	**Polish**	Poland	Europe	>40	**Indo-European**	Pagani *et al.* 2016
3	Hun1, Hun2, Hun4	**Hungarian**	Hungary	Europe	>40	**Uralic**	Pagani *et al.* 2016
3	Ger1, Ger2, Ger3	**German**	Germany	Europe	>40	**Indo-European**	Pagani *et al.* 2016
3	croat11, croat13, croat12	**Croatian**	Bosnia-Herzegovina	Europe	>40	**Indo-European**	Pagani *et al.* 2016
3	Iran1, Iran2, Iran3	**Iranian (Farsi)**	Iran	West Asia	>40	**Indo-European**	Pagani *et al.* 2016
3	Mari1, Mari2, Mari3	**Mari**	Russia	Europe	>40	**Uralic**	Pagani *et al.* 2016
3	Udmrd1, Udmrd2, Udmrd3	**Udmurt**	Russia	Europe	>40	**Uralic**	Pagani *et al.* 2016
3	Khant1, Khant2, Khant3	**Khanty**	Russia	Siberia	>40	**Uralic**	Pagani *et al.* 2016
3	Evnk2, Evk14, Evk16	**Evenki**	Russia	Siberia	>40	**Altaic**	Pagani *et al.* 2016

3	Bur2, Bur6, Bur11	**Buryat**	Russia	Siberia	>40	**Altaic**	Pagani *et al.* 2016
3	YakS4, YakK1, YakK3	**Yakut**	Russia	Siberia	>40	**Altaic**	Pagani *et al.* 2016
3	EvenM1, EvenM2, EvenM3	**Even**	Russia	Siberia	>40	**Altaic**	Pagani *et al.* 2016

Note: the three Russian individuals come from three different subsets.

1.3.3 Linguistic data and distances

For the linguistic distances, we use the PCM (Longobardi 2003, Guardiano and Longobardi 2005, Longobardi and Guardiano 2009), capitalizing on its latest elaboration and dataset (Ceolin, 2019): it consists of 97 binary parameters defining properties of nominal structures for 65 different languages. A principal component analysis (PCA) was obtained to understand the linguistic diversity of the studied populations. To perform the PCA the software PAST was used (Hammer *et al.*, 2001). Using the vegan package implemented in R, the Jaccard-Tanimoto distance was calculated (for comparing the similarity and diversity of sample sets) between pairs of languages. Therefore, two languages with fully identical features will have distance 0, two with all different feature values will display distance 1, all the other cases falling in between. A UPGMA tree was obtained from the syntactic distances calculated.

1.3.4 Principal Component Analysis

A PCA was obtained to understand the genetic diversity of the studied populations. To perform the PCA QTLtools was used (Delaneau *et al.*, 2017) (version v1.1) on scaled and centered genotype data on relatively independent (50 Kb distance) and frequent variants (minor allele frequency = 0.05; the minor allele is defined at the global level, and then its frequency is evaluated within each sample).

1.3.5 Genomic distances

To calculate the genomic distances (Weir & Cockerman's (1984)) between populations we used the *4P* software (Benazzo *et al.*, 2015) (version 1.0). Genomic regions that may be under selection were masked using bedtools subtract (version v2.26) and variants with a missing call rate exceeding 10% were excluded, resulting in a total of 9.881.752 autosomal SNPs.

1.3.6 ChromoPainter and fineSTRUCTURE

ChromoPainter (Lawson *et al.*, 2012) (version v2) allows to compare pairs of individuals and see how genetically close/distant they are. This method uses sampled chromosomes as "donors" and matchs (or "paints") other chromosomes to donors' DNA creating a cluster based on who shares many blocks of SNPs. Each individual is "painted" as a combination of all the other sequences. ChromoPainter outputs an heatmap in which each square is the number of DNA segments that each row (recipient) copies from each column (donor).

ChromoPainter output was used to cluster individuals into genetically homogeneous groups using fineSTRUCTURE (Lawson *et al.*, 2012) (version 2.1.3). fineSTRUCTURE is a powerful approach which infers fine-scale population structure from haplotype data. Each individual is presented as a matrix of non-recombining genomic chunks received from a set of multiple donor individuals. The patterns of similarities between these copying matrices are then used to cluster individuals into genetic groups using the Bayesian approach.

To perform this analysis, ChromoPainter was first run on a subset of the phased VCF files (i.e. only a subset of individuals and chromosomes), in which it was estimated the "switch" and "mutation (emission") rates using Expectation-Maximization (EM). Using the estimated parameters (average switch rate of 5.000 and a global mutation probability of 0.015), it was ran ChromoPainter again on all individuals and chromosomes. Lastly, fineSTRUCTURE was used with the output of ChromoPainter to cluster the individuals into genetic groups. The tree was plotted using FigTree (version 1.4.2).

1.3.7 Outgroup *f*3-statistics

We performed an *f*3 analysis using the *qp3Pop* package in ADMIXTOOLS (version 412). The outgroup *f*3-statistics (X, Y; Outgroup) is a function of shared branch length between X and Y in the absence of admixture with the outgroup. From a set of populations (Y) we wanted to find the most closely related to the population under examination (X). We used for all the analysis the African Yoruba as an outgroup that we assumed to be diverged from population X and all the other populations analysed. This analysis measured the amount of shared genetic drift of X and all the other populations and high values of *f3* indicate that X and Y are genetically closer.

The modern samples used in this study from Pagani *et al.* (2016) (Pagani *et al.*, 2016) were merged with the Yamnaya, Anatolia, Sintashta (the three representing ancient populations) and Nganasan (a modern population) individuals from Lazaridis *et al.* (2014), Mathieson *et al.* (2015), Allentoft *et al.* (2015) and Haak *et al.* (2015) and used as source populations (see Supplementary Tables S1.1 and S1.2). Variants with a missing call rate exceeding 10% were excluded. This resulted in 249.286 SNPs.

1.3.8 Modelling admixture

Using *qpAdm* package in ADMIXTOOLS (version 412) we estimated the proportions of ancestry in a *Test* population deriving from a mixture of three reference populations by leveraging shared genetic drift with a set of outgroup populations. Following Haak *et al.* (2015), where the method was first described, we used a core set of outgroups from worldwide modern human populations (Han, Mbuti, Karitiana, Ulchi and Mixe). The reference populations used were: Yamnaya, Anatolia and Nganasan (the latter used here as a Siberian proxy). The detail: YES parameter was set, which reports a normally distributed Z-score for the goodness of fit of the model (estimated with a Block Jackknife).

1.4 Results

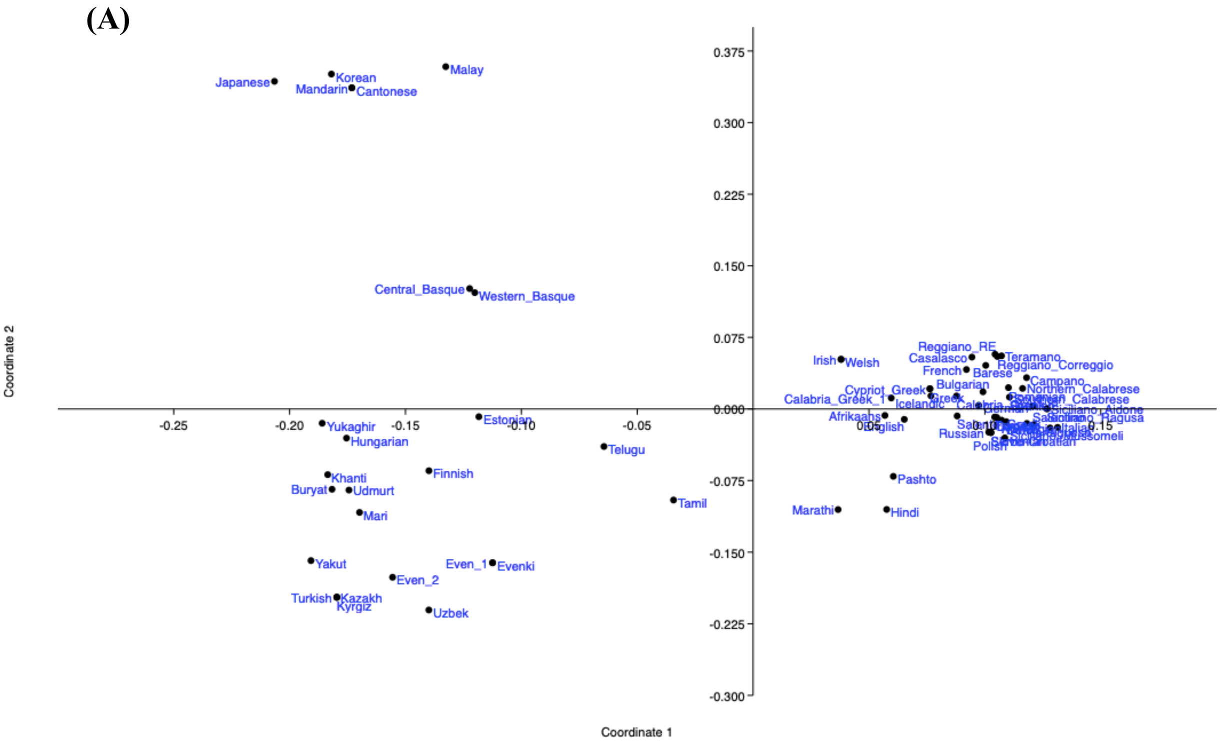
1.4.1 Linguistic comparison

Pairwise syntactic distances between languages

We used the dataset of 97 binary syntactic parameters (grammatical dimorphisms) valued in 65 modern Eurasian languages presented in Ceolin *et al.* (2019). We focussed on a subset of 40 languages from the IE, the UR and the genealogically controversial AL groupings (Longobardi and Guardiano, 2013): we then calculated parametric syntactic distances and inferred from them a UPGMA language tree (figure 1.11B).

All languages that belong to the same linguistic family appear in different quadrants of the PCA, except for Buryat, which displays salient proximity with some Uralic languages, probably revealing plausible contact effects of Mongolian with Hungarian and Udmurt.

In the tree, instead, languages from the same family neatly cluster together without exception (figure 1.11); in particular, all UR languages, including those (Hungarian, Finnish and Estonian) spoken in Europe, form a monophyletic cluster. However, in full agreement with traditional lexical scholarship, Hungarian clusters with Khanty (Ugric subfamily) and away from the Western (Finnish-Estonian) clade (Longobardi and Guardiano, 2013; Honkola *et al.*, 2013; Longobardi *et al.*, 2015; Tambets *et al.*, 2018).



(B)

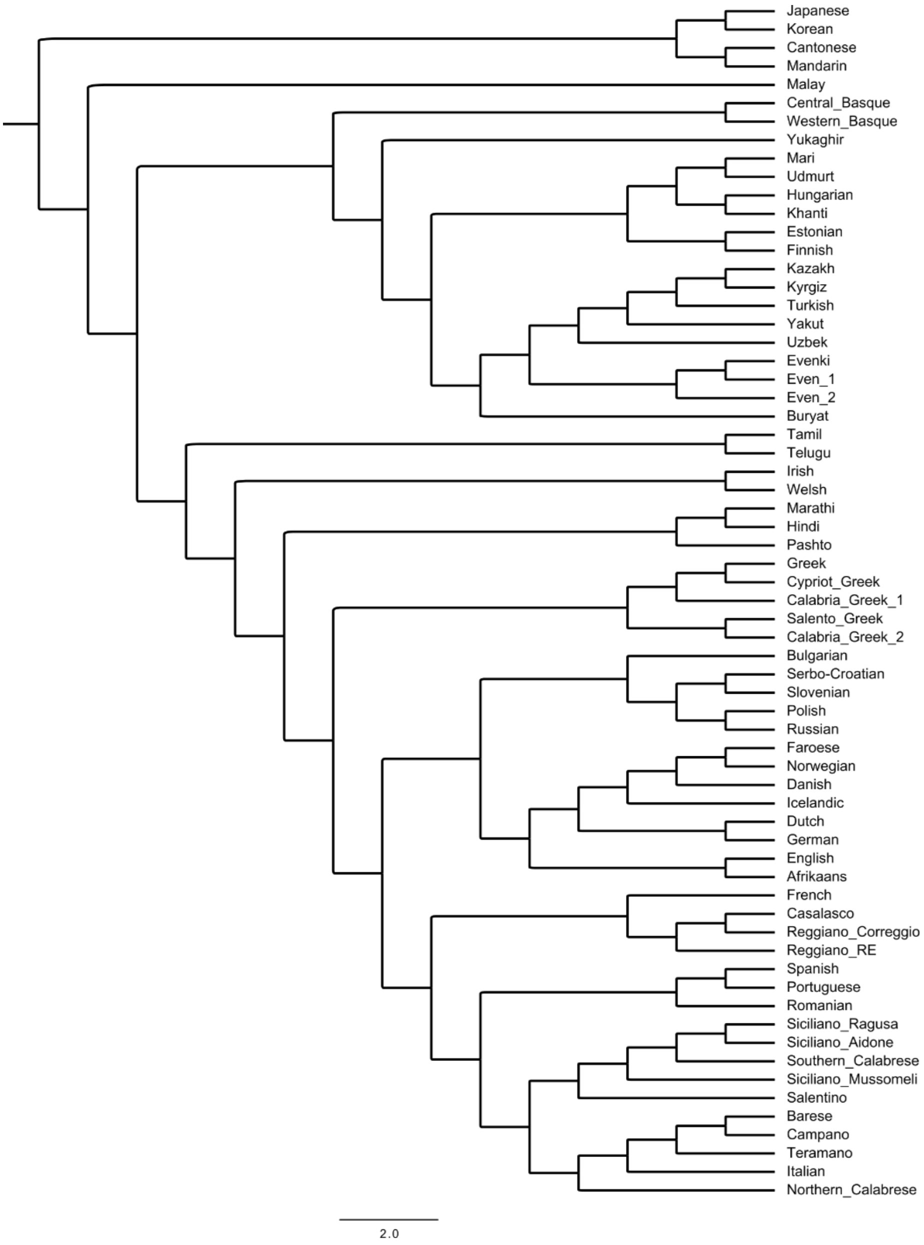


Figure 1.11. (A) PCA from the syntactic distances in 65 different languages. (B) UPGMA tree obtained from the syntactic distances in 65 languages (from Ceolin A., C. Guardiano, M. Irimia, G. Longobardi, *Formal syntax and deep history*, submitted).

The distribution of the syntactic distances can be observed in greater detail in the following heatmap (figure 1.12):

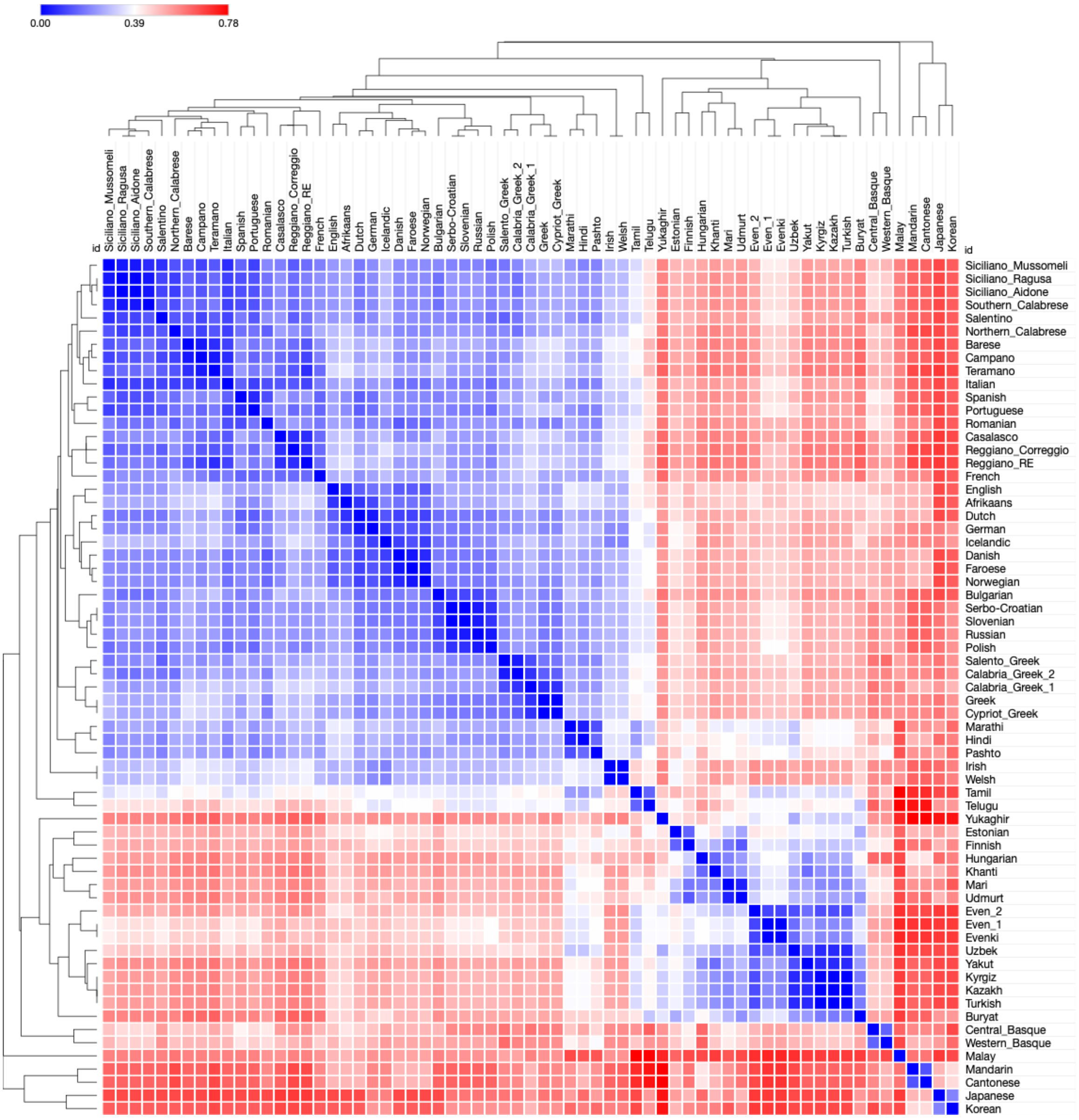


Figure 1.12. Heatmap of syntactic distances (from Ceolin A., C. Guardiano, M. Irimia, G. Longobardi, *Formal syntax and deep history*, submitted).

To sum up, syntax supports the main clades recognized by classical lexical approaches. Several results of this analysis, however could not be retrieved from traditional lexical data, namely: 1) AL (especially the union of Turkic and Tungusic) is more compact as a potential family than it appears lexically; 2) there is a UR-AL similarity much stronger than IE-UR or IE-AL; 3) within UR-AL, Balto-Finnic is clearly an outlier of UR, and Buryat (the only Mongolian of the sample) is an outlier of AL; 4) Balto-Finnic is slightly closer to IE than the rest of UR, including Hungarian (which is still connected to Khanty and less distinct from Mari and Udmurt), and less close to AL than the rest of UR; 5) this is particularly true of Estonian, which is clearly closer to IE (actually to Central-European: Slavic and Germanic) than Finnish; the latter has also, overall, drifted toward IE, but preserves some less sharp distances from the Eastern languages of the whole sample (the rest of UR and AL); 6) Mari, Udmurt, and Khanty are the least close to IE; once this is taken into account, it is remarkable that the IE languages they look less different from are not the European ones but the Indo-Iranian ones; 7) of these three Eastern UR languages, the Easternmost one, Khanty, is the most similar to Yakut, the Easternmost Turkic (AL) one.

1.4.2 Genetic comparison

Population structuring in Eurasia

We selected 17 populations, 7 in the IE, 6 in the UR and 4 in the AL language family, for which whole-genome (>11 million SNPs) data were available, all at a coverage >40 (figure 1.10; Table 1.1). As a first, exploratory step, we ran a PCA to investigate the genomic background of these populations (figure 1.13). The first component mostly reflects geography and separates Eastern from Western Eurasian populations, whereas the second component separates Western Eurasians along a north-south cline. The AL-speaking populations fall in a single cluster along the first PC axis. The European IE-speaking populations form a cluster along the PC2 axis, separated from the Iranians, the latter belonging to the Asian group of IE languages.

Conversely, the UR-speaking populations are scattered along the PC1 with Estonians falling within the IE diversity at the negative end of the X-axis, while Finns are placed in an intermediate position between the IE-speaking populations and the UR-speakers Udmurt (Votiaks) and Mari Cheremis), i.e. the modern populations that geographically closest to the region of the steppes (figure 1.13).

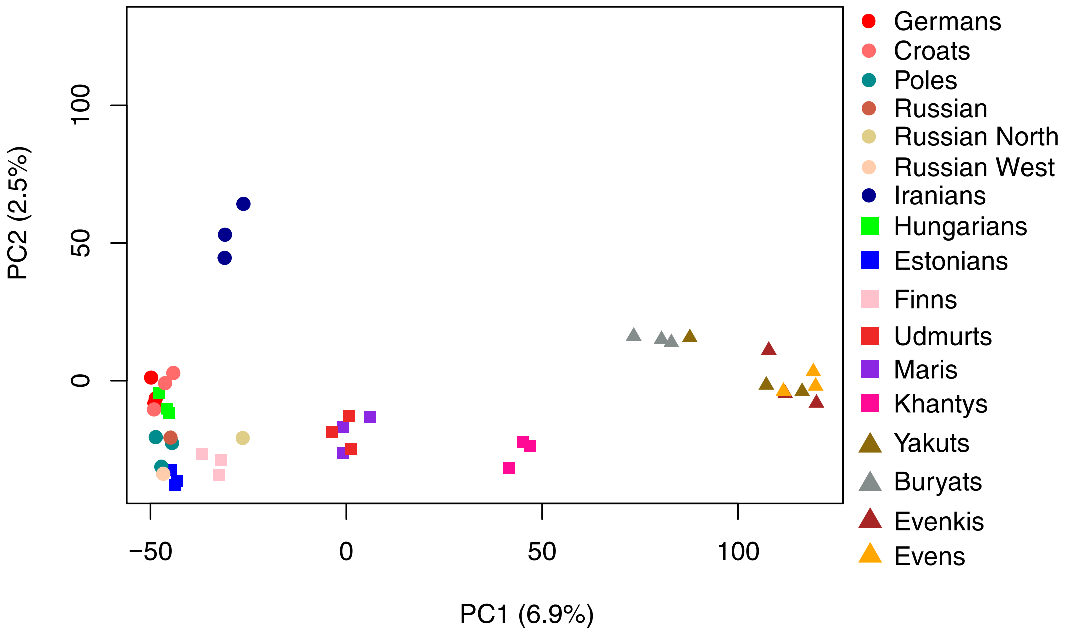


Figure 1.13. Principal component analysis (PCA). Populations speaking an IE, UR and AL language are represented by circles, squares and triangles, respectively. Projection on two dimensions of the main components (PCA) of genomic variation in IE, UR and AL speaking populations.

Genetic distances between the IE, Uralic and Altaic populations

Next, we calculated genetic distances (*Fst*) between pairs of populations (figure 1.14). All AL and IE speaking populations are genetically closer to other populations of their language family than to populations belonging to a different family. Instead, that is not the case for the UR-speakers; all of Estonians, Finns and Hungarians are genetically closer to their respective European neighbours speaking IE. In addition, among the Eastern populations, the Mari and Udmurt seem genetically more similar to the Europeans than to the AL-speakers. Exceptions are the Easternmost and Trans-Uralic Khanty (sometimes still called Ostyaks), which seem equally close to Mari, Udmurt and most of the AL speakers.

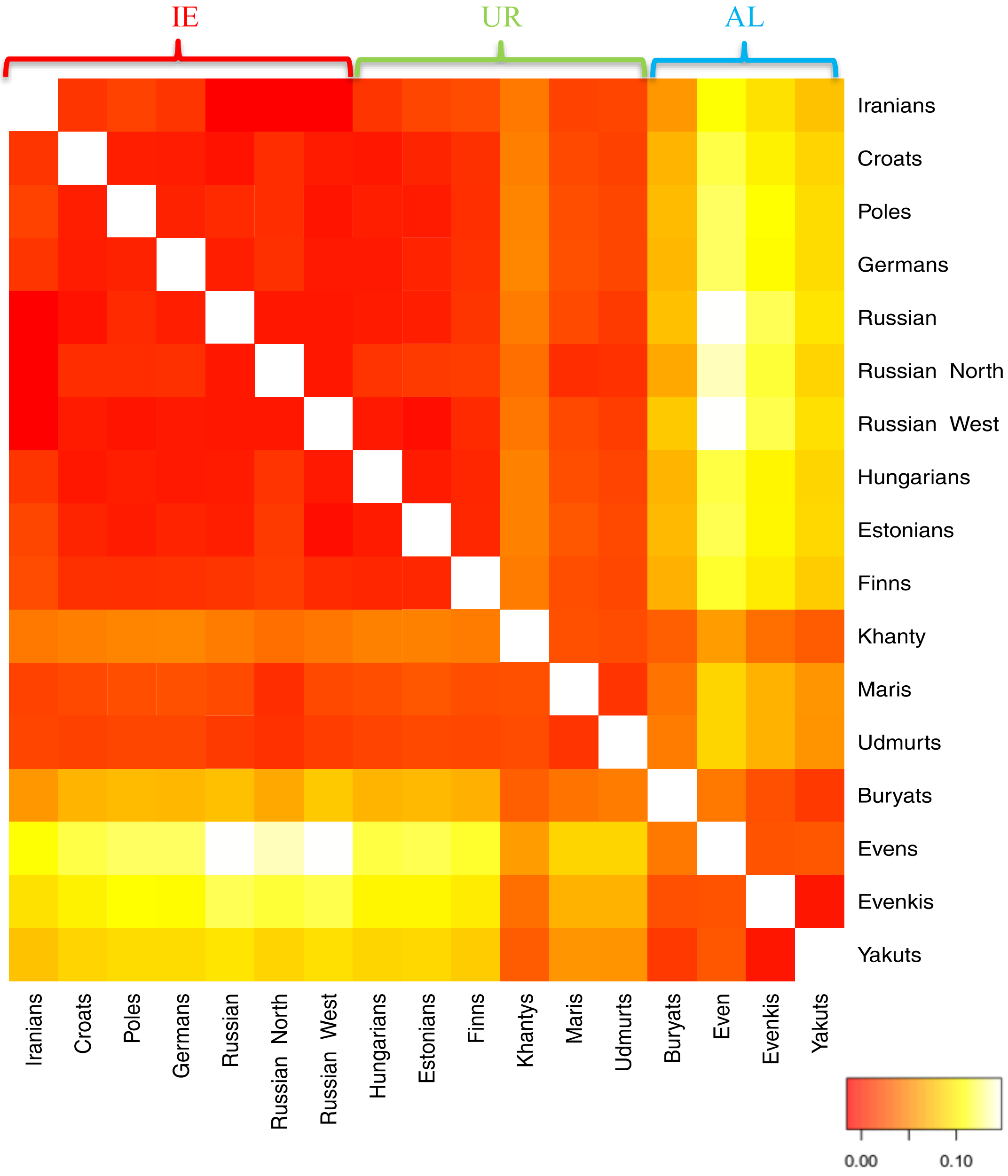


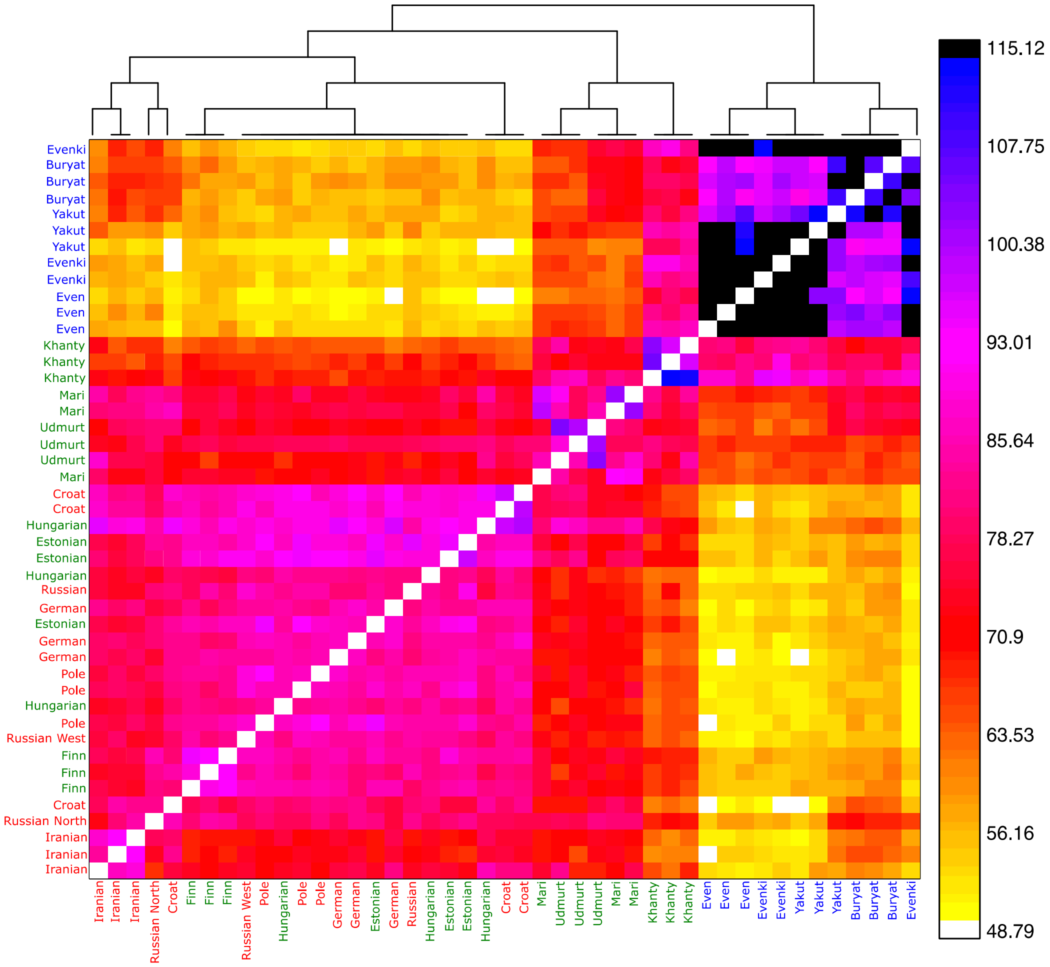
Figure 1.14. Pairwise genetic distances between Eurasian populations. Darker colours indicate that populations are genetically closer, while lighter colours indicate that populations are genetically distant.

Shared haplotypes

In the analysis of genetic distances, each single-nucleotide polymorphism is independently considered, regardless of its association with other polymorphisms. To analyse the patterns of population resemblance in finer detail, we thus moved to the haplotype level, using ChromoPainter and fineSTRUCTURE (figure 1.15). This approach does not depend on prior information on sample groupings and operates instead with data-driven natural groups defined by patterns of haplotype sharing.

This way, we identified three main genetic groups, once again broadly corresponding to the three main language families, IE, UR and AL. However, as already observed in the *Fst* analysis, there were exceptions. The Western UR-speaking populations (Estonians, Finns and Hungarians) seem to mainly share coancestry with other Europeans populations, regardless of the language spoken. Conversely, the Eastern UR-speakers, Udmurt, Mari and Khanty have a high level of haplotype sharing and form a clear cluster in the evolutionary tree inferred from these data (fineSTRUCTURE cluster analysis; figure 1.15B). That tree shows two deep splits, the first isolating all AL speakers, and the second separating Eastern UR speakers from a group composed by Western UR and IE speakers. The question then arises whether this points to different ancestries for the UR-speaking populations, with phenomena of horizontal language diffusion leading them to a shared linguistic identity. We searched for an answer in ancient DNA data.

(A)



(B)

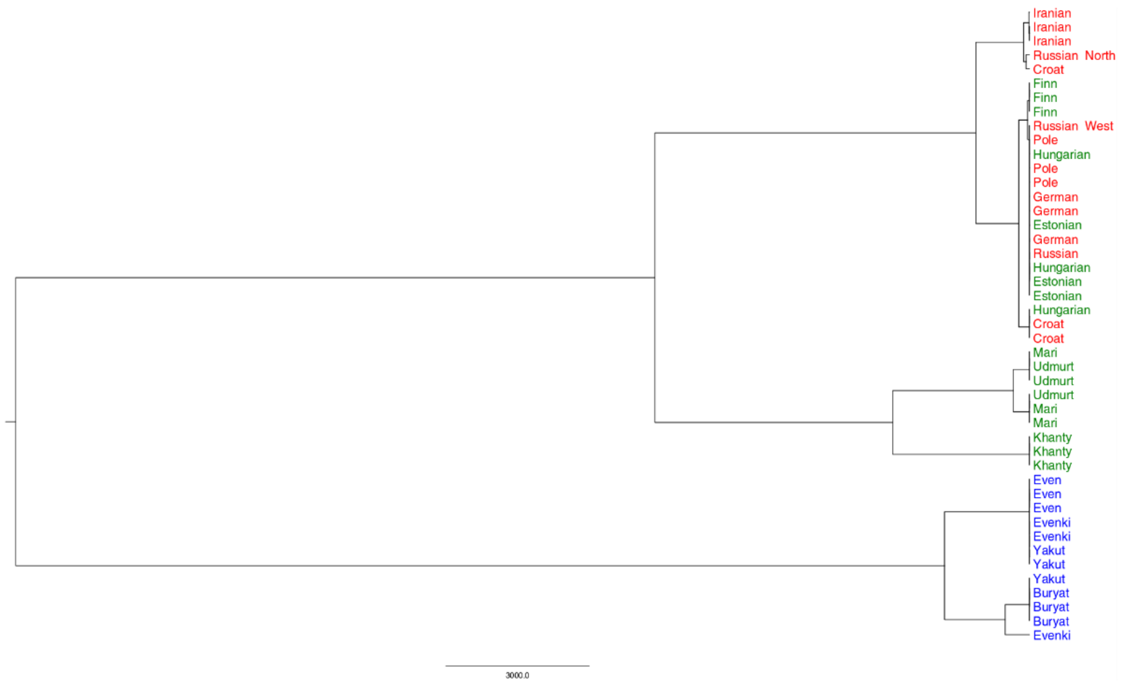


Figure 1.15. Estimates of shared ancestry between Eurasian individuals. (A) Coancestry heatmap. Each of the 45 individuals is represented as a row, where each pixel represents the level of coancestry shared with each of the other individuals. (B) fineSTRUCTURE cluster analysis obtained from the coancestry matrix. The tree is the same as the one shown on top of the heatmap in (A).

1.4.3 Affinities between modern and ancient populations

As previously remarked, Udmurt and Mari, despite being part of the Eastern UR-speaking populations, clearly show genetic affinities with Western UR speakers, Finns and Estonians. May this reflect descent from a common ancestor? If so, a natural candidate would be the Yamnaya populations, dwelling in the Bronze age around the Pontic-Caspian steppes, close to where Mari and Udmurt currently live. The Yamnaya expanded into Central and Western Europe around 4.500 yBP, contributing a Caucasian genomic component that nowadays is widespread in Europeans, especially in the Northeast (Haak *et al.*, 2015; Narasimhan *et al.*, 2019). We asked whether there is a genetic continuity between two ancient Steppe populations, Yamnaya (~4.700 yBP) and Sintashta (~3.900 yBP) on the one hand (Allentoft *et al.*, 2015; Haak *et al.*, 2015), and current populations on the other. An ancient Anatolian sample (Lazaridis *et al.*, 2016) was also included in our tests, potentially accounting for the genetic legacy of Anatolian farmers.

The outgroup *f*3-statistics estimates the length of the branch shared by two samples on the phylogenetic tree defined by an outgroup (the African Yoruba in our case) as a measure of the amount of genetic drift shared by the samples. We formulated 3 sets of outgroup *f*3-statistics of the form *f*3(AP, MP; Yoruba), where AP (ancient population) was represented in turn by Yamnaya, Sintashta and Anatolian farmers, and MP (modern population) was each of the modern samples in our dataset (figure 1.16 and Supplementary Figure S1.1). In general, we found all ancient samples to share more genetic drift with modern Europeans and Russians than with non-European populations. Among the Eastern populations, Udmurt and Mari (UR) are the ones sharing the most genetic drift with Yamnaya and Sintashta. Also, within the European populations the *f*3 values show opposite trends for the Anatolian and the Yamnaya/Sintashta, the former sharing more genetic drift with southern and central Europeans (Croats and Germans) and the latter being closer to Northeast Europeans, including the UR-speaking Estonians and Finns, once again in general agreement with previous findings (e.g. Haak *et al.* 2015). It is interesting to notice the peculiar behaviour of the Hungarians. They appear much closer to the ancient Anatolians than to the Yamnaya, which is common among southern European populations; however, they are the modern Europeans sharing most genetic drift with the Sintashta. This may be indicative of a relatively late genetic contact between them and the Steppe populations, i.e. after the process leading to the spread of the Yamnaya component in Europe.

In fact, the set of outgroup *f*3-statistics we calculated show that Mari and Udmurt are genetically close to the Yamnaya, but not as close as most European populations. Might this be due to the arrival in the Pontic Steppes of a post-Bronze-Age migration wave from the East, known to be widespread in contemporary Central and North Asian populations (Tambets *et al.*, 2018; Lamnidis *et al.*, 2018; Saag *et al.*, 2019; Jeong *et al.*, 2019)? If so, the Mari and Udmurt genomes should show that component, currently called the ‘Siberian’ component. We investigated its presence in our samples by modelling Nganasan, a population from the Taymyr Peninsula, as a proxy of the carriers of this Siberian component (as also done by Lamnidis *et al.* 2018 and Tambets *et al.* 2018). Besides showing a high level of Siberian ancestry in the AL samples, the outgroup *f*3-statistics of the form (Nganasan, MP/AP; Yoruba) showed that Udmurt and Mari are indeed closer to Nganasan than to Yamnaya. Figure 1.16B shows a very clear trend; the Nganasans share much greater genetic drift with all AL speakers, followed by Udmurt and Mari, and then by European populations, no matter if UR- or IE speakers.

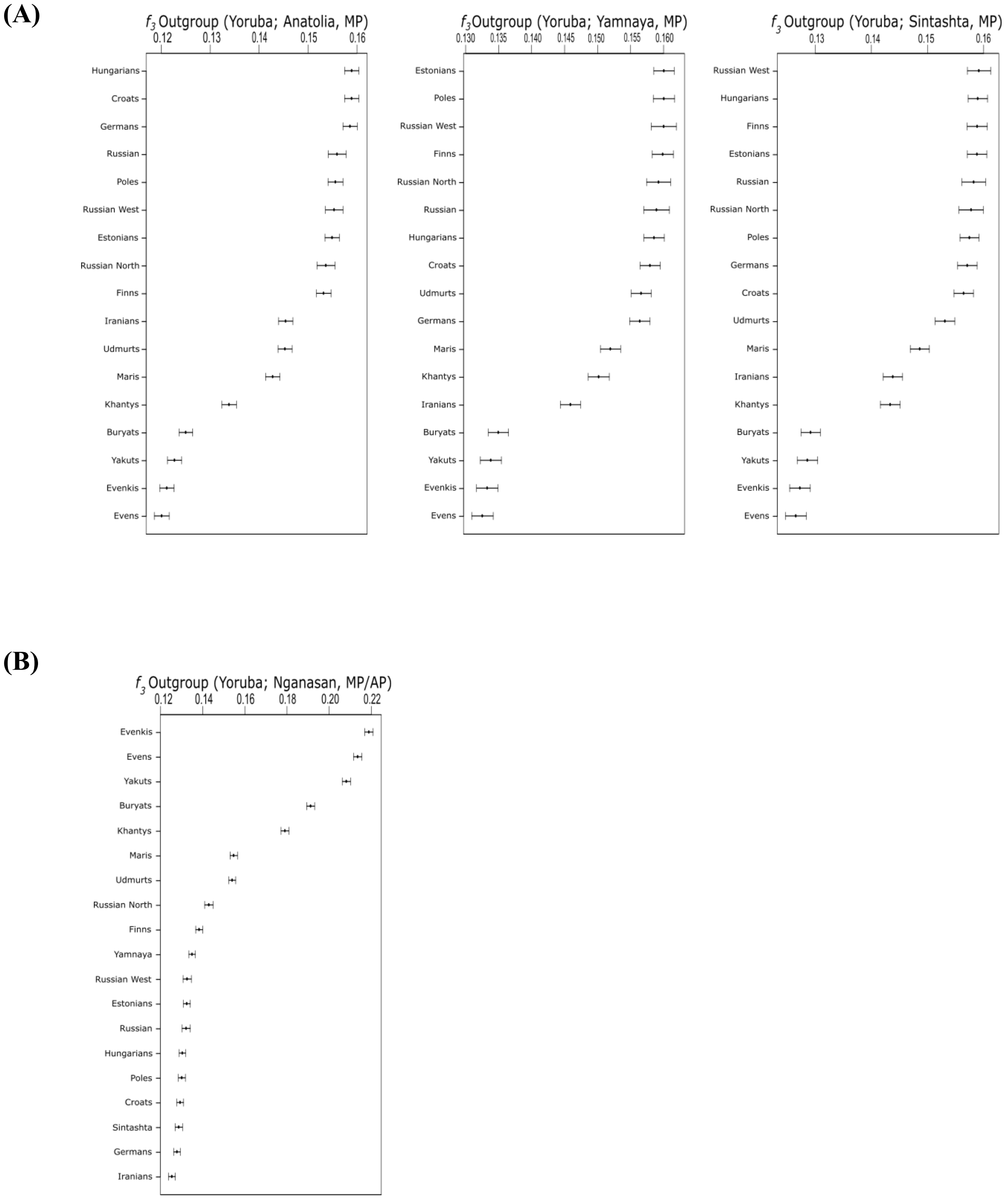


Figure 1.16. Outgroup *f3*-statistics analysis. Shared genetic drift between modern and ancient populations (MP and AP, respectively). (A) Shared genetic drift between ancient and modern populations. (B) Shared genetic drift between Nganasan and modern/ancient populations.

To further test whether the peculiar genetic position of the Udmurt and Mari is really associated with the higher presence of a Siberian genetic component in their genome, we ran a *qpAdm* analysis (figure 1.17 and Supplementary Table S1.3). With this method one can summarize information from multiple F-statistics, test whether an admixture model can account for the data and infer admixture proportions.

All the UR-speaking populations were successfully modelled as a mixture of Yamnaya, Anatolia and Nganasan-related ancestry, with the exception of the Khanty, who have no Anatolian ancestry. In particular, confirming the findings of the outgroup *f3*-statistics analysis, the Mari and Udmurt genomes do appear to contain a large component that can be related with a Siberian genetic ancestry, which is also present, at non-negligible percentages, in the Western UR-speaking Finns.

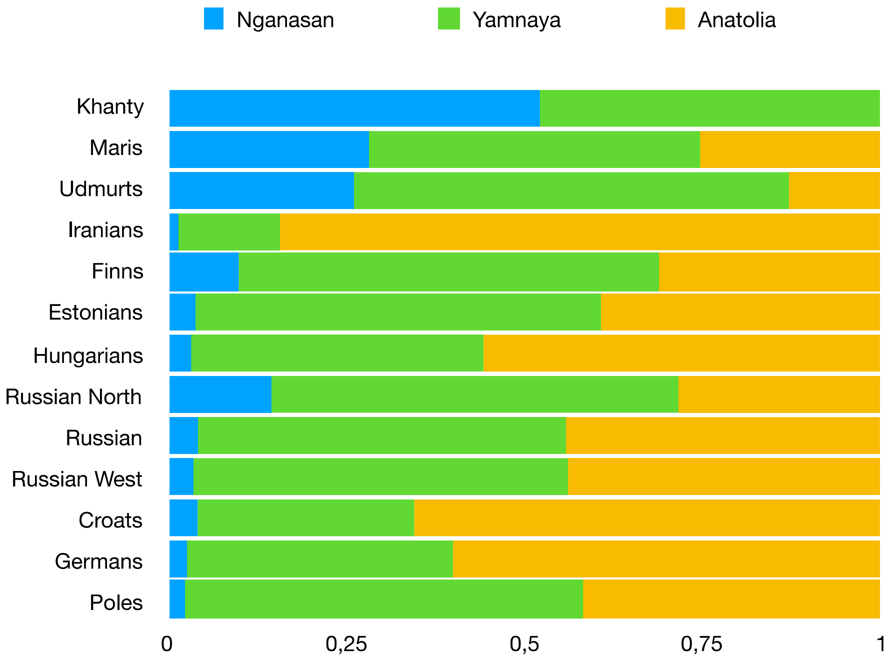


Figure 1.17. Admixture proportions from three sources estimated using *qpAdm*. Sources used were Nganasan (in blue), Yamnaya (in green) and Anatolia (in yellow) (percentages and chi-square values are shown in the Supplementary Table S1.3).

1.5 Discussion

1.5.1 Language diversity

In this study, we aim to take a quantitative approach on both sides of our comparisons, the genetic but also the linguistic one. The advantage of using the syntactic method is precisely that it allows us to quantify language distance with respect to the same generally defined grid of phenomena across the three distinct language families.

Syntax distinguishes IE, UR and AL languages quite well. The UR family (even limited to its Finno-Ugric branch, which we study here), however, turns out to be less compact than IE, in spite of the IE's greater geographic diffusion and size of the populations of speakers. It is also less compact than AL, which, at the lexical level, is less clearly established as a possible taxonomic unit (Georg *et al.*, 1999; Vovin, 2005; Robbeets, 2005; Ceolin, 2019).

Although all three groupings have outliers (Indo-Iranian and Celtic in IE, Buryat among the AL languages), the outlying position of Estonian among UR languages is very salient. Most importantly, observing the Heatmap of Figure 1.12., the whole family appears as scattered and in some structural contiguity with their Eastern and Western neighbours. Estonian has indeed the least contrast with the IE languages of Europe, followed by Finnish and then by Hungarian. Estonian has also maximum contrast with the AL languages, followed this time by Hungarian (in spite of the much more recent arrival of its speakers in Europe) and then by Finnish. Thus, Finnish shows both extreme Western and Eastern structural influences, Hungarian is less characterized by either (apart from its obviously genealogical affinities with Khanty; like the latter it has some similarity with Yakut, which could then date back to a common Ugric period), and Estonian is decidedly the most Westernized. On the other side, Khanty has clear similarities with the Easternmost Turkic language, Yakut. Combined with what we know about the dates of their presence in Europe, these crossfamily distances may tentatively suggest that Hungarian had obviously less time to be Westernized by European languages, but also that they must have been detached from Trans-Uralic Khanty for some time before it was imported into Europe; and that among the ones more anciently dwelling in Europe and grammatically more influenced by this geolinguistic experience, Finns have retained an Eastern syntactic component more than Estonians. Udmurt is syntactically closer to Balto-Finnic and Mari to Hungarian and Khanty; but both have similarities to Central and Western Turkic, and Udmurt also especially to Mongolian, but much less to Eastern Yakut.

1.5.2 Genome diversity

The genetic analysis shows that the AL-speaking populations do form a well-distinct cluster, confirming the results of the linguistic analysis. By contrast, UR and IE speakers fall in the same major cluster, which is further subdivided in two subclusters, comprising, respectively, Eastern UR-speakers (Khanty, Mari and Udmurt) on the one hand, and the Western UR-speakers, along with all IE speakers on the other.

We then asked to what extent such a genomic pattern may reflect differences in the genealogical relationships with prehistoric populations of the area. Although the sampling of such past populations is all but exhaustive, some features emerge from the *f3*-outgroup analysis. The Western UR speakers are indeed the populations showing the highest levels of relatedness with Yamnaya and Sintashta. Also, the populations currently closest to the region where these Bronze-Age groups are documented, Mari and Udmurt, are the ones showing the greatest resemblance with them, although along with several IE-speaking populations. Could that be due to the presence, among UR- but not IE-speakers, of another genomic component, presumably of Siberian origin? Indeed, Mari and Udmurt share with Siberians (along with AL-speaking populations) a greater amount of genetic drift than the bulk of the European populations. This may mean that the modern populations dwelling in the area close to the Steppes have incorporated a Siberian genome component, now widely spread among Asian populations. Because the Yamnaya show roughly the same, limited presence of such Siberian genome component as most current Europeans, it seems likely that this component reflects the effects of gene flow occurring after the Yamnaya Bronze-Age dispersal into Central Europe.

This result may have significant implications as for the origin and mechanisms of spread of both IE and UR languages. First, although the admixture proportion estimated by *qpAdm* should not be taken at face value, given the inevitable simplification imposed by the model, they indicate that the presence of an UR language in Europe is not necessarily correlated with the presence of a substantial Siberian component in the DNA of its speakers.

The genetic analysis shows that the three main linguistic groups are also biologically differentiated; in all analyses, IE, UR and AL samples form three distinct clusters, with just minor exceptions. Within the UR language family, however, a peculiar pattern emerges. While the Khanty show clear affinities with a well-differentiated cluster comprising all AL speakers, the other UR speakers appear to be part of a broad group, including all IE-speaking individuals.

In particular, the Western UR-speakers, namely Finns, Estonians and Hungarians, are genetically closer to IE populations in Europe than to the Asian UR-speaking populations.

The linguistic proximity between Estonian and Finnish speakers is also observed at the genetic-distance level, since both share more ancestry with each other than with the Hungarians. This genetic similarity can reflect: (i) a different source of steppe ancestry in the Hungarians (more closely related with the Sintashta) than in Finns and Estonians (genetically closer to the Yamnaya) (figure 1.16A); and/or (ii) a lower contribution of Siberian ancestors to the Hungarian genomes than to the Estonians and especially the Finns (figure 1.16B). Geographical proximity between Finland and Estonia is likely to also have played a role.

1.5.3 Reconstructing the history of UR speaking populations

Experts disagree about the geographic region proto-UR languages spread from, whether in the Volga river basin or further East, in Siberia. Studies supporting an origin in the Volga-Oka area date the initial proto-language around 6.000-7.000 yBP and estimate the first divergence of the Finno-Ugric family (including all UR languages except Samoyed) around 5300 yBP (Kallio, 2006; Janhunen, 2009b; Honkola *et al.*, 2013). The Volga basin comprises the area where the Yamnaya and Samara steppe cultures are documented, and their first expansion into Central Europe comes close to the time of diversification of the Finno-Ugric languages (around 4.800 yBP (Honkola *et al.*, 2013)). By contrast, UR language came to be spoken in current Hungary only at a much later time.

There is historical evidence that at the beginning of the modern era, the language spoken in nowadays Hungary was still Late Latin (at least as an official language), later subject to the effects of Slavic, Germanic and Avar invasions (Csányi *et al.*, 2008). The main linguistic shift can be approximately dated around 895-905 AD, when people coming from the East conquered Hungary, imposing their own language belonging to the UR family (Cavalli-Sforza, 1997). Ancient DNA studies of the invaders have shown that they were genetically related with the Sintashta, and apparently unrelated with Siberian ancestors (Neparáczki *et al.*, 2017), in fine agreement with our analysis. Although the admixture proportion estimated by *qpAdm* should not be taken at face value, given the inevitable simplification imposed by the model, they indicate that the presence of a UR language in Europe is not necessarily correlated with the presence of a substantial Siberian component in the DNA of its speakers.

Our genetic analyses are congruent with previous studies that identify the Finns as a peculiar genetic group (Lappalainen *et al.*, 2006; Salmela *et al.*, 2008; Lek *et al.*, 2016; Kerminen *et al.*, 2017). The fineSTRUCTURE analysis showed they form their own genetic cluster within the European clade, which we think could be related with the higher presence in their genomes of a Siberian component compared to other European populations, as shown in the *f3* analysis (figure 1.15B) and, to a lesser extent, by the Admixture graphs (Figure 1.17). Recent studies suggest that the Finns' genome diversity can be best explained by a model including four components: the Western hunter-gatherers, the Near East farmers, the Yamnaya and, indeed, a Siberian component (Tambets *et al.*, 2018; Lamnidis *et al.*, 2018).

When did this Siberian genome component reach Finland, then? The analysis of ancient remains from Fennoscandia and Estonia dated the Siberian ancestry to the Bronze Age/Iron Age transition (around 2.500 yBP) and suggested it could be related with the arrival of the Finnic language to the region (Lamnidis *et al.*, 2018; Saag *et al.*, 2019). However, this interpretation may be simplistic. Indeed, these ancient DNA studies identified the Siberian component in a 3.500 yBP individual from the Kola Peninsula, while there is a decrease of the Siberian component in Finnish samples from later periods (around 2.500 yBP) and a higher similarity, instead, to the Corded ware samples from Estonia (Lamnidis *et al.*, 2018). Altogether, these findings suggest a South-to-North, rather than North-to-South, gene flow. The same study interprets the presence of the Siberian component in the Kola Peninsula as pre-dating the spread of the extant UR languages to the area. In summary, archaeological, linguistic and genetic data may be reconciled assuming a Northward migration of people from Estonia into south Finland, after the development of the Neolithic in the coastal Baltic areas (around 2.000-1.600 yBP) (Kivikoski, 1961; Miettinen, 1996; Honkola *et al.*, 2013; Lamnidis *et al.*, 2018; Saag *et al.*, 2019). These people would be partly descended from the Baltic Corded Ware populations, i.e. the ones carrying the highest percentage of the Yamnaya genome component in Late Neolithic Europe (Jones *et al.*, 2017; Lamnidis *et al.*, 2018). The Baltic farmers would have migrated north through the Gulf of Finland, where they would have acquired the Siberian component by admixture with local populations inhabiting the South of Finland.

Therefore, based on the linguistic, archaeological and genetic data, we cannot reject the hypothesis that the Finno-Ugric languages spread into Northeast Europe along with the Yamnaya expansion. This migratory movement would have had linguistic consequences in areas where IE was not established because the transition to Neolithic societies either had not occurred yet, or had mostly been a cultural shift. For example, in Latvia, where ancient DNA data show that the Neolithic transition was not mediated by gene flow from Anatolia (Jones *et al.*, 2017), the extinct UR Livonian was replaced only recently by an IE language ((Honkola *et al.*, 2013); (Jones *et al.*, 2017)). Later on, the migration of these Baltic farmers toward Finland would have led to the acquisition of the Siberian ancestry and the further diversification of the Finnic peoples into the southern (including the Estonian) and the northern (Finnish) branches, in relatively recent times, around 1.500 yBP (Honkola *et al.*, 2013; Lehtinen *et al.*, 2014) . The syntactic similarity and the differences we observed between Estonian and Finnish and between them and other IE and UR languages (figure 1.11) seem in fine agreement with this dual-migration model, although by themselves they do not provide information on the direction of the cultural change, whether South-to-North or North-to-South.

In this context, the similarities between UR speakers and their geographical neighbours of different language families (IE in the West, AL in the East), repeatedly observed in this study both at the linguistic and genetic level, point to exchanges that entailed some level of biological admixture, because otherwise they would not have left such a mark in the genomes. The comparison of grammars allowed us to recognize a parallelism in patterns of genomic and linguistic variation, which would have been impossible to quantify, and perhaps even identify, using traditional approaches based on vocabulary comparisons. For a deeper understanding of these processes, a syntactic reconstruction of the language of the inhabitants of Northern Finland, the Sami, would be important.

1.5.4 A corollary on reconstructing aspects of the diffusion of IE into Europe

Among linguists, the timing and modes of spread of IE languages in Europe have long been debated. Among archaeologists, Gimbutas (Gimbutas, 1979) associated it with the Westward spread of the Kurgan culture, from the Pontic steppes during the Bronze age, whereas Renfrew (Renfrew, 1987) saw it as a consequence of the Neolithic farmers' demic diffusion from Anatolia (see also (Gray and Atkinson, 2003; Bouckaert *et al.*, 2012)). These alternatives (hereafter referred to as the Steppe and the Anatolian hypothesis, respectively) are paralleled at the genetic level, with studies supporting either an expansion and diversification of the IE languages with the Early Bronze Age migration of Yamnaya-related populations from the Russian steppes (Allentoft *et al.*, 2015; Haak *et al.*, 2015), or during the Neolithic transition

(Menozzi *et al.*, 1978; Cavalli-Sforza *et al.*, 1988; Sokal, 1988; Sokal *et al.*, 1990; Barbujani and Pilastro, 1993; de Barros Damgaard *et al.*, 2018).

If, as seems possible, the Yamnaya people introduced UR languages in Europe, it is difficult to imagine they were also the first IE speakers in Europe, as postulated by the Steppe hypothesis. In fact, there is ancient DNA evidence indicating that the arrival of the steppe component was not always accompanied by linguistic changes (Olalde *et al.*, 2019). In the present study, genome-wide and linguistic information were not sufficient to formally test these hypotheses.

Inference of complex processes, such as demographic and linguistic changes at the continental level, require the study of broader datasets than available for the present study. However, this study exemplifies how it is possible to analyse in depth language variation across different families, offering a novel insight into human past and paving the ground for comparative studies at larger geographical scales. We showed that, in Central/Western Eurasia, parallel linguistic and demographic change is a rule with some exceptions, and we identified cases in which secondary contacts between populations led to changes at both the linguistic and genetic level. Finally, we warn about the risk of over-interpreting the association between genetic and linguistic changes, especially in the absence of accurate dates of linguistic diversification and expansion, which at the moment are not yet well established.

CHAPTER II:

Integrating genomic and linguistic data through a new ABC framework - Explaining the Bantu expansion: Early or Late split hypothesis

2.1 Introduction

2.1.1 The Bantu population

Africa is one of the most culturally and genetically diverse regions in the world. Archaeological and genetic studies suggest an origin and diversification for the anatomically modern humans (AMH) within strongly subdivided populations possibly living in Africa, connected by occasional gene flow, approximately 200.000-150.000 yBP. Afterwards, a worldwide population expansion happened around 75.000-50.000 yBP (figure 2.1) (Watson *et al.*, 1997; Harpending *et al.*, 1998; Quintana-Murci *et al.*, 1999; Excoffier and Schneider, 1999; Underhill *et al.*, 2000; Ingman *et al.*, 2000; Thomson *et al.*, 2000; Salas *et al.*, 2002; Cavalli-Sforza and Feldman, 2003; Garrigan and Hammer, 2006; Patin *et al.*, 2009; Scerri *et al.*, 2018, 2019). However, several aspects of the history of the sub-Saharan African populations remain unclear. These are populations scattered over an entire continent, and displaying extremely high levels of cultural, linguistic, phenotypic and genetic diversity (Tishkoff and Williams, 2002; Patin *et al.*, 2009; Tishkoff *et al.*, 2009; Pickrell *et al.*, 2012; Schlebusch *et al.*, 2012, 2017; Barbieri, Güldemann, *et al.*, 2014; Skoglund *et al.*, 2017; Henn *et al.*, 2018; Schlebusch and Jakobsson, 2018; Fan *et al.*, 2019).

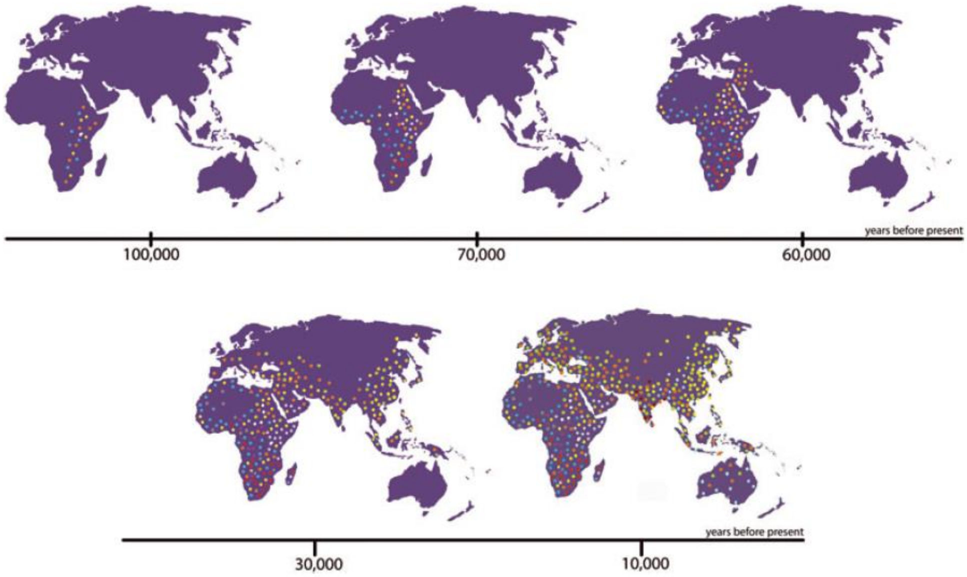


Figure 2.1. Schematic view of the evolution of human biodiversity in the last 100.000 years. Dots with different colours represent different genotypes. Approximate dates for the five panels: 100.000 years BP; 70.000 years BP; 60.000 years BP; 30.000 years BP; and 10.000 years BP (image from (Barbujani *et al.*, 2013)).

Multidisciplinary studies suggest a drastic shift in the population of sub-Saharan Africa during the Bantu expansion. This was one of the most significant moments in the African history, due to the sheer magnitude and relative rapidity it happened (Huffman, 1970; Vansina, 1995; Bostoen *et al.*, 2015). Based on archaeological and linguistic evidence, it has been suggested that the first Bantu speakers dispersed from West Central Africa (assumed homeland in North Cameroon) approximately 5.000 yBP (Vansina, 1995), most likely in more than a single wave of migrants (Beltrame *et al.*, 2016). These expansions were possibly accompanied by the spread of Bantu languages. The Bantu language family is the largest in Africa, with approximately 400 - 600 different languages occupying a vast territory and spoken by a high number of people (approximately 240 million) (Guthrie, 1962; Patin *et al.*, 2009; de Filippo *et al.*, 2012; Barbieri, Vicente, *et al.*, 2014; Grollemund *et al.*, 2015). Bantu languages are a subgroup of the Niger-Kordofanian linguistic division, one of the four independent major linguistic groups in Africa. The Bantu languages are divided into three major groups (figure 2.2): northwestern Bantu (subgroups A, B and C), eastern Bantu (subgroups E, F, G, J, N, P and S) and western Bantu (subgroups H, K, L, R, D and M) (Vansina, 1995; Holden, 2002). The close similarity among

Bantu languages and the vast geographical dispersion give support for a rapid and quite recent spread of these languages and have interested scholars for many years now (Huffman, 1970).

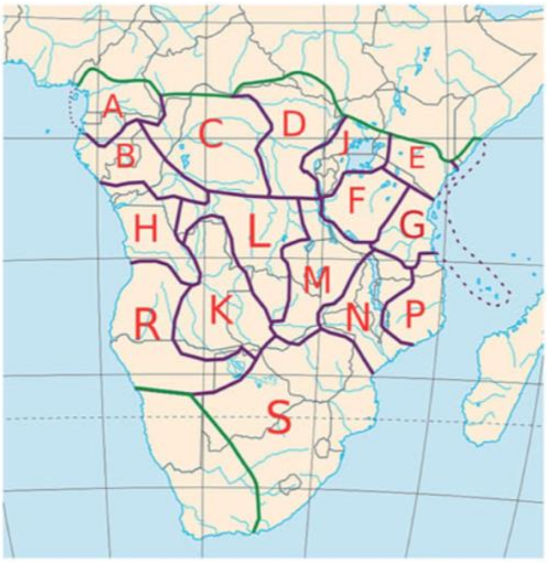


Figure 2.2. Map of sub-Saharan Africa illustrating the different Bantu language subgroups according to the Guthrie classification (image from (Li *et al.*, 2014)).

Archaeological, linguistic, and historical studies associated the Bantu expansion with the transition from hunter-gatherer societies to food producers that allowed populations to accumulate stored food and to increase in size, resulting in the expansion of farming populations at the expense of hunter-gatherer groups (Diamond and Bellwood, 2003; Patin *et al.*, 2009; de Filippo *et al.*, 2012; Barbieri, Vicente, *et al.*, 2014; Grollemund *et al.*, 2015). Therefore, it seems justified to draw a parallel between the processes of farming-related language change in Eurasia and in Africa, although these processes are documented at different moments in time.

Much as is the case for Europe, in principle the diffusion of a subsistence technology, farming and animal husbandry in this case, admits two kinds of explanation (and many intermediate possibilities). One can envisage either a process of cultural transmission, in which the technology was passed along among geographical neighbours with no migration, or a demic diffusion whereby migrants spread in parallel their technology, their language, and their genes. In practice, it is more than likely that both cultural contacts and migration played a role, but for the purposes of the analysis it seems better to compare extreme versions of the cultural and demic models, and test which one better accounts for the data. In the case of Africa, genetic data support the migration of people as the event originating the initial spread of Bantu culture (demic diffusion), instead of an horizontal transmission through culture and culture shift (Tishkoff *et al.*, 2009; Pakendorf *et al.*, 2011; de Filippo *et al.*, 2012; Bostoen *et al.*, 2015). Studies based on Y-chromosomal markers, indeed, found that modern-day Bantu speaking populations are characterized by a low diversity, with the majority of the ethnolinguistic groups showing only two Y chromosome haplogroups: E1b1a7a and E1b1a8 (Wood *et al.*, 2005; de Filippo *et al.*, 2010). The genetic diversity associated with both haplogroups does not decrease with the distance from the assumed homeland (North Cameroon) giving no evidence for a founder effect that one should expect if groups of people moved progressively through sub-Saharan Africa. Nonetheless, the overall genetic homogeneity and the widespread sharing of haplotypes in the Bantu-speaking populations does not support the hypothesis of simple cultural diffusion. Possible explanations are that the signal of an original founder event has been erased by later migrations or that the populations' genetic diversity was affected by drift during the Bantu migration. In addition, this expansion happened quite recently (3.000 – 5.000 yBP), thus possible preventing the accumulation of genetic variation and structure among populations (de Filippo *et al.*, 2010; Pakendorf *et al.*, 2011). On the other hand, Bantu populations show higher levels of mtDNA diversity. This can be interpreted as an indication of several waves of migrations, several ancestral populations, or matrilocality (Salas *et al.*, 2002; de Filippo *et al.*, 2010; Pakendorf *et al.*, 2011).

Studies based on autosomal data give contrasting results, motivating the debate regarding demic diffusion versus language shift (Sikora *et al.*, 2011). A weak point of most of these studies is that instead of focusing on patterns of diversity among Bantu-speaking populations, they focused on the comparison between farming and hunting-gathering populations (Pickrell *et al.*, 2012; Schlebusch *et al.*, 2012). Due to this fact, many aspects of the genetic history of Bantu populations remain unknown. Therefore, the internal structure of the Bantu-speaking populations remained relatively unexplored.

An important question about the expansion of the Bantu populations concerns their dispersal routes that might have been influenced by two major environmental events. Paleoenvironmental data indeed indicate that a crisis affected the central African forest block: a contraction of the Congo rainforest at its periphery about 4.000 yBP, followed by a second contraction, this time affecting the core of the rainforest about 2.500 yBP (figure 2.3) (Bostoen *et al.*, 2015; Grollemund *et al.*, 2015). This second event created patches of open forests and wooded or grassland savannahs, which merged into a corridor known as the "Sangha River Interval". This corridor may have facilitated the north-south migration of the Bantu populations (Maley, 2001; Bostoen *et al.*, 2015; Grollemund *et al.*, 2015).

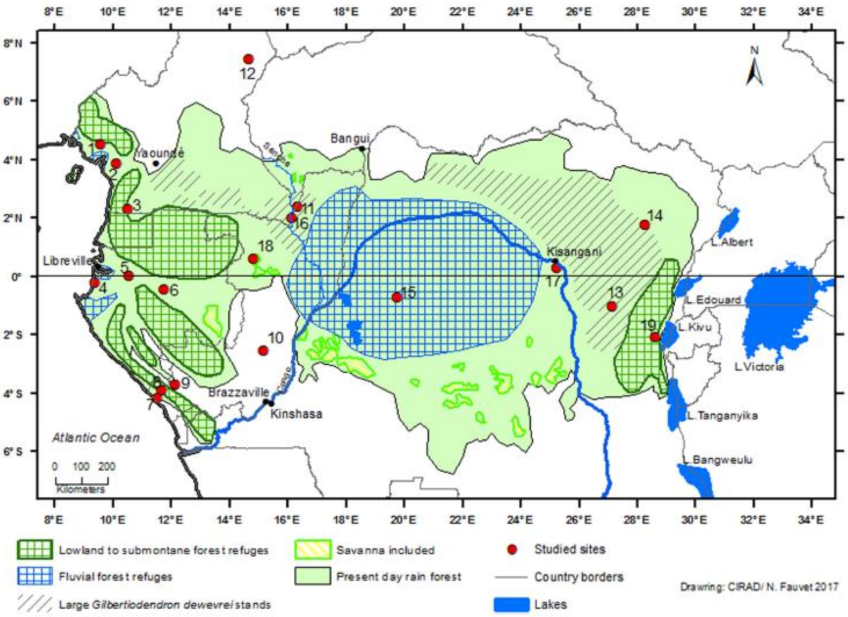


Figure 2.3. Schematic map of rainforest refugia across Central Africa approximately 2.500 – 2.000 yBP in the Equatorial Forest domain, and present-day area of monodominant forest with large *Gilbertiodendron dewevrei* stands (image from (Maley *et al.*, 2017).

2.1.2 Early or Late split hypothesis

The role of the Sangha river corridor is at the core of the debate on the mechanisms of spread of Bantu languages, with uncertainties regarding both their spatial and temporal dynamics. Two main hypotheses have been proposed. The Early split hypothesis (approximately 4.000 yBP), assumes a first split of Bantu farmers into Western and Eastern Bantu populations, at the north of the rainforest; the Later split hypothesis assumes that the Eastern group branches off the Western group (approximately 2.000 yBP) at a later stage, South of the rainforest (figure 2.4) (de Filippo *et al.*, 2012). Both routes are supported by several studies, but recent data showed that eastern and southeastern Bantu speakers are genetically closer to western Bantu speakers from the southern, rather than the northern region, giving greater support for the late-split model (Choudhury *et al.*, 2017; Patin *et al.*, 2017). In addition, archaeological data showed evidence of an arrival and spread of Bantu speakers in Eastern Africa around 2.500 yBP, again in better agreement with the late-split model (Clist, 1987; Ashley, 2010; Neumann *et al.*, 2012; Barbieri, Vicente, *et al.*, 2014). Recent studies have tried to shed light on the modality of the Bantu expansion combining data from different fields, such as genetics, linguistics and archaeology (de Filippo *et al.*, 2012; Grollemund *et al.*, 2015).

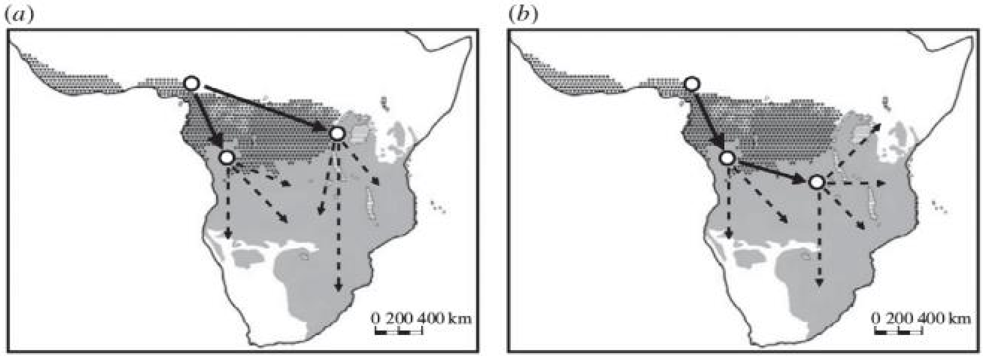


Figure 2.4. The two main models: (a) Early-split and (b) Late-split. In (a) the model assumes a first split of Bantu farmers into Western and Eastern Bantu populations, at the north of the rainforest; in (b) the model assumes that the Eastern group branches off the Western group at a later stage, south of the rainforest (image from (Pakendorf *et al.*, 2011)).

In 2012, de Filippo *et al.* used for the first-time genetic data together with linguistic data (lexical data) to study the two hypotheses and to test which one would provide a better fit for the observed pattern of genetic and linguistic variation. They also included a third model, a simple isolation by distance (IBD) model, to test whether the data depart enough from neutral expectations to justify more complex models (de Filippo *et al.*, 2012). They first observed that geographically distant Bantu-speaking population are closely related with each other, supporting the role of demic diffusion in the spread of Bantu languages. In addition, they found support for the late-split model, both in lexical and genetic data, although they could not reject the simple IBD model.

Grollemund *et al.* (2015) reconstructed a time-calibrated phylogenetic tree of Bantu languages to study the dynamics of expansion of Bantu speakers (figure 2.5). The generated tree has a clear comb-like structure, with the Eastern branch at the lower end of the comb, thus supporting the late-split hypothesis (Grollemund *et al.*, 2015).

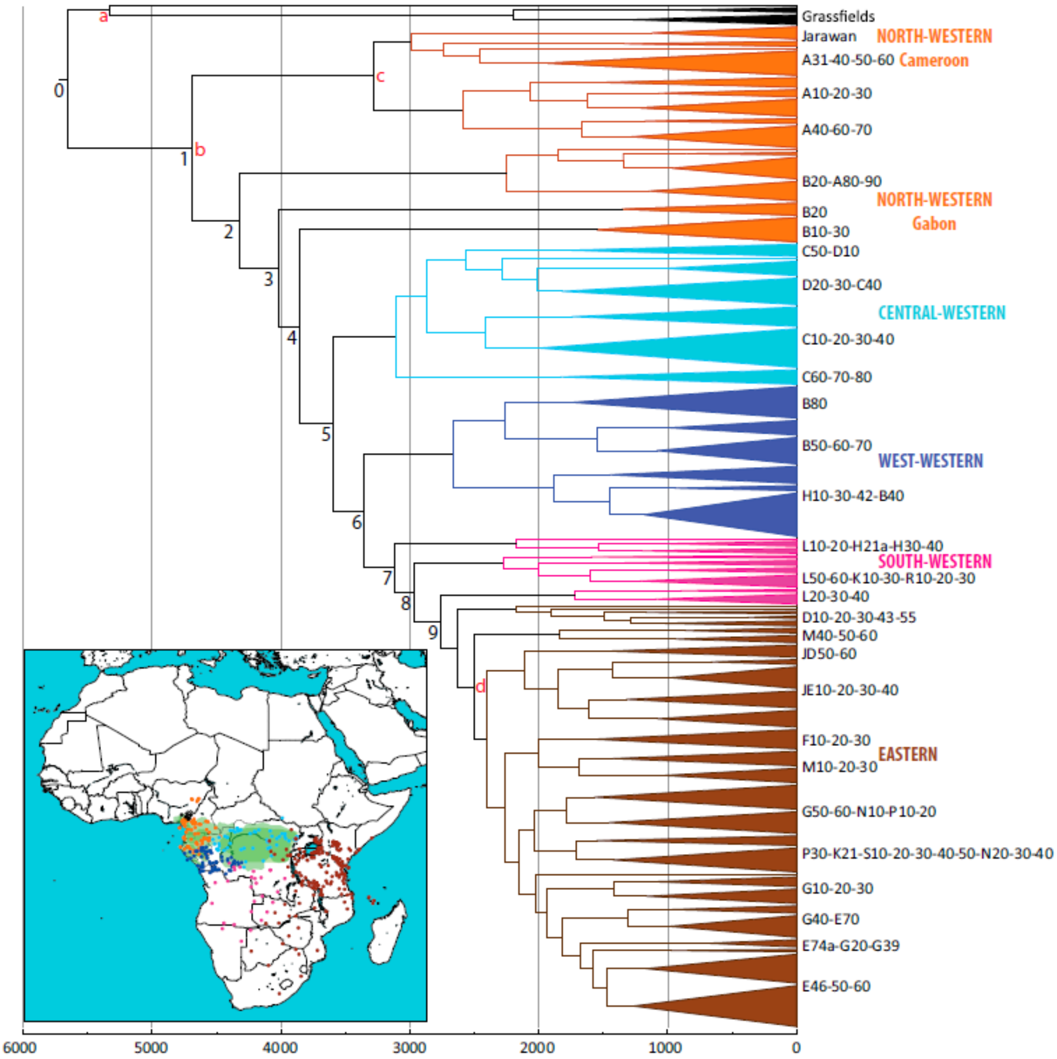


Figure 2.5. Consensus time tree of 424 Bantu languages, derived from 100 trees drawn from the Bayesian posterior distribution. Triangles are proportional to the number of languages in the group, and the labels are the codes used by Guthrie (Guthrie, 1962). The four calibrations used are identified by red letters (a) 5.000 yBP or older; b) 4.000 – 5.000 yBP; c) 3.000 – 3.500 yBP; and d) 2.500 yBP). (Inset) Map of Africa with coloured dots to represent the current location of the languages (image from (Grollemund *et al.*, 2015)).

One limitation of both cited works (Grollemund *et al.* (2015) and de Filippo *et al.* (2012)) is that their analyses focused on the reconstruction of the processes at the genetic level, without explicitly taking into account the dynamics of linguistic evolution. In addition, the models are not formally compared in these studies. As a consequence, the interpretation of the data and the choice of the best-fitting model are necessarily somewhat arbitrary (e.g. see de Filippo *et al.* (2012)). Also, when linguistic relationships are built based on phylogenetic trees, no level of horizontal transmission is considered, a rather strong assumption which seems in contradiction with the basic mechanism leading to shape the linguistic variability, in the Bantu families as well as in any set of languages. For example, the pattern found by Grollemund *et al.* (2015) where Eastern Bantu languages are closer to South-Western and West-Western Bantu (see figure 2.5) could have arisen not because they share a recent common ancestor, but because they were in relatively close geographic contact during the last stages of their history, as also suggested by the work by de Filippo *et al.* (2012). In population genetics, it is possible to explicitly account for different evolutionary histories comparing alternative demographic models (a representation of the evolutionary process that could have generated the data) by simulation of genetic data. It is possible to simulate the expected genetic variation under different evolutionary histories, thus explicitly estimating which scenario is able to generate a level of variation that is comparable with those observed in real populations. Several frameworks have been developed to estimate the most probable evolutionary model from genetic data, and a very powerful and flexible approach is that based on Approximate Bayesian Computation (ABC) methods. Extending the analysis to other variables, this method can allow the exploration of evolutionary histories, and the comparison of demographic models, with a high degree of resolution.

2.1.3 Approximate Bayesian Computation framework: integrating genomic and linguistic data

Methods for testing demographic/evolutionary hypotheses require the application of genealogical methods, either proceeding from the present towards the past (coalescent methods) or vice versa, from the past to the present (forward simulations). Genetic relationships between individuals can be described by their genealogies, the trajectories across which genes are transmitted through time. In the same way we can build genealogies of individuals, we can also build genealogies of genes, assuming that each gene is transmitted stochastically through a mechanism either of two alleles is passed on to the offspring.

Genealogies can be reconstructed from the past to the present, or from the present to the past. An advantage of the latter approach is that, by proceeding backwards, one must only take into account the individuals sampled and their ancestors. On the contrary, simulations proceeding forward must specify information for all members of a population, the vast majority of whom are irrelevant for testing purposes, since neither they nor their descendants are part of the sample. The backwards simulation process basically requires two steps: in the first one a complete genealogy is reconstructed, proceeding in time from the individuals sampled to their common ancestor (see below for details). In the second step, mutations are added to the tree according to some predefined model, to reproduce genetic change through time. Because one such tree is just one realization of a generally complex stochastic process, millions of trees are currently generated, thus exploring a broad range of possibilities. Note that, because of the stochasticity of both the genealogical and the mutational processes, there is no reason to expect that the genealogies of different genes will be the same.

Schematically, in each genealogy there are *n* external branches, one for each gene sampled in an individual. Going backwards in time, from the present to the past, whenever two lineages pick the same parent, their lineages coalesce, a process called a coalescence event (figure 2.6). The rate at which lineages coalesce depends on how many lineages are picking their parents and on the size of the population. The probability of coalescence is a direct function of the sample size (because a coalescence event is more likely when many lineages are present then when there are just a few) and an inverse function of the population size (because a coalescence event is less likely when two individuals can be descended from many than from a few potential parents). Each genealogy has *n*-1 coalescence events that end when it is reached the most recent common ancestor (MRCA) of all genes in the sample. Genealogies contain the information about the past population history. We can use the distribution of the time to the MRCA between two genes to infer the population size change over time. If we are studying a large population, the coalescent rate will be small, since it will take more time to find the most recent common ancestor; in contrast, in the case of a small population the coalescent rate will be bigger (Wakeley, 1999; Harpending and Rogers, 2000; Goldstein and Chikhi, 2002; Charlesworth *et al.*, 2003; Harding and McVean, 2004).

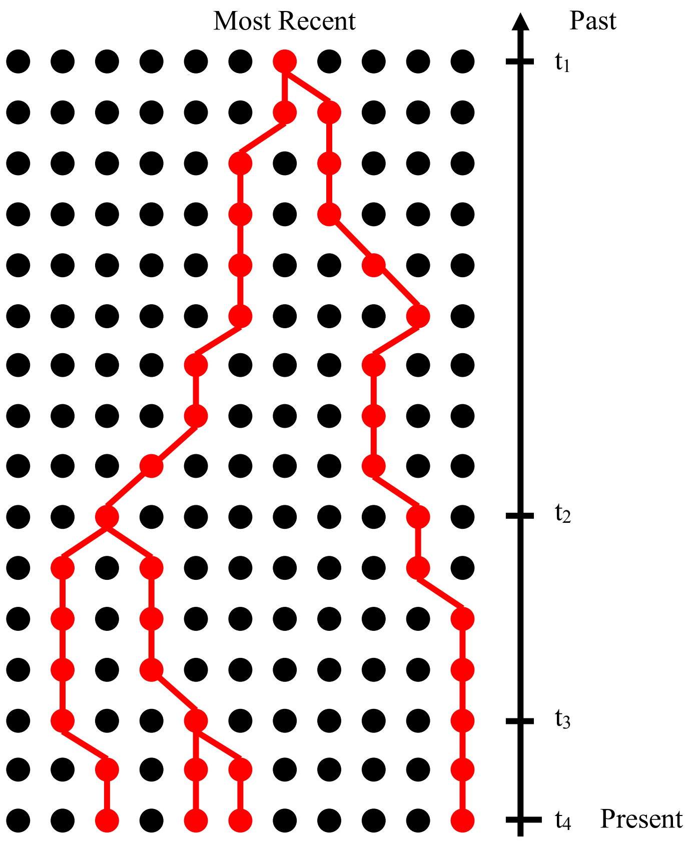
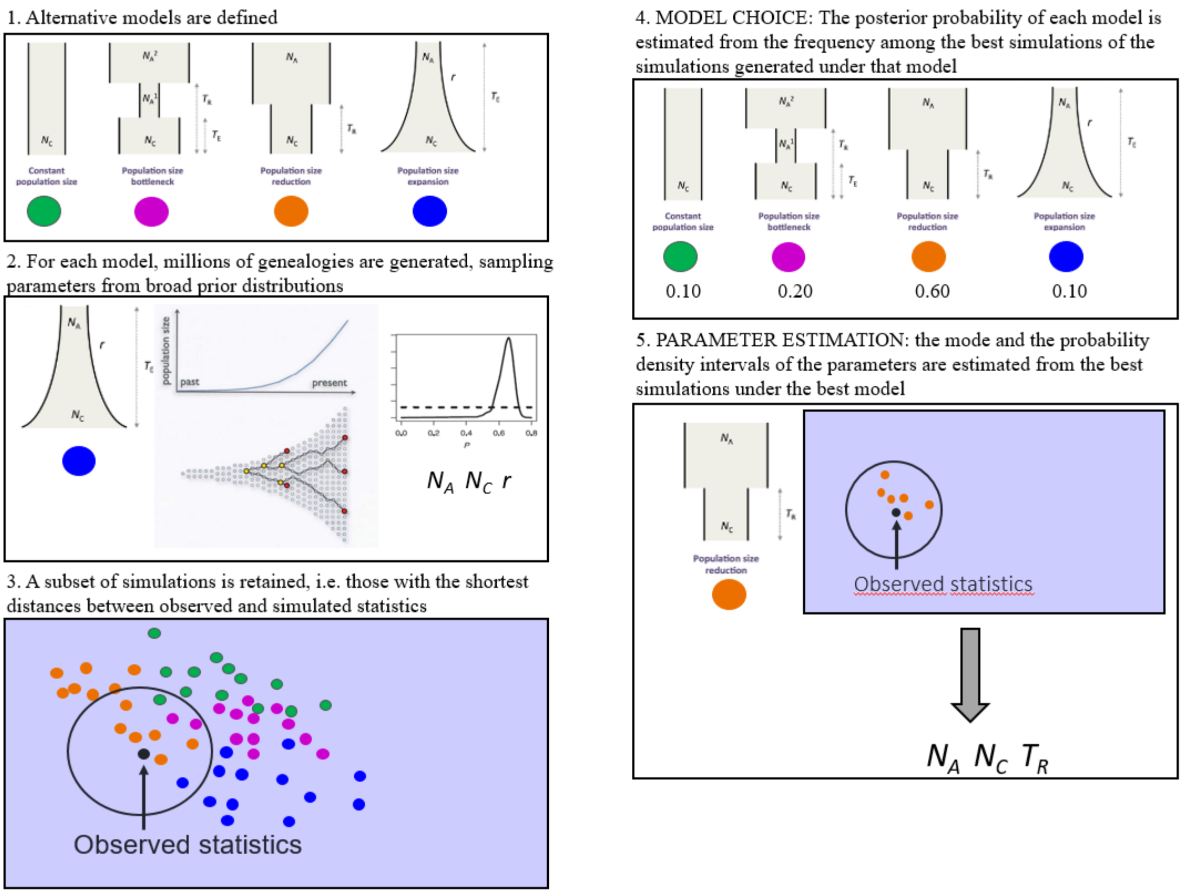


Figure 2.6. Genealogy of a sample. In the present (t_4) there are *n* external branches, one for each gene sampled in an individual. Going backwards in time, from the present to the past, whenever two lineages (here in red) pick the same parent, their lineages coalesce, a process called a coalescence event.

The coalescent process was formally described in mathematical form in 1982 by John Kingman (Kingman, 1982a, 1982b). Since then, the coalescent theory allowed a significant reduction in the computational costs of simulations and it allowed new insights in our understanding of the shapes of gene genealogies obtained from real species. In addition, it allowed the development of new methods to infer the demographic history of populations. Generating by simulation multiple datasets, one can compare them with models attempting to describe the process generating the observed genetic diversity. Each model will be defined by a number of parameters, such as population size, fluctuations of population sizes, timing of such fluctuations, mutations rates, etc.

ABC methods are a powerful and flexible way to quantitatively compare alternative models and perform model parameters estimation (figure 2.7) (Bertorelle *et al.*, 2010; Csilléry *et al.*, 2010). These methods use coalescent simulations across a vast range of parameter values within a demographic model to find the parameter values that match most closely those in the observed data. Likelihood functions do not need to be specified allowing to handle larger datasets under highly complex models, relying on simulations in which prior information is incorporated (Csilléry *et al.*, 2010; Schiffels and Durbin, 2014). Different demographic models can be simulated and tested against each other with respect to the observed data. Moreover, this method can be used to estimate the posterior distributions of the demographic parameters; once a model has been identified which best accounts for the observed data, the researcher can explore the parameter space leading to the simulations showing the closest match with the data. For that purpose, the datasets (simulated and observed) are summarized by the same set of Summary Statistics (SuSt), which are informative about the genealogical processes under investigation. The statistics chosen need to be efficient and sufficient to quantify the difference between the observed and simulated datasets (Bertorelle *et al.*, 2010; Estoup *et al.*, 2018). Each simulation is thus summarized by the set of SuSt chosen and then one can find the set of parameters that led to SuSt from simulations closest to the real data's SuSt.

ABC has proved a powerful statistical method, but it presents some limitations. The first limitation is the requirement of millions of simulations of samples of the same size as those observed, which is computationally expensive when analysing complete genomes or when the demographic models are complex. The second limitation is the choice of the SuSt describing both the observed and simulated data. The SuSt should be by definition sufficient, meaning that, in principle, they should summarize the data with minimum loss of information and without being redundant. There are methods for efficiently choosing the most suitable statistics to summarize the data, but they are computationally intensive (Joyce and Marjoram, 2008). In addition, to obtain statistically sound and reliable results, the amount of data needed to support the result grows exponentially with the dimensionality, i.e. the number of parameters of the model. All multivariate analyses point at identifying areas where objects form groups with similar properties; in high dimensional data, however, objects tend to be sparse and dissimilar in many ways, often preventing common data analysis strategies from being efficient. This problem is referred as the "curse of dimensionality" (Blum and François, 2010).



1. Alternative models are defined

2. For each model, millions of genealogies are generated, sampling parameters from broad prior distributions

3. A subset of simulations is retained, i.e. those with the shortest distances between observed and simulated statistics

4. MODEL CHOICE: The posterior probability of each model is estimated from the frequency among the best simulations of the simulations generated under that model

5. PARAMETER ESTIMATION: the mode and the probability density intervals of the parameters are estimated from the best simulations under the best model

Figure 2.7. ABC framework. Here the five main steps of the ABC method are outlined: (1) different demographic models are defined; (2) each demographic model previously defined is simulated millions of times; (3) the millions of simulations performed for each demographic model are tested against the observed data, and a subset of the simulations is retained (i.e. those simulations with the shortest distance between observed and simulated statistics); (4) the best-performing model is chosen (the posterior probability of each model is estimated from the frequency among the best simulations of the simulations generated under that model); (5) and lastly, once the model best accounting for the observed data has been identified, the researcher can explore the parameter space leading to the simulations showing the closest match with the data.

In 2015, a new ABC framework was developed; its features allow one to address most of the aforementioned problems. This framework is based on a machine-learning tool called Random Forests (ABC-RF) that constructs a classifier (a discrete-valued function used to assign categorical class labels to particular data points) from simulations from the prior distribution. Once the classifier is constructed and applied to the observed data, a secondary RF can produce an approximation of the posterior probability of the resulting model, which regresses the selection error over the statistics selected to summarize the data (Pudlo *et al.*, 2015). By this framework it is possible to simulate thousands instead of millions of realizations of the demographic models, and to select a set of SuSt, small but highly informative on the parameters defining the model (Pudlo *et al.*, 2015). These characteristics make the ABC-RF algorithm of particular interest for the statistical analysis of large amounts of genomic data.

From this perspective, the unfolded Site Frequency Spectrum (SFS) should be a proper statistic to summarize genomic data (Terhorst *et al.*, 2016; Lapierre *et al.*, 2017; Smith *et al.*, 2017). The SFS reports the number of derived alleles, over a certain number of genomic regions, segregating at different frequencies in a sample of n individuals. In these analysis, neutrality, i.e. no effect of natural selection, is generally assumed. The shape of the SFS is affected by the demographic history of the population under study. As has long been known, an expanding population carries an excess of low-frequency variants compared with the expectation under a constant size, and even more so compared with a shrinking population (Tajima, 1989). Specific demographic models can thus be tested based on the comparison of the observed SFS and the SFS estimated directly through coalescent simulations. However, we need to know the ancestral state of a SNP to construct the unfolded frequency spectrum. An uncertainty in the identification of the true ancestral state can cause a bias in the reconstruction of the spectrum and consequently in the inference of the demographic history (Hernandez *et al.*, 2007; Keightley and Jackson, 2018). In these cases, the folded version of the SFS (which takes into account the frequency of the minor allele) should be used, with inevitable loss of information (Keightley and Jackson, 2018). In addition, the SFS reliability should increase as the number of individuals analysed increase, since it is based on allele frequencies. However, the number of individuals available per population can be a limiting factor, as in the case of ancient data.

A recent work by Ghirotto *et al*. (2019) tested the power of ABC-RF to select the model with highest posterior probability using complete unphased genomes (Ghirotto *et al.*, 2019). The data are summarized by the genomic distribution of the four mutually exclusive categories of segregating sites (Frequency Distribution of Segregating Sites (*FDSS*) – figure 2.8) in two populations: (1) private polymorphisms in the first population; (2) private polymorphisms in the second population; (3) polymorphisms shared between populations; and (4) polymorphisms fixed for different alleles. This set of SuSt can be calculated from unphased genomes, do not need the information about the ancestral state of alleles, and are known to be informative about past evolutionary processes (Wakeley and Hey, 1997; Ghirotto *et al.*, 2019). This approach revealed that the ABC-RF coupled with the *FDSS* can indeed distinguish between different demographic histories, including cases in which when few individuals are sampled. This is an advantage compared with the other approaches previously mentioned, as we now can use the whole genome of one diploid individual per population to draw inferences about past demographic history (Ghirotto *et al.*, 2019). ABC methods have been successfully applied to the analysis of past population dynamics using genetics and genomic data. Recently, ABC methods have also been applied to the analysis of linguistic and genetic data (Thouzeau *et al.*, 2017). However, the real challenge in this field would be to be able to integrate in the framework the information coming from whole-genomes and linguistic data, so as to consider at the same time the biological and cultural factors that the studied evolutionary dynamics may have affected.

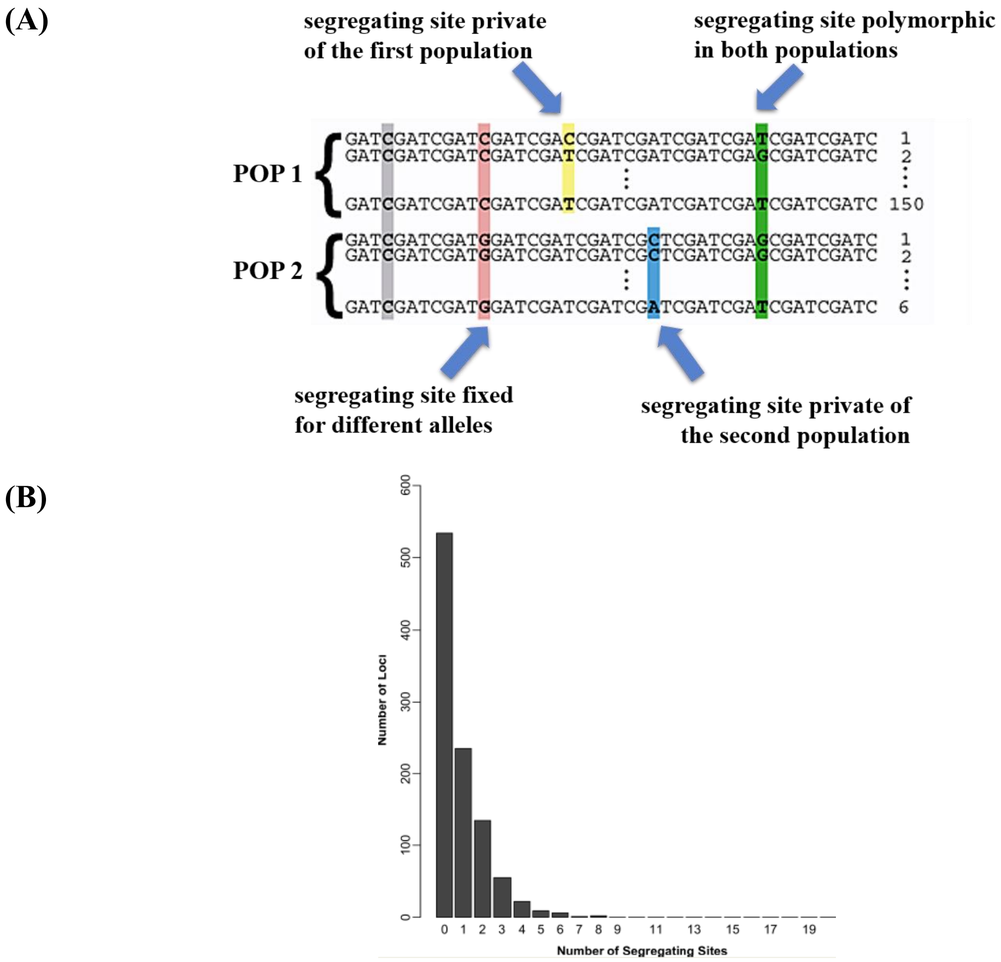
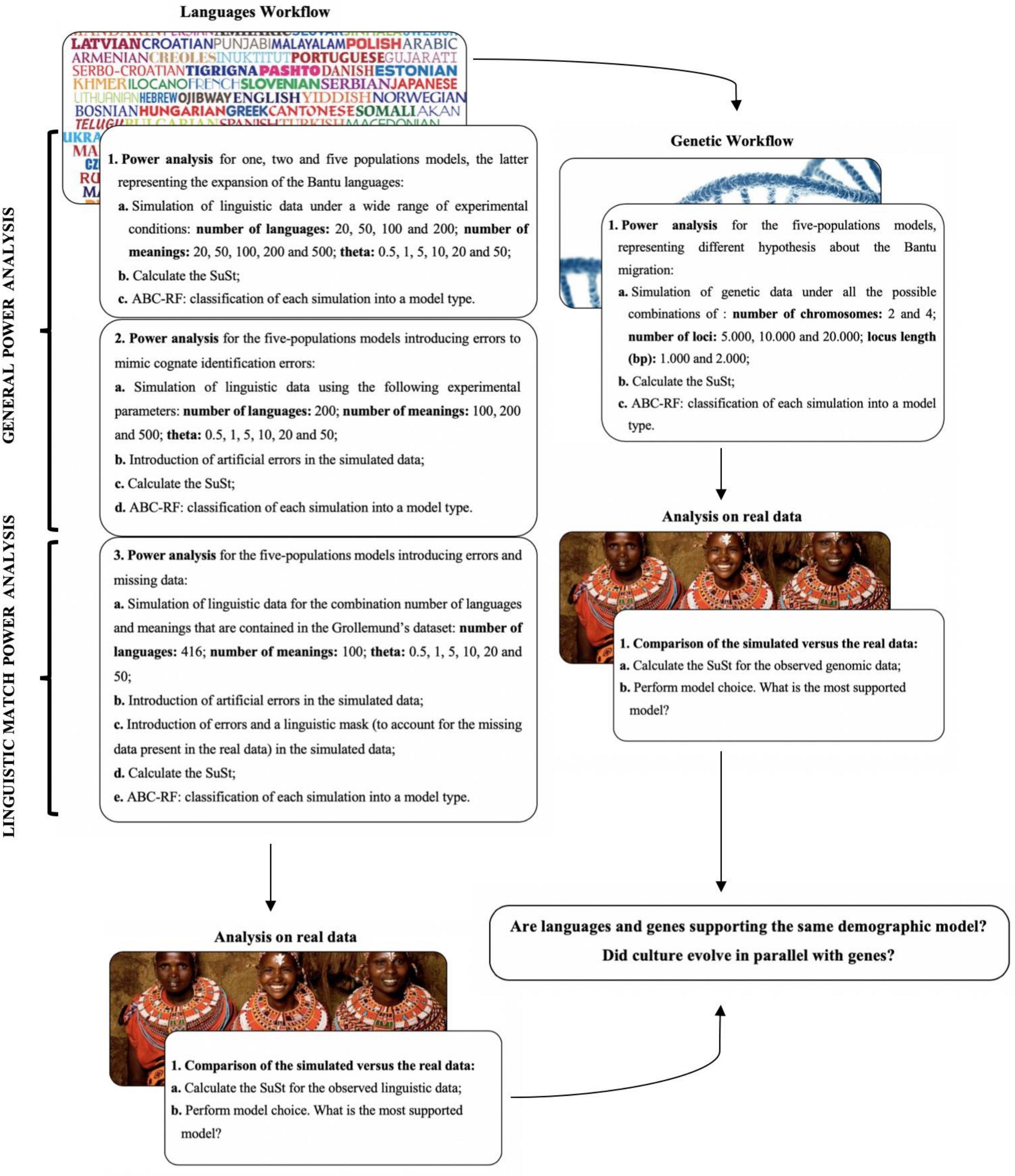

(A)
segregating site private of the first population
segregating site polymorphic in both populations
POP 1
GATCGATCGATCGATCGACCGATCGATCGATCGATCGATCGATC 1
GATCGATCGATCGATCGATCGATCGATCGATCGAGCGATCGATC 2
GATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATC 150
POP 2
GATCGATCGATGGATCGATCGATCGCTCGATCGAGCGATCGATC 1
GATCGATCGATGGATCGATCGATCGCTCGATCGAGCGATCGATC 2
GATCGATCGATGGATCGATCGATCGATCGATCGATCGATCGATC 6
segregating site fixed for different alleles
segregating site private of the second population
(B)
Number of Loci
0
100
200
300
400
500
600
Number of Segregating Sites
0 1 2 3 4 5 6 7 8 9 11 13 15 17 19

Figure 2.8. SuSt Genetics: (A) the four mutually exclusive categories of segregating sites; (B) Frequency Distribution of Segregating Sites (*FDSS*). In the x-axis it is the number of segregating sites, and in the y-axis the number of loci.

2.2 Aim

In this project, we propose the new ABC framework described above (figure 2.9) based on the analysis of linguistic data to perform explicit comparisons of demographic models. This framework would allow one to explore past demographic processes analysing simultaneously genomic and linguistic data, so as to make inference about cultural and biological evolution of populations. To this aim, we settled a collaboration with the Center for Advanced Studies "Words, Bones, Genes, Tools" at the University of Tübingen that developed the linguistic evolutionary models, which we provided to integrate in the classical ABC framework for model comparisons.

As a first step, we performed a power analysis, so as to understand whether, and to what extent, the proposed framework is actually able to correctly identify the true demographic history, using simulated linguistic data. Once assessed the power of the procedure, we applied this new ABC framework to the analysis and comparison of Bantu-expansion models, exploiting both linguistic and genetic variables. We tested both the Early and Late-Split hypothesis using for the first time whole-genome data from Bantu-speaking individuals, together with linguistic data from Grollemund *et al.*'s (2015) lexical dataset, including 416 Bantu languages. With these two extended datasets, combined with the power produced by the present ABC method, we expect to reveal details of the past history of Bantu population with an unprecedented definition.


Languages Workflow
LATVIAN CROATIAN PUNJABI MALAYALAM POLISH ARABIC
ARMENIAN CREOLES INUKTITUT PORTUGUESE GUJARATI
SERBO-CROATIAN TIGRIGNA PASHTO DANISH ESTONIAN
KHMER ILOCANO FRENCH SLOVENIAN SERBIAN JAPANESE
LITHUANIAN HEBREW OJIBWAY ENGLISH YIDDISH NORWEGIAN
BOSNIAN HUNGARIAN GREEK CANTONESE SOMALI AKAN
TELUGU BULGARIAN SPANISH TURKISH MACEDONIAN
GENERAL POWER ANALYSIS
LINGUISTIC MATCH POWER ANALYSIS
Genetic Workflow
Analysis on real data
Analysis on real data

1. Power analysis for one, two and five populations models, the latter representing the expansion of the Bantu languages:

a. Simulation of linguistic data under a wide range of experimental conditions: **number of languages:** 20, 50, 100 and 200; **number of meanings:** 20, 50, 100, 200 and 500; **theta:** 0.5, 1, 5, 10, 20 and 50;

b. Calculate the SuSt;

c. ABC-RF: classification of each simulation into a model type.

2. Power analysis for the five-populations models introducing errors to mimic cognate identification errors:

a. Simulation of linguistic data using the following experimental parameters: **number of languages:** 200; **number of meanings:** 100, 200 and 500; **theta:** 0.5, 1, 5, 10, 20 and 50;

b. Introduction of artificial errors in the simulated data;

c. Calculate the SuSt;

d. ABC-RF: classification of each simulation into a model type.

3. Power analysis for the five-populations models introducing errors and missing data:

a. Simulation of linguistic data for the combination number of languages and meanings that are contained in the Grollemund's dataset: **number of languages:** 416; **number of meanings:** 100; **theta:** 0.5, 1, 5, 10, 20 and 50;

b. Introduction of artificial errors in the simulated data;

c. Introduction of errors and a linguistic mask (to account for the missing data present in the real data) in the simulated data;

d. Calculate the SuSt;

e. ABC-RF: classification of each simulation into a model type.

1. Power analysis for the five-populations models, representing different hypothesis about the Bantu migration:

a. Simulation of genetic data under all the possible combinations of : **number of chromosomes:** 2 and 4; **number of loci:** 5.000, 10.000 and 20.000; **locus length (bp):** 1.000 and 2.000;

b. Calculate the SuSt;

c. ABC-RF: classification of each simulation into a model type.

1. Comparison of the simulated versus the real data:

a. Calculate the SuSt for the observed genomic data;

b. Perform model choice. What is the most supported model?

Are languages and genes supporting the same demographic model?

Did culture evolve in parallel with genes?

1. Comparison of the simulated versus the real data:

a. Calculate the SuSt for the observed linguistic data;

b. Perform model choice. What is the most supported model?

Figure 2.9. ABC workflow used for the study of languages and genes. In the left side of the image it is represented the workflow for the linguistic analysis and in the right side the workflow for the genetic analysis.

2.3 Methods

2.3.1 Power analysis

In the first phase of this project we assessed the power of this novel ABC framework to discriminate between simple evolutionary models using simulated data. For the simulated genetic data, the method was previously validated under a broad spectrum of experimental conditions (see Ghirotto *et al*., 2019), therefore we only performed the validation for the five-population models that mimics our Bantu expansion model. We did so for simulated linguistic data under a wide range of experimental conditions. Languages, like genes, can have different genealogies. Thus, in our linguistic simulations, we sampled languages (instead of genes) which may, or may not, have a phylogenetic relationship (cognate words). The idea is that each language is considered as a haploid individual and each meaning is a locus (figure 2.10A). For example: at the locus (i.e. meaning) for quick salutation formulae, the Italian language is a haploid individual and the word *ciao* the allele at that locus; the English language is a second haploid individual and the word *hello* another allele; etc. Words that have the same common ancestor (meaning that they are cognates) are classified as carrying the same allele. Moreover, unlike for genetic data, when a mutation creates a new segregating site at the locus, the entire allele changes. As a consequence, the mutation model is an infinite-allele model. We also set the recombination rate equal to zero.

We summarized the linguistic data by the mean, variance, skewness and kurtosis for each population and for the number of cognate sets shared between each pair of populations (figure 2.10B). To determine the power of the set of SuSt chosen in distinguishing among alternative evolutionary models, we simulated linguistic data considering different experimental conditions. We tested all the possible combinations of number of languages {20; 50; 100; 200}, number of meanings {20; 50; 100; 200; 500}, and different scaled mutations rates to account for the uncertainty linked to the linguistic mutation rate (theta) {0.5; 1; 5; 10; 20; 50}, for a total of 120 combinations of sampling conditions tested.

All the combinations of parameters just mentioned were used to generate datasets of simulated linguistic variability according to three sets of non-nested models of increasing complexity, namely, one-population models (six alternative models, figure 2.12), two-population models (four alternative models, figure 2.13) and five-population models (six alternative models, figure 2.14).

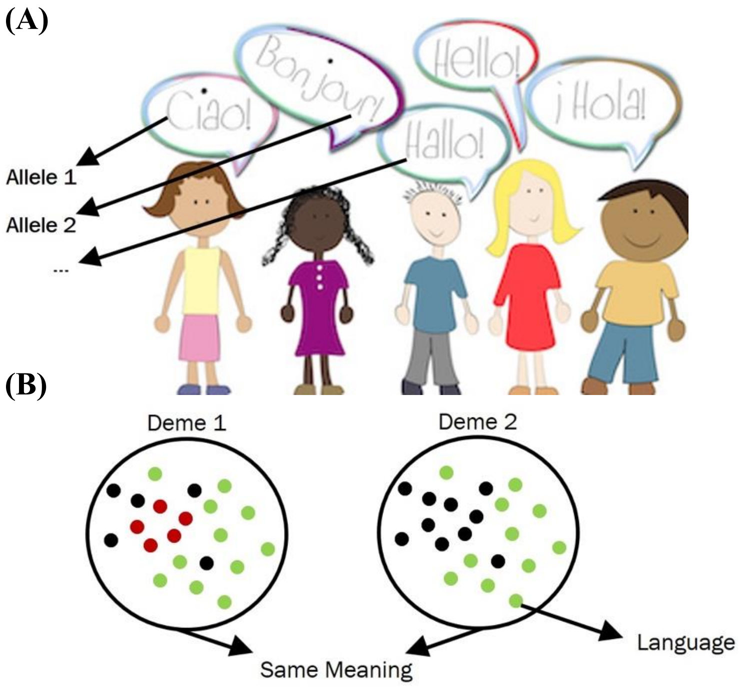


Figure 2.10. Genes vs Languages: (A) Language = one haploid individual; Meaning = Locus; (B) SuSt Languages – representation of one meaning analysed in two different demes, each of them containing several languages. We count the number of languages in each deme and shared by each deme. Each dot represent a language and equal colours represent the same language.

From the 50.000 simulated datasets, 1.000 pseudo-observed dataset (*pods* – figure 2.11) were selected for each model and combination of linguistic parameters. These *pods* were treated as if they were observed datasets, in a classical ABC analysis. The simulated data were then compared with each of the 1000 *pods*, thus calculating the rate of true positives, i.e., how many times each *pods* was correctly assigned to the demographic model that generated it. This rate represents the power to accurately discriminate among demographic models by means of the proposed ABC-RF framework. This way we could ask: (1) whether the set of SuSt chosen are sufficient to discriminate the different models and (2) which is the best combination of experimental conditions yielding the highest power of discrimination among models (number of languages, meanings and linguistic mutation rate).

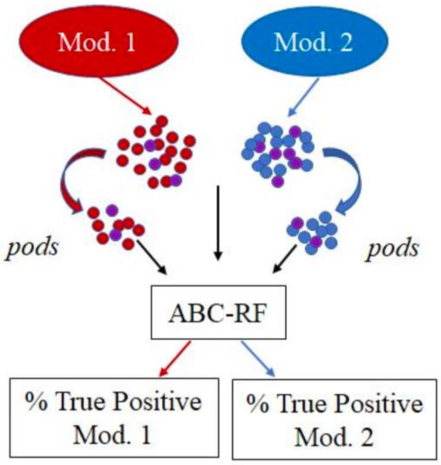


Figure 2.11. ABC-RF procedure. (1) 50.000 simulations of model 1 and 2. (2) Selection of 1.000 simulations as pseudo-observed datasets (pods) that are (3) analysed as the observed dataset from a classical ABC analysis. (4) The simulated data was then compared with each of the 1000 pods allowing to calculate the rate of true positives.

2.3.2 Demographic models

2.3.2.1 One-population models

We designed six one-population models (figure 2.12): (1) a population keeping constant size through time, (2) a bottleneck, (3) an exponential growth and (4) a structured population. For the structured population, we defined different migration rates: (4a) weak migration (sampling languages only in the first deme), (4b) strong migration (sampling languages only in the first deme), (4c) and strong migration (sampling languages randomly in all demes).

The demographic parameters associated with the models and their prior distributions are in table 2.1.

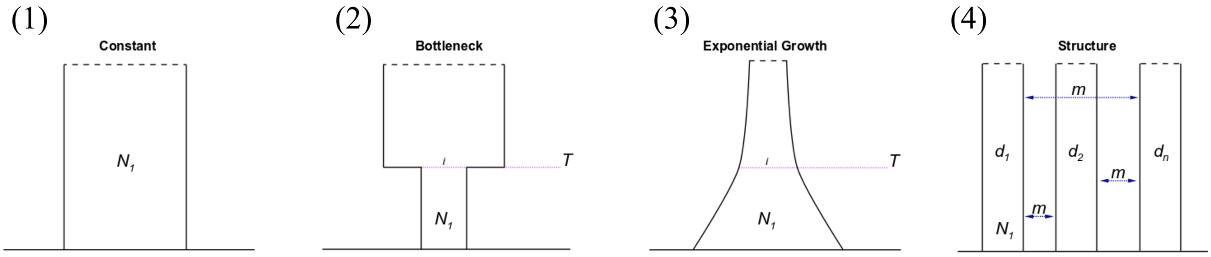


Figure 2.12. Demographic models: One-population models. (1) a population keeping constant size through time, (2) a bottleneck, (3) an exponential growth and (4) a structured population.

Table 2.1. Demographic parameters and prior distributions used for the linguistic simulations.

Demographic parameters	**Prior distribution**
Effective population size (N_0)	Uniform {500 - 50.000}
Bottleneck intensity	Uniform {10 - 100}
Exponential growth intensity	Uniform {10 - 100}
Time of the bottleneck (in *ms* time units)	Uniform {0.0001 - 0.05}
Time of the exponential growth (in *ms* time units)	Uniform {0.01 - 5}
Deme number	Uniform {2 - 10}
Migration rate (strong)	0.01
Migration rate (weak)	0.0001

2.3.2.2 Two-population models

We designed four alternative two-population models (figure 2.13): (1) a divergence model with isolation after the divergence (no gene flow), (2) a divergence model with continuous migration from the split until present times, and (3) a divergence model with admixture (a single event of bidirectional migration). In model (2) we defined two constant levels of migration rates: (2a) weak and (2b) strong migration.

In table 2.2 it is described the demographic parameters associated to the models and their prior distributions.

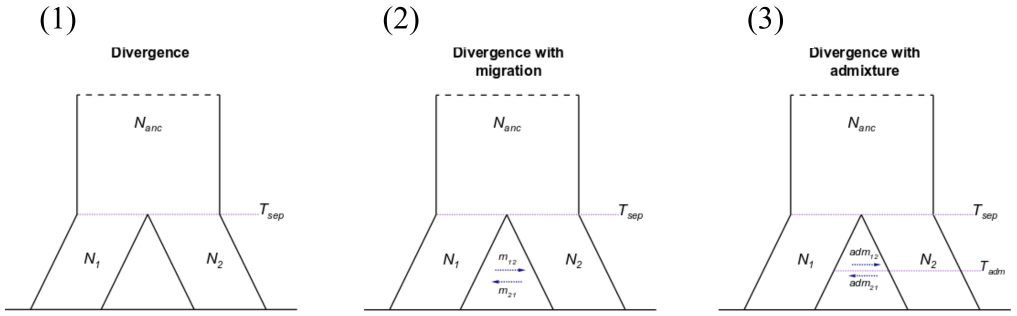


Figure 2.13. Demographic models: Two-population models. (1) a divergence model with isolation after the divergence (no gene flow), (2) a divergence model with continuous migration from the split until present times, and (3) a divergence model with admixture (a single event of bidirectional migration)

Table 2.2. Demographic parameters and prior distributions used for the linguistic simulations.

Demographic parameters	**Prior distribution**
Effective population size (N_0)	Uniform {500 - 50.000}
Split time (in *ms* time units)	Uniform {0.01 - 5}
Migration rate (strong)	0.01
Migration rate (weak)	0.0001
Admixture time (in *ms* time units)	Uniform {0 - time of the split event}
Admixture rate	Exponential {0.01 - 0.5}
Deme number	2

2.3.2.3 Five-population models - Languages

We then moved to the comparison of more realistic, and hence complex, demographic histories. To this end, we designed three main models (Early vs Late vs Very Late split), each of which was divided in two subgroups (river vs not river), for a total of six models (figure 2.14), representing some of the hypotheses proposed to explain the Bantu expansion. All the models consider five populations, each representing one of the five geographical groups of Bantu languages: North-Western (NW), West-Western (WW), Central-Western (CW), South-Western (SW) and Eastern (E). This division is based on the work by Nurse and Philippson (2003), which primarily uses gramattical, rather than lexical, features. Since our linguistic dataset is based on lexical data, using the Nurse and Philippson (2003) division we avoid circularity. In these models, we allowed for migration (or borrowing at the linguistic level) to take into account horizontal transmission of linguistic features. We designed models representing the two major hypotheses about the dynamics of the Bantu expansion: Early and Late split. We also added a third hypothesis, the Very Late split. The main difference among models is, under Early split, the E Bantu population originates from NW, the homeland group of Bantu; under Late split, the E Bantu originate from WW; and lastly, under the Very Late split, the E Bantu population stems from the SW groups. In addition, for each model (Early, Late and Very Late split) we added two different scenarios: river and not river. In the river scenario, the CW Bantu group penetrates the rainforest along Congo from the WW zone, while in the not river scenario the spread into the forests in the Congo starts from the NW zone.

The demographic parameters associated with the models and their prior distributions are in table 2.3.

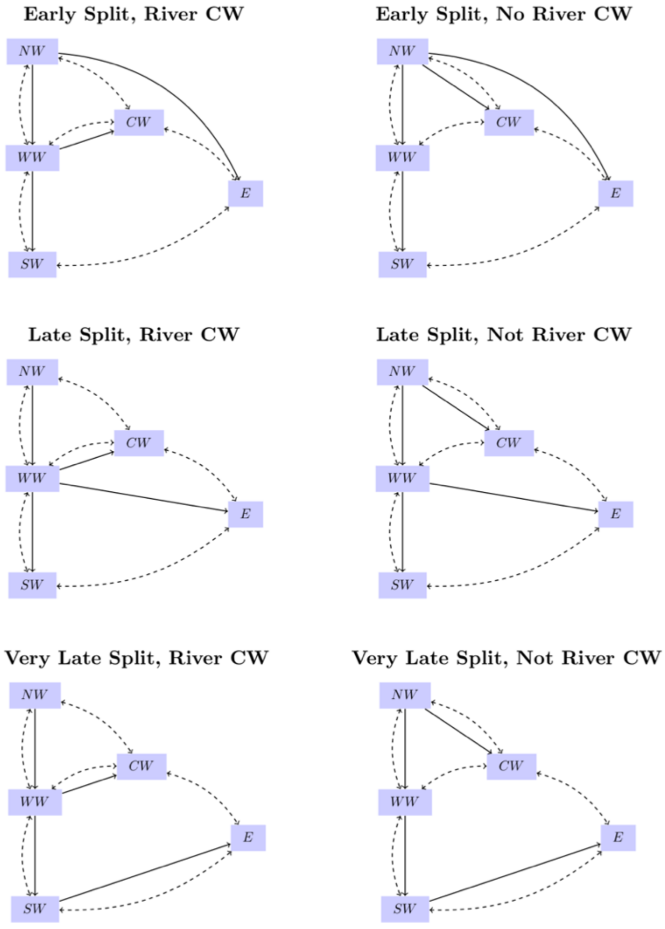


Figure 2.14. Demographic models: Five-population models. Arrows represent the population history graph; if there is an edge NW → WW that means that zone WW was empty until it was colonized by a group of migrants from NW. Dashed lines represent migrations links between populations.

Table 2.3. Demographic parameters and prior distributions used for the linguistic simulations.

Demographic parameters	**Prior distribution**
Effective population size (N_0)	Uniform {500 - 50.000}
Time of the events (in *ms* time units)	Uniform {0.01 - 5}
Base rate of migration (equall in all demes)	Uniform {10^{-10} - 10^{-2}}
Exponential growth intensity	100
Deme number	5
Current population share	Exponential {0.05 - 0.25}

2.3.2.3.1 Introducing artificial errors into linguistic simulated data to mimic cognate identification errors

The linguistic variables used for this study are sets of cognate data (i.e. words that share a recognizable common ancestor) (Grollemund *et al.*, 2015). Cognancy, the relationship between cognates, is usually estimated by linguists through the identification of phylogenetic relationships between words expressing the same meaning. This is a difficult process and the results may contain two kind of mistakes. First, linguistics may fail to recognize that two words share the same ancestor. For instance, English *death* and German *Tod*, meaning 'death' do not share any sounds in common, and yet they are actually descended from the same ancestral word, as established by careful historical-linguistic studies. Indeed, there are regular correspondences between English and German sounds: for example, English initial [d-] corresponds to German [t-] not only in *death-Tod*, but also in *daughter-Tochter* 'daughter', in *deep-tief* 'deep', *deer-Tier* 'animal' (in Old English, the ancestor of modern *deer* word also meant '(wild) animal'), *dew-Tau* 'dew', and others. Linguists making cognate judgements for understudied languages attempt to infer regular correspondences, but the less data, the less successful this is bound to be, and the probability of introducing errors in lexical cognacy datasets increases. Another category of mistakes regards the wrong assignment of a phylogenetic relationship between words that actually derive from two separate ancestors. For example, one might think that English *dry* and German *dürr* 'dry, withered' stem from the same ancestral word, since they sound similar, and this may appear to be too similar to arise by chance. But actually, these words are not related; German has another word inheriting the same root as English *dry*: it is *trocken* 'dry'. This example illustrates again the value of regular correspondences: if someone knows that with English initial *d-*, we expect to see German initial *t-* rather than *d-*, they would be less likely to be misled by the accidental similarity between *dry* and German *dürr*. Once again, the probability associated to this kind of mistakes increases for understudied languages.

Being associated to human errors, we can expect the presence of small amount of mistakes even in carefully prepared lexical cognacy dataset, including the one that we are using in this study (Grollemund dataset). For this reason, we checked how the presence of such mistakes would influence the whole performance of the framework proposed in this thesis. We thus introduced mistakes in the simulated data proceeding as follows:

1. We simulated a pseudodataset;
2. We introduced errors: with probability P_{error}, the actually simulated cognate class/allele is replaced by a wrong one. Replacement can be of two types. First, with $P_{create\ new}$, we replace the true class with a newly created cognate class not shared by any other language. Then, with $P_{shift} = 1-P_{create\ new}$, we replaced the true class with one of the other existing classes;

   The perturbation introduced in the dataset was P_{error} 0% (no error), 1% and 2% and the probability that the error is generating a new allele, $P_{create\ new}$, 0% (no error), 20%, 50% and 80%, weighting the error by languages and weighting the error by meaning.
3. Compute the SuSt.

In the observed data, the distribution of errors is likely non-random. Some languages may have inherent features (e.g., a large amount of historical sound change) that relatively hinder correct identification of cognates, while some words may have inherent features that make cognate identification difficult. As an example of the latter case, the same ancestral word may have experienced highly idiosyncratic developments in different branches of the family; one example would be words with suppletive roots, i.e. roots varying by grammatical form, which in subsequent history get more uniform in each language, but in different ways. Thus, we introduced a non-random patterning of errors by language and meanings in our linguistic simulated data, by which we can control if all languages and meanings are equally likely (or not) to have errors.

In summary, we performed new simulations for the five-populations models introducing errors, using the best combinations of parameters chosen from the power results obtained previously.

Specifically, we introduced errors as described above in simulations generated according to all the possible combinations of number of languages {200}, number of meanings {100; 200; 500}, and different scaled mutations rates {0.5; 1; 5; 10; 20; 50}, for a total of 18 combinations.

2.3.2.4 Five-population models - Masking linguistic simulated data to mimic missing data

The linguistic data are coded in a matrix in which rows correspond to languages and columns to distinct meanings/words that are coded by linguists for cognacy (sharing a common ancestral word). Whenever two words descend from a common ancestor they are classified in the same cognacy class. But this is not an easy task. In addition to the probability of introducing an error in a dataset compiled by hand, there is the difficult to correctly assign a common or a new ancestor for pairs of words, as previously decribed. In contrast, we have missing data when a language is understudied and the researchers may lack information on how a particular meaning is expressed. To take into account the presence of missing data in the observed dataset, and to make predictions about how much the observed level of missing data would affect the results, we introduced the missing data in the simulated dataset. As we will apply the framework to the study of the evolutionary dynamics of the Bantu languages, we created in the simulated data the same pattern of missing data observed in the Grollemund's Bantu dataset. To do this, we first masked the simulated data for the presence of missing data exactly in the same way as in those observed. We also allowed for a random inclusion of missing data, so as to make the procedure applicable to any type of dataset. One could then use 200 pseudo-loci (i.e. meanings) and apply to them the patterns from the 100 loci of the real Bantu data at random and test how it changes the power to discriminate among different evolutionary models.

Until now, we tested the models with all the possible combinations of parameters (general power analysis, see figure 2.9). For this step of our study, we performed a power analysis specific for the characteristics observed in the real linguistic data (linguistic match power analysis, see figure 2.9), i.e., we generated data with the same structure of those observed in terms of number of languages (i.e. 416 languages) and number of meanings (i.e. 100 meanings) considering all the different values of theta tested (0.5, 1, 5, 10, 20 and 50) and using the same prior distributions, for a total of 6 combinations of sampling conditions. We added to the simulated linguistic data the different probabilities of errors (previously tested) and a linguistic mask, so as to consider the same amount of missing data that is present in the observed linguistic dataset.

2.3.2.5 Five-population models – Genetics

Lastly, we assessed the power to discriminate among different evolutionary models using simulated genetic data, but we only tested the five-population models that mimic the Bantu migration, since the method was previously validated for generic genetic data in Ghirotto *et al.* (2019).

Since we do not have whole-genome data for all the five geographical areas corresponding to the five Bantu linguistic groups but only for three (see next sub-chapter 2.3.3), we wanted to understand how this loss of genetic data could affect the performance of the method. So, we performed the genetic simulations taking into account the information available for three out of the five populations (i.e. two populations are considered as ghost populations – figure 2.15), and, in addition, we simulated genetic data for the five populations model (figure 2.13). We tested the combinations of experimental parameters specific for the features observed in the real genomic data, which are number of chromosomes {2; 4}, number of loci {5.000; 10.000; 20.000}, and locus length (base pair (bp)) {1.000; 2.000}, for a total of 12 combinations of sampling conditions tested, to understand which is the best combination of experimental parameters to extract in the real genomic data. In table 2.4 the demographic parameters associated with the models and their prior distributions are described.

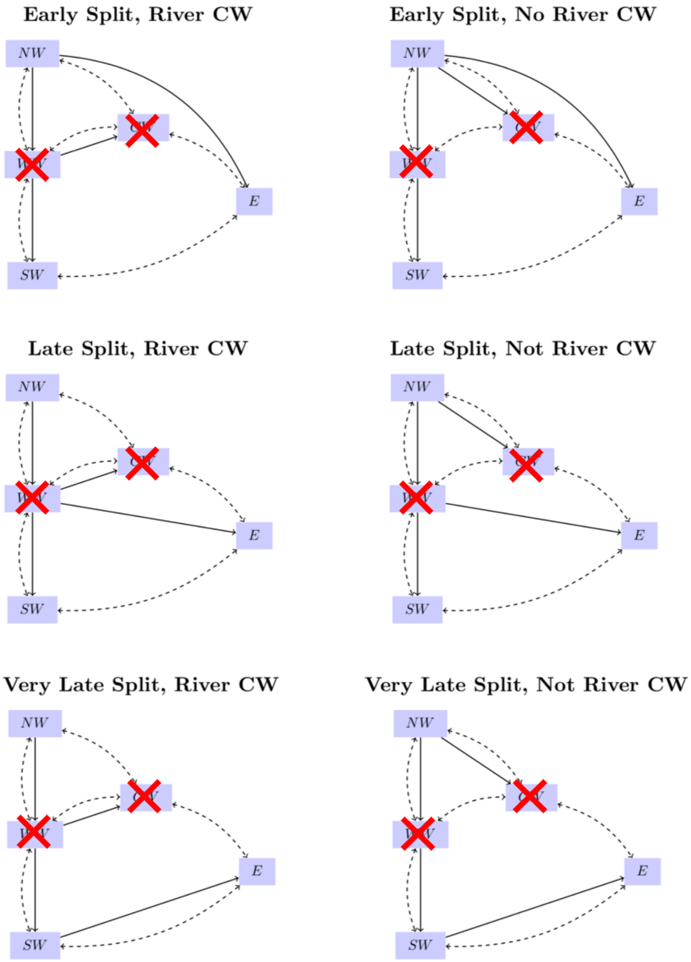


Figure 2.15. Demographic models: Five-population models considering two ghost populations. Simulation of genomic data taking into account that we only have available genomes from individuals belonging to the NW, SW and E populations.

Table 2.4. Demographic parameters and prior distributions used for the genetic simulations (generation time = 29 years).

Demographic parameters	Prior distribution
Effective population size (N_0)	Uniform {25.000 - 250.000}
Recombination rate	1.12e-08
Mutation rate	1.25e-08
Time of the events (in years BP)	Uniform {500 - 8.000}
Base rate of migration (equall in all demes)	Uniform {10^{-6} - 10^{-2}}
Exponential growth intensity	100
Deme number	5
Current population share	Exponential {0.05 - 0.25}

Each of the models mentioned above was simulated using the *ms* simulator (Hudson, 2002) that generates the data using a coalescent approach in which the random genealogy of the sample is first generated and then mutations are randomly place on the genealogy (Kingman, 1982a; Hudson, 1990; Nordborg, 2001). We ran 50.000 simulations per model and per combination of experimental parameters. For each of the evolutionary models described above, we tested the capacity of the framework to discriminate among the different scenarios simulated.

2.3.3 Application to the real case: The Bantu expansion dynamics

The linguistic Bantu data in our study were compiled by Rebecca Grollemund (Grollemund 2012, 2015, figure 2.16A). Grollemund's published data provide language names, wordlists, and expert judgements about the cognacy of words in the wordlists. It is the cognate judgements that we employed, treating each concept (i.e. meaning) as a site, and each set of cognates (i.e. words sharing a common ancestral word) as individual alleles at that site. We used a selection of 100 meanings, which are the best documented for the Bantu languages, as follows:

animal, arm, ashes, bark, bed, belly, big, bird, bite, blood, bone, breast, burn, child, cloud, come, count, dew, die, dog, drink, ear, eat, egg, elephant, eye, face, fall, fat/oil, feather, fingernail, fire, fire-wood, fish, five, fly, four, give, goat, ground/soil, hair, head, hear, heart, horn, house, hunger, iron, intestine, kill, knee, knife, know, leaf, leg, liver, louse, man, moon, mouth, name, navel, neck, night, nose, one, person, rain, road/path, root, 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, woman.

For each of these 100 lexical items (meanings), Grollemund *et al.* (2015) identified whenever possible cognate sets. Where it was not possible, they based the cognacy judgment on the principle of resemblance. In the end, they were able to identify 3859 cognate sets across the 100 meanings.

Our scenarios for the dynamics of the Bantu expansion (figure 2.4) were formulated in terms of five geographical groups of the Bantu languages: NW, WW, CW, SW and E (figure 2.11). There is no consensus on the internal classification of the Bantu family of languages (Bastin *et al.*, 1999; Nurse and Philippson, 2003; Grollemund *et al.*, 2015) and on the geographical features that can be used as proxies for regional groupings. While features such as the rainforest, the basin of the Congo river and the Great Lakes area do correlate with linguistic distributions, they do not however provide unambiguous boundaries between different Bantu zones. Any division of the considered set of languages would thus be somewhat arbitrary. Our selection is motivated by the discussion in Nurse and Philippson (2003) and relies largely on their groupings, which are however subtler than our crude division into five mega-zones. This way, we rely on groupings established based on grammatical features, as opposed to lexical data as in our linguistic dataset taken from Grollemund (2012; 2015), avoiding circularity. To calculate the observed SuSt (mean, variance, skewness and kurtosis) we used the dataset described above compiled by Grollemund et al (2015) containing information for 416 different Bantu languages and 100 lexical items. The SuSt obtained for the real linguistic data was then analysed through the ABC-RF model selection procedure.

The genomic dataset used were the modern high-coverage genomes from Mallick *et al.* (2016, Table 2.5) (Mallick *et al.*, 2016). We analysed whole-genome data from six Bantu individuals belonging to the NW (two individuals), SW (two individuals) and E (two individuals) populations (figure 2.16B). Until the end of the writing of this thesis, we did not find genomic data for two of the five populations that we were interested in studied.

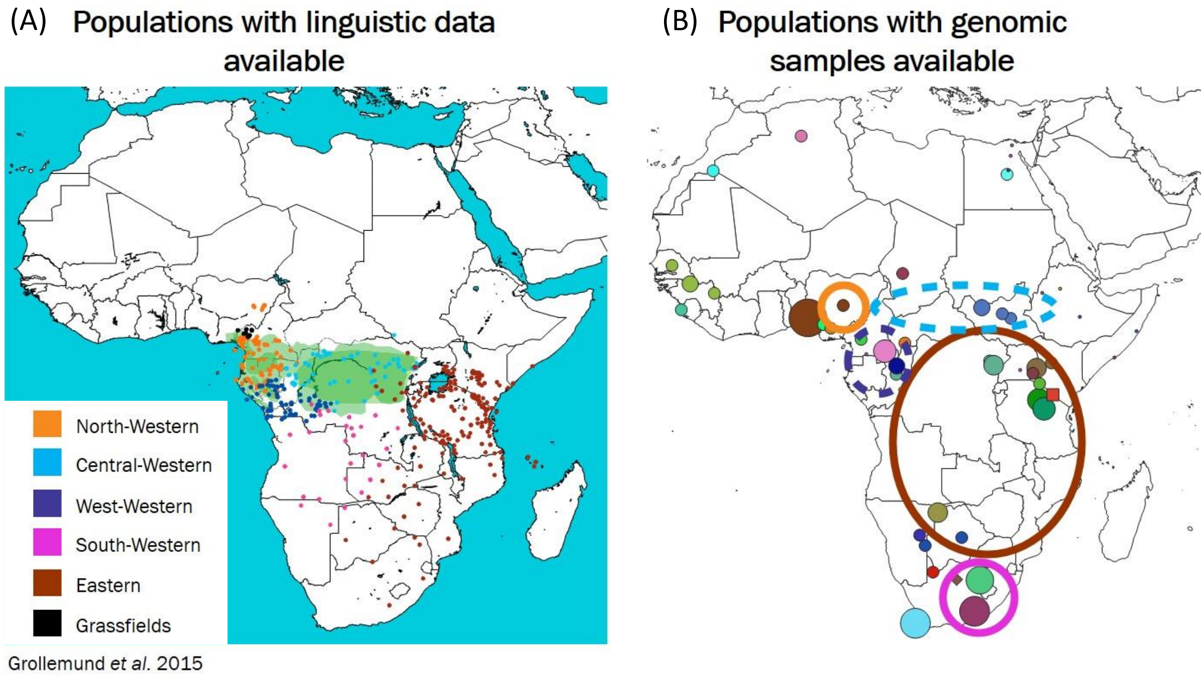


Figure 2.16. Geographical location of Bantu populations. (A) Bantu populations for which linguistic data is available (from Grollemund *et al.* 2015); (B) Bantu populations with genomic data available. In circles are the populations for which we have genomic data (from Mallick *et al.* 2016). Dashed circles represent the populations for which genomic data is missing.

Table 2.5. Whole-genome samples collected for the Bantu populations under study.

Sample ID	Population	Latitude	Longitude	Geographic group	Reference
LP6005677-DNA_C04	Lemande	4.5	11.1	NW	Mallick *et al.* 2016
LP6005677-DNA_D04	Lemande	4.5	11.1	NW	Mallick *et al.* 2016
LP6005443-DNA_E02	Bantu Herero	-22	19	SW	Mallick *et al.* 2016
LP6005441-DNA_F01	Bantu Herero	-22	19	SW	Mallick *et al.* 2016
LP6005443-DNA_A01	Bantu Kenya	-3	37	E	Mallick *et al.* 2016
LP6005441-DNA_B02	Bantu Kenya	-3	37	E	Mallick *et al.* 2016

All the individuals were mapped against the human reference genome hg19 build 37 and regions that did not pass a set of quality filters were eliminated (a detailed explanation of all the filters applied is described in Mallick *et al.* 2016). To study the demographic history of populations we should remove regions of the genome that can be under selection. Thus, to calculate the observed *FDSS* we selected autosomal regions found outside known and predicted genes (approximately 10.000 bp), outside CpG islands and repeated regions (as defined on the UCSC platform, (Hinrichs *et al.*, 2016). From these neutral regions, we extracted 20.000 independent loci of 1.000 bp length, separated by at least 10.000 bp. For each comparison, we used two individuals (four chromosomes) per population, starting from the NW, SW and E. The observed *FDSS* was then analysed through the ABC-RF model selection procedure.

2.4 Results

2.4.1 Power analysis

To determine the power of our ABC-RF framework in distinguishing among alternative evolutionary models, we simulated linguistic data considering different experimental conditions under simple demographic models, one and two populations models (figure 2.12 and 2.13, respectively), and a complex model, the five-population models (figure 2.14). The last model was designed based on a realistic scenario, the Bantu expansion. We tested all the possible combinations of number of languages (i.e. number of chromosomes) {20; 50; 100; 200}, number of meanings (i.e. number of loci) {20; 50; 100; 200; 500}, and different scaled mutations rates {0.5; 1; 5; 10; 20; 50}, for a total of 120 combinations of sampling conditions tested.

For the genetic data, the method was previously validated under the simple models (i.e. one and two populations models) (Ghirotto *et al.*, 2019), therefore, we only performed the validation considering the five-population models, which mimics the dynamics of the Bantu expansion. Since we do not have whole-genome data for two of the five populations that we were interesting on study, we did two different analysis: (1) we simulated genetic data sampling from all the five populations (figure 2.14) and (2) considering two populations for which there is missing data as ghost (figure 2.15). In addition, we only simulated genetic data considering two or four chromosomes, since we have a maximum of two individuals (four chromosomes) per population. In summary, we tested all the possible combinations of number of chromosomes {2; 4}, number of loci {5.000; 10.000; 20.000}, a locus length (bp) {1.000; 2.000}, for a total of 12 combinations of sampling conditions tested.

The simulated data was generated using the *ms* simulator (Hudson, 2002) according to each of these combinations for the models mentioned above. For each combination of parameters, we treated each simulation as *pod* to be compared within each set of models (one, two and five populations models) with the simulated data. This comparison was performed through the ABC-RF, computing the confusion matrices and the out-of-bag classification error (CE). Then, we calculated the proportion of True Positives (TP) for each comparison as 1-CE. The proportion of TP give us a measure of the power of the framework applied considering all its characteristics, like the model selection approach, alternative models compared, the SuSt that we chose to summarize the data, and the prior distributions used.

2.4.1.1 One-population models

To start, we compared six one-population models. Figure 2.17 shows the results for the power analysis. Each plot represents the proportion of TP for each combination of parameters, for each of the six models tested.

In general, the proportion of TP is high, around 80%, depending on the model and on the combination of parameters tested. We can observe that a high number of languages and meanings and a low value of theta improves the power. In general, however, the parameter that seems to be more associated with an increase of the power is when the value of theta decreases. The most identifiable model is the Bottleneck model, reaching almost 100% TP for all the combinations of parameters for a theta value of 0.5 and 1. On the other hand, when we simulated a Structured model with a high rate of gene flow, the power increases as the value of theta increases, regardless if we are sampling languages only in one or in all demes. In contrast to that, with a low migration rate the power increases as theta value decreases.

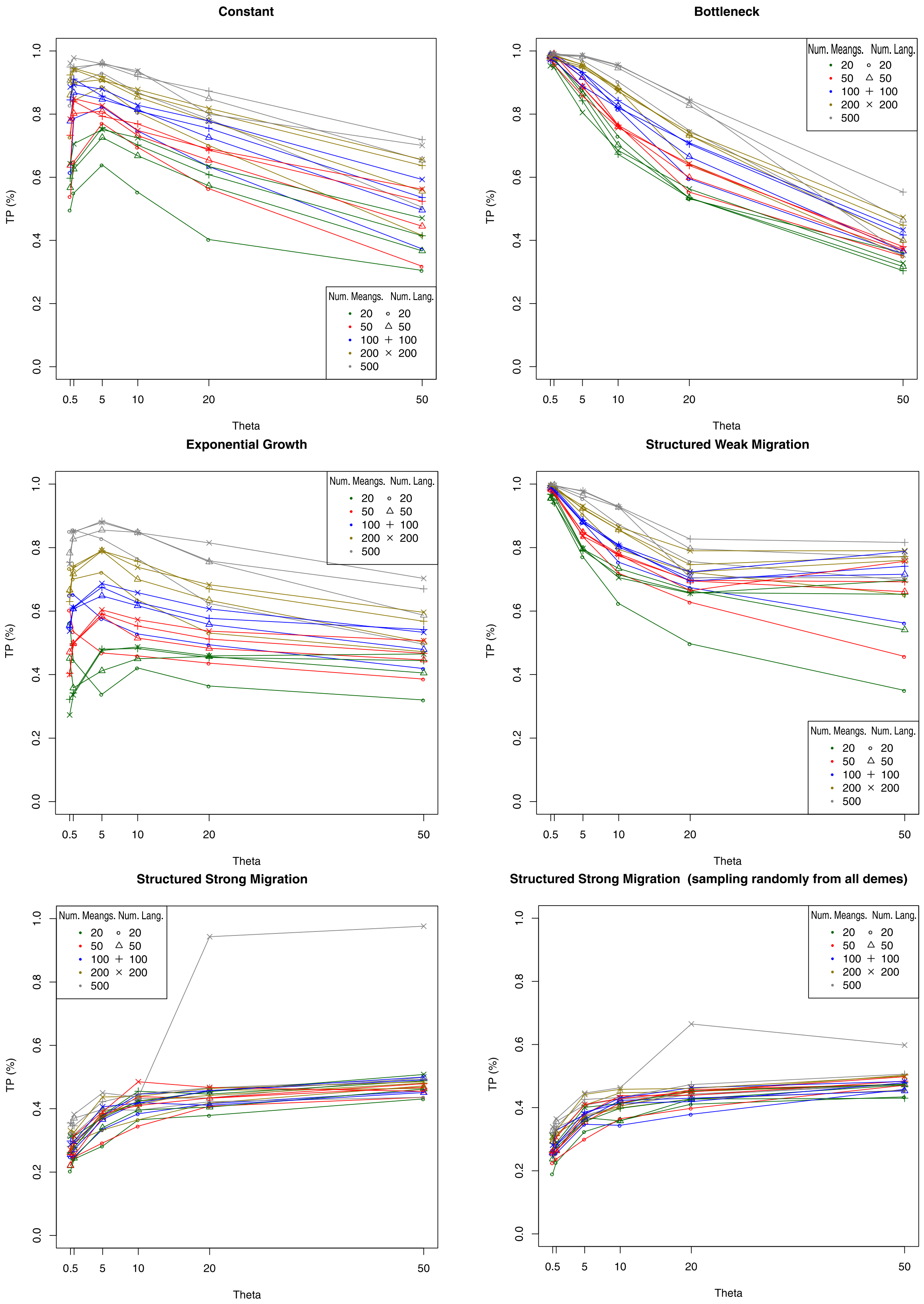


Figure 2.17. Proportion of True Positives (TP) for one-population models. The plots present the proportion of TP (y-axis) obtained analysing *pods* resulting from each of the six models under 120 combinations of experimental parameters. Different values of theta are in the x-axes, number of languages is represented by different symbols and number of meanings is represented by different colours. For example, the experimental condition of 200

languages and 500 meanings, which corresponds to a gray cross in each of the six the plots. Selecting a *pod* from a true model, the ABC-RF classifier can choose the right model (i.e. the true positive (TP)) or pick any one of the other five models (i.e. a false positive). So 1-TP is the combined rate with which the other, wrong, models are chosen.

2.4.1.2 Two-population models

The two-populations models include four alternative evolutionary scenarios (figure 2.13). The results for the power analysis are presented in figure 2.18.

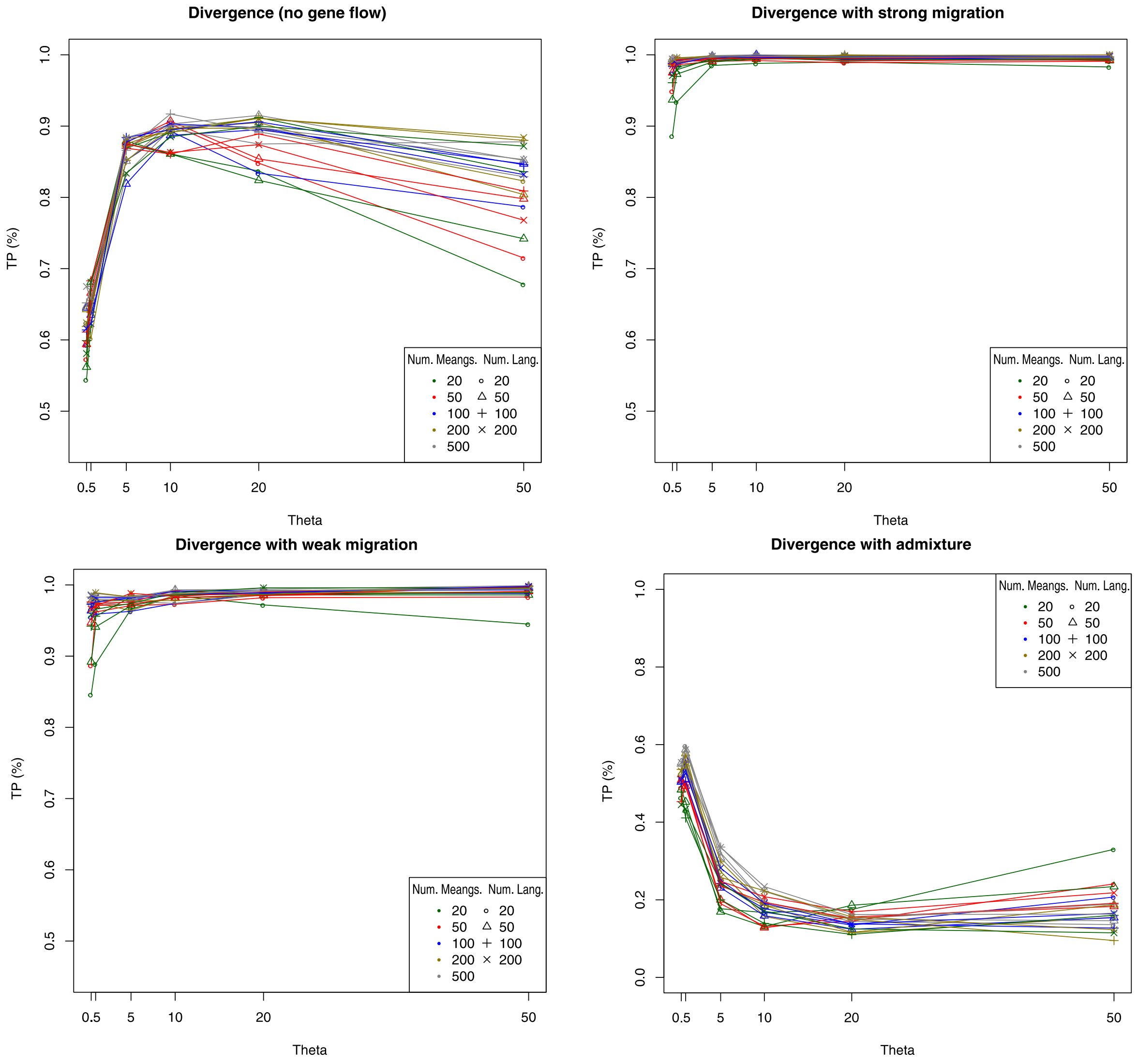


Figure 2.18. Proportion of True Positives (TP) for two-population models. The plots present the proportion of TP (y-axis) obtained analysing *pods* resulting from each of the four models under 120 combinations of experimental

parameters. Different values of theta are in the x-axes, number of languages is represented by different symbols and number of meanings is represented by different colours.

The most recognizable model is the Divergence with continuous migration (both strong and weak migration) from the split until present times, with a percentage of TP close to 100%. For the other two models, the power is low, with values ranging from 40% to 90%, for the Divergence model without gene flow, and from 10% to 60% for the Divergent model with a single event of admixture. When the pulse of admixture happens right after the divergence of the two populations, the Divergent model without gene flow and the Divergent model with a single event of admixture became undistinguishable, thus possibly affecting the model identifiability and the proportion of TP. To explicitly test this hypothesis, we subdivided the *pods* in eight categories, based on the time interval between the divergence and the admixture moment, and calculated separately the proportion of TP within each of these categories. As expected, when the time between the divergence and the admixture event increases, the proportion of TP increases (figure 2.19). This pattern is therefore influenced by the value of theta, how already seen indeed, when the theta is high the amount of TP remains low.

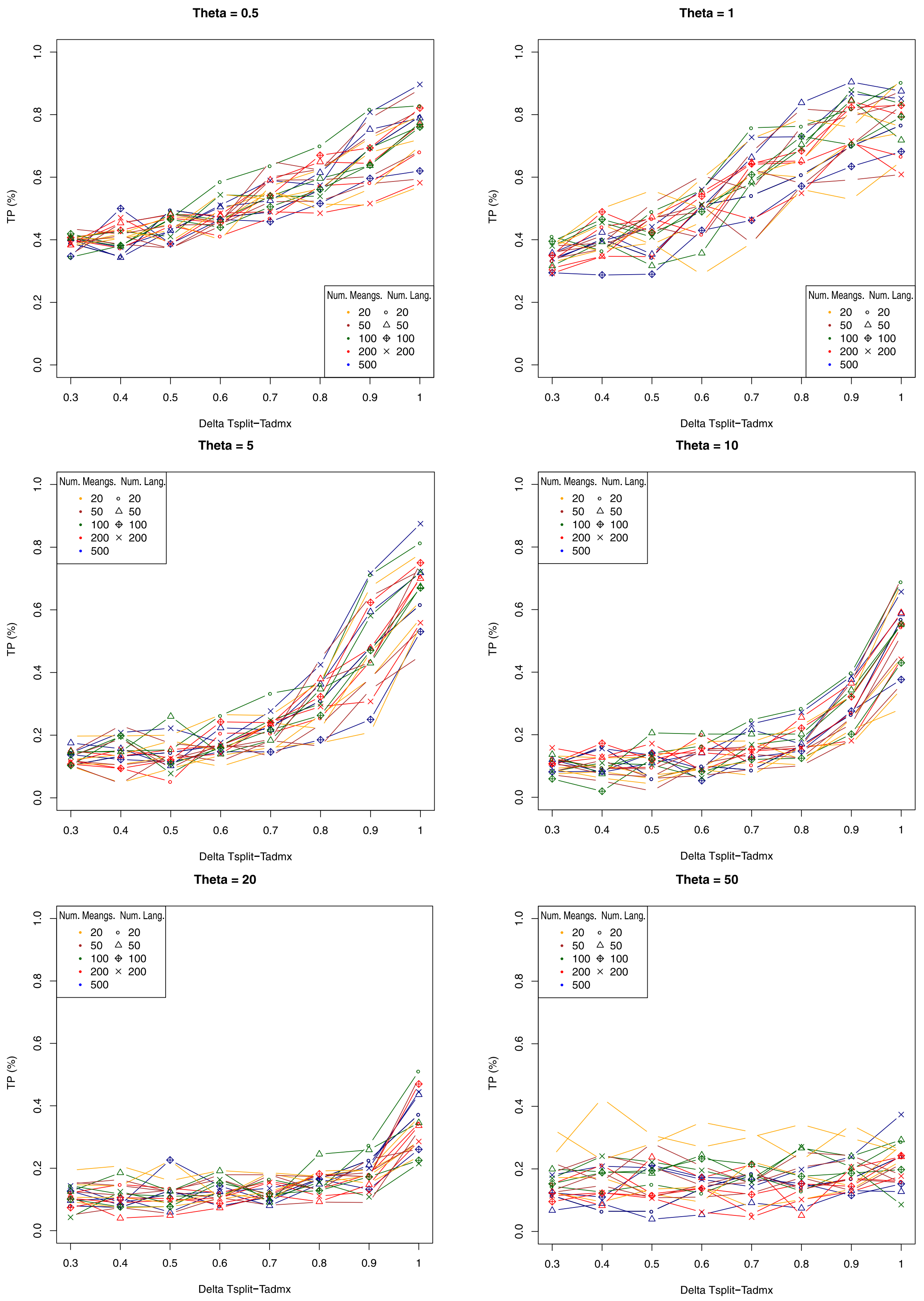


Figure 2.19. Proportion of True Positives (TP) for Divergent model with a single event of admixture model. The plots present the proportion of TP (y-axis) obtained analysing *pods* subdivided in eight categories, depending on

the time interval between the divergence and the admixture moment (Delta Tsplit – Tadmix). Different values of delta are in the x-axes, number of languages is represented by different symbols and number of meanings is represented by different colours.

2.4.1.3 Five-population models – Languages

There are two hypotheses about the dynamics of the Bantu expansion populations, Early and Late-split hypothesis. Based on both models, we designed six alternative scenarios (figure 2.14) to explain the evolution of the five geographical groups of the Bantu languages: NW, WW, CW, SW and E. We also introduced the possibility of borrowing, as a migration rate among the five groups. We started by analysing the performance of the method for all the combinations of parameters without the introduction of errors and the mask (figure 2.20).

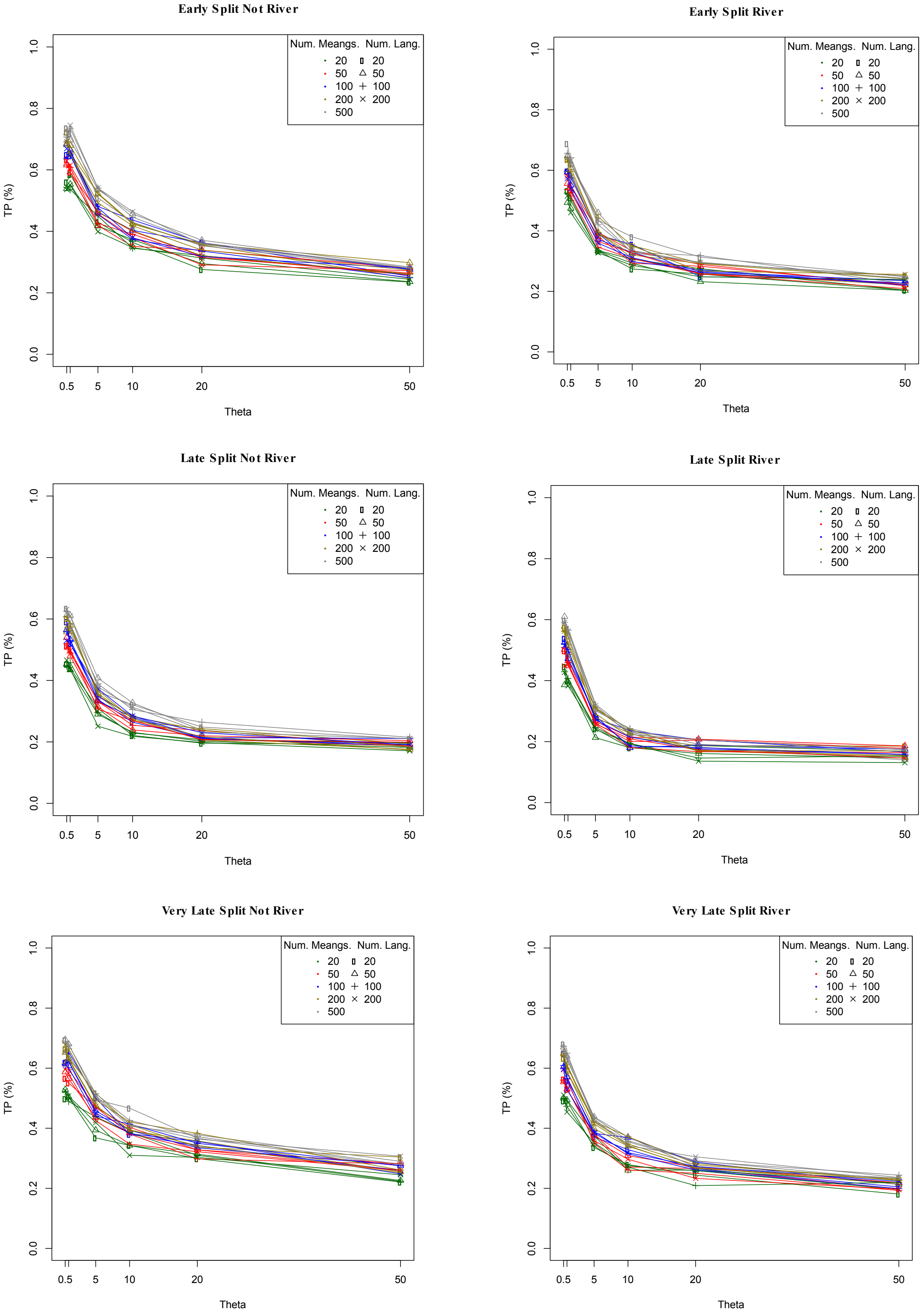
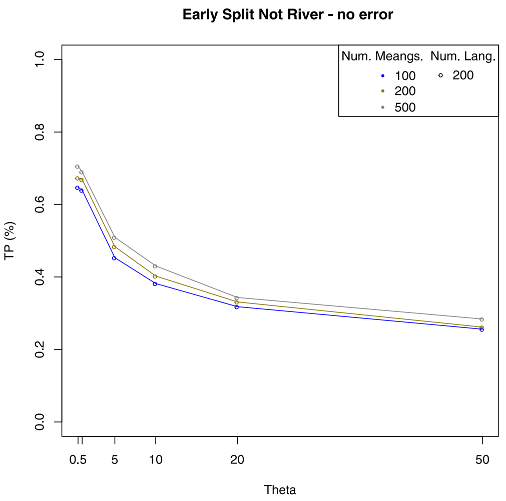
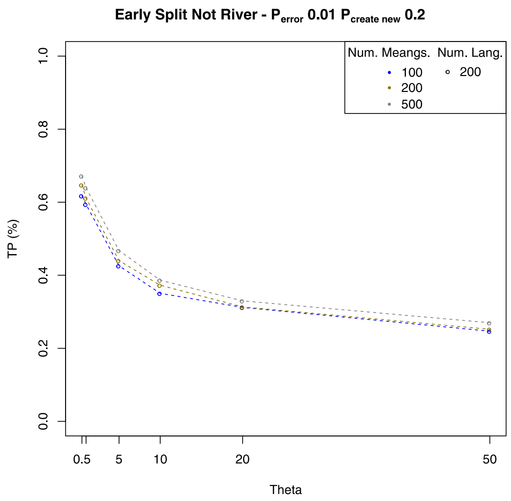


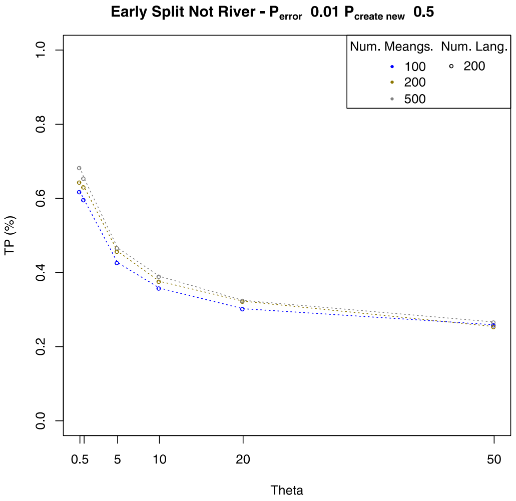
Figure 2.20. Proportion of True Positives (TP) for five-population models - Languages. The plots present the proportion of TP (y-axis) obtained analysing *pods* resulting from each of the six models under 120 combinations of experimental parameters. Different values of theta are in the x-axes, number of languages is represented by different symbols and number of meanings is represented by different colours.

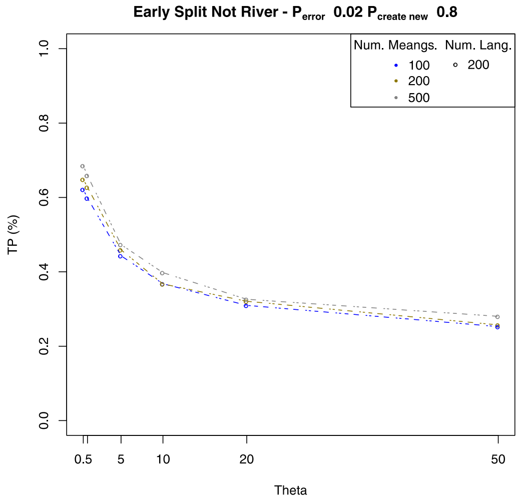
The proportion of TP ranges between 20% – 75% for all the six models. A high number of languages and meanings and a low value of theta allows to increase the power to distinguish among the different evolutionary models. The combination of experimental parameters showing the highest TP rate was 200 languages, 500 meanings and a value of theta of 0.5. Those values are the highest tested number of languages and meanings, and the lowest theta.

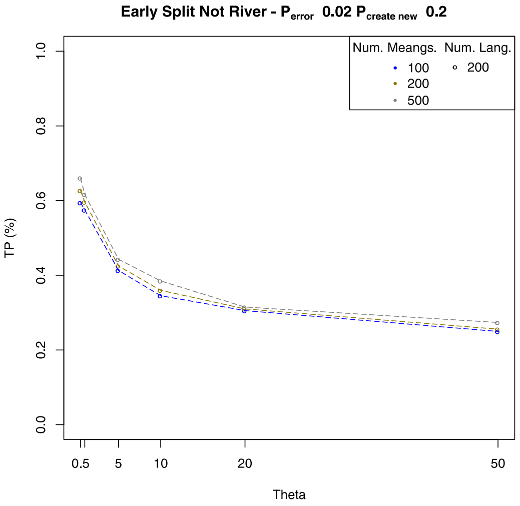
We then selected the best combinations of experimental parameters regarding the number of languages and meanings (number of languages 200 and number of meanings 100, 200 and 500), to perform the power analysis introducing different probability of errors (figure 2.21 and Supplementary Figure S2.1). The perturbation introduced in the dataset was P_{error} 0% (no error), 1% and 2%, and the probability that the error is generating a new allele, $P_{\text{create new}}$, 0% (no error), 20%, 50% and 80%, weighting the error by languages and meanings, which allows us to control if all languages and all meanings are equally likely or not to have errors. The idea was to verify how stable our method is when data contains different amount of errors.

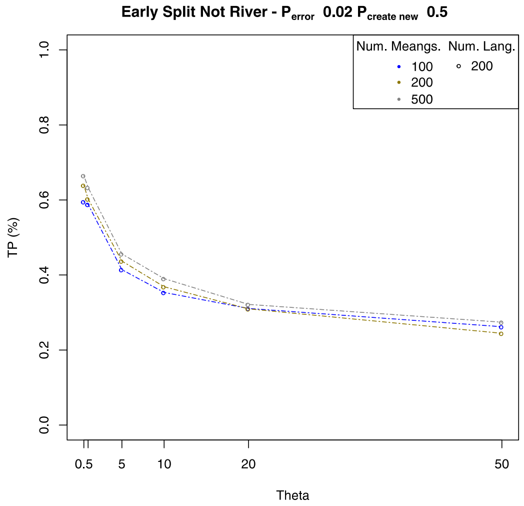

Early Split Not River - no error
Num. Meangs.
100
200
500
Num. Lang.
200
TP (%)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
5
10
20
50
Theta


Early Split Not River - P_{error} 0.01 $P_{create\ new}$ 0.2
Num. Meangs.
100
200
500
Num. Lang.
200
TP (%)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
5
10
20
50
Theta


Early Split Not River - P_{error} 0.01 $P_{create\ new}$ 0.5
Num. Meangs.
100
200
500
Num. Lang.
200
TP (%)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
5
10
20
50
Theta


Early Split Not River - P_{error} 0.02 $P_{create\ new}$ 0.8
Num. Meangs.
100
200
500
Num. Lang.
200
TP (%)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
5
10
20
50
Theta


Early Split Not River - P_{error} 0.02 $P_{create\ new}$ 0.2
Num. Meangs.
100
200
500
Num. Lang.
200
TP (%)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
5
10
20
50
Theta


Early Split Not River - P_{error} 0.02 $P_{create\ new}$ 0.5
Num. Meangs.
100
200
500
Num. Lang.
200
TP (%)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
5
10
20
50
Theta

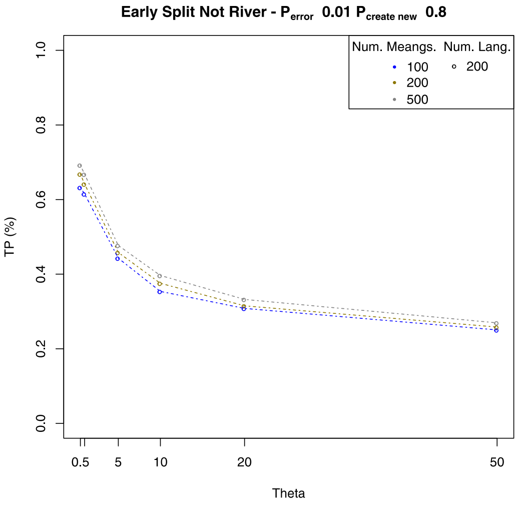


Figure 2.21. Proportion of True Positives (TP) for five-population models with introduction of errors on the simulated linguistic data. The plots present the proportion of TP (y-axis) obtained analysing *pods* resulting from the Early split not river model (see the other models in the Supplementary Figure S2.1) under 18 combinations of experimental parameters. Different values of theta are in the x-axes, number of languages is represented by different symbols and number of meanings is represented by different colours.

In general, the proportion of TP ranges between 16% – 71% for the simulated linguistic data without errors, and 14% – 70% for the simulated linguistic data with different probabilities of error. For the simulated data containing different levels of error, the results show that with a high probability of error (2%) and a low probability to create a new cognate class (20%), we obtained the lowest rate of TP. In contrast, when we introduced a low probability of error (1%) and a high probability to create a cognate class not shared by any other language (80%), the rate of TP is closer to the one without the introduction of errors. Nonetheless, there is not a significant decrease in the rate of TP between the data without errors and with different levels of errors.

2.4.1.4 Five-population models – Power analysis performed for linguistic simulated data with the same structure of those observed in the Grollemund's dataset

The last step to validate this new ABC linguistic framework was the simulation of linguistic data adding the different probabilities of errors (previously tested) and a linguistic mask, so as to consider the same amount of missing data that is present in the real dataset. We generated data with the same structure of those observed in terms of number of languages (i.e. 416 languages) and number of meanings (i.e. 100 meanings) considering all the different values of theta tested in the previous power analysis and using the same prior distributions. In figure 2.22 we can see the results of the power analysis for all the six scenarios. We saw that mask the simulated linguistic data does not change the power of the framework to distinguish among models. The proportion of TP in the simulated data containing errors and a linguistic mask do not change significantly when compared with the simulated data without errors and no mask (range for the former 15% – 63% and the latter 17% – 66%). Moreover, as previously seen, when we simulated linguistic data with 1% of probability of error and 80% of probability to create a new cognate class, the rate of TP is much closer to the simulated linguistic data without errors and with/without mask.

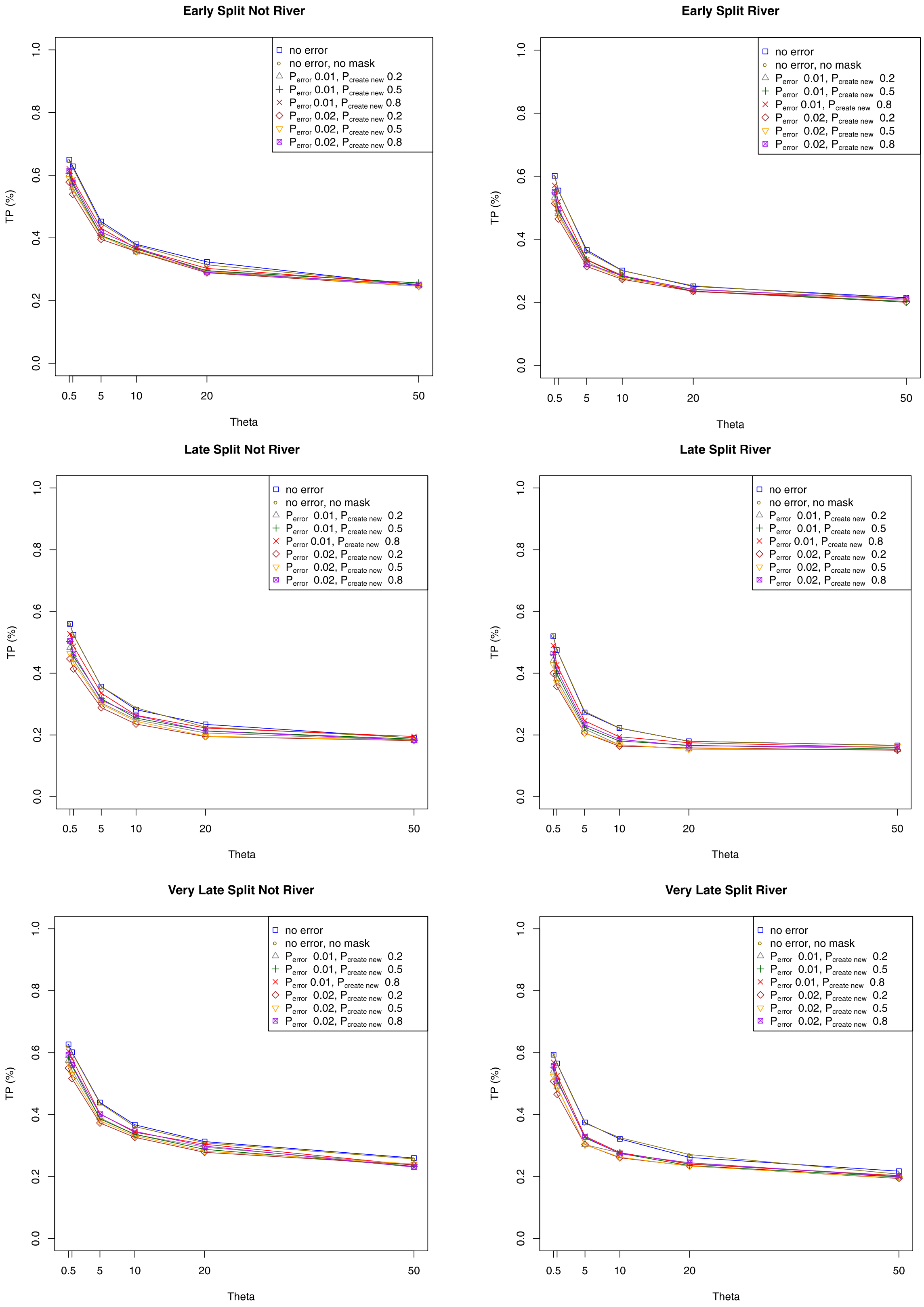


Figure 2.22. Proportion of True Positives (TP) for five-population models with introduction of errors and masking on the simulated linguistic data. The plots present the proportion of TP (y-axis) obtained analysing *pods* resulting

from each of the six models under six combinations of experimental parameters for 416 languages and 100 meanings. Different values of theta are in the x-axes, probability of error is represented by different symbols.

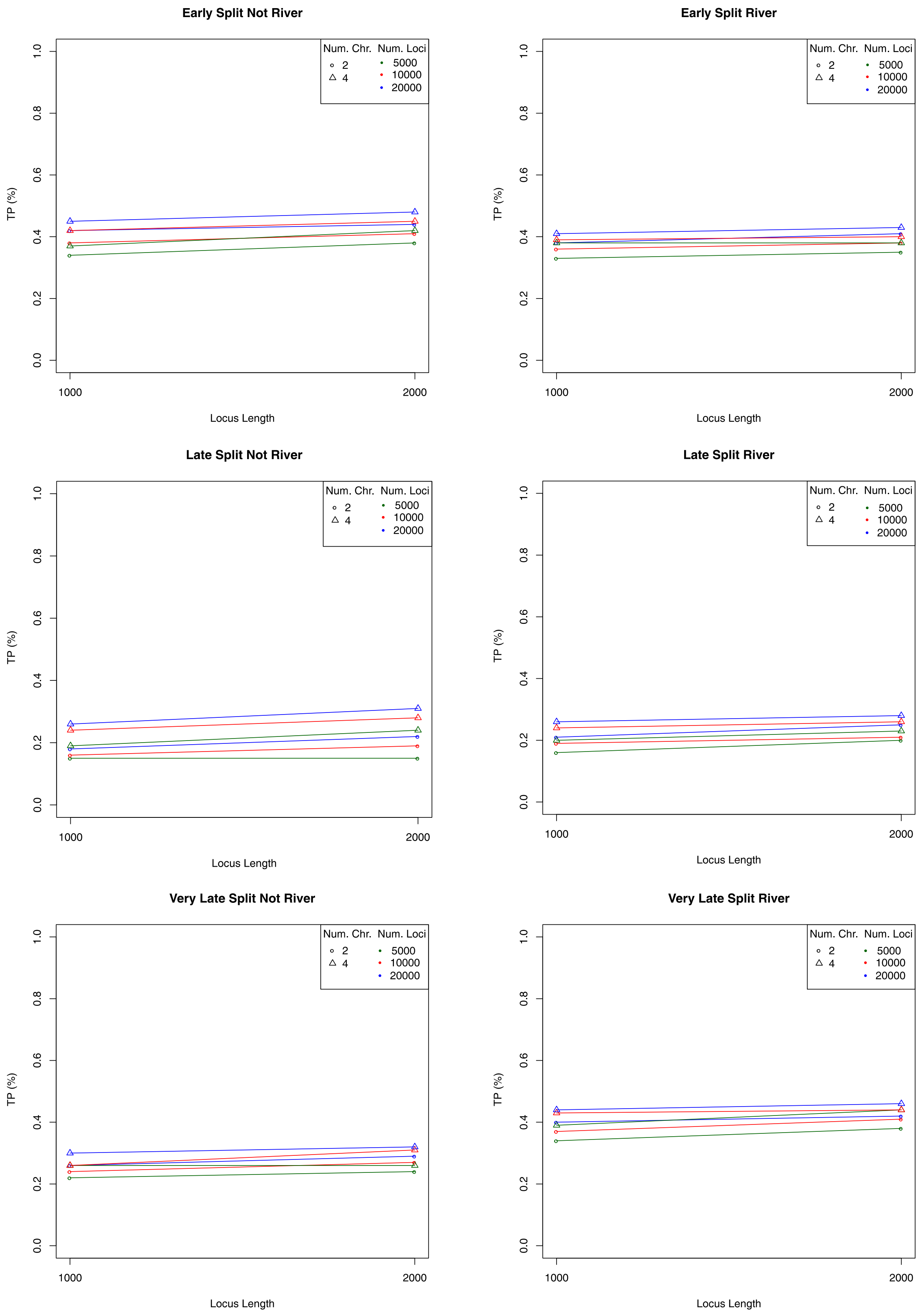
2.4.1.5 Five-population models – Genetic

We simulated genetic data under the same six models described above, in which the prior distributions were based on information found on the literature (table 2.4). As opposed to the analysis of languages, we just sampled (and hence compared) the genetic data from three out of five groups, so as to assess the power of the procedure in distinguish among the models of Bantu migrations with the amount of data available (i.e. NW, SW and E). We also tested different combinations of experimental parameters in terms of number and length of loci, and number of chromosomes sampled per group, so as to understand which would be the optimal set of parameters to extract in the real data. In general, sampling genetic data from three of the five groups, the proportion of TP is low, ranging values from 15% to 48% (figure 2.23A).

To better understand the pattern of the models selected in the power analysis, we evaluated the confusion matrix. The confusion matrix shows which is the model selected for each of the *pods* analysed, so as to understand how the False Positives (FP) are distributed among the alternative evolutionary histories (Supplementary Table S2.1). The simulated *pods* are correctly assigned to the model that generated them with a higher frequency, even if the rate of TP is low. However, it seems that when we have a low number of chromosomes and loci and a small locus length, the method is not able to distinguish the Early from the Late split models. For instance, in table S2.1, we can see that when we sample one individual (two chromosomes) per population and select 5000 loci with a locus length of 1000 bp, the Late split not river model (i.e. the true model) is wrongly assigned. Instead, it is selected the Early split not river model with a number of simulated *pods* of 2267 against the 1453 simulated *pods* for the Late split river model. A solution to increase the rate of TP could be ignore the typology river – not river merging the different scenarios, ending up with three main models, Early, Late and Very Late split. Moreover, we noticed that since we only have genetic information for three populations (NW, SW and E) the different topologies for the six models will have different weights. The different tree topologies generated under the Early split river and Early split not river models became undistinguishable, and the same happens for the Late and Very Late split river and Late and Very Late split not river models. We ended up with two big scenarios, the Early and Late split hypothesis, instead of the six models. Join the different models within these two big scenarios might allow us to increase the power of the framework, even if we only sample three groups.

The combination of parameters that maximizes the power to discriminate the different models is sampling two individuals (four chromosomes) per population and considering 20.000 loci, 2.000 bp length (figure 2.23A). We also tested the power we would reach considering to sample one or two individuals (i.e. two or four chromosomes) for each of the five groups analysed. When sampling genomic data from all the five populations the rate of TP increases and we have more power to discriminate among the different models, ranging values from 29% to 77% (figure 2.23B and Supplementary Table S2.2).

(A)


Early Split Not River
Early Split River
Late Split Not River
Late Split River
Very Late Split Not River
Very Late Split River
TP (%)
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
Locus Length
Num. Chr.
○ 2
△ 4
Num. Loci
• 5000
• 10000
• 20000

(B)

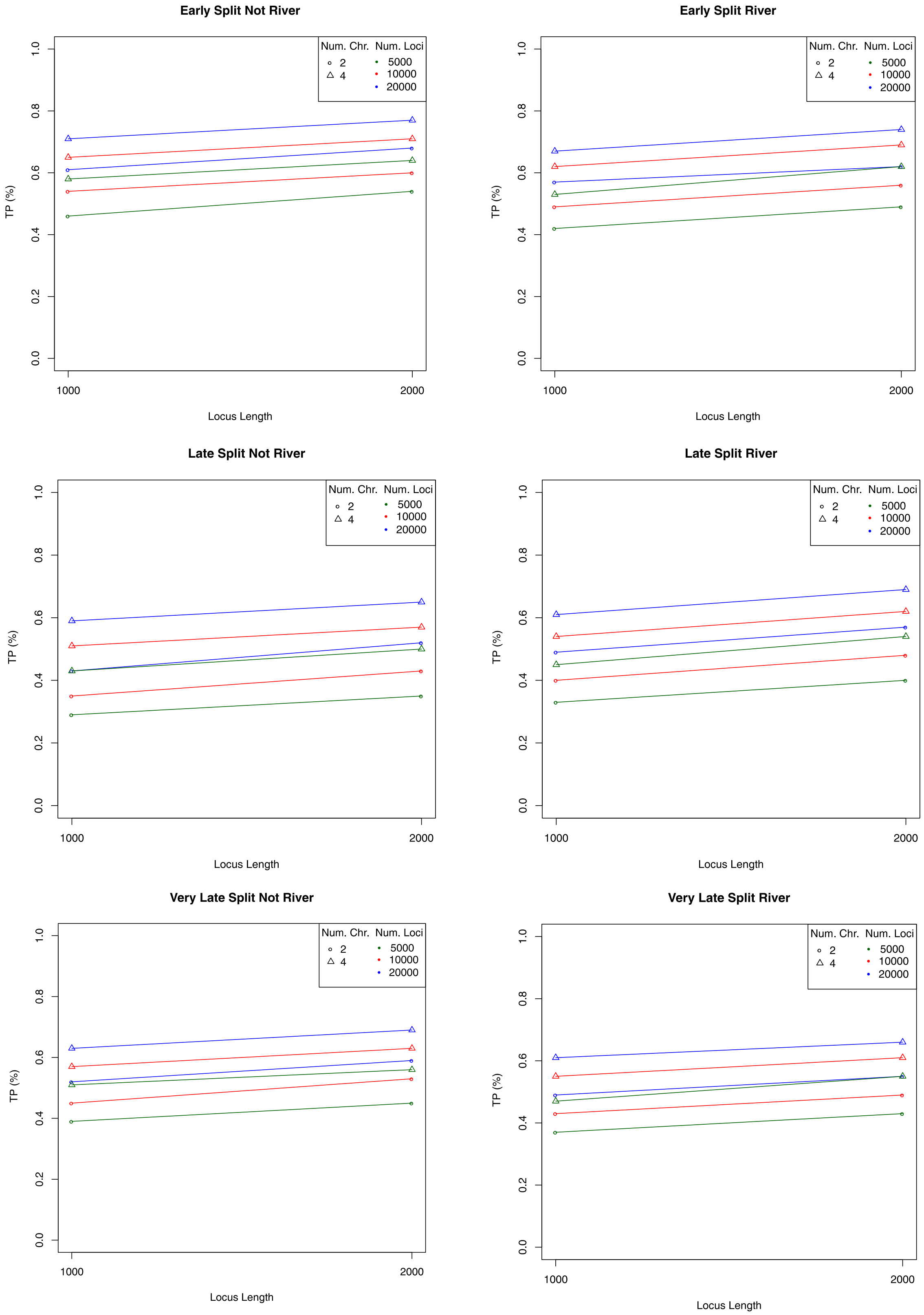


Figure 2.23. Proportion of True Positives (TP) for five-population models - Genetics. (A) Evolutionary models simulated considering only three populations. (B) Evolutionary models simulated considering all the five

populations. The plots present the proportion of TP (y-axis) obtained analysing *pods* resulting from each of the six models under 12 combinations of experimental parameters. Different values of locus length are in the x-axes, number of chromosomes is represented by different symbols and number of loci is represented by different colours.

2.4.2. Application to the real case

The simulations performed in the previous sub-chapters show that the evolutionary models representing different scenarios of the Bantu expansion can be distinguished with a certain degree of confidence. The power analysis also showed us that the confidence is different for genetic and linguistic data (15% – 48% for the former and 15% – 63% for the latter) and in some cases the similarity among different models may influence the ability of the method to recognize the true demographic history (e.g. Early vs. Late split models for genetic data). Keeping this in mind, as a final step, we applied the method to the analysis of real data so as to understand which set of demographic and evolutionary factors can best explain the linguistic and biological variation observed in Bantu populations. We applied the method separately to linguistic and genomic data and estimate the posterior probabilities of the alternative Bantu demographic dynamics.

We analysed the linguistic data generated by Grollemund *et al.* (2015) and calculated the SuSt (mean, variance, skewness and kurtosis of the number of cognate sets in each population and shared between each pair of populations). To take into account the uncertainty associated with the linguistic mutation rate, we compared the observed data with data simulated under a range of theta values (0.5, 1, 5, 10, 20 and 50). Our results indicate that simulating linguistic data considering a theta equal to 5 we obtain the combination of parameters closest to the observed data (table 2.6 and Supplementary Table S2.3). For this specific combination of parameters, the ABC-RF model selection procedure supported the Late split river hypothesis with a posterior probability of 54% (regardless of whether the probability of error in the simulated data was 1% or 2%, and a probability to create a new cognate class of 50% or 20%, respectively). We obtained the same probability even when the linguistic data did not contain errors or a linguistic mask. The posterior probability decreased to 53% when the dataset contained a linguistic mask but no errors. When we introduced a probability of error of 2% and a probability to create a new cognate class of 50% and 80%, the posterior probability changed slightly, becoming 52%. We observed the same results even when the probability of error is

1% and probability to create a new cognate class of 20%. Lastly, when we introduced a probability of error of 1% and a probability to create a new cognate class of 80%, the posterior probability fell slightly, to 51% (table 2.6).

Then, to understand if the simulated linguistic data do actually recreate the observed linguistic variability, we did a regression layer like linear discriminant analysis (LDA), and observed concordant results with the model selected (Supplementary Figure S2.2).

Table 2.6. Linguistic data: number of votes associated to each model under different probability of errors and masking the data by ABC-RF, and posterior probability of the most supported model (model 4). Model 1: Early split not river; Model 2: Early split river; Model 3: Late split not river; Model 4: Late split river; model 5: Very Late split not river; model 6: Very Late split river.

theta 5	selected model	votes model 1	votes model 2	votes model 3	votes model 4	votes model 5	votes model 6	post.proba
no error	4	81	104	81	105	57	72	0.531
no error; no mask	4	81	84	88	111	52	84	0.541
P_{error} 0.01; $P_{create\ new}$ 0.2	4	55	81	95	165	40	64	0.515
P_{error} 0.01; $P_{create\ new}$ 0.5	4	48	94	91	151	47	69	0.545
P_{error} 0.01; $P_{create\ new}$ 0.8	4	61	82	79	146	53	79	0.510
P_{error} 0.02; $P_{create\ new}$ 0.2	4	45	85	81	178	33	78	0.544
P_{error} 0.02; $P_{create\ new}$ 0.5	4	27	93	100	175	41	64	0.519
P_{error} 0.02; $P_{create\ new}$ 0.8	4	40	82	100	149	39	90	0.522

We ran a parallel analysis over genetic data. For this aim we considered genomic data from six Bantu individuals ((Mallick *et al.*, 2016), table 2.1) belonging to three different populations. We extracted from these genomes 20.000 shared independent fragments of 1.000 bp length and we considered four chromosomes per population. The ABC-RF model selection supported the Late split river hypothesis with a posterior probability of 51% (table 2.7), in agreement with the model selected using the linguistic data. However, from the LDA we can see that none of the simulated model is able to capture the variation observed in the real genomic data (Supplementary Figure S2.3).

Table 2.7. Genetic data: number of votes associated to each model under different probability of errors and masking the data by ABC-RF, and posterior probability of the most supported model (model 4). Model 1: Early split not river; Model 2: Early split river; Model 3: Late split not river; Model 4: Late split river; model 5: Very Late split not river; model 6: Very Late split river.

selected model	votes model 1	votes model 2	votes model 3	votes model 4	votes model 5	votes model 6	post.proba
4	9	19	58	183	76	155	0.505

2.5 Discussion

Reconstructing the evolutionary history of human populations is one of the main challenges in population genetics (Harpending and Rogers, 2000; Goldstein and Chikhi, 2002; Beaumont *et al.*, 2002; Hey and Machado, 2003; Li and Durbin, 2011; Liu and Fu, 2015; Scerri *et al.*, 2018). In the last years, there was a huge increase in the number of whole-genome data available and significant advances have been made in the development of inferential methods to extract as much information as possible from these data (Li and Durbin, 2011; Excoffier *et al.*, 2013; Schiffels and Durbin, 2014). However, there is the need to do more, especially in fitting more complex and realistic models.

Languages also keep a trace of demographic changes, and hence can be used to investigate the demographic history of human populations. They usually evolve at a higher rate than genes, being advantageous for inferences over shorter time scales (Grollemund *et al.*, 2015). To date, however, studies that explicitly use linguistic data to infer human demographic dynamics only considered and tested over-simplified version of the real evolutionary history (de Filippo *et al.*, 2012; Grollemund *et al.*, 2015; Thouzeau *et al.*, 2017). More often, linguistic data are used to formulate hypotheses that then will be tested using genetic data, thus assuming that biological evolution mirrors the cultural dynamics under investigation (which is not always the case) (Gray *et al.*, 2009; Amorim *et al.*, 2013; Thouzeau *et al.*, 2017). In this light, the real challenge would be to be able to integrate within the same framework the information coming from complete genomes and linguistic data, so as to consider at the same time biological and cultural variables that might have been influenced by the studied evolutionary dynamics.

Approximate Bayesian Computation represents a powerful method to draw inferences about past demographic history, compare different evolutionary models and estimate the model's probabilities (Beaumont, 2010) generally using genomic data. Only seldom could linguistic data be considered in the analysis (de Filippo *et al.*, 2012; Grollemund *et al.*, 2015), and, to our best knowledge, only once in the context of hypothesis testing (Thouzeau *et al.*, 2017). This method allows us to check whether the models being compared are distinguishable and quantify the reliability of the produced estimations (Csilléry *et al.*, 2010). Moreover, with the development of the machine learning algorithm Random Forest (Pudlo *et al.*, 2015), we can use a large number of SuSt and large genomic/linguistic datasets. A previous work (Ghirotto *et al.*, 2019) successfully applied this algorithm to test the flexibility of an ABC-based framework in comparing different demographic models summarizing the data through the *FDSS* (the complete genomic distribution of the four mutually exclusive categories of segregating sites for pairs of populations – (Wakeley and Hey, 1997)). Having this work as a reference, we developed a new framework in which we could use genomic and linguistic data to compare different evolutionary models, and potentially combine both types of data. We specifically applied this novel framework to the analysis of Bantu-expansion models. The modality and the timing of the Bantu expansion is still a matter of debate. Some genetic studies have been interpreted as showing that the spread of Bantu languages may have been the result of cultural diffusion and language shift without a significant role of migration (Sikora *et al.*, 2011); however, most genetic studies rather support a demic diffusion process for Bantu-speaking populations (Tishkoff *et al.*, 2009; de Filippo *et al.*, 2012; Verdu *et al.*, 2013; Bostoen *et al.*, 2015). The complex history of admixture with local populations, migration waves and the possible recent divergence times contributed to the development of peculiar genetic patterns (de Filippo *et al.*, 2012; Barbieri, Vicente, *et al.*, 2014; Grollemund *et al.*, 2015; Choudhury *et al.*, 2018; Beyer *et al.*, 2019). Therefore, it made sense to develop a framework in which genomic and linguistic data could be simultaneously considered in the analysis of complex demographic models to reveal details of the past history of Bantu population.

2.5.1 Power analysis

We started by doing an extensive analysis of the power of the framework in discriminate the true demographic history comparing set of models with increasing levels of complexity. We evaluated the power to distinguish among simple models (one and two populations) and complex models (five populations) simulating linguistic and genetic data under a broad spectrum of experimental parameters.

Considering first the results obtained in the linguistic analysis, for the one-population models, we were able to well distinguish and recognize all the six models with quite high power and posterior probabilities (figure 2.17). However, there is a decrease in power for the combination of experimental parameters in which we have high rates of linguistc change (theta). One possible explanation is that with high theta value, the high rate of change leads to some sort of "saturation" of the variation, thus making the models not distinguishable. This seems to happen whenever theta is higher than 5. Exceptions are when we analyse the Structured models with strong migration, in which as theta value increases the power increases.

We tested two different sets of Structured models: (1) a Structured model with weak migration, and (2) a Structured model with strong migration. For the model with high migration rate, we followed two different sampling schemes. In the first scheme, languages were sampled only from the first deme (as in the Structured model with weak migration), and in the second scheme we sampled the languages randomly from all demes. The latter approach allowed us to simulate the condition in which we sample from a meta-population as if it were panmictic, while in fact it is substructured, so, the overall size was Ne as in the constant model, and we divided it into several equal-size subparts and then sampled randomly from each of them. It is clear from the results that the model with weak migration can be effectively distinguished from the model with strong migration. The different sampling scheme in the strong migration model does not affect the power to discriminate both of them (sampling in one vs all demes). In a constant population size model, drift would affect the variation at a certain rate. If we consider a structured population with low migration and then we sample from a single deme, we would expect a little bit more variation than in the Constant model, owing to the new (few) lineages brought by migration, explaining the higher rate of TP for the Structured model with low migration rate versus the Constant model. Instead, if migration is strong and we sample a single deme, the reduction of variation expected by the amount of drift affecting a population of size *N* is more counterbalanced by the migrants, resulting in higher variation respect to the low migration scenario. If we sample from multiple demes, with a higher probability we would sample different alleles, thus increasing further the variation. This increase in variation might be influenced by the rate of migration: the higher the rate of migration the lower the differences between the two sampling schemes for the strong migration model, because we would expect to sample the same alleles in one deme and in the whole meta-population. This may be also the reason why there is an opposite behaviour with higher theta values increasing the rate of TP.

The higher variation introduced from the strong migration rate combined with the high value of theta make the model more distinguishable from all the remaining simulated scenarios. Nevertheless, there is a clear pattern as the rate of TP increases it also increases the number of languages and meanings sampled combined with a low value of theta.

Then we moved on to testing the power of the framework for the two-population models (figure 2.18). The framework is able to distinguish with a high probability, between 85% and 100%, the Divergence model with continuous migration (both strong and weak). However, for the Divergence without gene flow and Divergence with a single event of admixture models our method is able to discriminate the different scenarios with probabilities varying over an extremely broad range, between 10% and 90%. This reduction in the power to discriminate among the different scenarios is probably due to the nature of the demographic models that they describe: the first model (Divergence without gene flow) describe two populations that separate from the ancestral population and remain isolated forever; the second model (Divergence with a single event of admixture) describe two populations that separate from the ancestral population, but after the split exchange genes in a single moment of admixture between the two populations. If the rate of admixture is quite low or if the admixture event happens at a time close to the time of divergence from the ancestral population, the two models become hard, or impossible, to distinguish (figure 2.19). Besides, as we previously saw for the one population models, even for these scenarios the probability to distinguish among different evolutionary models increases when we increase the number of languages and meanings sampled, combined with a low value of theta.

Taken all together, these results show that the method and statistics chosen allow to distinguish among different simple demographic models. However, when we have a low number of languages and meanings and a high value of theta the probability to discriminate the models decreases. This was predictable, since we cannot expect to get a good signal with any amount of data. With small amounts of data, i.e., low number of languages and meanings, the accuracy of the reconstructions is doomed to be low.

After testing simple, and probably unrealistic, models, we introduced in our analysis a more complex model that could in fact describe the Bantu demographic history. Previous studies on Bantu populations only tested the Early versus Late split hypothesis, disregarding the question where the central Bantu languages, in the rainforest and largely in the inner Kongo basin, came from, and not taking into account linguistic migration (borrowing) between populations (de Filippo *et al.*, 2012; Grollemund *et al.*, 2015). Here, we established six different scenarios (figure 2.13) taking into account all the genetic and linguistic variability present in Bantu groups: (1) Early Split Not River, (2) Early Split River, (3) Late Split Not River, (4) Late Split River, (5) Very Late Split Not River and (6) Very Late Split River.

We first analysed the performance of the framework in distinguishing the six mutually exclusive scenarios under a wide range of experimental parameters, identifying the combination of parameters that maximizes the power to discriminate among models (high number of languages and meanings, and a low value of theta) (figure 2.20). However, one cannot expect that linguistic cognacy datasets would contain perfectly exhaustive information. There are difficulties in identifying identical-by-descent words, and there are also human errors. It is expected that approximately 2% of linguistic “identical-by-descent judgements” might be erroneous in empirical datasets (data not published). For the simple scenarios, this issue was not very relevant; however, as we move to complex contact scenarios there will be an amount of languages and meanings shared between populations that might be significantly affected by the presence of errors. Consequently, it is important to check the effect of small mistakes in the determination of linguistic “haplotypes”. Therefore, we selected the best combinations of experimental parameters (number of languages: 200, number of meanings: 100, 200 and 500, theta: 0.5, 1, 5, 10, 20 and 50) and performed a new power analysis introducing two probabilities of errors in the simulated linguistic dataset to verify how stable our method is when data contain different amounts of errors (figure 2.21 and Supplementary Figure S2.1). Our method revealed to be stable even when the data contain errors, with a probability to recognize the true model falling between 14% and 70%. One interesting thing to notice is that with a low probability of error (1%) and a high probability to create a new cognate (80%) the power does not decrease significantly with respect to the case in which there are no errors. This is the situation that we expect to be closest to the real case, i.e., when a linguist asks whether or not two words are cognates, the probability of saying that one of the words belongs to a new cognate class is higher than making a mistake and say that both words are cognates.

Subsequently, we analysed the power of our framework generating linguistic data with the same structure of those observed in the Grollemund’s dataset in terms of number of languages (i.e. 416 languages) and number of meanings (i.e. 100 meanings) to later compare with our real linguistic dataset. We added different probabilities of errors (as previously tested) and a linguistic mask, so as to consider the same amount of missing data that is present in the real dataset. Once again, there was not a significant change in the power to discriminate between the different evolutionary models, which range between 15% and 66% (figure 2.22). These three different analyses (general power, power for simulated linguistic data with introduction of errors, and power with specific parameters that match the observed data) allow us to explore a broader range of possible scenarios to compare with the Bantu linguistic data, for which we do not have an idea about the amount of errors that the data can contain as well the generation and mutation rate of these languages.

Finally, the results provided for the genetic analysis were not so encouraging when only genomic data from three out of the five populations are available. The ability to identify the correct model among the different proposed models varied between 15% to 48% when we have genetic data for three populations (figure 2.23A), but when we have data for the five populations the power increase ranging between 29% - 77% (figure 2.23B). Due to the lack of genetic data for two populations, we noticed that we should use more than one individual to increase the accuracy of the model selection, contrasting with the results from Ghirotto *et al.* (2019) in which they showed that the accuracy of the method seemed to be more dependent on the number of loci and locus length considered than on the number of individuals sampled per population (Ghirotto *et al.*, 2019).

To understand if the model selected falls anyway within the main model even when the true model is not favoured (for instance, Late split river vs Late split not river), i.e., the proportion of false positives for each combination of experimental parameters, we calculated the proportion of TP and FP (i.e. a confusion matrix) for the six "main models" (Supplementary Table S2.1 and S2.2). When we sample one individual per population combined with a low number of loci and a small locus length, the Early split and Late split model are indistinguishable. We saw that since we only have complete genome data for three populations, the different topologies for the six models will have different weights and they will not allow to distinguish between the river – not river hypotheses. If we disregard the information for the topology river – not river and merge both scenarios for the Early split model and the Late plus the Very Late split scenarios, we can increase the power to distinguish the two big scenarios (Early and Late split). However, this procedure originates different scientific questions. If we only take into account the main models and ignore the topology river – not river, we are ignoring where the central Bantu languages, in the rainforest and largely in the inner Kongo basin, came from. This was the question already made by de Filippo *et al.* (2012). Another question can arise if we merge three river models into one, and three not river models into another, supposing we do not care much about the Eastern group, we can ask where the central languages came from.

2.5.2 Application to a real case: The Bantu expansions

The population history of the Bantu was doubtless a very complex process, with different expansionary phases, contractions in the last few centuries, explosive growth in post-colonial times. But we cannot capture all of that in a model, partly because the details are unknown, partly because it would blow up the complexity (de Filippo *et al.*, 2012; Barbieri, Vicente, *et al.*, 2014; Grollemund *et al.*, 2015; Choudhury *et al.*, 2018; Beyer *et al.*, 2019). So instead, we assigned a constant exponential-growth rate parameter to each of the five zones. The growth rate can be thought of as the natural rate induced by the environmental conditions of the geographical area. We also took into account migration (borrowing in the linguistic case) between each of the five zones.

Considering first the results for the linguistic data curated from Grollemund *et al.* (2015), we compared our observed linguistic data with the simulated data. Since we do not know the mutation rate for languages, we compared the real data against the simulated data including all theta values (0.5, 1, 5, 10, 20 and 50). We found support for the Late split river model with good posterior probabilities (0.51 to 0.55 – table 2.6) in agreement with previous studies (de Filippo *et al.*, 2012; Grollemund *et al.*, 2015; Patin *et al.*, 2017). However, our results present a somewhat ambiguous answer (see Supplementary Table S2.3). First, it is clear that the inference is not very strong, and second, the data suggests only mild preferences for some scenarios over others. It is possible that the amount of migration/borrowing on Bantu languages was large enough to almost erase the traces of the initial splits, at least as far as the lexical data is concerned (de Filippo *et al.*, 2012; Grollemund *et al.*, 2015; Beyer *et al.*, 2019). This is in agreement with previous findings using linguistic, genetic and environmental data in which it was hypothesized that the weaker signal found using different datasets is probably due to later contact among people and languages and the rapid nature of the expansion (de Filippo *et al.*, 2012; Barbieri, Vicente, *et al.*, 2014; Grollemund *et al.*, 2015; Choudhury *et al.*, 2018; Beyer *et al.*, 2019).

We then moved to investigating the Bantu demographic history using genomic data. We analysed the modern genome data of six Bantu individuals belonging to three different populations (NW, SW and E). We restricted the analysis to DNA stretches of 1.000 bp length, 20.000 loci and four chromosomes per population. Our genetic analysis supported the model Late split river (table 2.7) in agreement with the model chosen using linguistic data. However, when we plot the observed data against the simulated data, the observed data fall far away from the simulations (Supplementary Figure S2.3) meaning that it is necessary to adjust the demographic models and the prior distributions used, if we are to be closer to the real Bantu case. Analysing the observed versus the simulated data, we noticed that the genetic variability is higher in the real genomic data and that the simulated data generally fail to fall close to the observed ones. Since in our model we are introducing high migration and contact between populations, it is possible that we are forcing our simulated genetic data to be more homogeneous than they are in the reality. We should also take into consideration interactions with populations already present in the region, which contribute to the extensive genetic diversity observed in the current Bantu populations, as found in the works by Li *et al.* (2014) and Choudhury *et al.* (2017). Moreover, human genes carry genetic diversity generated over longer periods of time, while languages do not preserve well old signals. The demographic models that we designed generated genetic variation until 8.000 yBP, but the genetic variation present in Bantu population is for sure much older. To account for that, we should model the ancestral Bantu population (NW) as having been a part of a much larger human population.

Once we can solve this issue, we may be able to show that this novel ABC framework allows us to support with a high probability one of the models of expansion for the Bantu populations and also to better understand the importance of migration/borrowing between the different Bantu zones and give clues about the linguistic demographic parameters. Nonetheless, we need to take into account the fact that we do not have genetic data for the CW and WW populations, which can affect the power of inference of the demographic model for the Bantu expansion.

In conclusion, our linguistic and genomic analyses seem to support the Late split river model. However, they also point to the existence of confounding factors, which may complicate the analyses of these, as well as other, data. At the very start of the dispersal process, all members of the original population must be assumed to be speaking the same language; however, their DNAs are not identical, because they accumulated diversity through millennia of mutation and recombination processes. In other words, the linguistic diversity begins building up at a stage in which genetic diversity is certainly present, and conceivably extensive, especially if the expanding population occupies a broad territory and so has developed a spatial structure. As a consequence, the model best accounting for the linguistic patterns of variation may not necessarily be exactly the same that best accounts for patterns of variation at the genomic level. Among the possible solutions to this problem there is a change in the simulation strategy, allowing some time for genetic (but not linguistic) diversity to accumulate. This work is currently in progress, but many details are yet to be fixed.

Further analysis are needed to confirm these findings, especially in the light of the incomplete genomic sampling of the main population groups. The statistics chosen to summarize the genomic and linguistic data and the ABC-RF framework have demonstrated to be robust to discriminate simple evolutionary scenarios. We also emphasize the importance of performing a power analysis of the model's tested to be aware of the level of uncertainty and the importance of confirmed the results through different approaches (model selection and LDA).

This newly developed framework shows interesting results for the languages, but we need a stronger signal to be able to distinguish the main Bantu theories. Given the properties of the evolutionary process that we discovered with the help of our ABC-RF approach, we do not expect the lexical data to contain a strong signal of the original splits of populations. This is an objective limit of what we can do with the available linguistic data; moving from vocabulary comparisons to structural comparisons of grammar and syntax may be a way to increase the power of the tests (Longobardi and Guardiano, 2009; Longobardi *et al.*, 2015). However, as we have seen in the first part of this thesis, this approach requires collecting a large amount of complex linguistic information, a task that, to our knowledge, no one has even started in Africa so far. Potential genetic and linguistic signals might be erased by admixture, later contact among people and languages or maybe the divergences happened in a quite recent time. However, if we are able to detect through ABC the effect of "linguistic migration" (borrowing) between the different Bantu zones we will make a step forward in the linguistic field.

In the future, we will proceed with the parameter estimation and a posterior predictive control to validate that our real data looks like the simulated data. For the genetic framework, we need to find the demographic model that recreates the variability present in the real genetic data. After that, the idea is to join both linguistic and genomic data and perform the model choice. This framework is not limited to the analysis of the models proposed in this thesis, it can be applied to other studies regarding our species for which we have available cultural and genomic data.

General Conclusions

There is a link, if elusive, between the spread of languages and the genetic differences between the peoples who speak them. Traditional methods of comparative historical linguistics can help us understand the evolution of human populations and contacts that occurred between populations, but only up to a point. Combining genomics with linguistic data may turn out to be the best way to decipher subtle aspects of human biological and cultural evolution. Comparative analyses may help us identify the nature, whether reproductive or purely cultural, of contacts between populations, reveal discrepancies between genetic and language relationships indicative of recent language drift and infer the demographic history of populations from the recent past until ancient times.

In this thesis, we presented two different multidisciplinary studies combining linguistic and genomic data. We claim that for the comparison of languages belonging to different linguistic families, using syntactic data is not only indispensable, but also opens new research avenues that at present, are almost completely unexplored (chapter I). Conversely, for studies regarding a single language family, where etymological relationships among words can be safely established, lexical data are suitable, and certainly easier to collect than data on the deep structure of languages (chapter II). Both studies showed how it is possible to analyse in depth language variation obtaining new insight into the human past, coupled with the genomic information to reach beyond the time range attainable just through the linguistic information. In the course of these studies, we became aware that data quality is a key factor, both in terms of coverage of the geographical region considered, and of the sheer reliability of the information that we then process statistically. In addition, it is of fundamental importance to define evolutionary models that fit well with all the information available about the populations sampled and their history. This may be a hard task when we are dealing with populations that underwent a complex demographic history; but, on the other hand, it seems likely that all populations did. The rapid development of new and ever more efficient genomic technologies is leading to a fast accumulation of publicly available complete genomes, which in turn will help to improve our understanding of the complex nature of human cultural and genetic evolution.

Bibliography

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs R a, *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–73.

Abondolo D (1998). *The Uralic languages* (D Abondolo, Ed.). New York, London: Routledge.

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, *et al.* (2015). Population genomics of Bronze Age Eurasia. *Nature* **522**: 167–172.

Ammerman AJ, Cavalli-Sforza LL (1973). *A population model for the diffusion of early farming in Europe*. Duckworth.

Ammerman AJ, Cavalli-Sforza LL (1984). *The Neolithic Transition and the Genetics of Populations in Europe*. Princeton University Press.

Amorim CEG, Bisso-Machado R, Ramallo V, Bortolini MC, Bonatto SL, Salzano FM, *et al.* (2013). A bayesian approach to genome/linguistic relationships in native South Americans. *PLoS One* **8**: e64099–e64099.

Ashley CZ (2010). Towards a socialised archaeology of ceramics in great Lakes Africa. *African Archaeol Rev* **27**: 135–163.

Atkinson QD, Gray RD (2006). How Old is the Indo-European Language Family? Progress or more moths to the flame? In: Forster P, Renfrew C (eds) *Phylogenetic Methods and the Prehistory of Languages*, Cambridge, UK: McDonald Institute for Archaeological Research, pp 91–109.

Barbieri C, Güldemann T, Naumann C, Gerlach L, Berthold F, Nakagawa H, *et al.* (2014). Unraveling the complex maternal history of Southern African Khoisan populations. *Am*

J Phys Anthropol **153**: 435–448.

Barbieri C, Vicente M, Oliveira S, Bostoen K, Rocha J, Stoneking M, *et al.* (2014). Migration and interaction in a contact zone: mtDNA variation among Bantu-speakers in Southern Africa. *PLoS One* **9**.

Barbujani G (1997). DNA variation and language affinities. *Am J Hum Genet* **61**: 1011–1014.

Barbujani G (2013). Genetic Evidence for Prehistoric Demographic Changes in Europe. *Hum Hered* **76**: 133–141.

Barbujani G, Ghirotto S, Tassi F (2013). Nine things to remember about human genome diversity. *Tissue Antigens* **82**: 155–164.

Barbujani G, Pilastro A (1993). Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. *Proc Natl Acad Sci U S A* **90**: 4670–4673.

Barbujani G, Sokal RR, Oden NL (1995). Indo-European origins: A computer-simulation test of five hypotheses. *Am J Phys Anthropol* **96**: 109–132.

de Barros Damgaard P, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, *et al.* (2018). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**.

Bastin Y, Coupez A, Mann M (1999). Continuity and Divergence in the Bantu Languages: Perspectives from a Lexicostatistic Study. *Tervuren, Belgium Ann Sci Hum R Museum from Cent Africa*.

Beaumont MA (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annu Rev Ecol Evol Syst* **41**: 379–406.

Beaumont MA, Zhang W, Balding DJ (2002). Approximate Bayesian Computation in Population Genetics. *Genetics* **162**: 2025 LP – 2035.

Belle EMS, Barbujani G (2007). Worldwide analysis of multiple microsatellites: Language diversity has a detectable influence on DNA diversity. *Am J Phys Anthropol* **133**: 1137–1146.

Beltrame MH, Rubel MA, Tishkoff SA (2016). Inferences of African evolutionary history from genomic data. *Curr Opin Genet Dev* **41**: 159–166.

Benazzo A, Panziera A, Bertorelle G (2015). 4P: Fast computing of population genetics statistics from large DNA polymorphism panels. *Ecol Evol* **5**: 172–175.

Bertorelle G, Benazzo A, Mona S (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Mol Ecol* **19**: 2609–2625.

Beyer R, Singarayer JS, Stock JT, Manica A (2019). Environmental conditions do not predict diversification rates in the Bantu languages. *Heliyon* **5**: e02630.

Blum MGB, François O (2010). Non-linear regression models for Approximate Bayesian Computation. *Stat Comput* **20**: 63–73.

Bostoen K, Clist B, Doumenge C, Grollemund R, Hombert JM, Muluwa JK, *et al.* (2015). Middle to late holocene paleoclimatic change and the early bantu expansion in the rain forests of Western Central Africa. *Curr Anthropol* **56**: 354–384.

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko A V., Drummond AJ, *et al.* (2012). Mapping the origins and expansion of the Indo-European language family. *Science (80-)* **337**: 957–960.

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, *et al.* (2009). Genetic discontinuity between local hunter-gatherers and central Europe’s first farmers. *Science (80-)* **326**: 137–140.

Cann RL, Stoneking M, Wilson AC (1987). Mitochondrial DNA and human evolution. *Nature* **325**: 31–36.

Cavalli-Sforza LL (1997). *Geni, popoli e lingue*.

Cavalli-Sforza LL, Feldman MW (2003). The application of molecular genetic approaches to the study of human evolution. *Nat Genet* **33**: 266–275.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993). Demic expansions and human evolution. *Science (80-)* **259**: 639–646.

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci U S A* **85**: 6002–6.

Ceolin A (2019). Significance testing of the Altaic family. *Diachronica* **36**: 299–336.

Charlesworth B, Charlesworth D, Barton NH (2003). The Effects of Genetic and Geographic Structure on Neutral Variation. *Annu Rev Ecol Evol Syst* **34**: 99–125.

Chen J, Sokal R, Ruhlen M (1995). Worldwide analysis of genetic and linguistic relationships of human populations. *Hum Biol* **67**: 595–612.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G (1998). Clines of nuclear DNA markers suggest a largely Neolithic ancestry of the European gene pool. *Proc Natl Acad Sci* **95**: 9053–9058.

Chikhi L, Nichols RA, Barbujani G, Beaumont MA (2002). Y genetic data support the Neolithic demic diffusion model. *PNAS* **99**.

Choudhury A, Aron S, Sengupta D, Hazelhurst S, Ramsay M (2018). African genetic diversity provides novel insights into evolutionary history and local adaptations. *Hum Mol Genet* **27**: R209–R218.

Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, *et al.* (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat Commun* **8**: 1–12.

Clist B-O (1987). A Critical Reappraisal of the Chronological Framework of the Early Iron Age Urewe Industry. *MUNTU*: 35–62.

Colonna V, Boattini A, Guardiano C, Dall'Ara I, Pettener D, Longobardi G, *et al.* (2010). Long-range comparison between genes and languages based on syntactic distances. *Hum Hered* **70**: 245–254.

Consortium T 1000 GP, Auton A, Abecasis GR, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, *et al.* (2015). A global reference for human genetic variation. *Nature* **526**: 68.

Creanza N, Ruhlen M, Pemberton TJ, Rosenberg NA, Feldman MW, Ramachandran S (2015). A comparison of worldwide phonemic and genetic variation in human populations. *Proc Natl Acad Sci U S A* **112**: 1265–72.

Crystal D (2011). *A Dictionary of Linguistics and Phonetics*, 6th edn. Blackwell Publishing.

Csányi B, Bogácsi-Szabó E, Tömöry G, Czibula Á, Priskin K, Csõsz A, *et al.* (2008). Y-Chromosome Analysis of Ancient Hungarian and Two Modern Hungarian-Speaking Populations from the Carpathian Basin. *Ann Hum Genet* **72**: 519–534.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol* **25**: 410–418.

Delaneau O, Marchini J, Zagury JF (2012). A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**: 179–181.

Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET (2017). A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8**: 1–7.

Diamond J, Bellwood P (2003). Farmers and Their Languages: The First Expansions. *Science (80-)* **300**: 597 LP – 603.

Dunn M (2005). Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science (80-)* **309**: 2072–2075.

Estoup A, Raynal L, Verdu P, Marin J (2018). Model choice using Approximate Bayesian Computation and Random Forests : analyses based on model grouping to make inferences about the genetic history of Pygmy human populations. *J la Société Française Stat* **159**: 167–190.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**: e1003905–e1003905.

Excoffier L, Schneider S (1999). Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc Natl Acad Sci U S A* **96**: 10597–10602.

Fan S, Kelly DE, Beltrame MH, Hansen MEB, Mallick S, Ranciaro A, *et al.* (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol* **20**: 82.

de Filippo C, Barbieri C, Whitten M, Mpoloka SW, Gunnarsdóttir ED, Bostoen K, *et al.* (2010). Y-Chromosomal Variation in Sub-Saharan Africa: Insights Into the History of Niger-Congo Groups. *Mol Biol Evol* **28**: 1255–1269.

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

Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, *et al.* (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nat Commun* **5**: 1–9.

Garrigan D, Hammer MF (2006). Reconstructing human origins in the genomic era. *Nat Rev Genet* **7**: 669–680.

Georg S, Michalove PA, Ramer AM, Sidwell PJ (1999). *Telling General Linguists about Altaic*.

Ghirotto S, Vizzari MT, Tassi F, Barbujani G, Benazzo A (2019). Distinguishing among complex evolutionary models using unphased whole-genome data through Approximate Bayesian Computation. *pre-print*.

Gimbutas M (1979). The Three Waves of Kurgan People into Old Europe, 4500–2500 BC. *Arch suisses d'anthropologie genérale* **43**: 113–137.

Goldstein DB, Chikhi L (2002). HUMAN MIGRATIONS AND POPULATION STRUCTURE: What We Know and Why it Matters. *Annu Rev Genomics Hum Genet* **3**: 129–152.

González-Fortes G, Jones ER, Lightfoot E, Bonsall C, Lazar C, Grandal-d'Anglade A, *et al.* (2017). Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. *Curr Biol* **27**: 1801-1810.e10.

Gray RD, Atkinson QD (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**: 435–439.

Gray RD, Drummond AJ, Greenhill SJ (2009). Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science (80-)* **323**: 479–483.

Greenberg JH (2000). *Indo-European and Its Closest Relatives: The Eurasiatic Language Family, Volume 1, Grammar*. Stanford University Press: Stanford.

Greenberg JH (2002). *Indo-European and Its Closest Relatives: The Eurasiatic Language Family, Volume 2, Lexicon*. Stanford University Press: Stanford.

Greenhill SJ, Atkinson QD, Meade A, Gray RD (2010). The shape and tempo of language evolution. *Proc R Soc B Biol Sci* **277**: 2443–2450.

Greenhill SJ, Wu C-H, Hua X, Dunn M, Levinson SC, Gray RD (2017). Evolutionary dynamics of language systems. *Proc Natl Acad Sci*: 201700388.

Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proc Natl Acad Sci U S A* **112**: 13296–13301.

Guthrie M (1962). Some Developments in the Prehistory of the Bantu Languages. *J Afr Hist* **3**: 273–282.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, *et al.* (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**: 207–211.

Hammer O, Harper D, Ryan P (2001). PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontol Electron* **4**: 1–9.

Harding RM, McVean G (2004). A structured ancestral population for the evolution of modern humans. *Curr Opin Genet Dev* **14**: 667–674.

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998). Genetic traces of ancient demography. *Proc Natl Acad Sci* **95**: 1961 LP – 1967.

Harpending H, Rogers A (2000). Genetic Perspectives on Human Origins and Differentiation. *Annu Rev Genomics Hum Genet* **1**: 361–385.

Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, *et al.* (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* **8**.

Henn BM, Steele TE, Weaver TD (2018). Clarifying distinct models of modern human origins in Africa. *Curr Opin Genet Dev* **53**: 148–156.

Hernandez RD, Williamson SH, Bustamante CD (2007). Context Dependence, Ancestral Misidentification, and Spurious Signatures of Natural Selection. *Mol Biol Evol* **24**: 1792–1800.

Hey J, Machado CA (2003). The study of structured populations — new hope for a difficult and divided science. *Nat Rev Genet* **4**: 535–543.

Heyer E, Mennecier P (2009). Genetic and linguistic diversity in Central Asia. In D'Errico F., Hombert J-M., eds. Becoming eloquent. In: *Becoming Eloquent: Advances in the emergence of language, human cognition, and modern cultures*, John Bejamins Puslishing Company, pp 163–180.

Hinrichs AS, Raney BJ, Speir ML, Rhead B, Casper J, Karolchik D, *et al.* (2016). UCSC Data Integrator and Variant Annotation Integrator. *Bioinformatics* **32**: 1430–1432.

Holden CJ (2002). Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proc R Soc B Biol Sci* **269**: 793–799.

Honkola T, Vesakoski O, Korhonen K, Lehtinen J, Syrjänen K, Wahlberg N (2013). Cultural and climatic changes shape the evolutionary history of the Uralic languages. *J Evol Biol* **26**: 1244–1253.

Hubby JL, Lewontin RC (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. *Genetics* **54**: 577–594.

Hudson RR (1990). *Gene genealogies and the coalescent process*.

Hudson RR (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.

Huffman TN (1970). The Early Iron Age and the Spread of the Bantu. **25**: 3–21.

Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* **408**: 708–713.

Janhunen J (2009a). *Proto-Uralic: what, where, and when? In The quasquicentennial of the Finno-Ugrian Society*. Helsinki: Suomalais-Ugrilainen Seura.

Janhunen J (2009b). Proto-Uralic–what, where, and when? *Mémoires la Société Finno-Ougrienne* **258**: 57–78.

Jeong C, Balanovsky O, Lukianova E, Kahbatkyzy N, Flegontov P, Zaporozhchenko V, *et al.* (2019). The genetic history of admixture across inner Eurasia. *Nat Ecol Evol* **3**: 966–976.

Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC, Cavalli-Sforza LL (1983). Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. *J Mol Evol* **19**: 255–271.

Jones ER, Zarina G, Moiseyev V, Lightfoot E, Nigst PR, Manica A, *et al.* (2017). The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. *Curr Biol* **27**: 576–582.

Joyce P, Marjoram P (2008). Approximately Sufficient Statistics and Bayesian Computation. *Stat Appl Genet Mol Biol* **7**.

Kaiser M, Shevoroshkin V (1988). Nostratic. *Annu Rev Anthropol* **17**: 309–329.

Kallio P (2006). Suomen kantakielen absoluuttista kronologiaa. *Virittäjä* **110**: 2–25.

Keightley PD, Jackson BC (2018). Inferring the Probability of the Derived &lt;em&gt;vs.&lt;/em&gt; the Ancestral Allelic State at a Polymorphic Site. *Genetics* **209**: 897 LP – 906.

Kerminen S, Havulinna AS, Hellenthal G, Martin AR, Sarin A-P, Perola M, *et al.* (2017). Fine-Scale Genetic Structure in Finland. *G3 (Bethesda)* **7**: 3459–3468.

Kingman JFC (1982a). The coalescent. *Stoch Process their Appl* **13**: 235–248.

Kingman JFC (1982b). On the genealogy of large populations. *J Appl Probab* **19**: 27–43.

Kivikoski E (1961). Suomen Historia I: Suomen Esihistoria. *Werner-Söderström Oy, Porvoo*.

Kraaijenbrink T, van der Gaag KJ, Zuniga SB, Xue Y, Carvalho-Silva DR, Tyler-Smith C, *et al.* (2014). A linguistically informed autosomal STR survey of human populations residing in the greater Himalayan region. *PLoS One* **9**: e91534–e91534.

Lamnidis TC, Majander K, Jeong C, Salmela E, Wessman A, Moiseyev V, *et al.* (2018). Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat Commun* **9**.

Lapierre M, Lambert A, Achaz G (2017). Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population. *Genetics* **206**: 439–449.

Lappalainen T, Koivumäki S, Salmela E, Huoponen K, Sistonen P, Savontaus M-L, *et al.*

(2006). Regional differences among the Finns: a Y-chromosomal perspective. *Gene* **376**: 207–15.

Lawson DJ, Hellenthal G, Myers S, Falush D (2012). Inference of population structure using dense haplotype data. *PLoS Genet* **8**: 11–17.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, *et al.* (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**: 419–424.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, *et al.* (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**: 409–413.

Lehtinen J, Honkola T, Korhonen K, Syrjänen K, Wahlberg N, Vesakoski O (2014). Behind Family Trees: Secondary Connections in Uralic Language Networks. *Lang Dyn Chang* **4**: 189–221.

Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291.

Levinson SC, Gray RD (2012). Tools from evolutionary biology shed new light on the diversification of languages. *Trends Cogn Sci* **16**: 167–173.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, *et al.* (2007). The diploid genome sequence of an individual human. *PLoS Biol* **5**: 2113–2144.

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

Li H, Durbin R (2012). Inference of Human Population History From Whole Genome Sequence of A Single Individual. *Nature* **475**: 493–496.

Li S, Schlebusch C, Jakobsson M (2014). Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proc R Soc B Biol Sci* **281**.

Liu X, Fu Y-X (2015). Exploring population size changes using SNP frequency spectra. *Nat Genet* **47**: 555–559.

Longobardi G (2003). Methods in parametric linguistics and cognitive history. *Linguist Var*

Yearb **3**: 101–138.

Longobardi G, Ghirotto S, Guardiano C, Tassi F, Benazzo A, Ceolin A, *et al.* (2015). Across language families: Genome diversity mirrors linguistic variation within Europe. *Am J Phys Anthropol* **157**: 630–640.

Longobardi G, Guardiano C (2009). Evidence for syntax as a signal of historical relatedness. *Lingua* **119**: 1679–1706.

Longobardi G, Guardiano C (2013). Toward a syntactic phylogeny of modern Indo-European languages. *J Hist Linguist* **3**: 122–152.

Maley J (2001). La destruction catastrophique des forêts d'Afrique Centrale survenue il y a environ 2500 ans exerce encore une influence majeure sur la répartition actuelle des formations végétales. *Syst Geogr Plants* **71**.

Maley J, Doumenge C, Giresse P, Mahé G, Philippon N, Hubau W, *et al.* (2017). Late Holocene forest contraction and fragmentation in central Africa. *Quat Res (United States)* **89**: 43–59.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, *et al.* (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, *et al.* (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**: 499–503.

Menozzi P, Piazza A, Cavalli-Sforza L (1978). Synthetic Maps of Human Gene Frequencies in Europeans. **201**: 786–792.

Miettinen T (1996). *Suomenlahden ulkosaarten esihistoriaa. In: Suomenlahden ulkosaaret: Lavansaari, Seiskari, Suursaari, Tytärsaari (R. Hamari, M. Korhonen, Miettinen Timo & I. Talve, eds)*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, *et al.* (2019). The formation of human populations in South and Central Asia. *Science (80-)* **365**: eaat7487.

Neparáczki E, Kocsy K, Tóth GE, Maróti Z, Kalmár T, Bihari P, *et al.* (2017). Revising mtDNA haplotypes of the ancient Hungarian conquerors with next generation sequencing. *PLoS One* **12**.

Neumann K, Bostoen K, Höhn A, Kahlheber S, Ngomanda A, Tchiengue B (2012). First Farmers in the Central African Rainforest: a View from Southern Cameroon. *Quat Int* **249**: 53–62.

Newberry MG, Ahern CA, Clark R, Plotkin JB (2017). Detecting evolutionary forces in language change. *Nature* **551**: 223–226.

Nordborg M (2001). Coalescent theory. Handbook of statistical genetics. In: Balding D, Bishop M& Cannings C, pp 179-212.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, *et al.* (2008). Genes mirror geography within Europe. *Nature* **456**: 98–101.

Nurse D, Philippson G (2003). *The Bantu Languages*. Taylor & Francis.

Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, *et al.* (2019). The genomic history of the Iberian Peninsula over the past 8000 years. *Science (80-)* **363**: 1230–1234.

Olalde I, Schroeder H, Sandoval-Velasco M, Vinner L, Lobón I, Ramirez O, *et al.* (2015). A common genetic origin for early farmers from mediterranean cardial and central european LBK cultures. *Mol Biol Evol* **32**: 3132–3142.

Pagani L, Lawson J, Jagoda E, Mörseburg A, Clemente F, Hudjashov G, *et al.* (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*.

Pakendorf B, Bostoen K, De Filippo C (2011). Molecular Perspectives on the Bantu Expansion: A Synthesis. *Lang Dyn Chang* **1**: 50–88.

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, *et al.* (2009). Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* **5**.

Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, *et al.* (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science (80-)* **356**: 543–546.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, *et al.* (2012). Ancient

Admixture in Human History. *Genetics* **192**: 1065 LP – 1093.

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, *et al.* (2012). The genetic prehistory of southern Africa. *Nat Commun* **3**: 1–6.

Poloni ES, Passarino G, Santachiara-Benerecetti AS, Langaney A, Excoffier L, Poloni E (1997). *Human Genetic Affinities for Y-Chromosome P49a,f/Taql Haplotypes Show Strong Correspondence with Linguistics*.

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP (2015). Reliable ABC model choice via random forests. *Bioinformatics* **32**: 859–866.

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet* **23**: 437–441.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, *et al.* (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**: 87–91.

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, *et al.* (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science (80-)* **349**: aab3884.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, *et al.* (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**: 757–762.

Renfrew C (1987). *Archaeology and Language: The Puzzle of Indo-European Origins*.

Renfrew C (1991). Before Babel: Speculations on the Origins of Linguistic Diversity. *Cambridge Archaeol J* **1**: 3–23.

Renfrew C (1992). Archaeology, Genetics and Linguistic Diversity. *R Anthropol Inst Gt Britain Irel* **27**: 445–478.

Robbeets M (2005). *Is Japanese related to Korean, Tungusic, Mongolic and Turkic?* Wiesbaden: Harrassowitz.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, *et al.* (2002). Genetic Structure of Human Populations. *Science (80-)* **298**: 2381 LP – 2385.

Rowe BM, Levine DP (2015). *A Concise Introduction to Linguistics*. Routledge.

Ruhlen M (1991). Survey of the world's languages.

Ruhlen M (1994). *The origin of language*. John Wley & Sons.

Saag L, Laneman M, Varul L, Malve M, Valk H, Razzak MA, *et al.* (2019). The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Curr Biol* **29**: 1701-1711.e16.

Salas A, Richards M, De la Fe T, Lareu M-V, Sobrino B, Sánchez-Diz P, *et al.* (2002). The making of the African mtDNA landscape. *Am J Hum Genet* **71**: 1082–1111.

Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A, *et al.* (2008). Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS One* **3**.

Scerri EML, Chikhi L, Thomas MG (2019). Beyond multiregional and simple out-of-Africa models of human evolution. *Nat Ecol Evol* **3**: 1370–1372.

Scerri EML, Thomas MG, Manica A, Gunz P, Stock JT, Stringer C, *et al.* (2018). Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends Ecol Evol* **33**: 582–594.

Schiffels S, Durbin R (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**: 919–925.

Schlebusch CM, Jakobsson M (2018). Tales of Human Migration, Admixture, and Selection in Africa. *Annu Rev Genomics Hum Genet* **19**: 405–428.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, *et al.* (2017). Ancient genomes from southern Africa pushes modern human divergence beyond 260,000 years ago. *bioRxiv*: 145409.

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Blum MGB, Soodyall H, *et al.* (2012). Genomic Variation in Seven Khoe-San. **1187**: 374–379.

Schraiber JG, Akey JM (2015). Methods and models for unravelling human evolutionary history. *Nat Rev Genet* **16**: 727–740.

Shijulal NS, List JM, Geisler H, Fangerau H, Gray RD, Martin W, *et al.* (2011). Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proc R Soc B Biol Sci* **278**: 1794–1803.

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.

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, *et al.* (2015). Genetic evidence for two founding populations of the Americas. *Nature* **525**: 104–108.

Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, *et al.* (2017). Reconstructing Prehistoric African Population Structure. *Cell* **171**: 59-71.e21.

Smith ML, Ruffley M, Espíndola A, Tank DC, Sullivan J, Carstens BC (2017). Demographic model selection using random forests and the site frequency spectrum. *Mol Ecol* **26**: 4562–4573.

Sokal RR (1988). Genetic, geographic, and linguistic distances in Europe. *Proc Natl Acad Sci* **85**: 1722–1726.

Sokal RR, Menozzi P (1982). Spatial Autocorrelations of HLA Frequencies in Europe Support Demic Diffusion of Early Farmers. *Univ Chicago Press Am Soc Nat* **Vol. 119**: 1–17.

Sokal RR, Oden NL, Legendre P, Fortin M-J, Kim J, Thomson BA, *et al.* (1990). Genetics and Language in European Populations. : 157–175.

Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585 LP – 595.

Tambets K, Yunusbayev B, Hudjashov G, Ilumäe AM, Rootsi S, Honkola T, *et al.* (2018). Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol* **19**.

Tennessen JA, Bigham AW, O’Connor TD, Fu W, Kenny EE, Gravel S, *et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**: 64–69.

Terhorst J, Kamm JA, Song YS (2016). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49**: 303–309.

Terhorst J, Kamm JA, Song YS (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49**: 303–309.

Thomason SG, Kaufman T (1988). *Language Contact, Creolization, and Genetic Linguistics*. University of California Press.

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW (2000). Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc Natl Acad Sci* **97**: 7360 LP – 7365.

Thouzeau V, Mennecier P, Verdu P, Austerlitz F (2017). Genetic and linguistic histories in Central Asia inferred using approximate Bayesian computations. *Proc R Soc B Biol Sci* **284**: 20170706.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, *et al.* (2009). The genetic structure and history of Africans and African Americans. *Science* **324**: 1035–1044.

Tishkoff SA, Williams SM (2002). Genetic analysis of African populations: Human evolution and complex disease. *Nat Rev Genet* **3**: 611–621.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, *et al.* (2000). Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**: 358–361.

Vansina J (1995). New Linguistic Evidence and 'the Bantu Expansion'. **36**: 173–195.

Verdu P, Becker NSA, Froment A, Georges M, Grugni V, Quintana-Murci L, *et al.* (2013). Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Mol Biol Evol* **30**: 918–937.

Vovin A (2005). The end of the Altaic controversy: In memory of Gerhard Doerfer. **49**: 71–132.

Wakeley J (1999). Nonequilibrium migration in human history. *Genetics* **153**: 1863–1871.

Wakeley J, Hey J (1997). Estimating Ancestral Population Parameters. *Genetics* **145**: 847 LP – 855.

Watson E, Forster P, Richards M, Bandelt HJ (1997). Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* **61**: 691–704.

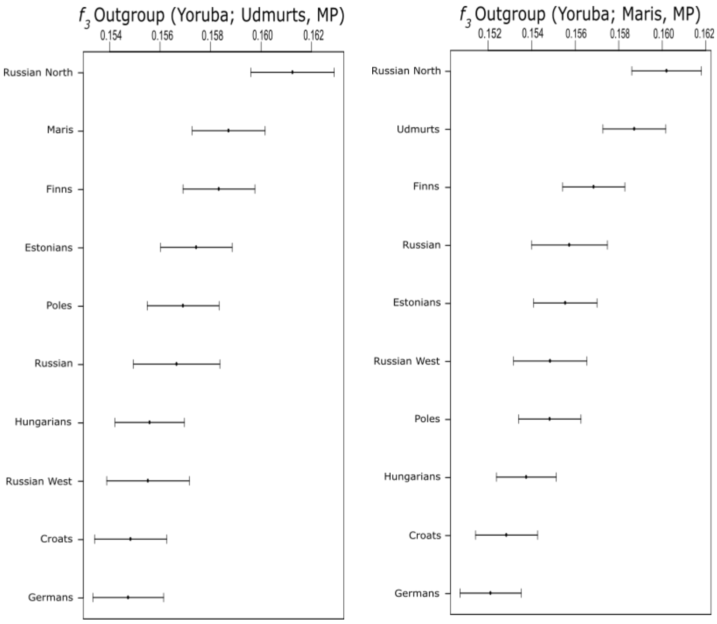
Wheeler D a, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, *et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–6.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, *et al.* (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet* **13**: 867–876.

Supplementary Materials for

CHAPTER I: Parallel signatures of past human migration in linguistic and genomic diversity of Western/Central Eurasia

Supplementary Figures



Supplementary Figure S1.1. Outgroup *f3*-statistics analysis. Shared genetic drift between modern Pontic steppes populations and modern European populations (MP).

Supplementary Tables

Supplementary Table S1.1. Ancient DNA samples used in this study.

Sample ID	Archaeological Culture	Date	Country	Region	Contributor
I1100	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1102	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1099	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1103	Anatolia_Neolithic	6400-5600 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1101	Anatolia_Neolithic	6400-5600 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1097	Anatolia_Neolithic	6400-5600 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0744	Anatolia_Neolithic	6400-5600 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1579	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1581	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1096	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1580	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1098	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1585	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0708	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0745	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015

I0746	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I1583	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0707	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0709	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0725	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0727	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0724	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0736	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0723	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0726	Anatolia_Neolithic	6500-6200 calBCE	Turkey	Barcın	Mathieson *et al.* 2015
I0231	Yamnaya	2910-2875 calBCE	Russia	Ekaterinovka, Southern Steppe, Samara	Haak *et al.* 2015
I0357	Yamnaya	3090-2910 calBCE	Russia	Lopatino I, Sok River, Samara	Haak *et al.* 2015
I0370	Yamnaya	3500-2700 calBCE	Russia	Ishkinovka I, Eastern Orenburg, Pre-Ural steppe, Samara	Haak *et al.* 2015
I0429	Yamnaya	3339-2917 calBCE	Russia	Lopatino I, Sok River, Samara	Haak *et al.* 2015
I0438	Yamnaya	3021-2635 calBCE	Russia	Luzkhi I, Samara River, Samara	Haak *et al.* 2015
I0439	Yamnaya	3305-2925 calBCE	Russia	Lopatino I, Sok River, Samara	Haak *et al.* 2015
I0441	Yamnaya	3010-2622 calBCE	Russia	Kurmanaevka III, Buzuluk, Samara	Haak *et al.* 2015
I0443	Yamnaya	3335-2912 calBCE	Russia	Grachevka II, Sok_River, Samara	Haak *et al.* 2015

I0444	Yamnaya	3335-2881 calBCE	Russia	Kutuluk I, Kutuluk River, Samara	Haak *et al.* 2015
RISE386	Sintashta	2298-2045 calBCE	Russia	Bulanovo	Allentoft *et al.* 2015
RISE391	Sintashta	2120-1887 calBCE	Kazakhstan	Tanabergen II	Allentoft *et al.* 2015
RISE392	Sintashta	2126-1896 calBCE	Russia	Stepnoe VII	Allentoft *et al.* 2015
RISE394	Sintashta	1949-1754 calBCE	Russia	Bulanovo	Allentoft *et al.* 2015
RISE395	Sintashta	1960-1756 calBCE	Russia	Bol'shekaraganskii	Allentoft *et al.* 2015

Supplementary Table S1.2. Human Origins data on present-day humans used in this study.

Sample ID	Population	Country	Region	Contributor
Nov_005	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00514	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00513	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00509	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00512	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00504	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00507	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00511	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00510	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00508	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
ADR00515	Nganasan	Russia	Central Asia Siberia	Lazaridis *et al.* 2014
HGDP00774	Han	China	East Asia	Patterson *et al.* 2012
HGDP00775	Han	China	East Asia	Patterson *et al.* 2012
HGDP00776	Han	China	East Asia	Patterson *et al.* 2012
HGDP00777	Han	China	East Asia	Patterson *et al.* 2012
HGDP00779	Han	China	East Asia	Patterson *et al.* 2012
HGDP00780	Han	China	East Asia	Patterson *et al.* 2012
HGDP00781	Han	China	East Asia	Patterson *et al.* 2012
HGDP00782	Han	China	East Asia	Patterson *et al.* 2012
HGDP00783	Han	China	East Asia	Patterson *et al.* 2012
HGDP00784	Han	China	East Asia	Patterson *et al.* 2012
HGDP00785	Han	China	East Asia	Patterson *et al.* 2012
HGDP00786	Han	China	East Asia	Patterson *et al.* 2012
HGDP00811	Han	China	East Asia	Patterson *et al.* 2012
HGDP00812	Han	China	East Asia	Patterson *et al.* 2012

HGDP00813	Han	China	East Asia	Patterson *et al.* 2012
HGDP00814	Han	China	East Asia	Patterson *et al.* 2012
HGDP00815	Han	China	East Asia	Patterson *et al.* 2012
HGDP00817	Han	China	East Asia	Patterson *et al.* 2012
HGDP00818	Han	China	East Asia	Patterson *et al.* 2012
HGDP00819	Han	China	East Asia	Patterson *et al.* 2012
HGDP00820	Han	China	East Asia	Patterson *et al.* 2012
HGDP00821	Han	China	East Asia	Patterson *et al.* 2012
HGDP00822	Han	China	East Asia	Patterson *et al.* 2012
HGDP00971	Han	China	East Asia	Patterson *et al.* 2012
HGDP00972	Han	China	East Asia	Patterson *et al.* 2012
HGDP00973	Han	China	East Asia	Patterson *et al.* 2012
HGDP00974	Han	China	East Asia	Patterson *et al.* 2012
HGDP00975	Han	China	East Asia	Patterson *et al.* 2012
HGDP00976	Han	China	East Asia	Patterson *et al.* 2012
HGDP00977	Han	China	East Asia	Patterson *et al.* 2012
HGDP01021	Han	China	East Asia	Patterson *et al.* 2012
HGDP01023	Han	China	East Asia	Patterson *et al.* 2012
HGDP01024	Han	China	East Asia	Patterson *et al.* 2012
HGDP00449	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00462	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00463	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00467	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00474	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00476	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00478	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00982	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00984	Mbuti	Congo	Africa	Patterson *et al.* 2012

HGDP01081	Mbuti	Congo	Africa	Patterson *et al.* 2012
HGDP00995	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP00999	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01001	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01003	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01006	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01010	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01012	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01013	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01014	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01015	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01018	Karitiana	Brazil	America	Patterson *et al.* 2012
HGDP01019	Karitiana	Brazil	America	Patterson *et al.* 2012
Ul5	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul31	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul65	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul6	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul33	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul71	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul10	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul43	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul72	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul44	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul74	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul19	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul24	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul59	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul56	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov

Ul55	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul16	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul69	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul1	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul36	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul25	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul52	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul70	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul51	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
Ul39	Ulchi	Russia	Central Asia Siberia	Rem Sukernik / Stanislav Dryomov
mixe0029	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0030	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0015	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0035	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0018	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0026	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0027	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0028	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0007	Mixe	Mexico	America	William Klitz / Cheryl Winkle
mixe0009	Mixe	Mexico	America	William Klitz / Cheryl Winkle

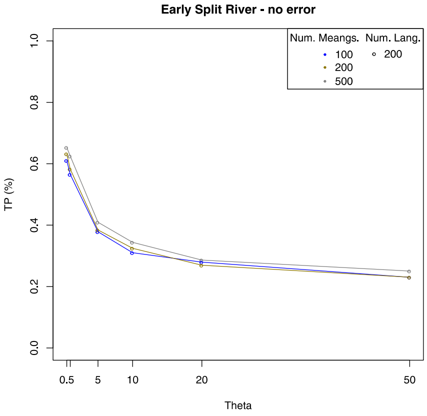
Supplementary Table S1.3. qp*Adm* models.

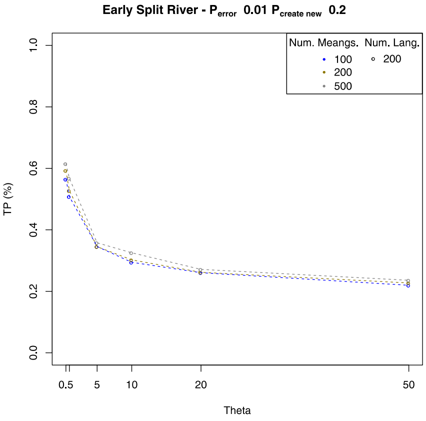
Test	Outgroup set	Nganasan	Yamnaya	Anatolia	chi-square
Khanty	Han, Mbuti, Karitiana, Ulchi and Mixe	0.521	0.479	0	10.056
Maris	Han, Mbuti, Karitiana, Ulchi and Mixe	0.281	0.465	0.254	10.493
Udmurts	Han, Mbuti, Karitiana, Ulchi and Mixe	0.261	0.611	0.128	9.032
Iranians	Han, Mbuti, Karitiana, Ulchi and Mixe	0.016	0.141	0.843	7.053
Finns	Han, Mbuti, Karitiana, Ulchi and Mixe	0.101	0.589	0.31	6.623
Estonians	Han, Mbuti, Karitiana, Ulchi and Mixe	0.04	0.568	0.391	8.095
Hungarians	Han, Mbuti, Karitiana, Ulchi and Mixe	0.032	0.412	0.556	8.398
Russian North	Han, Mbuti, Karitiana, Ulchi and Mixe	0.144	0.571	0.284	2.255
Russian	Han, Mbuti, Karitiana, Ulchi and Mixe	0.042	0.517	0.441	5.389
Russian West	Han, Mbuti, Karitiana, Ulchi and Mixe	0.035	0.526	0.440	0.819
Croats	Han, Mbuti, Karitiana, Ulchi and Mixe	0.042	0.303	0.655	4.373
Germans	Han, Mbuti, Karitiana, Ulchi and Mixe	0.027	0.373	0.6	5.999
Poles	Han, Mbuti, Karitiana, Ulchi and Mixe	0.023	0.561	0.416	2.634

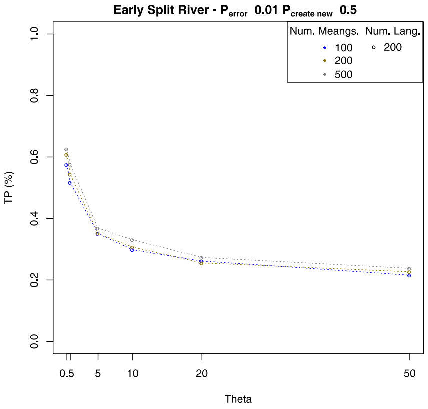
Supplementary Material for

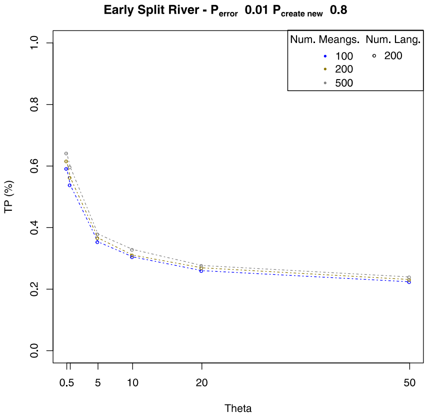
CHAPTER II: Integrating genomic and linguistic data through a new ABC framework - Explaining the Bantu expansion: Early or Late split hypothesis

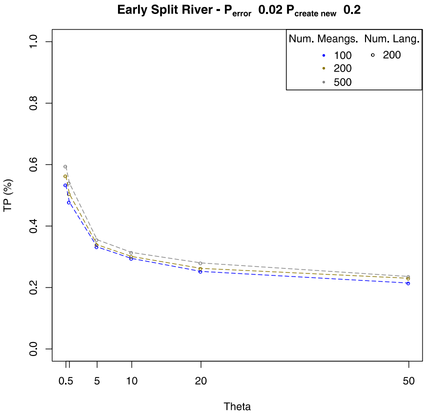
Supplementary Figures

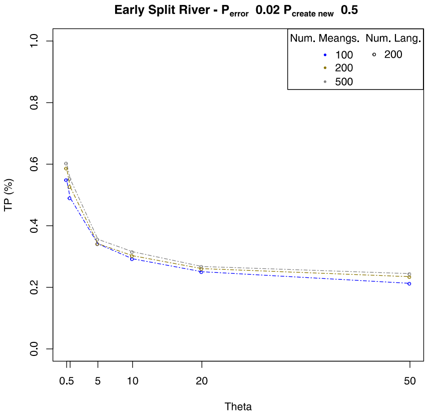


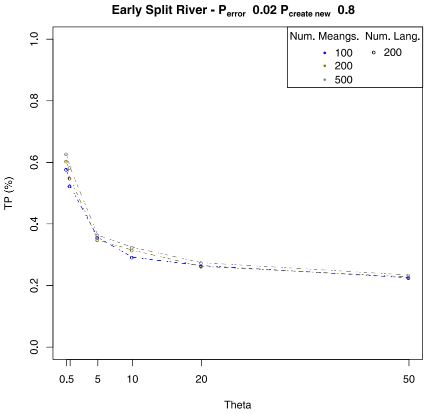


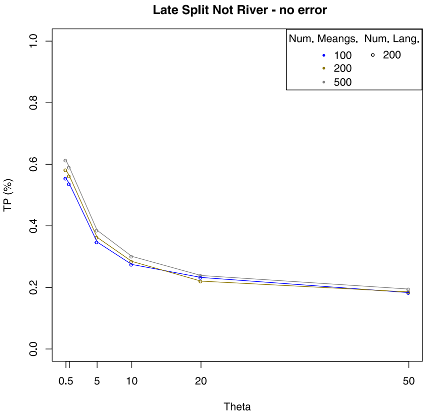


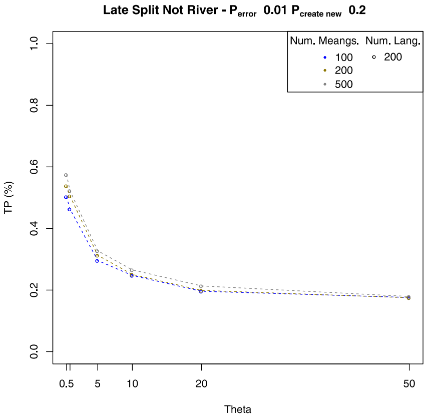




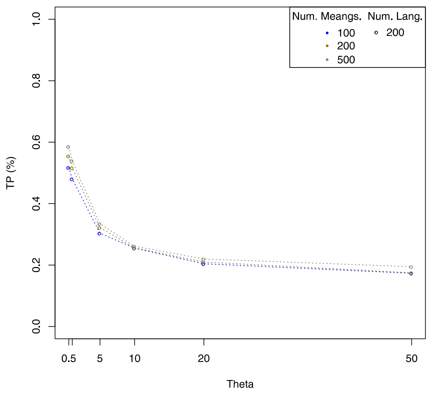




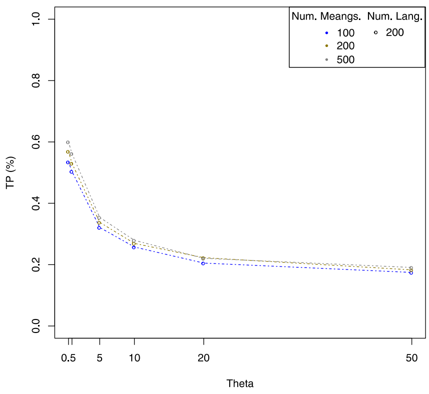




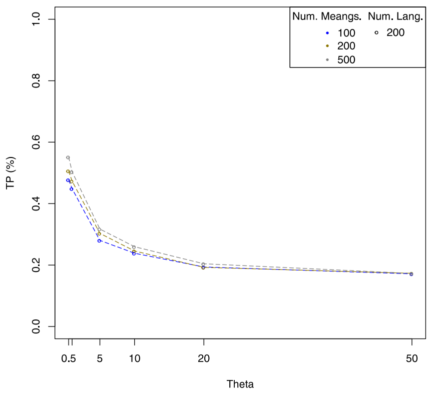
Late Split Not River - P_{error} 0.01 $P_{create\ new}$ 0.5



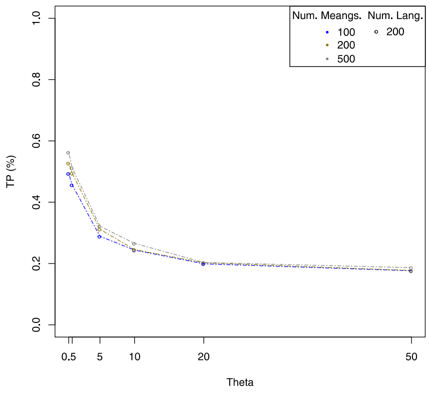
Late Split Not River - P_{error} 0.01 $P_{create\ new}$ 0.8



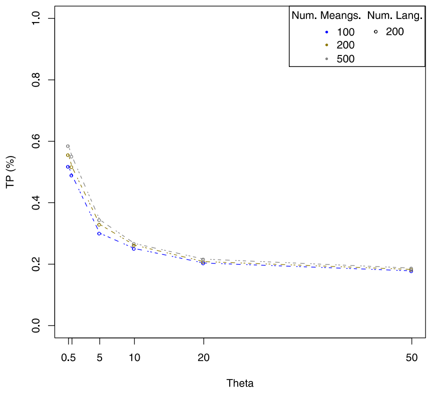
Late Split Not River - P_{error} 0.02 $P_{create\ new}$ 0.2



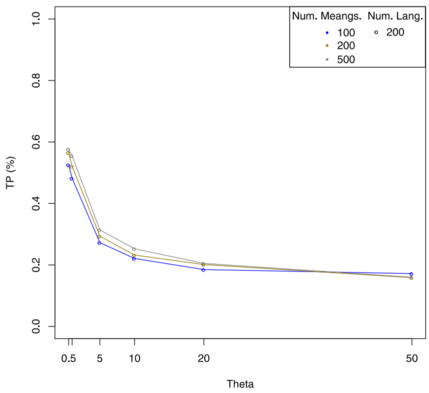
Late Split Not River - P_{error} 0.02 $P_{create\ new}$ 0.5



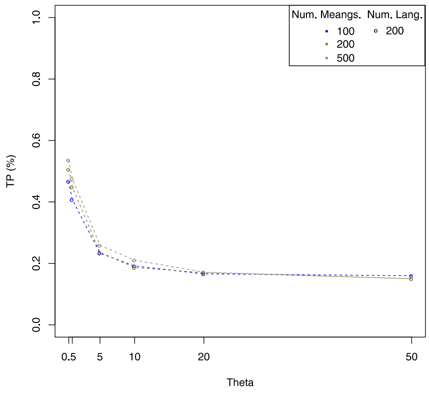
Late Split Not River - P_{error} 0.02 $P_{create\ new}$ 0.8



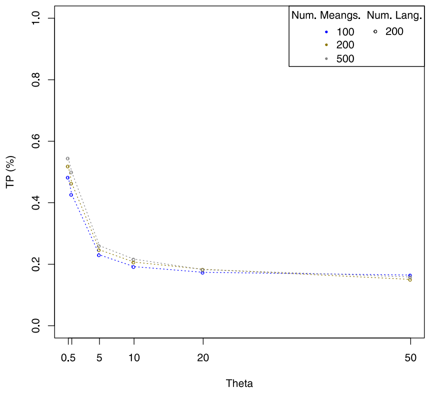
Late Split River - no error



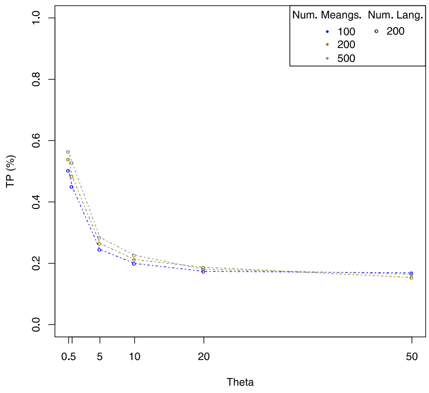
Late Split River - P_{error} 0.01 $P_{create\ new}$ 0.2

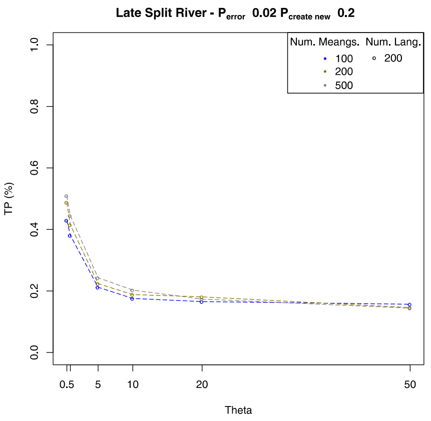


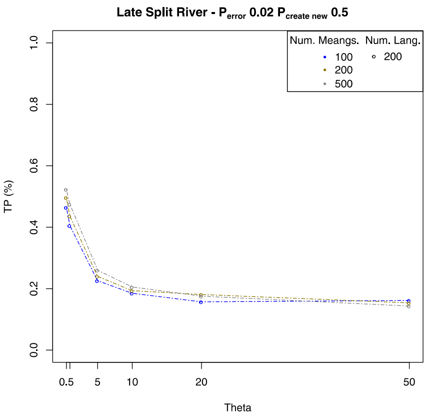
Late Split River - P_{error} 0.01 $P_{create\ new}$ 0.5

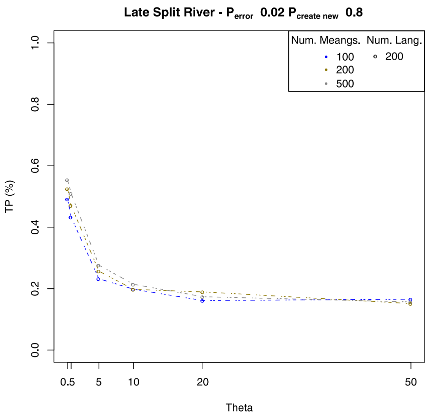


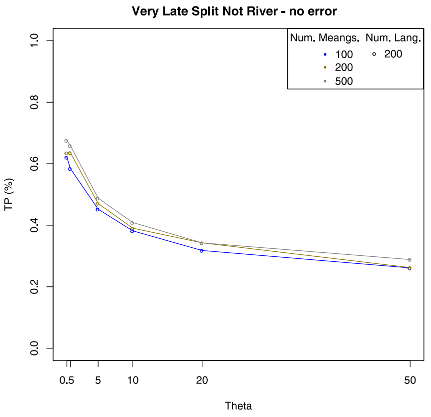
Late Split River - P_{error} 0.01 $P_{create\ new}$ 0.8

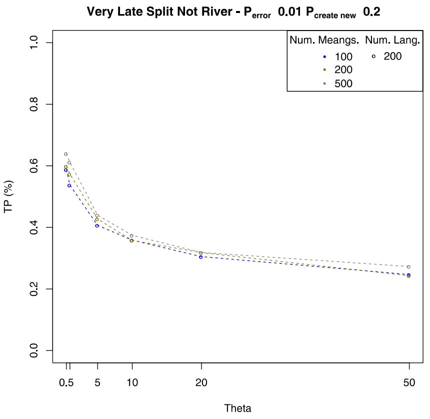


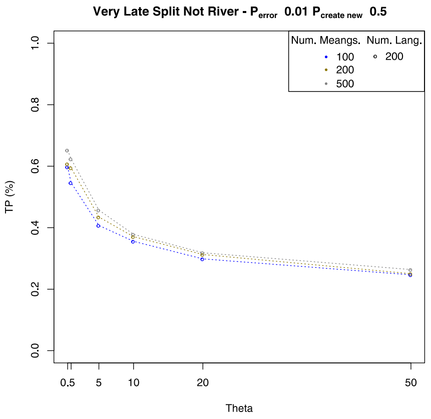

Late Split River - P_{error} 0.02 $P_{create\ new}$ 0.2

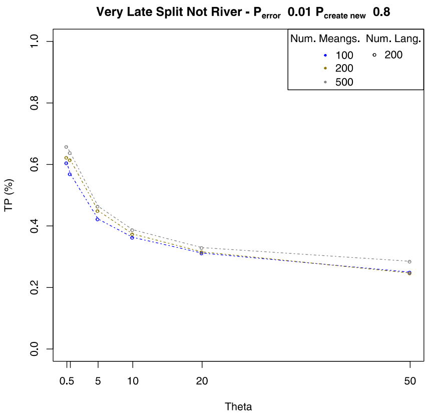

Late Split River - P_{error} 0.02 $P_{create\ new}$ 0.5

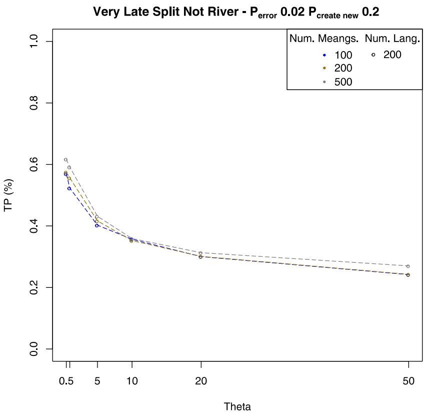

Late Split River - P_{error} 0.02 $P_{create\ new}$ 0.8

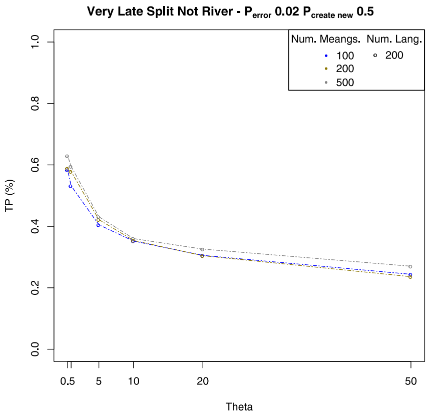

Very Late Split Not River - no error

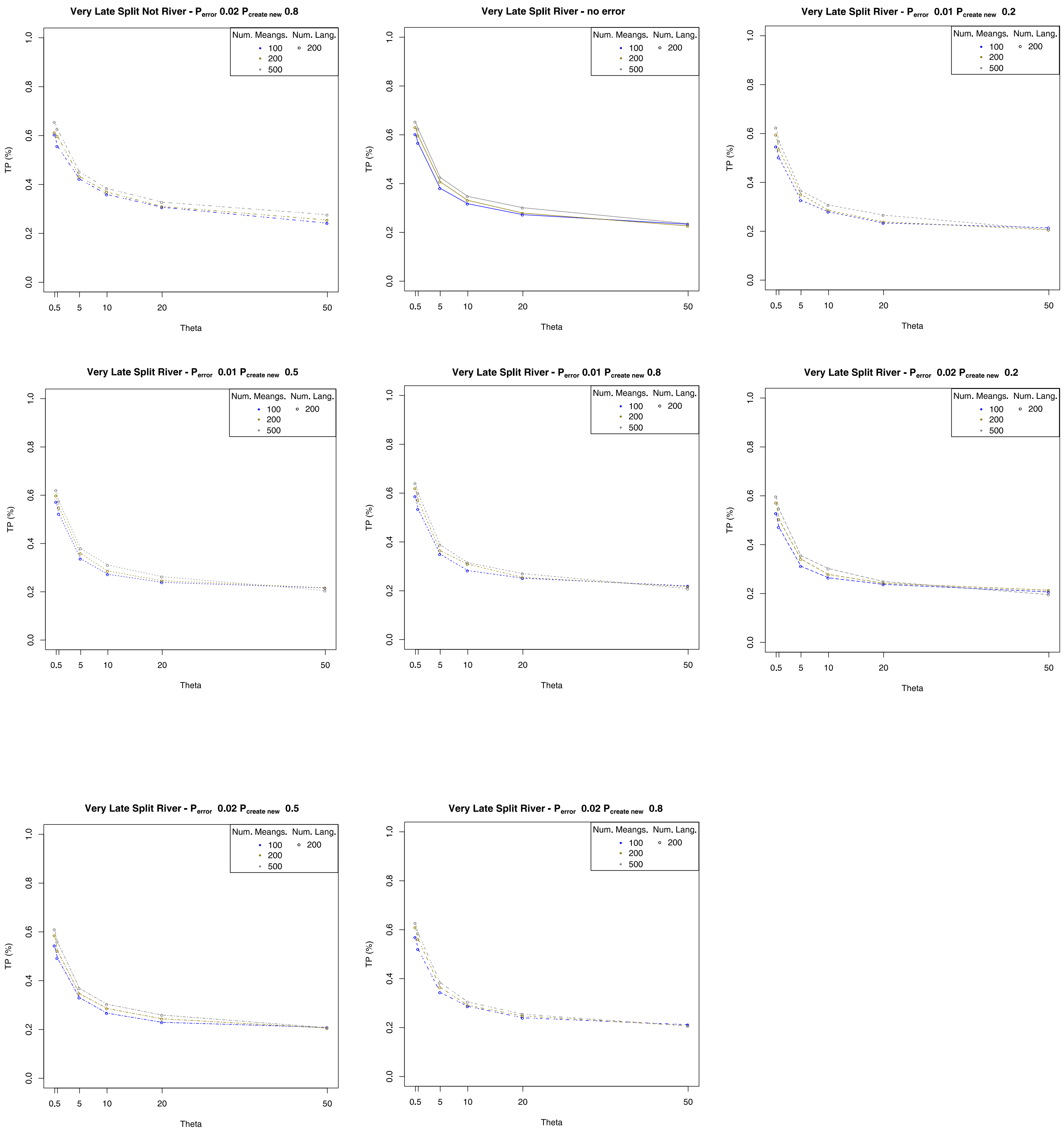

Very Late Split Not River - P_{error} 0.01 $P_{create\ new}$ 0.2


Very Late Split Not River - P_{error} 0.01 $P_{create\ new}$ 0.5

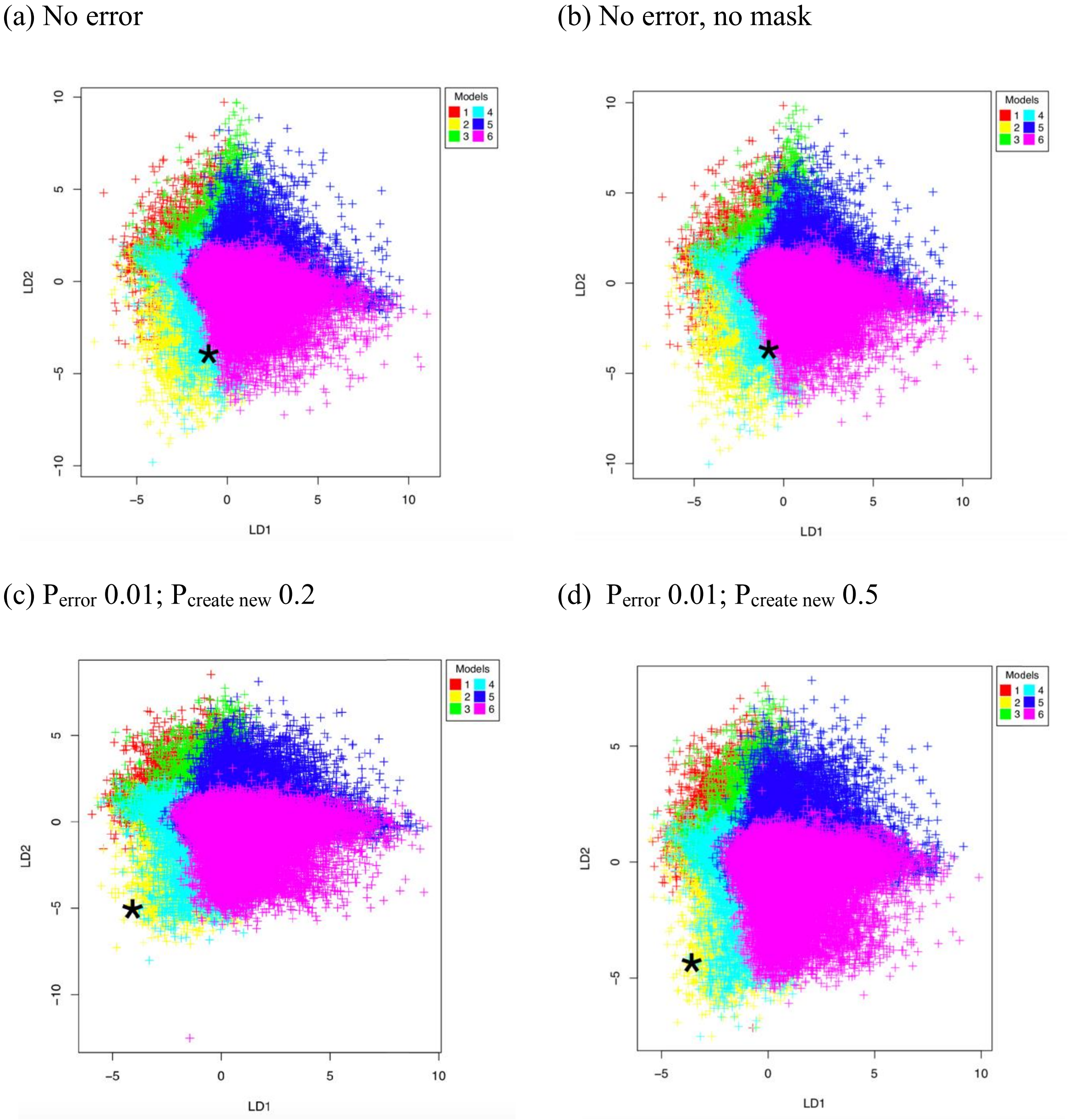

Very Late Split Not River - P_{error} 0.01 $P_{create\ new}$ 0.8


Very Late Split Not River - P_{error} 0.02 $P_{create\ new}$ 0.2


Very Late Split Not River - P_{error} 0.02 $P_{create\ new}$ 0.5



Supplementary Figure S2.1. Proportion of True Positives (TP) for five-population models with introduction of errors on the simulated linguistic data. The plots present the proportion of TP (y-axis) obtained analysing pods resulting from the five models (see the sixth model in the figure 2.17) under 18 combinations of experimental parameters. Different values of theta are in the x-axes, number of languages is represented by different symbols and number of meanings is represented by different colours.

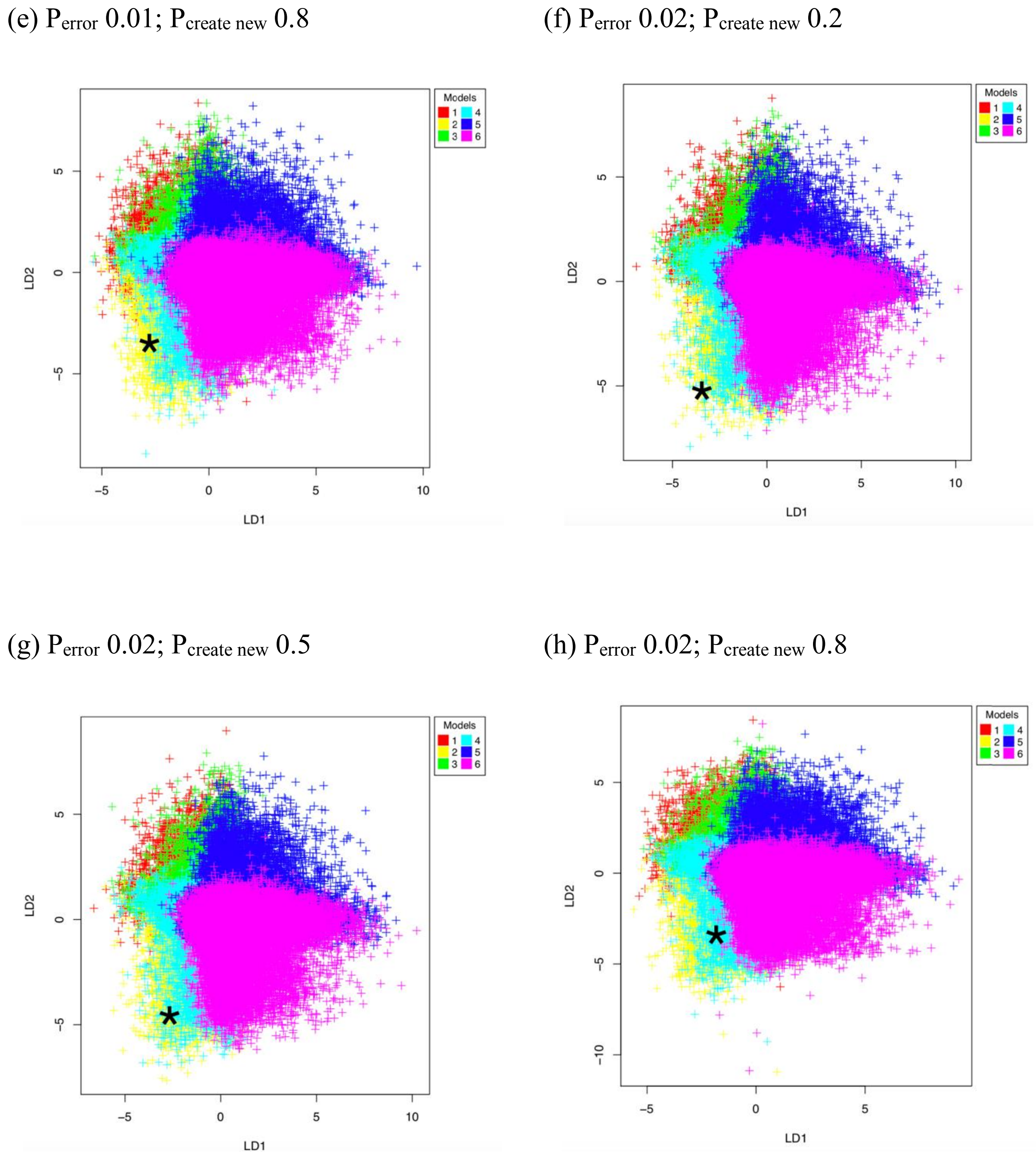


(a) No error

(b) No error, no mask

(c) P_{error} 0.01; $P_{create\ new}$ 0.2

(d) P_{error} 0.01; $P_{create\ new}$ 0.5



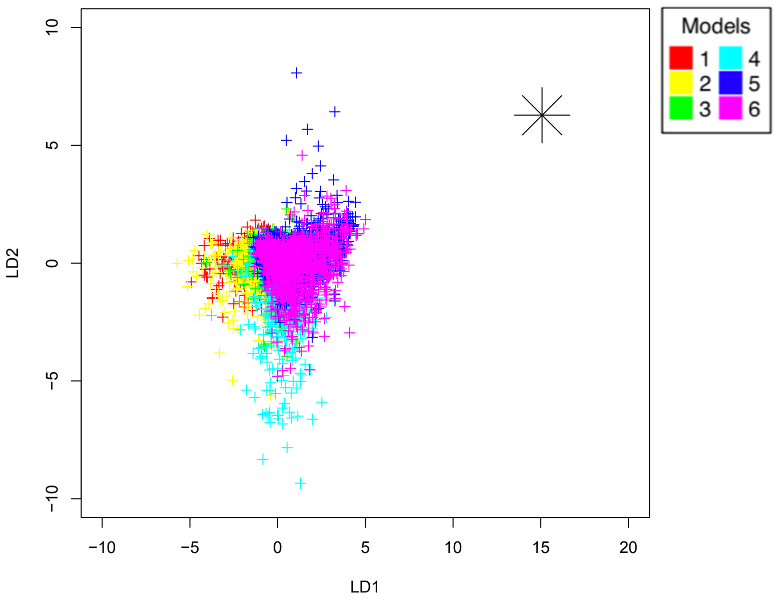
(e) P_{error} 0.01; $P_{\text{create new}}$ 0.8

(f) P_{error} 0.02; $P_{\text{create new}}$ 0.2

(g) P_{error} 0.02; $P_{\text{create new}}$ 0.5

(h) P_{error} 0.02; $P_{\text{create new}}$ 0.8

Supplementary Figure S2.2. Linguistic simulated (theta = 5) and observed data. Projection of the reference table on the first two LDA axes. Colours correspond to the model indexes. Model 1: Early split not river; Model 2: Early split river; Model 3: Late split not river; Model 4: Late split river; model 5: Very Late split not river; model 6: Very Late split river. The location of the observed data is indicated by the asterisk.



Supplementary Figure S2.3. Genetic simulated and observed data. Projection of the reference table on the first two LDA axes. Colours correspond to the model indexes. Model 1: Early split not river; Model 2: Early split river; Model 3: Late split not river; Model 4: Late split river; model 5: Very Late split not river; model 6: Very Late split river. The location of the observed data is indicated by the asterisk.

Supplementary Table S2.1. Confusion Matrix for the five-population models considering only three out of five populations. Number of genetic simulations assigned to each of the six models.

		Late Split Not River	Late Split River	Early Split Not River	Early Split River	Very Late Split Not River	Very Late Split River
nc6_nl5000_ll1000_recrate1.12e-08	**Late Split Not River**	1453	1471	2267	2014	1356	1439
	Late Split River	1364	1595	2024	1909	1372	1736
	Early Split Not River	1268	1012	3358	3035	703	624
	Early Split River	1207	1126	3056	3258	732	621
	Very Late Split Not River	1171	1148	1379	1115	2155	3032
	Very Late Split River	1056	1156	1250	939	2174	3425
nc6_nl5000_ll2000_recrate1.12e-08	**Late Split Not River**	1492	1707	2368	1825	1307	1301
	Late Split River	1415	2030	1927	1598	1450	1580
	Early Split Not River	1089	900	3781	3307	576	347
	Early Split River	1151	997	3472	3481	527	372
	Very Late Split Not River	1127	1196	1239	861	2425	3152
	Very Late Split River	947	1179	973	708	2407	3786
nc6_nl10000_ll1000_recrate1.12e-08	**Late Split Not River**	1579	1669	2247	1830	1402	1273
	Late Split River	1571	1910	1784	1642	1475	1618

	Early Split Not River	1131	869	3814	3259	571	356
	Early Split River	1104	954	3454	3574	568	346
	Very Late Split Not River	1169	1136	1216	857	2436	3186
	Very Late Split River	1035	1168	986	675	2408	3728
nc6_nl10000_ll2000_recrate1.12e-08	**Late Split Not River**	1911	1840	2088	1729	1288	1144
	Late Split River	1778	2141	1703	1433	1391	1554
	Early Split Not River	1230	723	4072	3405	372	198
	Early Split River	1121	820	3717	3800	323	219
	Very Late Split Not River	1151	1067	970	609	2661	3542
	Very Late Split River	997	1136	767	501	2486	4113
nc6_nl20000_ll1000_recrate1.12e-08	**Late Split Not River**	1836	1862	2220	1622	1340	1120
	Late Split River	1695	2144	1777	1457	1412	1515
	Early Split Not River	1181	734	4172	3292	397	224
	Early Split River	1127	829	3650	3787	381	226
	Very Late Split Not River	1226	991	1013	640	2623	3507
	Very Late Split River	1012	1046	741	509	2670	4022
nc6_nl20000_ll2000_recrate1.12e-08	**Late Split Not River**	2157	2132	1972	1519	1298	922
	Late Split River	1941	2508	1514	1255	1390	1392

	Early Split Not River	1074	613	4355	3592	231	135
	Early Split River	1099	744	3760	4056	217	124
	Very Late Split Not River	1275	1078	750	400	2850	3647
	Very Late Split River	974	1071	528	347	2840	4240
nc12_nl5000_ll1000_recrate1.12e-08	**Late Split Not River**	1917	1848	2078	1658	1324	1175
	Late Split River	1736	2034	1685	1487	1465	1593
	Early Split Not River	1275	921	3681	3415	400	308
	Early Split River	1233	952	3395	3750	417	253
	Very Late Split Not River	1248	1266	812	691	2582	3401
	Very Late Split River	1052	1261	730	541	2528	3888
nc12_nl5000_ll2000_recrate1.12e-08	**Late Split Not River**	2350	2006	1940	1458	1185	1061
	Late Split River	2060	2259	1531	1333	1387	1430
	Early Split Not River	1179	704	4205	3506	236	170
	Early Split River	1077	747	3939	3834	230	173
	Very Late Split Not River	1317	1249	632	474	2602	3726
	Very Late Split River	1021	1152	477	384	2599	4367
nc12_nl10000_ll1000_recrate1.12e-08	**Late Split Not River**	2387	1980	1971	1470	1133	1059
	Late Split River	2070	2390	1500	1263	1296	1481

	Early Split Not River	1219	656	4178	3549	246	152
	Early Split River	1182	767	3702	3918	250	181
	Very Late Split Not River	1362	1227	687	525	2597	3602
	Very Late Split River	1048	1221	485	363	2626	4257
nc12_nl10000_ll2000_recrate1.12e-08	**Late Split Not River**	2751	2239	1661	1301	1181	867
	Late Split River	2431	2576	1303	1021	1406	1263
	Early Split Not River	1068	546	4539	3611	154	82
	Early Split River	1050	666	4053	4010	131	90
	Very Late Split Not River	1341	1121	455	316	3082	3685
	Very Late Split River	1038	1119	306	238	2923	4376
nc12_nl20000_ll1000_recrate1.12e-08	**Late Split Not River**	2595	2265	1753	1299	1160	928
	Late Split River	2345	2608	1294	1124	1268	1361
	Early Split Not River	1098	520	4517	3631	148	86
	Early Split River	992	660	4028	4109	122	89
	Very Late Split Not River	1318	1177	538	328	2957	3682
	Very Late Split River	1006	1114	382	243	2833	4422
nc12_nl20000_ll2000_recrate1.12e-08	**Late Split Not River**	3073	2358	1568	1130	1087	784
	Late Split River	2666	2843	1083	916	1357	1135

Early Split Not River	975	447	4781	3682	74	41
Early Split River	927	492	4137	4330	73	41
Very Late Split Not River	1370	1126	344	183	3184	3793
Very Late Split River	979	1028	248	121	2982	4642

Supplementary Table S2.2. Confusion Matrix for the five-population models. Number of genetic simulations assigned to each of the six models.

		Late Split Not River	Late Split River	Early Split Not River	Early Split River	Very Late Split Not River	Very Late Split River
nc10_nl5000_ll1000_recrate1.12e-08	**Late Split Not River**	2851	1114	2500	1162	1670	703
	Late Split River	1170	3320	1623	1754	809	1324
	Early Split Not River	1398	908	4600	1862	846	386
	Early Split River	866	1271	2554	4241	550	518
	Very Late Split Not River	1574	681	1546	763	3866	1570
	Very Late Split River	802	1534	1104	948	1930	3682
nc10_nl5000_ll2000_recrate1.12e-08	**Late Split Not River**	3477	950	2489	958	1587	539
	Late Split River	1110	4049	1430	1518	657	1236
	Early Split Not River	1318	725	5392	1712	623	230
	Early Split River	665	1197	2538	4881	401	318
	Very Late Split Not River	1611	580	1349	528	4539	1393
	Very Late Split River	804	1570	819	695	1797	4315
nc10_nl10000_ll1000_recrate1.12e-08	**Late Split Not River**	3480	946	2520	945	1558	551
	Late Split River	1102	3962	1484	1546	608	1298

	Early Split Not River	1207	770	5397	1772	574	280
	Early Split River	643	1104	2643	4882	362	366
	Very Late Split Not River	1623	616	1324	581	4506	1350
	Very Late Split River	726	1567	819	742	1813	4333
nc10_nl10000_ll2000_recrate1.12e-08	**Late Split Not River**	4316	899	2287	702	1351	445
	Late Split River	964	4847	1225	1349	462	1153
	Early Split Not River	1165	639	6026	1561	448	161
	Early Split River	588	979	2386	5567	247	233
	Very Late Split Not River	1642	455	1040	363	5312	1188
	Very Late Split River	596	1542	640	535	1804	4883
nc10_nl20000_ll1000_recrate1.12e-08	**Late Split Not River**	4305	932	2219	735	1400	409
	Late Split River	1015	4877	1198	1329	455	1126
	Early Split Not River	1259	582	6080	1490	425	164
	Early Split River	561	959	2327	5662	236	255
	Very Late Split Not River	1630	505	1027	388	5157	1293
	Very Late Split River	678	1603	653	547	1651	4868
nc10_nl20000_ll2000_recrate1.12e-08	**Late Split Not River**	5187	772	2026	536	1191	288
	Late Split River	855	5657	989	1141	371	987

	Early Split Not River	1052	442	6801	1310	289	106
	Early Split River	504	857	2118	6238	141	142
	Very Late Split Not River	1606	331	807	242	5919	1095
	Very Late Split River	535	1641	461	373	1511	5479
nc20_nl5000_ll1000_recrate1.12e-08	**Late Split Not River**	4252	990	2136	794	1341	487
	Late Split River	1133	4535	1215	1401	539	1177
	Early Split Not River	1226	693	5800	1591	447	243
	Early Split River	599	1087	2333	5331	322	328
	Very Late Split Not River	1581	550	977	459	5052	1381
	Very Late Split River	714	1640	647	599	1679	4721
nc20_nl5000_ll2000_recrate1.12e-08	**Late Split Not River**	5025	823	1945	655	1186	366
	Late Split River	994	5385	998	1297	352	974
	Early Split Not River	1089	567	6379	1394	386	185
	Early Split River	471	863	2055	6196	186	229
	Very Late Split Not River	1621	403	817	302	5560	1297
	Very Late Split River	587	1490	482	445	1482	5514
nc20_nl10000_ll1000_recrate1.12e-08	**Late Split Not River**	5084	865	1934	574	1126	417
	Late Split River	1021	5409	959	1282	337	992

	Early Split Not River	1090	541	6474	1370	340	185
	Early Split River	479	901	2045	6191	172	212
	Very Late Split Not River	1647	412	739	292	5659	1251
	Very Late Split River	547	1562	431	429	1495	5536
nc20_nl10000_ll2000_recrate1.12e-08	**Late Split Not River**	5734	763	1693	511	1051	248
	Late Split River	912	6189	753	1071	244	831
	Early Split Not River	931	447	7132	1184	220	86
	Early Split River	388	725	1765	6870	122	130
	Very Late Split Not River	1546	346	595	222	6259	1032
	Very Late Split River	451	1519	280	301	1387	6062
nc20_nl20000_ll1000_recrate1.12e-08	**Late Split Not River**	5888	747	1681	417	977	290
	Late Split River	929	6101	811	1035	245	879
	Early Split Not River	993	483	7144	1053	232	95
	Early Split River	420	759	1830	6710	137	144
	Very Late Split Not River	1539	311	534	172	6346	1098
	Very Late Split River	475	1483	268	259	1376	6139
nc20_nl20000_ll2000_recrate1.12e-08	**Late Split Not River**	6490	629	1478	388	808	207
	Late Split River	840	6933	562	877	182	606

Early Split Not River	826	326	7672	982	139	55
Early Split River	337	588	1512	7386	90	87
Very Late Split Not River	1476	209	348	139	6890	938
Very Late Split River	385	1372	193	187	1243	6620

Supplementary Table S2.3. Model choice for the linguistic framework. Number of votes associated to each model, under different probability of errors and masking the data, by ABC-RF, and posterior probability of the most supported model (model 4). Model 1: Early split not river; Model 2: Early split river; Model 3: Late split not river; Model 4: Late split river; model 5: Very Late split not river; model 6: Very Late split river.

theta 0.5	**selected model**	**votes model 1**	**votes model 2**	**votes model 3**	**votes model 4**	**votes model 5**	**votes model 6**	**post.proba**
no error	4	105	77	83	116	59	60	0.571
no error, no mask	1	109	96	87	95	61	52	0.535
P_{error} 0.01; $P_{\text{create new}}$ 0.2	4	75	82	101	117	63	62	0.548
P_{error} 0.01; $P_{\text{create new}}$ 0.5	2	68	101	98	100	67	66	0.572
P_{error} 0.01; $P_{\text{create new}}$ 0.8	4	79	79	96	109	74	63	0.549
P_{error} 0.02; $P_{\text{create new}}$ 0.2	4	45	101	83	146	51	74	0.534
P_{error} 0.02; $P_{\text{create new}}$ 0.5	4	62	91	80	114	56	97	0.584
P_{error} 0.02; $P_{\text{create new}}$ 0.8	4	79	95	92	103	64	67	0.588

theta 1	**selected model**	**votes model 1**	**votes model 2**	**votes model 3**	**votes model 4**	**votes model 5**	**votes model 6**	**post.proba**
no error	2	88	95	87	88	87	55	0.565
no error, no mask	3	86	92	102	86	71	63	0.577
P_{error} 0.01; $P_{\text{create new}}$ 0.2	4	68	94	78	116	63	81	0.510
P_{error} 0.01; $P_{\text{create new}}$ 0.5	4	92	73	90	105	73	67	0.551

P_{error} 0.01; $P_{create\ new}$ 0.8	2	82	102	81	81	88	66	0.555
P_{error} 0.02; $P_{create\ new}$ 0.2	4	51	89	87	141	51	81	0.515
P_{error} 0.02; $P_{create\ new}$ 0.5	3	80	90	102	97	55	76	0.562
P_{error} 0.02; $P_{create\ new}$ 0.8	3	95	78	97	86	71	73	0.589
theta 10	**selected model**	**votes model 1**	**votes model 2**	**votes model 3**	**votes model 4**	**votes model 5**	**votes model 6**	**post.proba**
no error	2	80	107	63	106	59	85	0.514
no error, no mask	4	53	116	59	120	51	101	0.531
P_{error} 0.01; $P_{create\ new}$ 0.2	4	55	123	35	160	38	89	0.524
P_{error} 0.01; $P_{create\ new}$ 0.5	4	43	126	44	145	40	102	0.497
P_{error} 0.01; $P_{create\ new}$ 0.8	2	60	126	55	118	54	87	0.492
P_{error} 0.02; $P_{create\ new}$ 0.2	4	42	116	51	167	36	88	0.504
P_{error} 0.02; $P_{create\ new}$ 0.5	4	45	128	45	156	38	88	0.540
P_{error} 0.02; $P_{create\ new}$ 0.8	2	54	134	45	122	41	104	0.488
theta 20	**selected model**	**votes model 1**	**votes model 2**	**votes model 3**	**votes model 4**	**votes model 5**	**votes model 6**	**post.proba**
no error	2	53	115	60	113	48	111	0.547
no error, no mask	6	45	103	55	117	56	124	0.525
P_{error} 0.01; $P_{create\ new}$ 0.2	4	42	99	42	151	31	135	0.541
P_{error} 0.01; $P_{create\ new}$ 0.5	4	36	123	44	143	25	129	0.500

P_{error} 0.01; $P_{\text{create new}}$ 0.8	6	37	129	39	128	36	131	0.476
P_{error} 0.02; $P_{\text{create new}}$ 0.2	6	20	72	57	128	43	180	0.522
P_{error} 0.02; $P_{\text{create new}}$ 0.5	6	57	94	33	127	46	143	0.494
P_{error} 0.02; $P_{\text{create new}}$ 0.8	6	43	96	46	130	45	140	0.436

theta 50	selected model	votes model 1	votes model 2	votes model 3	votes model 4	votes model 5	votes model 6	post.proba
no error	4	36	113	49	131	47	124	0.501
no error, no mask	6	42	111	46	108	75	118	0.460
P_{error} 0.01; $P_{\text{create new}}$ 0.2	6	25	104	28	121	47	175	0.450
P_{error} 0.01; $P_{\text{create new}}$ 0.5	6	32	105	23	132	60	148	0.444
P_{error} 0.01; $P_{\text{create new}}$ 0.8	2	34	144	27	128	61	106	0.469
P_{error} 0.02; $P_{\text{create new}}$ 0.2	6	20	72	28	126	61	193	0.515
P_{error} 0.02; $P_{\text{create new}}$ 0.5	6	26	92	38	111	73	160	0.469
P_{error} 0.02; $P_{\text{create new}}$ 0.8	6	29	127	22	113	55	154	0.426

Acknowledgments

I would like to thank my PhD supervisor Guido Barbujani for giving me the opportunity to collaborate within his group during my doctoral degree, for all the help and support during these 3 years.

I want to thank Silvia Ghirotto for her supervision, assistance, support and, more importantly, for all her patience while teaching me and helping me to achieve more and more each day. I also want to thank her for entrusting me with important research and giving me the opportunity to travel abroad and start new collaborations. I also want to thank Gloria Gonzalez-Fortes for all the hours spent teaching me, for all the support, for all the patience and, specially, for her friendship. Both their contributions were essential to the development of these projects and thesis. You have made me stronger and wiser, allowing me to consolidate and acquire new skills both academic and at a personal level.

A special thanks to Andrea Benazzo and Igor Yanovich for all the support and time spent teaching me. To Teresa Vizzari who helped me a lot with the ABC-RF.

To the Population Genetics group in Ferrara for all their friendship and support. To Hugo Reyes-Centeno and all the “Words, Genes, Bones and Tools” group at the University of Tübingen for receiving me.

I also want to thank all the people with whom I had the opportunity to collaborate in different projects in the last 3 years. To Prof. Giuseppe Longobardi, Dr. Cristina Guardiano, Dr. Guido Cordoni and Dr. Andrea Ceolin for the linguistic analyses.

To the “rallentati di UNIFE” with whom I shared a lot of good moments. To the “cineforum” group that helped me to improve my Italian culture.

A special thanks (and a really important one) to the best housemates that I could ever have asked, who also taught me how to cook real Italian food, Patricia, Valentina and Davide. To all my friends, especially to Ana Rebelo, Nádia Rocha, Rodrigo Godinho and Marcos Chuva, for always being there for me and for all the support and friendship, even at long distance.

Last but not least, a special thanks to my family who have always supported me in every decision made. Without them I could not have achieved half of what I have already.