# GEO-GENETICS PATTERNS IN BAMBARA GROUNDNUT: INVESTIGATING THE ROLE OF GEOGRAPHY IN THE DISTRIBUTION OF GENETIC VARIATION

Roberto Oliveira Santos

*Thesis submitted to the School of Geography, University of Nottingham for the degree of Doctor of Philosophy*

August 14, 2018

# Abstract

The study of the geo-genetic structure of human-associated plant species can help us to understand the ways that the landscape, environment, and socio-cultural factors influence these populations. This research project aims to identify spatial-related factors (distance, environment, culture) that affect the genetic structure of Bambara groundnut, an underutilised crop mainly cultivated in sub-Saharan Africa. I used a genetic database of 33 landraces of Bambara groundnut, 128 samples in total, genotyped using 19 microsatellite markers. I tested for the presence of genetic structure and genetic barriers using allele frequencies, ordination, and clustering analyses. Next, I explored how geographic distance, environment, and human-mediated dispersal influence genetic diversity by investigating the relationship between four measures of dissimilarity. I analysed these relationships using Mantel tests and structural equation modelling (SEM). Finally, I examined the relationship between genetic variation and soil water content using Moran's I statistics and generalised least squares (GLS) regression analyses. I found genetic structure among different landraces. The clustering analyses result indicated strong differentiation between landraces populations, with clusters mainly associating landraces to their geographic origins. The SEM results suggested that genetic similarity among landraces of Bambara is related to

geographic proximity and, to a lesser extent, linguistic similarity between human communities. There was no relationship between genetic similarity and environmental differences. I also discussed the challenges and possible solutions for integrating distinct discipline-datasets. Overall, this thesis highlighted the importance of cultural aspects and social networks in the genetic structure of human-associated plant species.

# Scientific communication

Since early 2014, a number of scientific communications related to this PhD project have been conducted in the format of talks, extended abstracts and the co-authorship of peer reviewed paper[1].

Leibovici, D.G., Anand, S., Santos, R. et al. (2017). "Geospatial binding for transdisciplinary research in crop science: the GRASPgfs initiative". In *Open Geospatial Data, Software and Standards*, 2(1), p.20. DOI: 10.1186/s40965-017-0034-3.

Santos R, Algar A, Field R, Mayes S. (2017). "Integrating GIScience and Crop Science datasets: a study involving genetic, geographic and environmental data". In *PeerJ Preprints*. DOI: 10.7287/peerj.preprints.2248v4.

Santos R, Huynh D, Anand S, Ray RV, Mayes S, Leibovici D. (2016). "A geoprocessing modelling interoperable framework for AgriGIS using open data and open standards". In *PeerJ Preprints*. DOI: 10.7287/peerj.preprints.2136v2

[1]Further papers are currently being prepared for submission to the Journal of Biogeography and other journals.

# Acknowledgements

While this thesis is the result of my work, it would not have been possible without the help, guidance, and support of my supervisors and the financial support from the Brazilian funding agency (CAPES), the University of Nottingham, the School of Geography, and the Crops for the Future in Malaysia.

I would like to thank my supervisors who provided support during these four years. I am sincerely in debt with Adam Algar, Richard Field, Sean Mayes, and Jeremy Morley for the opportunities and guidance during my training.

I had the opportunity to meet and work with a fantastic group of people in many places during this research. In the Nottingham Geospatial Institute, my home during the first year, thanks to Didier Leibovici and Suchith Anand for the continuous support and encouragement. In the School of Geography, thanks to James Hicks and Simon Tarr for the many comments on drafts and good laughs during our meetings. In the Crops for the Future, thanks to Presidor Kendabie, Razlin Azman, Faraz Khan, and Dai Huynh for the constant help and attention.

Thanks to the School of Geography, for accepting my transference in the middle of my project, providing a friendly and stimulating environ-

ment. Thanks to Ian Conway and the fantastic office colleagues Prasujya, Paphaphit, Han, Joe, and Ben.

Thanks to my dear friends for all the support and love.

To my family Roberto, Romilda, Lela, Rafa, Limi, and Lêle, for your all your love and support. To my in-laws, for the love and support, in particular, Bia, Gabriel, David, and Barbara.

To Thomas, for your unconditional love, and beautiful smiles.

To Elaine, for the most import yes of my life. I love you.

# Contents


**1 Introduction** **13**
1.1 Contribution of the research . . . . . 14

**2 Literature Review** **17**
2.1 Food security and agricultural challenges . . . . . 17
2.2 Bambara groundnut . . . . . 20
2.3 Crop genetics . . . . . 23
2.3.1 Genetic variation . . . . . 23
2.3.2 Molecular markers . . . . . 28
2.4 GIScience in interdisciplinary projects . . . . . 29

**3 Geographic patterns of genetic variation in an indigenous crop legume** **31**
3.1 Introduction . . . . . 32
3.2 Material and methods . . . . . 36
3.2.1 Genetic data . . . . . 36
3.2.2 Georeferencing the genetic data . . . . . 38
3.2.3 Population structure . . . . . 39
3.3 Results . . . . . 42
3.4 Discussion . . . . . 44
3.4.1 Genetic diversity . . . . . 46
3.4.2 Genetic sub-populations . . . . . 46
3.4.3 Integrating geographic and molecular data in analysis 49
3.5 Conclusion . . . . . 49

**4 Human migration history is reflected in the geo-genetic variation in Bambara groundnut in sub-Saharan Africa** **52**
4.1 Introduction . . . . . 53
4.2 Materials and methods . . . . . 56
4.2.1 Study species . . . . . 56
4.2.2 Estimating genetic distance between landraces . . . . 57
4.2.3 Geo-referencing and estimating geographic distance between landraces . . . . . 58
4.2.4 Estimating linguistic similarity among sites of Bambara groundnut . . . . . 59
4.2.5 Estimating climatic similarity . . . . . 62
4.2.6 Analysis . . . . . 62
4.3 Results . . . . . 64
4.4 Discussion . . . . . 65

4.5 Conclusion . . . . . 69

**5 Spatial regression captures signals of water stress influence on the spatial patterns of genetic variation in a human-associated crop species** **71**
5.1 Introduction . . . . . 72
5.2 Material and methods . . . . . 75
5.2.1 Study species . . . . . 75
5.2.2 Genetic and environmental datasets . . . . . 76
5.2.3 Analysis . . . . . 77
5.3 Results . . . . . 79
5.4 Discussion . . . . . 83
5.5 Conclusion . . . . . 85

**6 Conclusion** **87**
6.1 Suggested further research . . . . . 91

# List of Figures

2.1 (a) DNA double helix. Both sequences are linked individually by a sugar-phosphate backbone, and together by hydrogen bonds. DNA sequences orientation is usually defined by the use of 3' and 5', three and five primers, respectively. This means that at one end there is 5' phosphate group and at the other end there is 3' hydroxyl group. (b) Single-stranded DNA. This shows the complementary sequences. Single-stranded sequences are also called denatured DNA, and it is usually the result of heat application. Source: Rowe et al. (Rowe et al. 2017) . . . . . 24

3.1 Bambara groundnut plants cultivated in the Crops for the Future glasshouses at Sutton Bonington campus, University of Nottingham. Pictures taken on 2016. . . . . 36

3.2 Origin or local of the collection of the germplasm (seeds) of the landraces of Bambara groundnut. Symbols (plus) identify the location of the landraces. Names and numbers identify the landraces according to the IITA database naming system. 40

3.3 Ordination (PCA) and clustering (K-means) results plotted using PC1 and PC2 (a), and longitude and latitude (b). Symbols identify landraces origin location aggregate by African regions: Western (plus), Eastern (square), Southern (cross), and Central (circle). Colours represent the cluster attribution as informed by K-means results for K = 4. . . . . 43

3.4 Monmonier's algorithm genetic boundaries identification. The black circles indicate landraces location, the black edges between the nodes indicates connection according the Gabriel's network methods, and the blue lines with arrows indicates genetic barrier. . . . . 45

3.5 Ordination (PCA) and clustering (K-means) results plotted using PC1 and PC2 (a), and longitude and latitude (b). Symbols identify landraces origin location aggregate by African regions: Western (plus), Eastern (square), Southern (cross), and Central (circle). Colours represent the cluster attribution as informed by K-means results for K = 2 . . . . . 50

3.6 Ordination (PCA) and clustering (K-means) results plotted using PC1 and PC2 (a), and longitude and latitude (b). Symbols identify landraces origin location aggregate by African regions: Western (plus), Eastern (square), Southern (cross), and Central (circle). Colours represent the cluster attribution as informed by K-means results for K = 3. . . . . 51

4.1 Overview of the area of study. The red and black circles indicate the origin and the destinations used in the least cost path analysis. The solid lines show least cost path to move between pairwise locations. The cost of movement between two adjacent cell is the inverse of the difference of the altitude (see equation 4.1). In this figure, the red circle represents landrace 32. . . . . 60

4.2 Overview of the area of study. The black crosses indicate accessions. Circles around the landraces represent the coverage area and have a fixed radius of 1.5 degrees. Grey squares indicate the location of languages available in the ASJP database. Circles around the grey squares indicate the coverage area of each language, calculated as the area required to host the number of speakers of the respective language and average human density in sub-Saharan Africa (34 people/km2). Inset with focus on landrace AHM968 (blue triangle), that exemplifies the procedure. There are six languages located around the landrace, and three of them (dhm, lnb, kua) have their coverage areas intersecting with the landrace's coverage area. I used these three languages to calculate the linguistic distance between AHM968 and other landraces. Source: author. 61

4.3 Schematic design of the pipeline implemented to run the analysis in the R software. The pipeline starts at the top right of the diagram, outputting on the left (black circles). Small rectangles represent the data sources and intermediary results. Large rectangles represent the processes and algorithms, and arrows provide information about the flow of data in the pipeline. Solid lines represent a connection between processes and dashed lines represent inputs and outputs of these processes. . . . . 64

4.4 Structural equation model of the influence of geographic distance and environmental and linguistic similarities on genetic similarity. . . . . 66

5.1 Location of the 33 landraces plotted over the distribution of soil water content monthly average (swc mean). Due to the broad range of environments where Bambara groundnut is cultivated, it is possible to identify landraces on environment conditions varying from very wet (swc mean above 80%) in the Central Africa next to the Atlantic coast, to very dry (swc mean below 40%) in the North Africa and border between Angola and Namibia. . . . . 77

5.2 Two ordinary least squares regressions were calculate to predict the PCA component scores of the genetic data (based on the landraces allele frequency calculated in Chapter 3) based on the soil water content. Regression of PC1 (a) results were slope 0.05 and p-value 0.0696, and regression of PC2 (b) results were -0.06 and p-value 0.0129. . . . . 80
5.3 I tested the PCA result (based on landraces' allele frequency) for spatial autocorrelation using Moran's I. Figure (a) and (b) show correlogram plots of the coefficient Moran's I for PC1 and PC2 respectively, as a function of distance (km). PC1 correlogram indicates strong spatial clustering, while PC2 correlogram indicates week or absence of spatial clustering. I calculated the spatial autocorrelation coefficient Moran's I using package NCF - function 'correlog' with parameters increment = 500km and 999 permutations under the null hypothesis to assess the level of significance. Black filled circle represent correlations with p-value $< 0.05$. . . . . 82

6.1 Least cost path analysis results for landraces 9, 32, 89, 118, 144, and 231. . . . . 94
6.2 Least cost path analysis results for landraces 286, 289, 329, 335, 385, and 390. . . . . 95
6.3 Least cost path analysis results for landraces 476, 484, 536, 696, 754, and 790. . . . . 96
6.4 Least cost path analysis results for landraces 792, 793, 1276, AHM753, AHM968, and BOTS 1. . . . . 97
6.5 Least cost path analysis results for landraces BOTS 5, DOD R, Mahenene Black, S-1913, S19-3, and SB4-2. . . . . 98
6.6 Least cost path analysis results for landraces SB16 5A, UNIS R, and VSSP 6. . . . . 99

# List of Tables


2.1 Allelic variation of a microsatellite marker based on the DNA sequence "CA". Each table line contains a different allele due to the number of adjacent repetitions. . . . . 29

3.1 List of all the 19 primers used in the genotyping using SSR. 37
3.2 List of all the 33 landraces used in the analyses. . . . . 39
3.3 The top 20 loadings for the components (PC1 and PC2) of the PCA based on allele frequency. . . . . 48

5.1 Pearson's correlation between four soil water content based variables. The water stress duration variables (swc duration 30, swc duration 50 and swc duration 70) showed an expected high positive (lower value was 0.84) correlation since their formulation only differ on the threshold definition of water stress (30, 50, and 70% of the soil capacity). They also showed a high negative correlation (higher value was -0.95), also expected since values of swc mean close to 1 would mean more water available in the soil, and a likely indication of shorter periods of water stress. . . . . 79
5.2 Correlogram plot data source, containing n (the number of pairs within each distance class), mean of class (the actual average of the distances within each distance class), correlation (Moran's I), and p-value (the permutation two-sided p-value for each distance class). . . . . 81
5.3 Moran's I statistic results based on PC1, PC2, and the respective residuals of ordinary least squares regressions. . . . 81
5.4 Generalised least squares regression results. Exp. Var. - Explanatory variable, Struct - Spatial Structure, E - Exponential, G - Gaussian, L - Linear, R - Rational quadratic, S - Spherical. * The regression of PC1 with linear structure did not converge. . . . . 86

# List of abbreviations

| | |
|---|---|
| AFLP | Amplifed Fragment Length Polymorphism |
| AIC | Akaike information criterion |
| ASJP | Automated Similarity Judgement Program |
| BIC | Bayesian information criterion |
| DArT | Diversity Arrays Technique |
| DNA | Deoxyribonucleic acid |
| FAO | Food and Agriculture Organization |
| GADM | Global Administrative Areas |
| GIScience | Geographic Information Science |
| GLS | Generalised least squares |
| IBD | Isolation by distance |
| IITA | International Institute of Tropical Agriculture |
| mRNA | messenger Ribonucleic acid |
| NGS | Next Generation Sequencing |
| OLS | Ordinary least square |
| PCA | Principal component analysis |
| RNA | Ribonucleic acid |
| SDGs | Sustainable Development Goals |
| SEM | Structural Equation Modelling |
| SNP | simple nucleotide polymorphism |
| SSR | Simple Sequence Repeats |
| SWC | Soil water content |
| UN | United Nations |
| WGS84 | World Geodetic System 1984 |

# Chapter 1

# Introduction

In the last sixty years the world's urban population has more than quadrupled, growing from 732 million to 3.4 billion. By 2050, 6.3 billion people will live in urban areas, an expected increase of about 2.9 billion. Combining both urban and rural areas, a world population of more than nine billion is expected (UN 2015). How to produce food to feed all this population, not only in terms of quantity, but quality, is a complex and interlinking challenge (Foley et al. 2011).

Research challenges in the 21st century require an interdisciplinary approach, to advance fundamental understanding or to tackle problems whose solutions are beyond the scope of a single discipline or body of knowledge (Academies 2004). Sharing and reusing datasets from different disciplines is common practice and aims to strengthen the research by adding new dimensions to the data available or verifying the results obtained from a different perspective. However, integrating discipline-associated datasets is not a smooth process and is subject to varying concepts of quality and

abundance (Dyer 2015).

Geospatial information science (GIScience) is inherently interdisciplinary, providing tools and data that can be useful to solve problems within other disciplines (Joost et al. 2016). Disciplines such as ecology, biogeography, and more recently, landscape genetics (Manel et al. 2003) have used and contributed to geospatial data and methods while investigating spatial-associated research problems. An example is population genetics in crop science, where genetic data of plants can be linked to geographic coordinates expanding the opportunities to link the genetic dataset to other spatial-associated datasets, such as weather and soil data (Joost et al. 2016). Therefore, GIScience has the potential to facilitate and support new types of hypotheses and analysis (Manel et al. 2003) and to be the enabler for multidisciplinary research (Leibovici et al. 2017).

## 1.1 Contribution of the research

This thesis contributes to knowledge by investigating the relationship between human-associated plant species cultivated in human-dominated landscape and anthropogenic-environmental factors. These plants can lie on the boundary between non-domesticated and domesticated species and thus can supply unique insight into the simultaneous effects of environment and human actions on crop evolution. I focus on Bambara groundnut (*Vigna subterranea* (L.) Verdc.), a legume crop classed as an underutilised crop, a broad group of plant species that have potential to help achieving food security currently and in the future.

Through the next chapters, I explore the challenges of integrating data from distinct disciplines, in particular georeferencing genetic samples, and its effects on the further analyses. I also test the genetic structure of Bambara groundnut, and its spatial distribution. I test the relationship among genetic, environmental and geographic factors by analysing several measures of distance among landraces. I also test the direct relationship between genetic variation and soil water content, as it is a key factor potentially influencing crop yield. This thesis is organised into this introduction, followed by a literature review outlining the theoretical framework, and then presentation of three data-based chapters which are aimed toward publication. The data chapters are followed by a synthetic discussion of the results of all chapters, concluding statements and prospects for future work.

This thesis is important because it unravels the direct and indirect effects of geographic distances, landscape characteristics and human actions in the determining patterns of genetic variation of an underutilised crop. It explores the spatial distribution of genetic material of Bambara groundnut throughout sub-Saharan Africa, i.e. its geo-genetics, and their relationship with environmental, geographic factors and human factors.

The key research question is what drives the spatial distribution of the genetic variation of Bambara groundnut. I hypothesise that the current spatial distribution of the genetic variation of Bambara groundnut can be explained by three processes acting over recent centuries: (H1) genetic similarity positively correlated with environmental similarity where this crop has been cultivated, (H2) isolation by distance among the populations that

have genetically drifted apart, and (H3) human movement and cultural connections that have helped the dispersal of the seeds and increased gene flow of Bambara groundnut throughout sub-Saharan Africa. I also hypothesise that (H4) there is correlation between genetic variation and soil water content, an important environmental factor influencing crop yield.

# Chapter 2

# Literature Review

## 2.1 Food security and agricultural challenges

Current estimates suggest it will be necessary to double the current crop production to meet expected future demands (Licker et al. 2010; Foley et al. 2011). Traditional approaches to tackle the problem, such as expanding agricultural activity to new areas and improving major crop yields, may not be enough. Agricultural activity has already covered most of the suitable and available lands (Licker et al. 2010) and further expansion would require introducing extensive and intensive farming systems to marginal areas usually characterised by poor soils and abiotic stress or natural environments, particularly protected natural and indigenous areas that are already threatened (Foley et al. 2011).

Expansion into these areas would also require a closing of the yield gap, i.e. the difference between the potential yield of a crop and yields obtained in the fields (Licker et al. 2010). Advances in breeding programs in re-

cent decades have allowed major crops to approach their potential yield in favourable environments (usually a combination of adequate soil, water availability and farming practices) (Anderson et al. 2010; Mueller et al. 2012; Sentelhas et al. 2015), but the yield gap remains larger in marginal agricultural lands characterised by poor soils, water scarcity and lack of financial support to farmers to invest in inputs and machinery (Van Ittersum et al. 2013). Thus, even expansion into such currently unfavourable lands would have limited returns. The lack of diversity and limited growth potential of established, dominant, crops, threaten future food security under a growing human population and rapid environmental change (Galluzzi and Noriega 2014, Cheng et al. 2017, Massawe et al. 2016).

According to FAO (2001), "Food security exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life". This definition enumerates four dimensions of food security: availability, access, utilisation and stability. Availability is related to existence of sufficient quantities of food of appropriate quality, supplied through domestic production or imports. Access is related to individuals with access to adequate resources for acquiring appropriate foods for a nutritious diet. Utilisation is concerned with the use of food through adequate diet, clean water, sanitation and health care to reach a state of nutritional well-being where all physiological needs are met. Stability is the concept of continuous access to adequate food at all times by a population, household or individual.

Neglected and underutilised crops, also known as minor crops or orphan crops, have potential to contribute to future food security (Padulosi et al. 2014, Cheng et al. 2017, Massawe et al. 2016) and are an alternative to expanding intensive farming of major crops. Historically, the term "underutilised crops" has been used to describe a broad range of plant species, usually of indigenous origin, that have potential to play an important role securing food security, particularly in poor and vulnerable communities (Azam-Ali et al. 2001). They are considered a reliable option where environment and management practices are not favourable to staple crops, although yields are usually lower than the former (Karunaratne 2009). Generally cultivated for subsistence in marginal environments using low-input agricultural systems, these species may already contain or have the potential to develop desirable agronomic characteristics to cope with biotic and abiotic stress, such as drought tolerance, disease resistance or low demand for inputs (Galluzzi and Noriega 2014). However, these species commonly lack desirable commercial characteristics such as high yields (given high chemical usage), that are likely to create resistance among farmers that can afford inputs and aim for high yields (Kahane et al. 2013; Barbieri et al. 2014).

Underutilised crops are mainly cultivated using landraces under traditional farming systems and initially developed by indigenous people, have undergone centuries of cultural and biological evolution (Galluzzi and Noriega 2014). Landraces are local varieties that have been developed over time, adapting to a number of constraints, including physical constraints

such as agro-meteorological conditions or cultural constraints such as taste, colour and size. Also, landraces are usually more genetically and physically diverse than formal breeds.

A number of studies contend that neglected and underutilised crops have potential to help improve food security. This is supported by several studies related to aspects of neglected and underutilised crops in regions where food security is vulnerable (Mayes et al. 2012, Jacobsen et al. 2013, Kahane et al. 2013, Kunwar et al. 2012, Padulosi et al. 2011). Currently, estimates are that over 7000 edible plant species are available in different stages of domestication throughout the globe.

## 2.2 Bambara groundnut

Bambara groundnut (*Vigna subterranea* (L.) Verdc.) is a grain legume species mainly cultivated in sub-Saharan Africa and classed as an underutilised crop (Azam-Ali et al. 2001, Basu et al. 2007, Jørgensen et al. 2010). The crop has nutritional value, good tolerance to adverse environmental conditions, in particular, water stress and heat, and usually grown for local consumption and subsistence (Azam-Ali et al. 2001, Karunaratne 2009). Bambara groundnut can be found in a wide range of climatic zones and can be cultivated from sea level to up to 1600m in altitude, with annual rainfall as low as 200 mm in Botswana to as high as 1250 mm in Swaziland (Mwale et al. 2007). The cultural association of the crop with food for poor people and the absence of promotion in both local and international markets constrain its cultivation to marginal areas with reduced or no inputs.

However, Bambara groundnut has potential to alleviate hunger, contribute to a more nutritious diet, be a source of income, and be a reliable crop in an uncertain environment subject to climate change (Azam-Ali et al. 2001, Hillocks et al. 2012).

The precise location of Bambara's origin is still debated (Olukolu et al. 2012), but it is believed to be in West Africa, in the region between Nigeria and Cameroon. Currently found throughout the African continent, the International Institute of Tropical Agriculture (IITA) germplasm database contains thousands of seeds with origin or collection location including most of the African countries. The expansion of Bambara groundnut is hypothesised to be due to human migration to East Africa and Madagascar, then later to South and South-East Asia (Ntundu et al. 2006). Despite the importance of understanding the origin and dispersal of this crop to plan conservation measures and evaluating the diversity and potential use of the genetic resources for crop improvement, I am not aware of any recent research examining past movement of Bambara seeds.

Smallholding farmers cultivate Bambara using traditional farming practices. The crop's drought tolerance is a relevant agronomic characteristic and offers an advantage over other legumes in low rainfall areas (Karunaratne 2009). Other abiotic stress factors such as heat, cold, and photoperiod affect its growth and development. In the absence of water stress, temperature has a primary influence on developmental process within the plant (Karunaratne 2009). Rain-fed agriculture is common in Africa, despite its vulnerability to climate change (Rurinda et al. 2013; Salmon et al. 2015). In regions that

have the combination of this type of agriculture, unreliable rainfall, poor soils and poor crop productivity, the production of more drought-tolerant crops, such as Bambara groundnut, are encouraged (Molosiwa et al. 2015).

Over the centuries, farmers have acquired vast amounts of knowledge about Bambara groundnut and cultivated a set of landraces. Bambara groundnut plays a role in cropping systems in semi-arid Africa, due to its stable and reliable, though low, yield (Azam-Ali et al. 2001). It takes part in intercropped systems, with cereals, tuber crops, vegetables and other legumes (Karunaratne 2009). Mainly female farmers cultivate Bambara groundnut as food for human consumption and subsistence under traditional low input agricultural systems (Azam-Ali et al. 2001). The fact that women traditionally grow it gives it less value and less priority in land allocation. The high nutritional value of Bambara groundnut provides a cheap source of protein to poorly resourced farmers in semi-arid areas, making it a supplement to cereal-based diets (Molosiwa 2012).

Currently, there are no cultivars or varieties of Bambara groundnut explicitly bred for particular agro-ecological conditions or farming systems, but according to Massawe et al. (2005), there are substantial Bambara groundnut genetic resources available for improvement of the crop. Recently, as part of a set of efforts to create breeding programs, studies have been conducted into the genetics of Bambara groundnut. This approach is cost-effective due to the development of technologies like the Next Generation Sequencing (NGS). These technologies have increased the ability of researchers to generate sequence data, allowing molecular tools, such as DNA

markers, to be produced at a reasonable cost (Mayes et al. 2012). These sequences and tools can be used to compare genetic information from Bambara groundnut with information available in staple crops, which do have significant databases of genetic and trait information. The possibility of using knowledge derived from a staple crop in underutilised crops could mean an advance in the understanding of the relation between genotype-phenotype with relatively low investment (Ho et al. 2017).

Bambara groundnut can be considered an option in the food security policy for Africa since it addresses some of the dimensions of the food security definition. The reasons for this are: first, women in domestic production traditionally cultivate Bambara groundnut; second, its characteristics, in particular, resilience and above average yield in drought environments; third, its nutritional characteristics, which means it can help provide a nutritious diet together with other food.

## 2.3 Crop genetics

### 2.3.1 Genetic variation

Adenine (A), guanine (G), cytosine (C), and thymine (T) are the four chemical bases used to store genetic information in the Deoxyribonucleic acid (DNA). These bases, usually connected in pair to a correspondent base, A to T and C to G, form the double helix structure of DNA. The order, or sequence, of bases provides the information for building and maintaining individual organisms. DNA contains genes that codes for proteins, which

are created through the processes of transcription and translation. During transcription, DNA is translated to messenger Ribonucleic acid (mRNA) and during translation the sequence of mRNA is read. Each sequence of three bases codes for an amino acid and mRNA is converted to a protein, one amino acid at a time. Over all, these long sequences of DNA are very similar among individuals of the same species, however, the small differences, known as genetic variation, are important (Alberts et al. 2014, p. 175).

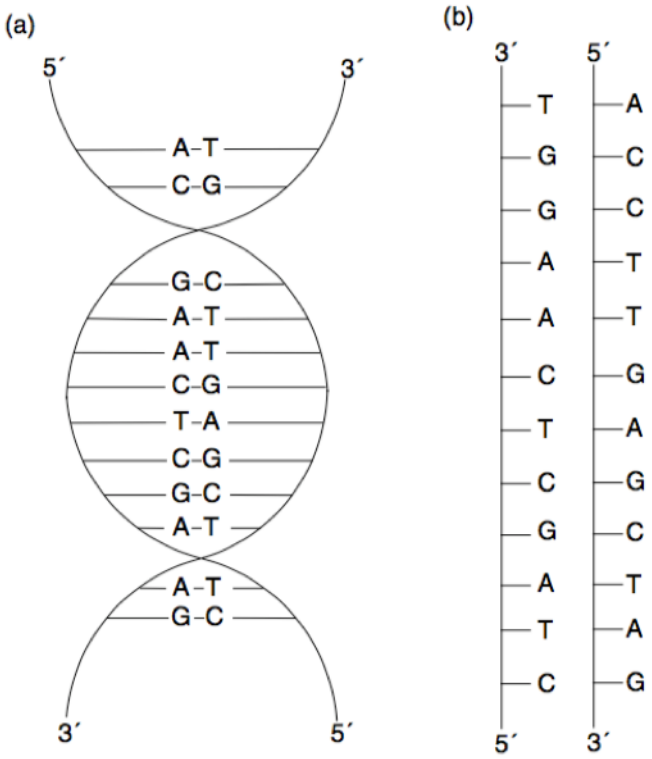



Figure 2.1: (a) DNA double helix. Both sequences are linked individually by a sugar-phosphate backbone, and together by hydrogen bonds. DNA sequences orientation is usually defined by the use of 3' and 5', three and five primers, respectively. This means that at one end there is 5' phosphate group and at the other end there is 3' hydroxyl group. (b) Single-stranded DNA. This shows the complementary sequences. Single-stranded sequences are also called denatured DNA, and it is usually the result of heat application. Source: Rowe et al. (Rowe et al. 2017)

The genetic variation of a population or species is influenced by four

main processes: mutation, selection, gene flow and genetic drift. Mutation is the process that creates genetic variation or alleles due to error in DNA replication. Selection impacts genetic variation by increasing the frequency of alleles that are favourable in a particular environment and decreasing the frequencies of alleles that are less favourable in that environment. If a locus or a mutation is neutral, selection does not directly influence the frequency of alleles. Gene flow occurs as a result of migration (dispersal and subsequent reproduction) and this process moves alleles between populations and tends to make them more similar genetically. Genetic drift is the change in allele frequency due to random sampling effects as alleles are passed on from one generation to the next. Genetic drift is thus influenced by the number of breeding individuals and the variance in reproductive success among breeders (Hedrick 2011).

Two aspects of a population are of particular interest in population genetics studies: the amount of variation (genetic diversity) and its distribution (genetic structure).

Heterozygosity is a commonly used approach to measure genetic diversity in populations. Heterozygosity is a measure of genetic diversity that uses the Mendelian model to calculate the expected heterozygosity in a population and compare against the observed heterozygosity. Consider a locus of a diploid species that segregates in two alleles, $a1$ and $a2$. Given that diploid species carry two copies of DNA, the possible allele combinations are $(a1, a1)$, $(a1, a2)$, and $(a2, a2)$. The observed heterozygosity ($H_o$) is the proportion of individuals that are heterozygous at a particular locus, while

expected heterozygous ($H_e$) is calculated based on the allele frequencies, assuming Hardy-Weinberg Equilibrium (Hartl and Clark 2007).

The Hardy-Weinberg genetic model states that under specific conditions, a diploid organism with two different alleles at a particular locus will have the allele frequencies defined as $p^2 + 2 * pq + q^2 = 1$, with $p^2$ representing homozygous $p$, (both alleles equal to $p$), $2*pq$ representing the heterozygote frequency, and $q^2$ representing homozygous q. The Hardy-Weinberg equilibrium rules state that a population which 1) is infinite in size, 2) has no selection, 3) has no mutation, 4) has no migration, and 5) has random mating, only requires one generation of random mating to achieve the above described equilibrium. Despite the unrealistic expectations, in particular for natural populations, evaluating the Hardy-Weinberg equilibrium is a common starting point in the investigation of forces influencing the genetic variation in populations (Hartl and Clark 2007).

$$F_{is} = 1 - \frac{H_o}{H_e} \tag{2.1}$$

The inbreeding parameter ($F_{is}$) shown in equation 2.1 and derived from Wright's F-statistics compares the level of observed heterozygosity to the expected under Hardy-Weinberg equilibrium.

Genetic structure is a synonymous of differentiation within a population, leading to sub-populations. There are numerous factors that can influence this division, such as landscape, geographic distance, environmental aspects, among others.

A traditional measure of genetic structure is the Wright's F-statistics

family of indexes. These indices describe the partition of genetic variation with populations (Wright 1943).

$$F_{st} = \frac{H_t - H_s}{H_t} \tag{2.2}$$

where 0 means panmixia and 1 means complete genetic isolation.

Genetic distance is the degree of genetic difference between individuals or populations. The Nei's distance is a traditional genetic distance metric and is estimated as:

$$I = \frac{\sum_l \sum_u X_u Y_u}{\sqrt{(\sum_l \sum_u X_u^2)(\sum_l \sum_u Y_u^2)}} \tag{2.3}$$

$$D_{Nei} = -ln(I) \tag{2.4}$$

The Nei's distance ($D_{Nei}$) assumes a drift-mutation equilibrium and the genetic distance between two populations is directly proportional to the distance value.

Identifying genetic barriers can help better understand the differentiation process on subpopulations (Manel et al. 2003, Storfer et al. 2007, Storfer et al. 2010). Barriers also are important because they decrease gene-flow and promote genetic isolation, and loss of genetic diversity due to genetic drift. Edge detection methods are a particularly important class of techniques in spatial analyses used to detect barrier among groups of genotypes.

### 2.3.2 Molecular markers

Molecular markers usually aim to detect signals of genetic variation. DNA-based molecular markers use sequences of DNA, and the genetic analyses outputs typically inform the presence or absence of a marker or its variation. They represent an important tool in breeding programmes of major crops and have potential to help the improvement of underutilised crops.

The molecular markers currently available have a range of uses and characteristics. They differ with respect to: abundance in the genome, level of detected polymorphism, locus, reproducibility, or technical requirements (Karp 1996). The different types of molecular marker do not necessary imply superiority among them, but they offer distinct particularities that allow the investigation of different scientific questions. The choice of the most appropriate molecular marker used in a scientific project is often defined by a combination of scientific question or goal, time constraint, technical limitations, and financial budget.

The selection of DNA sequence to be used as a molecular marker usually involves a polymorphism assessment. Polymorphism is variation that might occur in a particular locus, identified through the analysis of a sequence of DNA. When a diploid individual has two identical copies of a DNA sequence at a given locus, they are called a homozygote, while individuals with two distinct copies are heterozygotes. An important characteristic of a molecular marker is to be highly polymorphic, which allows the identification of parental lineage and the detection of genetic diversity.

Microsatellite markers have been used in studies of genetic diversity

due to their polymorphism and relative abundance over the genome. A microsatellite is a sequence of one to six nucleotide repeats, in which the repetitions are adjacent, and each distinct amount of repetition identifies a particular allelic variation (see table 2.1). It is highly polymorphic in relation to the amount of repetitions that is useful for measuring genetic relationships between individuals (Kalia et al. 2011). Bambara groundnut is a diploid crop species (Ochatt et al. 2011), thus every cell contains two copies of the DNA information, and the results of the genotyping procedure can indicate the presence of the same allelic variation in both copies (homozygous), or the presence of different variations in each copy (heterozygous).

Table 2.1: Allelic variation of a microsatellite marker based on the DNA sequence "CA". Each table line contains a different allele due to the number of adjacent repetitions.

| Allele |
|---|
| ATTTCAGGCACACACACACAGTCAATCC |
| ATTTCAGGCACACACACACACACACAGTCAATCC |
| ATTTCAGGCACACACACACACACACACACACAGTCAATCC |

## 2.4 GIScience in interdisciplinary projects

Geographic Information Science (GIScience) explores the location property of entities such as objects, events and processes, associating them with coordinates, such as longitude and latitude (Goodchild 2010; Stevens and Pfeiffer 2015). This process may seem straightforward, given the increasing presence of location-based sensors and mapping technologies in our daily lives. However, it is still a challenge because of the nature and contexts of

the facts examined. In museums and herbaria, artefacts and specimens have usually been collected over decades or centuries and their finding location is often ambiguous or very imprecise (Erp et al. 2015). In Health Science, investigation of historical records of disease occurrence in individuals relies on the names of locations and vague addresses. The lack of historical address databases undermines the potential use of those health data (Lash et al. 2012). In Crop Science, breeding programs rely on the collections of seeds available to germplasm banks. Knowledge about the origin of these seeds is necessary to characterise the environment around their collection location, and to go beyond basic measurements of diversity. However, again, location information about their origin may be vague or associated with markets where the seeds were obtained instead of where they grew or were originally sourced from (Richards 2011).

# Chapter 3

# Geographic patterns of genetic variation in an indigenous crop legume


**Abstract**

*Aim*

Genetic variation in plant genetic resources represents an opportunity for development of new crop varieties or improvement of the existing ones. Integrating geographic and crop genetic-associated datasets can strengthen multidisciplinary research by adding new dimensions to the data available and expanding their application to a broader range of research problems. Here, I use Bambara groundnut (*Vigna subterranea* (L.) Verdc.) as a case study, integrating microsatellite data and geographic coordinates, to investigate potential spatial patterns of genetic variation.

*Location*

Sub-Saharan Africa.

*Methods*

I used a molecular marker dataset covering 33 landraces of Bambara groundnut, 128 samples in total, genotyped using 19 microsatellites markers. I measured allele frequency within landraces and performed ordination and clustering analyses to identify groups of genetically similar landraces. I applied the Monmonier's algorithm to identify genetic boundaries and potential biogeographical barriers.

*Results*

I found spatial genetic structure in Bambara groundnut dataset. PCA and clustering analyses identified genetic groups reflecting the known geographical origin of the landraces. Locations of potential genetic boundaries var-

ied depending on the cut-off values used, but there were consistently clear boundaries between the Eastern and Western African landraces.

*Main conclusions*

I conclude that geographic distance is important in determining the genetic structure of Bambara groundnut. I also highlight the challenges of integrating discipline-associated datasets, in particular geographic and genetic data, and its effects on analyses and interpretation of the results.

## 3.1 Introduction

Diversity, through genetic variation, is an important characteristic of plant genetic resources because it provides a pool of genetic material to develop and improve crops with desirable agronomic traits (FAO 2010). Genetic variation is the result of processes such as mutation, selection, gene flow, and genetic drift. The effect of these processes, individually and combined, varies in shape and intensity from one place to another, leading to differentiation among geographically separated populations (Jarvis et al. 2005). Given the spatial context where these processes occur, space and spatial information are a potentially important element in the genetic variation of plants and can aid in conserving these genetic resources and making sustainable use of these resources to guarantee food security.

The access to, and use of, location-based data in agriculture have been fundamentally changed by technologies for collection and dissemination of geolocated information (Jackson et al. 2011). In parallel, an increasing amount of geospatial data has been made available through governmental agencies, research centres, and citizen science projects. Integrating these datasets in interdisciplinary projects can strengthen research by adding new

dimensions to the data available (Santos et al. 2016). In particular, integration of geographic and genetic-associated datasets can help 1) to identify geographic, environmental and anthropogenic barriers to gene flow, 2) to infer evolutionary processes that have influenced the geographic patterns in these crop species, 3) to investigate conservation of vulnerable genetic material, and 4) to help understanding how these human-associated species adapt to different environments.

Integrating geographic and genetic data may seem straightforward, given the increasing presence of location-based sensors and mapping technologies in our daily lives. However, it is still a challenge because of the nature and contexts of the facts examined. In museums and herbaria, specimens have usually been collected over decades or centuries and their finding location is often ambiguous or very imprecise (Erp et al. 2015). In Health Science, investigation of historical records of disease occurrence in individuals relies on the names of locations and vague addresses. The lack of historical address databases undermines the potential use of those health data (Lash et al. 2012). In Crop Science, breeding programs rely on the collections of seeds available to germplasm banks. Knowledge about the origin of these seeds is necessary to characterise the environment around their collection location, and to go beyond basic measurements of diversity. However, again, location information about their origin may be vague or associated with markets where the seeds were obtained instead of where they grew or were originally sourced from Richards (2011).

Geographical patterns of genetic variation in indigenous and under-

utilised crops (plant species that usually live in the interface between natural and agricultural systems) are the outcome of evolutionary processes and human demographic events (Manni et al. 2004). These crops represent a large group of plant species, whose development occurred over years of traditional small-scale farming practices, and are usually cultivated from landraces (Villa et al. 2005). Landraces can be defined as unimproved domesticated varieties of plant species, that have historical origin and identity. These varieties have not undergone formal breeding improvement, which restricts their capacity to achieve higher yields. They are often genetically diverse, thought to be adapted to local environmental conditions, and related to traditional farming systems. Farmers typically rely on landraces because of their resilience in adverse environments. However, the lack of better understanding about their geographic origin and distribution, genetic characterisation, and role in farming community culture, restricts their development. There is also a risk of their extinction due to the increasing adoption of staple crops, usually associated with higher income. They usually lack formal research, breeding programs, and commercial varieties (Galluzzi and Noriega 2014).

In the last 20 years, various molecular marker systems have helped in understanding the genetic diversity in Bambara groundnut. Pasquet et al. (1999) used isozymes to investigate the genetic diversity among 79 accessions of domesticated and 21 accessions of wild populations. Massawe et al. (2003) used Amplified Fragment Length Polymorphism (AFLP) to analyse 16 landraces, and then cluster analysis that grouped the landraces in 3

clusters mainly according to their places of collection or geographic origin. Ntundu et al. (2004) also used AFLP to analyse landraces in Tanzania, identifying two clusters associated with geographic origin. To my knowledge, Basu et al. (2007) were the first to use microsatellite markers, but with no further exploration of spatial patterns. Olukolu et al. (2012) used Diversity Arrays Technique (DArT) to evaluate seeds representing 25 African countries, and explored allele frequency and diversity indices by region. Siise and Massawe (2012) and Molosiwa et al. (2015) examined genetic diversity in landraces from throughout sub-Saharan Africa and Ghana respectively, using microsatellite markers. More recently, studies have investigated Bambara groundnut genetic association with low temperatures (Bonthala et al. 2016), water stress (Khan et al. 2017), and relationship with other related legumes (Ho et al. 2017). Despite the increasing interest in the genetic structure and diversity of Bambara groundnut, there is a gap in our knowledge of the spatial patterns of genetic variation and their relationship with biotic and abiotic factors.

In this study, I conduct an exploratory analysis of the genetic variation in Bambara groundnut (*Vigna subterranea* (L.) Verdc.) (see Figure 3.1), investigating its patterns of variation and clustering in genetic/Euclidian space, and mapping those results in geographic space. I aim to use these results to have a better understanding of the genetic variation available in the crop and develop hypotheses about the factors that influence its geographic distribution.



Figure 3.1: Bambara groundnut plants cultivated in the Crops for the Future glasshouses at Sutton Bonington campus, University of Nottingham. Pictures taken on 2016.

## 3.2 Material and methods

### 3.2.1 Genetic data

I quantified the geo-genetic structure of Bambara groundnut throughout sub-Saharan Africa using a genetic dataset composed of genotyping information based on presence or absence of 19 microsatellite markers (Molosiwa et al. 2015). The origin of this dataset back to 2008 and 2009, when Molosiwa et al. (2015) used seeds of 33 landraces to cultivate Bambara groundnut in a glasshouse experiment at the University of Nottingham, Sutton Bonington, UK, and in a field experiment conducted at the Botswana College of Agriculture, Gaborone, Botswana.

These plants were genotyped using microsatellite markers (see Table 3.1), also known as simple sequence repeats (SSR). Bambara groundnut is a diploid species (Ochatt et al. 2011), thus every cell contains two copies of the DNA information, and the results of the genotyping procedure can indicate the presence of the same allelic variation in both copies (homozygous), or the presence of different variations in each copy (heterozygous).

Table 3.1: List of all the 19 primers used in the genotyping using SSR.

| Marker | Sequence |
|---|---|
| primer 7 | GTAGGCCCAACACCACAGTT |
| primer 10 | TCAGTGCTTCAACCATCAGC |
| primer 15 | AGGAGCAGAAGCTGAAGCAG |
| primer 16 | CCGGAACAGAAAACAACAAC |
| primer 19 | AGGCAAAAACGTTTCAGTTC |
| primer 21 | CAAACTCCACTCCACAAGCA |
| primer 23 | CAGTAGCCATAATTTGCTATGAACA |
| primer 30 | AATGCAAGATTTTGGCTTGG |
| primer 31 | GCTAAGGTGGAGTGGTGGAA |
| primer 32 | TTCACCTGAACCCCTTAACC |
| primer 33 | ACGCTTCTTCCCTCATCAGA |
| primer 37 | CCGATGGACGGGTAGATATG |
| primer 44 | TGTGGGCGAAAATACACAAA |
| primer d8 (G372AB4-D8) | GCATCTTTACAGCAAGAGTTTCAA |
| primer d11 (G180B2-D11) | GAGGAAATAACCAAACAAACC |
| primer d14 (G240-9-B2-D14) | GAACGAAGCCAGGATAATGATAGT |
| primer d15 (G358B2-D15) | TGACGGAGGCTTAATAGATTTTTC |
| primer e7 | CATGATTTGTTGTGATGATGAT |
| mbam2co80 | GAGTCCAATAACTGCTCCCGTTTG |

Molosiwa et al. (2015) used seeds from IITA's germplasm database, that also provides information about the geographic origin or locale of collection of those seeds. For the seeds used in this research, the geographic information is available only for the landraces, groups of similar genetic material, so I aggregated the plants in populations according to the respective landraces (see Figure 3.2). In the Molosiwa et at. (2015) dataset, each landrace population is composed of three or four plants.

### 3.2.2 Georeferencing the genetic data

The plants are the basic unit of genetic information in the dataset, but I aggregate them at the landrace level, given that the geographical reference was only available at this aggregation level. I applied quality control procedures on the coordinates in order to validate the geographical information. Firstly, coordinates were mapped against the African land area, identifying landraces falling outside the African continent. A common reason for the type of problem is the inversion of the longitude and latitude values. Secondly, I validated the country of each landrace, as identified in the attributes of the dataset, with the country that contains the coordinates. This test can identify errors associated with datum and projection mistakes, as well as information collected from devices with low accuracy or information identified in maps with low resolution. Given that the geographical information associated with each landrace indicates either original cultivation area or markets and other places that exchange or trade the seeds, there is potential uncertainty in the geographical information, as well as potential to identify multiple landraces collected from same area (in particular trade markets).

Each Bambara groundnut landrace's geographic location was defined according to information available at the International Institute of Tropical Agriculture (IITA) germplasm database. The geographic coordinates were available in the WGS84 geodetic reference system, in decimal degrees(see Figure 3.2). All geographic coordinates were validated using the variables

Table 3.2: List of all the 33 landraces used in the analyses.

| Landrace | Longitude | Latitude | Landrace | Longitude | Latitude |
|---|---|---|---|---|---|
| 289 | 2.00 | 7.10 | S19-3 | 14.46 | -17.26 |
| BOTS 5 | 25.54 | -24.39 | AHM968 | 14.46 | -17.26 |
| BOTS 1 | 27.30 | -21.10 | SB16 5A | 27.05 | -26.42 |
| VSSP 6 | 11.31 | 3.52 | SB4-2 | 27.05 | -26.42 |
| 536 | 13.23 | 8.29 | 9 | 3.54 | 7.29 |
| 476 | 13.24 | 9.15 | 32 | 3.54 | 7.29 |
| 484 | 13.22 | 9.26 | 329 | 12.06 | 8.26 |
| 231 | -15.58 | 13.22 | 335 | 11.28 | 9.05 |
| 144 | -0.11 | 5.60 | 286 | 4.50 | 9.20 |
| 118 | -3.58 | 7.04 | 390 | 30.13 | 12.51 |
| 1276 | -5.32 | 7.32 | UNIS R | 31.18 | -26.28 |
| 793 | 34.32 | 0.10 | 385 | 35.39 | -7.43 |
| 792 | 34.32 | 0.10 | DOD R | 35.44 | -6.10 |
| 89 | -4.22 | 13.48 | 696 | 32.15 | -14.10 |
| S-1913 | 14.46 | -17.26 | 790 | 31.30 | -11.40 |
| Mahenene Black | 14.46 | -17.26 | 754 | 32.40 | -10.25 |
| AHM753 | 14.46 | -17.26 | | | |

available in the WorldClim[1] database and the GADM[2] database of Global Administrative Areas. The use of both datasets turned out to be efficient in detecting anomalies such as landraces located outside the shorelines (located at the sea) and locations with no data in the WorldClim (e.g. water bodies). I used the landrace locations as the basic geographic unit for attribute collection (environmental used in Chapters 4 and 5, and linguistic used in Chapter 4).

### 3.2.3 Population structure

An important question in population genetics is what factors or phenomena influence the genetic structure of a population. In the context of population genetics, genetic structure in a population means consistent allele frequency difference among subpopulations. I used descriptive statistics, principal

[1] www.worldclim.org
[2] gadm.org

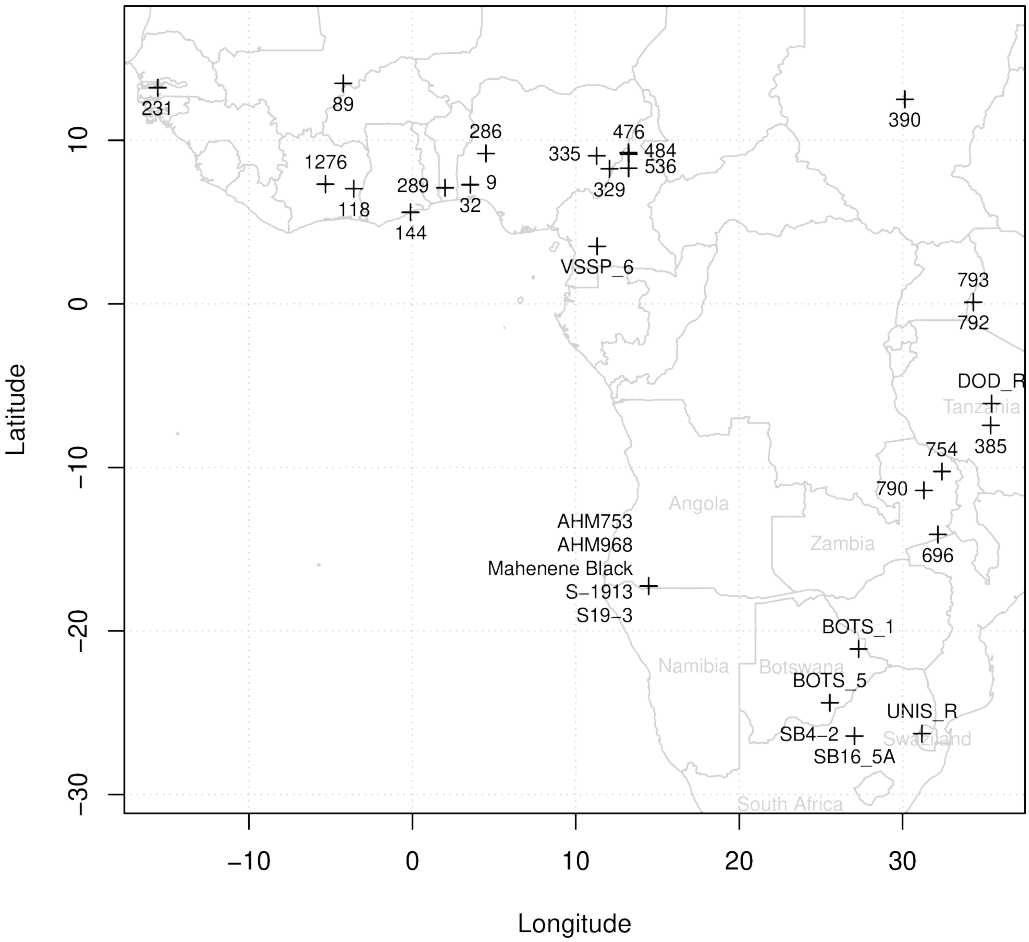



Figure 3.2: Origin or local of the collection of the germplasm (seeds) of the landraces of Bambara groundnut. Symbols (plus) identify the location of the landraces. Names and numbers identify the landraces according to the IITA database naming system.

component analysis, and cluster analysis to measure the level of genetic differentiation among Bambara groundnut landraces.

In the population genetics context, descriptive statistics, such as the F-statistics, aim to estimate the partitioning of genetic information within and between subpopulations. $F_{ST}$ and $F_{IS}$ are two popular indices, usually applied to populations and individuals, respectively. $F_{IS}$ is an inbreeding coefficient of individuals within their population, and is based on the homozygosity excess relative to Hardy-Weinberg equilibrium expectations. $F_{ST}$ measures the degree of inbreeding of sub-populations relative to the

total population (Rowe et al. 2017).

I calculated the allele frequency aggregated by landraces using the package ADEGENET (Jombart and Ahmed 2011) and conducted a principal component analysis, using the function prcomp, for the allele frequency matrix to obtain a smaller set of independent variables that capture the overall genetic variation in the dataset. Then, I applied clustering algorithms (K-means) to partition the landraces by genetic similarity (based on the allele frequencies) using the function kmeans. Clustering algorithms applied to genetic data provide a valuable method to explore genetic structure and are an important tool in defining genetic conservation strategies and population genetic studies (Hubisz et al. 2009). In the basic model, these algorithms aim to partition a set of X individuals into Y populations (clusters) that can be either known or not known, a priori. I used K-means to partition the landraces into discrete clusters and evaluate the spatial distribution of similar genetic landraces. The number of clusters is an a priori requirement for K-means, and the estimation of the number of groups is the subject of numerous debates (Jain 2010). I then ran the clustering analysis several times, varying the number of clusters with K values from 2 to 9.

I tested for the existence of potential genetic barriers using the Monmonier's maximum difference algorithm (Monmonier 1973) using the function optimized.monmonier. The spatially explicit method aims to highlight boundaries within a dataset. It creates a graph based on the sample locations and defines the weights of the connection as a measure of genetic dissimilarity. The boundaries are determined by identifying abrupt value

differences between nodes in the graph.

I used the statistical software R version 3.3.0 (R Core Team 2017) for the analyses in this study.

## 3.3 Results

Values for the F statistic indicate genetic structure in the Bambara groundnut's dataset, with $F_{IS}$ (0.80) and $F_{ST}$ (0.43) in the total population. Pairwise $F_{ST}$ across landraces varied from 0.24 to 0.66 with mean 0.43, indicating the presence of genetic differentiation among the landraces. As for the principal component analysis, I retained the first two components that represented 14.22% (7.97% and 6.25%, respectively) of the genetic variation of the dataset. The low variance explained by the PCA suggests low covariance among alleles measured here. Figure 3.3 shows a significant split between landraces along PC1, which suggests a potential latitudinal gradient. The top three loadings were primers e7 allele 85 (0.21), e7 allele 79 (-0.20), and 44 allele 201 (-0.17) for PC1, and primers 19 allele 270 (-0.23), d15 allele 196 (0.17), primer 16 allele 186 (0.17) for PC2 (see table 3.3 for top 20 contributors).

As landraces constitute the main elements to be analysed, and up to five landrace locations in the IITA database coincided, it was necessary to add a very small amount of noise to some geographic coordinates, in order to allow some of the spatial analysis algorithms to run. This noise did not affect the extraction of information from WorldClim and other databases, since the range of noise was smaller than the minimum area of information

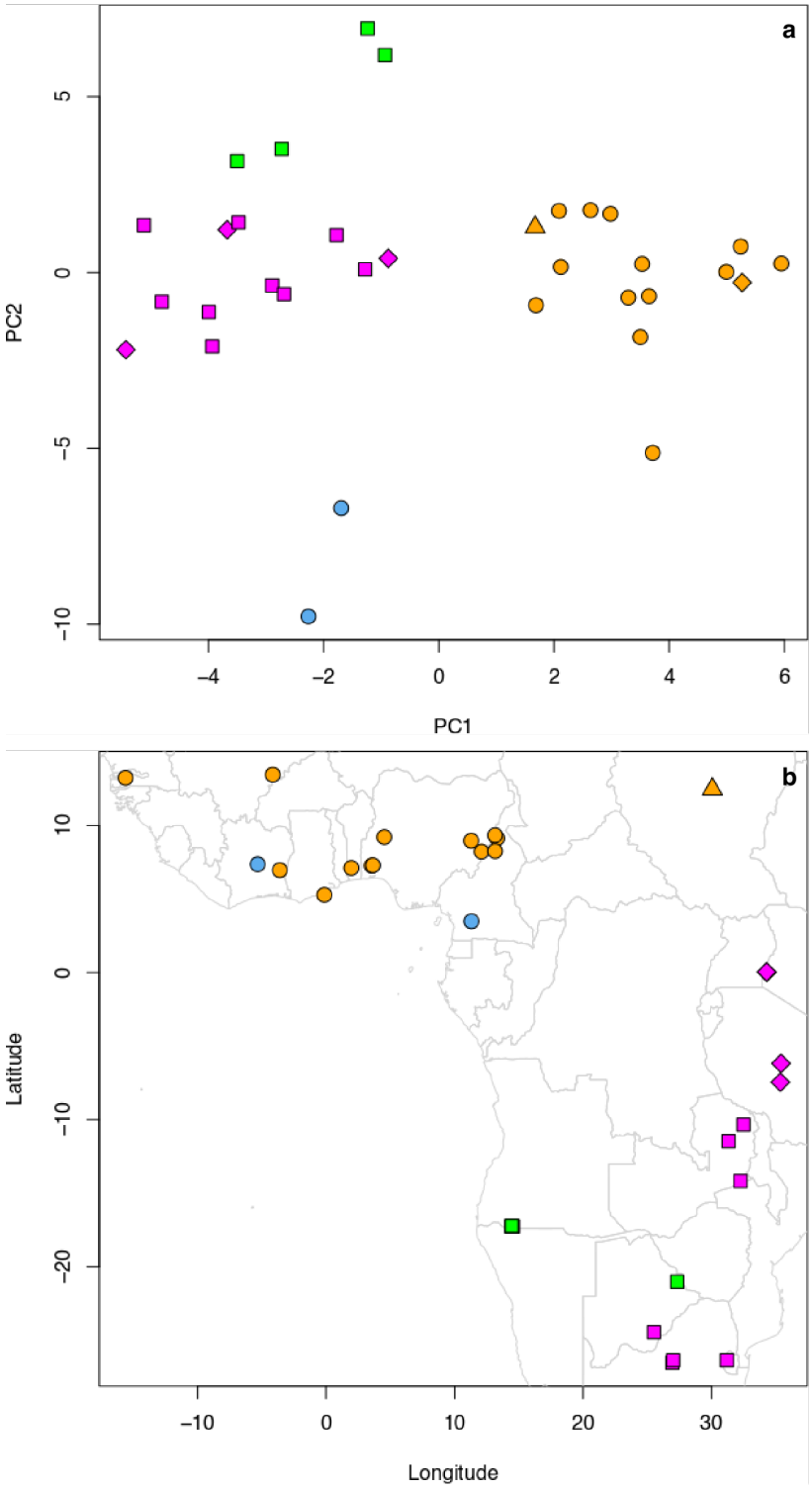



Figure 3.3: Ordination (PCA) and clustering (K-means) results plotted using PC1 and PC2 (a), and longitude and latitude (b). Symbols identify landraces origin location aggregate by African regions: Western (plus), Eastern (square), Southern (cross), and Central (circle). Colours represent the cluster attribution as informed by K-means results for K = 4.

in those databases, but allowed the use of algorithms that rely on distances between elements of the analyses being greater than zero.

I mapped the results of the clustering analyses, which reinforced the results of the PCA when K = 4 (I ran the K-means algorithm with K values from 2 to 9), showing four distinct geographic groups: Eastern, Central, Western, and Southern Africa. I defined a priori K=4 that reflects the number of Africa divisions identified in the IITA database. However, the clustering results showed similar patterns (see Figures 3.5 and 3.6 for K=2 and K=3). Genetic clusters reflected the known geographical origin of the landraces. Placement of genetic boundaries in sub-Saharan Africa varied significantly, depending on the cut-offs values. However, in most of the simulated values, it was possible to identify well-defined boundaries around East African landraces, supporting the results of the PCA and clustering analysis and highlighting the genetic coherence of this geographic group. Monmonier's results also supported the results of the PCA and clustering analysis, identifying genetic boundaries and potential biogeographical barriers between the four geographic groups mentioned earlier in the paragraph (see Figure 3.4).

## 3.4 Discussion

I investigated the genetic structure of Bambara groundnut in sub-Saharan Africa. A population is considered genetically structured when a spatially heterogeneous distribution of alleles occurs, usually leading to subpopulations. This imbalance in the genetic variation can be the result

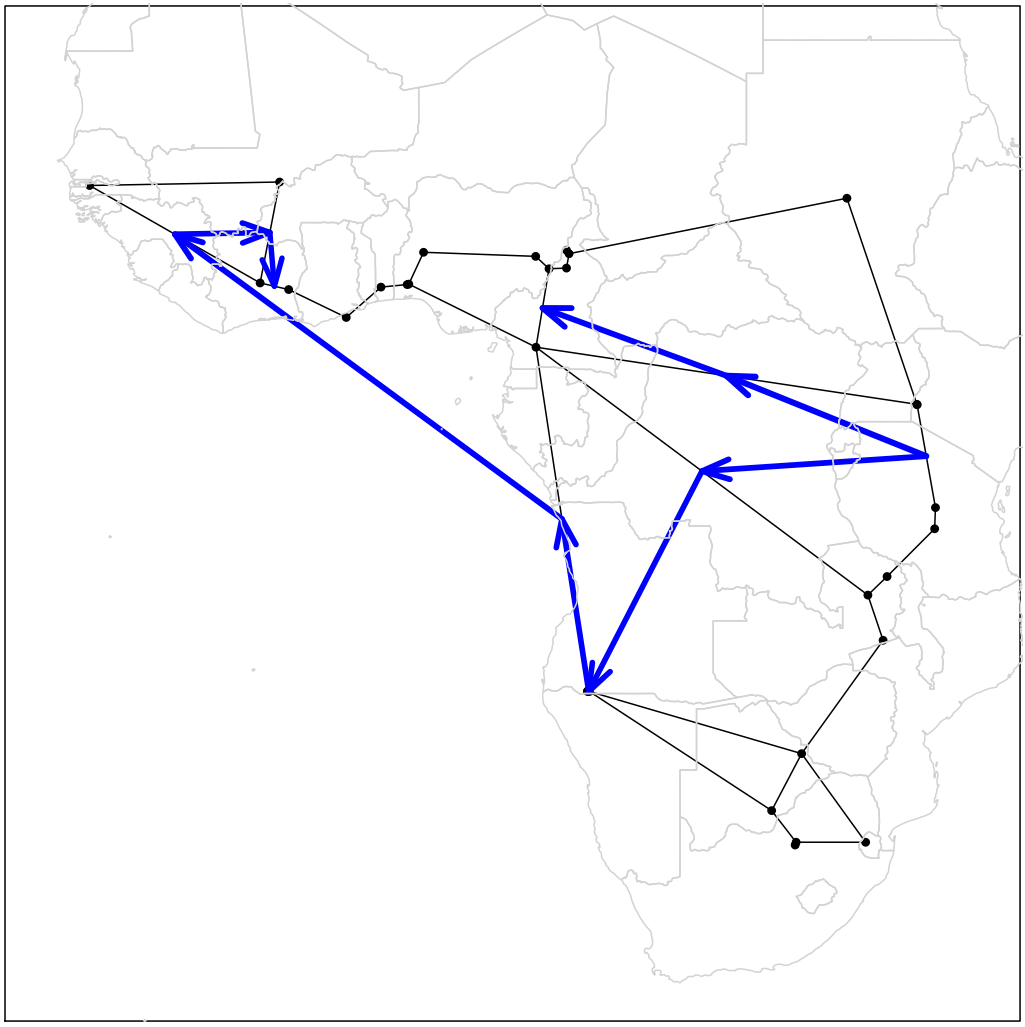


Figure 3.4: Monmonier's algorithm genetic boundaries identification. The black circles indicate landraces location, the black edges between the nodes indicates connection according the Gabriel's network methods, and the blue lines with arrows indicates genetic barrier.

of different processes happening in the geographic surrounds of the sub-populations, driven by environmental and/or anthropogenic forces. Here, I tested for the existence of genetic structure in the Bambara groundnut's dataset using descriptive statistics, and secondly explored the spatial distribution of sub-populations using ordination and clustering analyses, and finally, explored the existence of genetic barriers using Monmonier's algorithm.

### 3.4.1 Genetic diversity

The results seem to be consistent with others' research which found genetic structure in Bambara groundnut throughout sub-Saharan Africa (Molosiwa et al. 2015). A possible explanation for this might be that distance has negatively affected the exchange of genetic material between landraces (gene flow) while promoting differentiation through genetic drift and local adaptation. The origin of the seeds used in this study, and more generally, the origin of Bambara groundnut seeds available in germplasm databases such as the one at IITA, cover most of sub-Saharan Africa. This area contains a number of distinct environments with particular characteristics related to climate, weather, and topography. There are reports of Bambara groundnut being cultivated in a wide range of climatic zones; it has the potential to be cultivated from sea level to up to 1600m in altitude, with annual rainfall as low as 200 mm in Botswana to 1250 mm in Swaziland. It is possible that, over time, exposure to distinct environmental conditions has benefitted alleles that increase fitness in these environments, fixing these alleles in the landraces (sub-populations) while alleles that did not increase fitness in the specific environment were lost.

### 3.4.2 Genetic sub-populations

Clusters of genetic data reflect landrace origins. Multiple clustering analysis using microsatellites consistently assigned landraces in groups that are spatially close. It is therefore possible that the molecular markers captured signals of genetic difference between landraces in the Western Africa

and Southern/Eastern Africa. Given the significant distance among those groups, a possible explanation might be that (1) the genetic material gradually migrated from its commonly accepted origin, Cameroon and Nigeria, to Eastern Africa and Western/Southern Africa. Further movement of genetic material did not increase the gene flow, allowing processes such as local adaptation to act on these sub-populations. This hypothesis of genetic structure linked to natural or artificial movement of genetic material (seeds) reflects the Vavilov (Harlan 1971) concept of centres of diversity. Another possible explanation for the genetic structure is (2) the existence of second centre of biodiversity and domestication in Eastern or Southern Africa. The PCA analysis showed relatively low variation explained by the first components, which could be explained by low collinearity between markers. This inconsistency may be due to the choice of molecular markers, which would aim for highly polymorphic markers without linkage association, that would strengthen the results of genotyping analysis.

Monmonier's algorithm detected a genetic barrier in the border between the two main clusters, Western Africa and East/Southern Africa (see Figure 3.4). The low density of samples in that region may have influenced the result, but there are alternative explanations such as (3) the existence of environmental and topographic barriers that overlap the genetic barrier. The African Rift Valley is a natural barrier, that splits the populations, and might have acted as a barrier to the exchange of genetic material in the past. In Western Africa, the tropical forest was a natural barrier to the artificial movement of seeds (through human migration) and natural

movement due to the low height of Bambara groundnut, which restricts the access to sunlight in areas of high vegetation such as tropical forests. Monmonier's algorithm is sensitive to a number of parameters, and several simulations using the algorithm failed to identify genetic barriers. A possible explanation could be (4) the effect of isolation by distance (IBD) in Bambara groundnut dataset, where a genetic gradient is present, with no particular sharp difference in the genetic distance among the landraces. These findings may be somewhat limited by the relatively low sample density, which may not have detected small-scale genetic barriers. It is possible to hypothesise that these conditions are less likely to occur with further analyses including a higher density of plant samples.

Table 3.3: The top 20 loadings for the components (PC1 and PC2) of the PCA based on allele frequency.

| | PC1 | | | PC2 | | |
|---|---|---|---|---|---|---|
| | Primer | Allele | loading | Primer | Allele | Loading |
| 1 | primer e7 | 85 | 0.0200 | primer 19 | 270 | 0.0244 |
| 2 | primer e7 | 79 | 0.0191 | primer d11 | 202 | 0.0214 |
| 3 | primer 30 | 248 | 0.0158 | primer 23 | 236 | 0.0195 |
| 4 | primer d15 | 186 | 0.0152 | primer 32 | 253 | 0.0195 |
| 5 | primer 44 | 201 | 0.0149 | primer d11 | 188 | 0.0195 |
| 6 | primer 15 | 212 | 0.0149 | primer d15 | 196 | 0.0192 |
| 7 | primer 21 | 250 | 0.0141 | primer 31 | 193 | 0.0190 |
| 8 | primer d14 | 154 | 0.0134 | primer 33 | 166 | 0.0188 |
| 9 | primer 16 | 190 | 0.0131 | primer d15 | 200 | 0.0179 |
| 10 | primer 44 | 200 | 0.0129 | primer 44 | 198 | 0.0170 |
| 11 | primer 30 | 246 | 0.0123 | primer d11 | 204 | 0.0167 |
| 12 | primer d8 | 197 | 0.0122 | primer d14 | 148 | 0.0166 |
| 13 | primer d14 | 138 | 0.0120 | primer d11 | 196 | 0.0163 |
| 14 | primer d15 | 196 | 0.0119 | primer 19 | 273 | 0.0163 |
| 15 | primer d8 | 200 | 0.0118 | primer d14 | 150 | 0.0160 |
| 16 | primer 32 | 222 | 0.0118 | primer 7 | 216 | 0.0154 |
| 17 | primer 16 | 186 | 0.0109 | primer 16 | 186 | 0.0153 |
| 18 | primer 44 | 202 | 0.0109 | primer 15 | 238 | 0.0153 |
| 19 | primer 37 | 262 | 0.0105 | primer d14 | 138 | 0.0146 |
| 20 | primer 15 | 236 | 0.0105 | primer 16 | 188 | 0.0145 |

### 3.4.3 Integrating geographic and molecular data in analysis

Integrating geographic and genetic datasets is a common requirement to investigate various research problems that exist beyond disciple boundaries. The geo-referencing of germplasm databases after the collection of genetic material brings a series of potential issues for further analyses. I identified inconsistencies in the geographic coordinates of some of the landraces, although for a continental area of study, such as sub-Saharan Africa, manual adjustments are likely to not affect the spatial analysis results. However, the data must be interpreted with caution and conclusions must take into account the broad spatial scale of the study and the number of sampled locations.

## 3.5 Conclusion

By investigating the genetic diversity and structure of a dataset of Bambara groundnut throughout Sub-Saharan Africa, I conclude that geographic distance plays an important role in determining the genetic structure of this species. These results guide the development of hypotheses that I present and test in Chapters 3 and 4. I also highlight the challenges of integrating discipline-associated datasets, in particular, geographic and genetic data, and its effects on analyses and interpretation of the results.

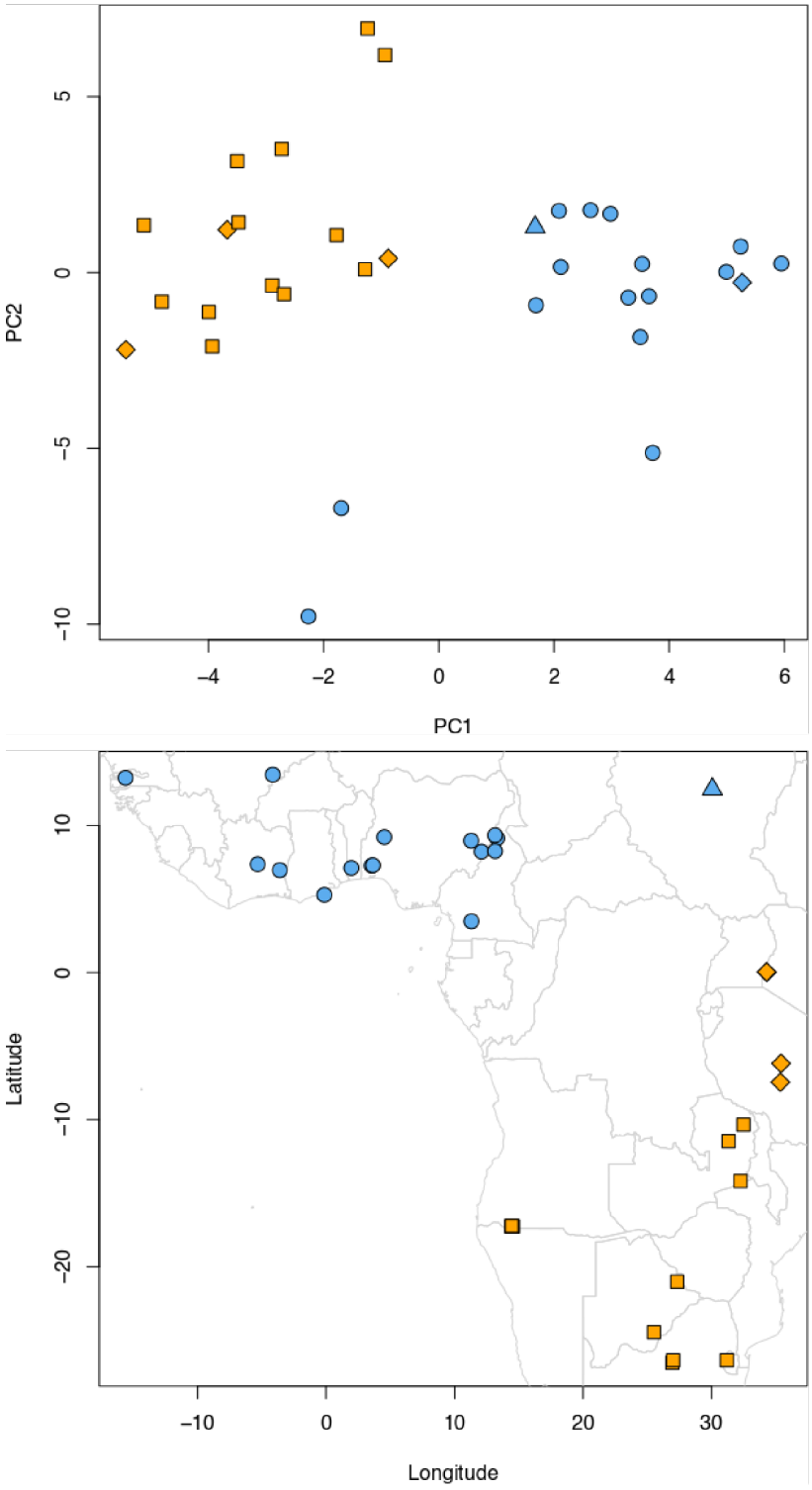



Figure 3.5: Ordination (PCA) and clustering (K-means) results plotted using PC1 and PC2 (a), and longitude and latitude (b). Symbols identify landraces origin location aggregate by African regions: Western (plus), Eastern (square), Southern (cross), and Central (circle). Colours represent the cluster attribution as informed by K-means results for K = 2

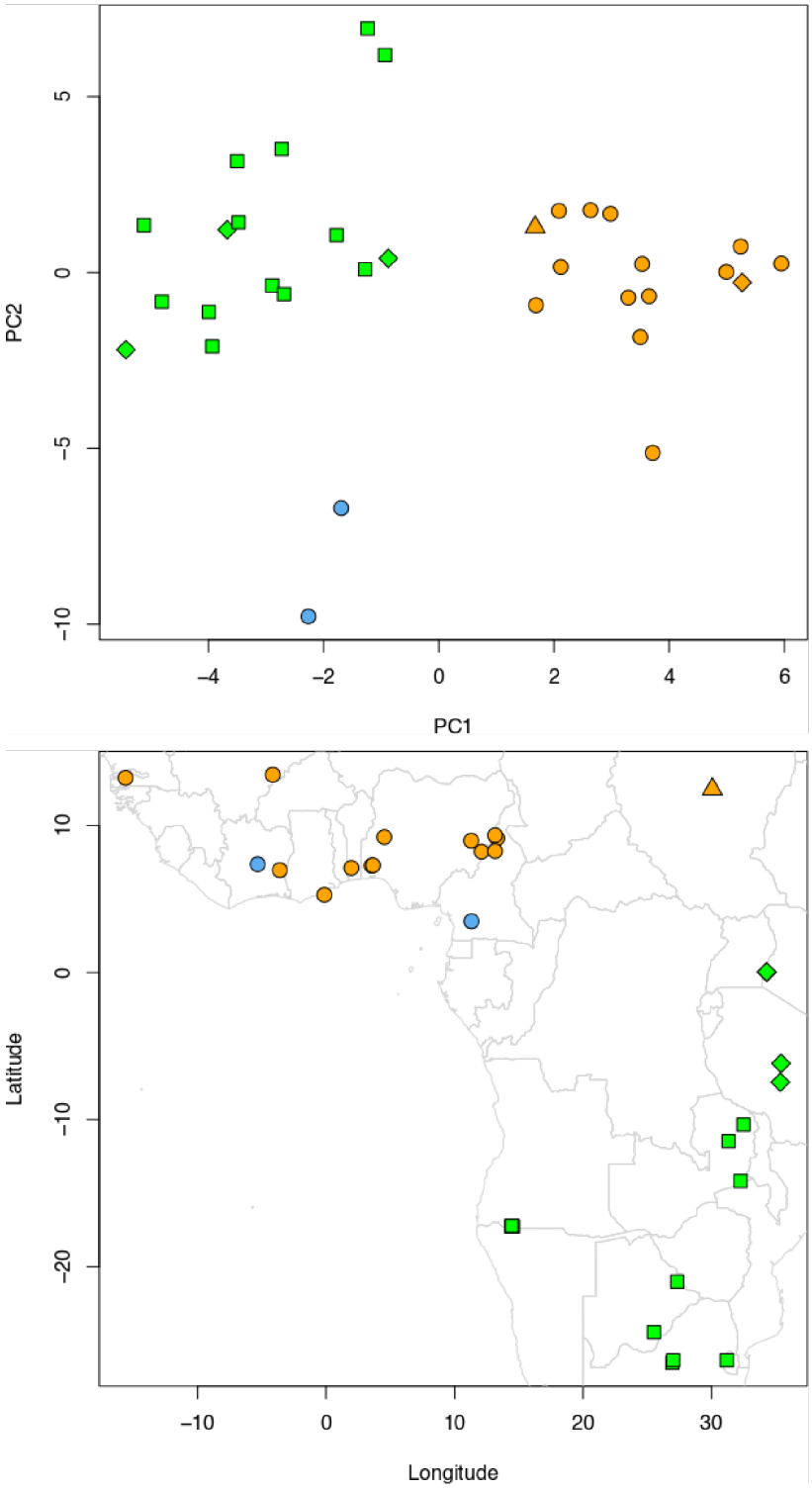



Figure 3.6: Ordination (PCA) and clustering (K-means) results plotted using PC1 and PC2 (a), and longitude and latitude (b). Symbols identify landraces origin location aggregate by African regions: Western (plus), Eastern (square), Southern (cross), and Central (circle). Colours represent the cluster attribution as informed by K-means results for K = 3.

# Chapter 4

# Human migration history is reflected in the geo-genetic variation in Bambara groundnut in sub-Saharan Africa

**Abstract**

*Aim*

Human-associated plant species have the potential to help support food security in the most vulnerable human populations. Understanding the ecological and anthropogenic processes that influence their genetic variation is an essential step in changing farming practices from landraces to varieties bred for desirable agricultural traits. I test whether environmental differences, the distance among populations, and human activity influence the geo-genetic structure and distribution of Bambara groundnut (*Vigna subterranea* (L.) Verdc.), an underutilised crop, throughout sub-Saharan Africa. I test three hypotheses which may explain spatial genetic variation observed in this species; (H1) genetic similarity positively correlated with environmental similarity; (H2) isolation by distance among populations; (H3) human movement as a dispersal vector.

*Location*

Sub-Saharan Africa.

*Methods*

I used a database of 33 landraces of Bambara groundnut, 128 samples in total, and measured the distance or similarity among the landraces using four distinct dimensions: genetic, geographic, environmental, and linguistic. I tested the hypotheses using Mantel tests and Structural Equation Modelling (SEM).

*Results*

Genetic similarity among landraces of Bambara is most strongly related to geographic proximity and, to a lesser extent, linguistic similarity between human communities. There was no relationship between genetic similarity and environmental differences.

*Main conclusions*

Distance decay and cultural and social practices are important in shaping the geo-genetic structure of human-associated plant species. However, I found no correlation between genetic similarity and environmental similarity, that might suggest the absence of local climate adaptation. These findings illuminate the role of these species in the cultural identity of distinct ethnolinguistic human populations, a key factor if we, as a scientific community, want to tackle the food security problem respecting human idiosyncrasies.

## 4.1 Introduction

Modern agriculture faces the challenge of keeping crop production at levels that can meet the demands of a growing human population. The problem is not limited just to increasing production, but also to the pursuit of balance between demands for energy, water, food security, meat and dairy products and constraints such as the limited amount of land available and suitable for agriculture, uncertainty about the effects of climate change on agricultural practices, and conservation of biodiversity (Foley et al. 2011; Ray et al. 2013; Godfray and Garnett 2014).

In the last decades, a relatively small group of crops of high market value and high productivity have dominated most of the land dedicated to agriculture, potentially narrowing the options for nutritious food, undermining food security of vulnerable groups and posing a threat to biodiversity

(Beddington 2010). In 2013, thirty crops accounted for 90% of the global harvested area. Only three major crops, wheat (*Triticum spp.*), rice (*Oryza sativa*) and maize (*Zea mays*), occupied respectively 16.3%, 13.8%, and 12.3% of the total area (FAOSTAT 2017). The lack of diversity and limited growth potential of established, dominant, crops, threaten future food security under a growing human population and rapid environmental change (Galluzzi and Noriega 2014).

Underutilised crops, also known as minor crops or orphan crops, have potential to contribute to future food security (Padulosi et al. 2014, Cheng et al. 2017, Massawe et al. 2016). These underutilised crops species, mainly cultivated under traditional farming systems and initially developed by indigenous people, have undergone centuries of cultural and biological evolution and are available in different stages of domestication throughout the globe (Galluzzi and Noriega 2014). They usually exist in the form of landraces (locally developed mixtures of genotypes) rather than cultivars (genetically pure varieties bred specifically for particular agro-ecological conditions or production systems) (Zeven 1998).

The very different domestication and evolutionary history of these underutilised crops, compared to dominant monocultures, means they could have unique adaptations that better equip them for environmental change, but little is known about the factors that determine genetic diversity in these crops. Understanding how geographic, environmental and anthropogenic factors contribute to the genetic variation of underutilised crops can provide useful information for agricultural management and conserva-

tion practices (Wang et al. 2013, Li et al. 2008). Moreover, because underutilised crops exist at the interface between ecological and cultural systems, they can also aid our understanding of how human-plant associations have evolved in early human-associated species in human-dominated landscapes.

Current research on the genetics of underutilised crops is at different stages, but there is considerable focus on exploring their genetic diversity and genetic structure, aimed at conserving genetic material (Galluzzi and Noriega 2014, Galluzzi et al. 2015, Padulosi et al. 2011, Padulosi et al. 2002, Camacho-Henriquez et al. 2015) and developing varieties (Mayes et al. 2012, Ebert et al. 2014, Chivenge et al. 2015, Abberton et al. 2016, Barbieri et al. 2014). Better understanding from such research may make it possible to identify conservation priorities to reduce or stop genetic erosion (Wouw et al. 2010) and maintain the genetic diversity of these species. The association between genotype and phenotype can also be explored, potentially supporting selection of genetic material or accessions for breeding programs. Recent research has identified that underutilised crops usually harbour high levels of genetic diversity, which may reflect their existence in the form of landraces. This may facilitate their adaptation to a broader range of environments (Camacho-Henriquez et al. 2015; Galluzzi and Noriega 2014).

Here, I focus on the spatial genetic diversity of Bambara groundnut (*Vigna subterranea* (L.) Verdc.), a grain legume classed as an underutilised crop (Azam-Ali et al. 2001). This crop is mainly cultivated in sub-Saharan Africa and may have potential to nourish vulnerable human communities

and provide an extra source of income for small farmers. I hypothesise that the current spatial distribution of the genetic variation of Bambara groundnut, explored in Chapter 3, can be explained by three processes acting over recent centuries: (H1) genetic similarity positively correlated with environmental similarity where this crop has been cultivated, (H2) isolation by distance among the populations that have genetically drifted apart, and (H3) human movement and cultural connections that have helped the dispersal of the seeds and increased gene flow of Bambara groundnut throughout sub-Saharan Africa. I test the role of these processes by analysing the relationship between landraces using four measures of distance: genetic, geographic (isolation by distance and genetic drift), environmental and linguistic (human-mediated gene flow).

## 4.2 Materials and methods

### 4.2.1 Study species

Bambara groundnut is a legume and a food crop, cultivated throughout sub-Saharan Africa, and it is thought to have origins in the border between Nigeria and Cameroon (Azam-Ali et al. 2001). Mainly grown by small-scale farmers as a subsistence crop, Bambara groundnut helps to provide food security and can also provide an alternative source of income in poor communities. The high nutritional value of Bambara groundnut provides a cheap source of protein, making it a good supplement to a cereal-based diet (Karunaratne et al. 2011). Farmers often prefer its low, but stable and

reliable, yield to high-yielding groundnuts that are usually associated with more yield volatility (Azam-Ali et al. 2001). Due to this physiological characteristic, there is an agronomic advantage in growing Bambara groundnut in low rainfall areas compared to other legumes.

Bambara groundnut is present in a wide range of climatic zones and can be cultivated from sea level to up to 1600 m in altitude, and areas with annual rainfall as low as 200 mm in Botswana to 1250 mm in Swaziland (Azam-Ali et al. 2001). Currently, there are no bred varieties of Bambara groundnut, and farmers rely on the use of landraces (Aliyu et al. 2016). Landraces are a locally developed mixture of genotypes that may have been affected by physical constraints such as agro-meteorological conditions or cultural preferences such as taste, colour and size. Landraces are more genetically and physically diverse than formal breeds, and as such, have a broad range of genotypic variation, with the potential to harbour beneficial genetic information that can be useful for the development of new varieties.

### 4.2.2 Estimating genetic distance between landraces

I quantified the geo-genetic structure of Bambara groundnut throughout sub-Saharan Africa using a genetic dataset composed of genotyping information based on presence or absence of 20 microsatellite markers, also known as simple sequence repeats (SSR)(Molosiwa et al. 2015). A microsatellite marker is a sequence of one to six nucleotide repeats, in which the repetitions are adjacent, and each distinct amount of repetition identifies a particular allelic variation. It has desirable genetic characteristics for plant breeding

programs such as relative abundance over the all extension of the genome in coding and non-coding regions, and is highly polymorphic in relation to the amount of repetitions that is useful to measure genetic relationships between individuals (Kalia et al. 2011). Bambara groundnut is a diploid species (Ochatt et al. 2011), thus every cell contains two copies of the DNA information, and the results of the genotyping procedure can indicate the presence of the same allelic variation in both copies (homozygous), or the presence of different variations in each copy (heterozygous). Molosiwa et al. (2015) genotyped samples from 128 plants of Bambara groundnut, with three or four samples from each of 33 landraces. I calculated the genetic distance between landraces using Nei's distance (Nei 1972), using the function 'dist.genpop' in the package ADEGENET (Jombart and Ahmed 2011) for R version 3.3.0 (R Core Team 2017). The output of this procedure is a matrix containing the pairwise genetic distances among all landraces.

### 4.2.3 Geo-referencing and estimating geographic distance between landraces

I used Bambara groundnut germplasm database[1] published by the International Institute of Tropical Agriculture (IITA) to provide geographical information about the origin or locale of collection of the landraces, available as longitude and latitude, using datum WGS84 (see figure 3.2). I explored measures of geographical distance among the locations of pairs of landraces. I initially calculated the direct line distance between landraces;

[1] http://my.iita.org/accession2/collection.jspx?id=8

however, this assumes that gene flow occurs in a straight line between samples, which is unlikely given variation in geographical barriers such as water bodies or mountain ranges. To resolve this, I conducted a least-cost path analysis using resistance surfaces based on the topographic relief of sub-Saharan region, and used the cost of moving between two landrace locations as a measure of the distance between them (see figure 4.1). The cost of movement between cells was defined as the inverse of the altitude difference between two adjacent cells (see equation 4.1). This analysis was conducted using the package GDISTANCE (Etten 2017).

$$cost = \frac{1}{cell_1 - cell_2} \tag{4.1}$$

### 4.2.4 Estimating linguistic similarity among sites of Bambara groundnut

I estimated the pairwise linguistic similarity among peoples inhabiting the regions around landraces using the geo-referenced dataset of idioms and a method of calculating linguistic distance from the Automated Similarity Judgement Program (ASJP). ASPJ maintains a database of languages with the geographical location and the estimated number of speakers. It uses an adapted Levenshtein distance method (Bakker et al. 2009) to calculate the similarity between two languages based on a vocabulary of 40 standard words (Wichmann et al. 2016).

To calculate the linguistic distance between two landraces $A$ and $B$, here identified as $linguisticDistance$, I defined a two-step procedure. First,

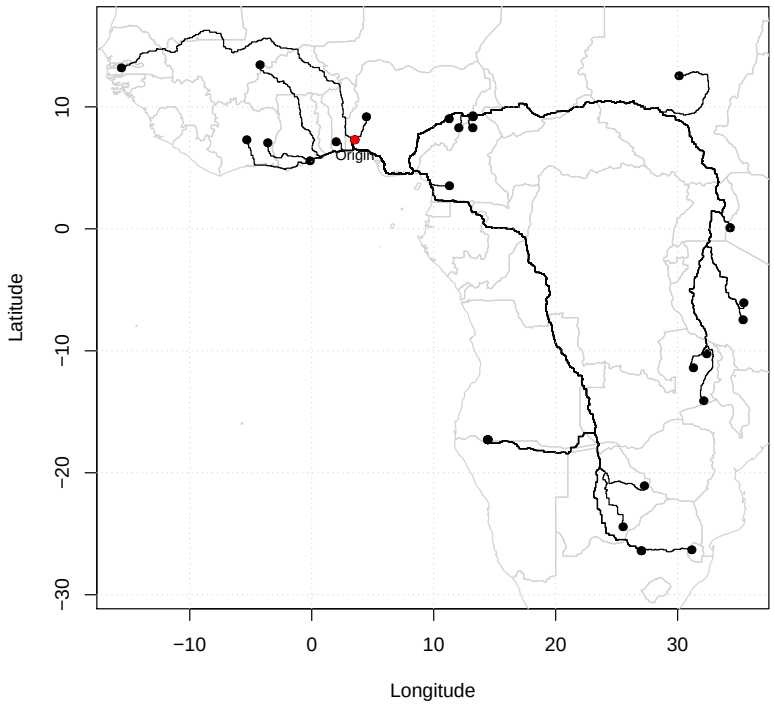



Figure 4.1: Overview of the area of study. The red and black circles indicate the origin and the destinations used in the least cost path analysis. The solid lines show least cost path to move between pairwise locations. The cost of movement between two adjacent cell is the inverse of the difference of the altitude (see equation 4.1). In this figure, the red circle represents landrace 32.

I identified the set of languages ($L_X$) with potential to influence a particular landrace ($X$). I calculated the average population density ($populationDensity$), in the area of the study, $34\frac{people}{km^2}$. Then, using the number of the speakers of each language ($numberOfSpeaker_L$) available in the ASJP dataset, I calculated the area required to host the speakers of $L$, $areaL$, as in the equation 4.2.

$$area_L = \frac{numberOfSpeakers_L}{averagePopulationDensity} \tag{4.2}$$

Then, I generated a buffer around each language ($L$) location with area

equal to $areaL$. I defined a fixed buffer radius of 1.5 degrees around the location of each landrace ($X$), and established the overlap between languages and landraces buffers (see figure 4.2), obtaining $L_x$.

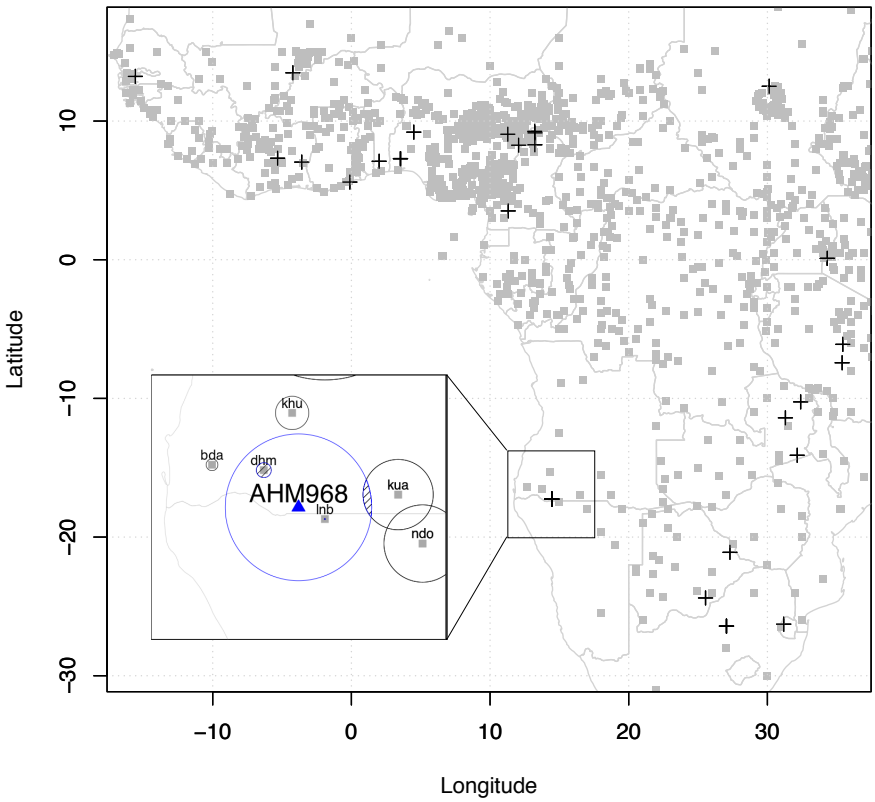



Figure 4.2: Overview of the area of study. The black crosses indicate accessions. Circles around the landraces represent the coverage area and have a fixed radius of 1.5 degrees. Grey squares indicate the location of languages available in the ASJP database. Circles around the grey squares indicate the coverage area of each language, calculated as the area required to host the number of speakers of the respective language and average human density in sub-Saharan Africa (34 people/km2). Inset with focus on landrace AHM968 (blue triangle), that exemplifies the procedure. There are six languages located around the landrace, and three of them (dhm, lnb, kua) have their coverage areas intersecting with the landrace's coverage area. I used these three languages to calculate the linguistic distance between AHM968 and other landraces. Source: author.

Second, I defined $L_A$ (set of languages influencing landrace $A$), $L_B$ (set of languages influencing landrace $B$), and $distanceASJP$ (ASJP linguistic distance method). Thus, $linguisticDistance(A, B)$ is the smallest linguistic distance calculated using $distanceASJP$ among all possible pairs of lan-

guages given by the Cartesian product between $L_A$ and $L_B$ (see equation 4.3).

$$linguisticDist(A, B) = min(distance_{asjp}(L_A * L_B)) \tag{4.3}$$

I used the program for calculating ASJP distance matrices version 2.1[2] in this analysis and package SP (Bivand et al. 2013).

### 4.2.5 Estimating climatic similarity

Given the potential importance of climate change for future food security (Foley et al. 2011), I used altitude and four climate variables as estimates of environmental similarity: maximum temperature of the warmest month, minimum temperature of the coldest month, precipitation of the wettest month and precipitation of the driest month, respectively BIO 5, BIO 6, BIO 13 and BIO 14 (Hijmans et al. 2005). All variables are available in GeoTIFF format with 30 arc second resolution at the WorldClim database. I extracted values for each variable at every landrace location using package RASTER (Hijmans et al. 2016b) and performed a principal components analysis using function 'prcomp'.

### 4.2.6 Analysis

To evaluate the relationships between genetic distance of landraces and geographic, environmental and linguistic distances, I implemented a pipeline (see figure 4.3) for integrating the multiple data sources and executing the

[2] http://asjp.clld.org/software

analyses. I first assessed the pairwise correlation among the four distance measures using Mantel tests and the global covariance structure among them using structural equation modelling. I used the Mantel permutation test (Mantel 1967) to evaluate the correlation between predictor variables (geographical, environmental and linguistic distances) and the genetic similarity of the landraces of Bambara groundnut. The Mantel test aims to assess the association between distinct measures of distance or similarity between individuals or populations. It is commonly used in studies investigating spatial genetic variation where measures of genetic similarity are analysed with measures of geographic distance or cost of movement. Mantel tests have been also used in studies involving environmental and linguistic data.

I used structural equation modelling (SEM) (Bollen 2005) to investigate the effects of multiple predictors and responses variables in the context of the hypotheses H1, H2, and H3. SEM allows the design of models based on hypothesised causal relationships, considering both direct and indirect effects. Direct effects use the standardised partial regression coefficients between a predictor and response variable, while indirect effects can be calculated by adding the products of all standardised partial regressions coefficients along all paths between predictor and response (Grace 2006).

In order to assess the statistical significance of the path coefficients from the SEM, I implemented a permutation test based on shuffling the genetic distance values and rerunning the SEM analysis. I repeated this process 9999 times to compute p-values and the confidence intervals (95%) following

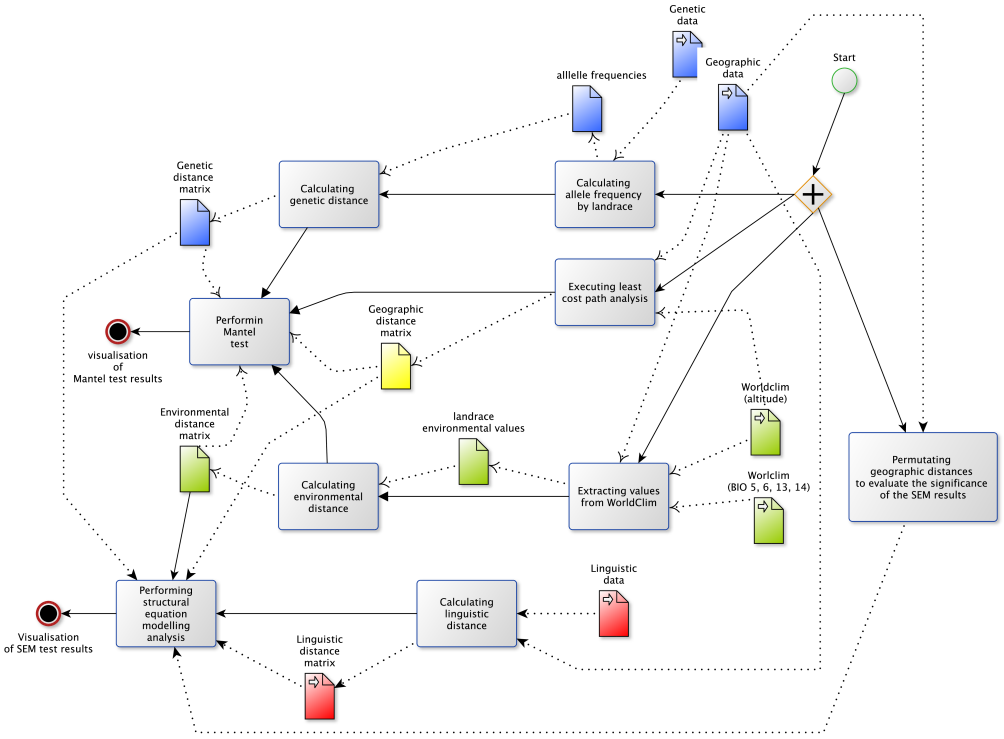



Figure 4.3: Schematic design of the pipeline implemented to run the analysis in the R software. The pipeline starts at the top right of the diagram, outputting on the left (black circles). Small rectangles represent the data sources and intermediary results. Large rectangles represent the processes and algorithms, and arrows provide information about the flow of data in the pipeline. Solid lines represent a connection between processes and dashed lines represent inputs and outputs of these processes.

Ruxton and Neuhäuser (2013), with the observed value included in both the numerator and denominator of p-value calculations.

## 4.3 Results

I calculated four distance measures: genetic, geographic (see figure 4.1), environmental and linguistic, producing symmetrical matrices with a total of 528 distinct pairwise distances (lower triangle of the distance matrix excluding diagonal) between 33 landrace populations for each type of distance.

In relation to the environmental distance, I selected the first principal component, whose loadings mainly represented altitude and temperature,

and the second principal component, mainly precipitation, for subsequent analyses. Together, these principal components accounted for 82.1% of dataset variability. I calculated the Euclidean distance, in climate space, between the landraces to create a single environmental distance matrix.

I calculated the pairwise relationship between genetic similarity and geographic distance (Mantel's r = 0.3467, p = 0.001), genetic similarity and linguistic similarity (Mantel's r = 0.2856, p = 0.001) and genetic similarity and environmental similarity (Mantel's r = 0.1268, p = 0.012).

I analysed the combined relationships between the different measures of similarity or distance using SEM. I calculated the direct and indirect effects of geographic, environmental and linguistic distances or similarities on genetic similarity (see figure 4.4). I tested the model fitness against the observed data performing the chi-square test (p = 0.044). The three direct effects were: geographic distance on genetic similarity (estimate = 0.31, $p < 0.001$), linguistic similarity on genetic similarity (estimate = 0.09, p = 0.0457), and environmental similarity on genetic similarity (estimate = -0.06, p = 0.0913). The two indirect effects were: geographic distance on linguistic similarity (estimate = 0.68, p = 0.2998) and geographic distance on environmental similarity (estimate = 0.50, p = 0.1745).

## 4.4 Discussion

SEM results indicate that genetic similarity among landraces of Bambara groundnut is related to the geographic proximity and, to a lesser extent, linguistic similarity between human communities. The landscape genetics

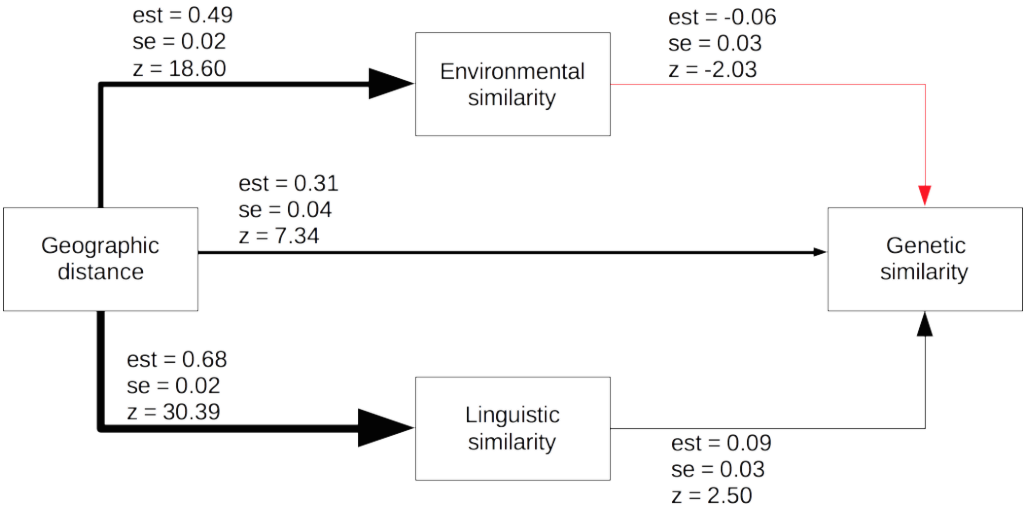



Figure 4.4: Structural equation model of the influence of geographic distance and environmental and linguistic similarities on genetic similarity.

theory predicts the relationship between geographic distance and genetic similarity (Balkenhol et al. 2015). It suggests that gene flow between two populations gradually decreases with the geographic distance, and limited gene flow scenario enhances genetic drift (Hardy and Vekemans 1999). The connectivity between sample sites or individuals can influence the genetic structure of a species or population. Instead of estimating the cost of movement between landraces using Euclidean distance, a measure that hardly would represent the cost of movement in the nature, I used a resistance surface to estimate the ability of genes to move though the landscape elements between landraces (Spear et al. 2010). Bambara groundnut geographic origin is thought to be close to the border between Nigeria and Cameroon (Azam-Ali et al. 2001), and its current distribution throughout sub-Saharan Africa is not completely understood. The least-cost path analysis produced possible routes for human movements in the past (see figure 4.1) which, when compared with routes proposed by recent research on the dispersal of Bantu groups (Russell et al. 2014), shows reasonable similarity.

Given the proximity between the geographic origin of Bambara groundnut and the origin of the first Bantu groups before expansion, a possible explanation for the isolation by distance results may be that initial dispersal of Bantu groups was responsible for carrying landraces of Bambara groundnut to West, East and South. This process is likely to have occurred over centuries, and would have led to the isolation of genetic material, as new areas were colonised and there was little exchange of seeds with other landraces on the continental scale. This also suggests that the landraces have been cultivated around their current locations for a reasonable period of time, long enough to allow local landraces to differentiate from their ancestors.

The significant relationship between the genetic and linguistic similarity suggests there are human/cultural-mediated effects on the geo-genetic diversity of Bambara groundnut and that humans have acted as vectors of gene flow in this species across Africa. The "farming language dispersal hypothesis" (Diamond 2003) posits that the advent of early agriculture around 10,000 years before today conferred to the farmers' communities an advantage over huntergatherer groups and triggered the dispersal of the former groups, with their languages and lifestyles. This process is likely to have promoted the expansion of farmers from their origins, acting as a dispersal vector for various human-associated plant species (Kemp et al. 2010). The cultural structure of the area of the study, sub-Saharan Africa, has been greatly influenced by the Bantu expansion. The Bantu expansion, the dispersal of farmers speaking early Bantu languages from West Africa to the east, and south of sub-Saharan Africa, represents the major human move-

ment in the African continent. The results reflect this structure by showing a strong association between geographic distance and linguistic similarity. It is possible therefore that distinct cultural practices in the use of plants, and possible language barriers, had an important role in the seed exchange network for Bambara groundnut.

Seed exchange networks are an important factor for crop diversity patterns in small-scale farming systems (Thomas et al. 2011). They provide a pathway for gene flows, diversifying the pool of genetic material or recovering the loss of particular alleles, and are usually based on relations of trust such as kinship, ethnolinguistic groups or neighbourhood (Labeyrie et al. 2016, Westengen et al. 2014, Deletre et al. 2011). The significant relationship between genetic and linguistic distance in the SEM results indicates that anthropogenic factors may have been playing a major role in the geo-genetic structure of human-associated plant species such as Bambara groundnut. Seed exchange among speakers of Bantu derived languages, the primary language group spoke in the study area (Currie et al. 2013, Russel et al. 2014), may have shaped the borders of seed distribution in sub-Saharan Africa, as has been tested on other crops cultures (Deletre et al. 2011, Labeyrie et al. 2016, Perales et al. 2005, Carrasco et al. 2016). It is also likely that the seed exchange through kinship, more commonly marriages where partners take seeds of landraces with them to the new home, promoted the diffusion of Bambara groundnut through singular pathways. It is important to acknowledge that these pathways can have acted in ambivalent mode, initially enabling gene flow between particular communities

and improving the diversity in the genetic pool, but also limiting the exchange of genetic material with those that did not fit the concept of trust, limiting farmers' access to landraces with the potential to contribute to the global fitness of local landraces. This process of gene flow highly associated with social and cultural similarity could be defined as isolation by culture.

One unanticipated finding was the weak the association between genetic similarity and environmental distance. Several factors could explain this result. Firstly, the fact that the spatial and temporal resolution of the WorldClim climate means data, 30 arc-seconds ($\tilde{1}$km) collected between 1960 and 1990, may not reflect the adaptation of landraces to the local environments. Secondly, the molecular markers used to calculate the genetic similarity may not cover regions of the genome sensitive enough to the environmental pressure. Thirdly, the effects of isolation by distance and social/cultural barriers override the environmental effects on Bambara groundnut. The last explanation could indicate the existence of potential mismatch between genotype and environment, meaning that some landraces may be better equipped to grow in areas dominated by other landraces.

## 4.5 Conclusion

I tested three hypotheses for the factors affecting the geo-genetic variation in Bambara groundnut: (H1) genetic similarity positively correlated with environmental similarity, (H2) isolation by distance among the populations (genetic drift), and (3) human movement as a dispersal vector and agent of gene flow. The SEM results indicated that genetic similarity among lan-

draces of Bambara groundnut is related to the geographic proximity and, to a lesser extent, linguistic similarity between human communities. It shows that ecological and anthropogenic factors may have shaped the geo-genetic structure of Bambara groundnut, in large geographic scale by the diffusion of plants through human dispersal, and in finer grained scale through cultural and social practices related to seed exchange networks.

My work suggests that cultural and social practices may be important in shaping the geo-genetic structure of human associated plant species. It brings light to the role of these species in the cultural identity of distinct ethnolinguistic human populations, a key factor if we, as a scientific community, want to tackle the food security problem while respecting human idiosyncrasies.

# Chapter 5

# Spatial regression captures signals of water stress influence on the spatial patterns of genetic variation in a human-associated crop species

**Abstract**

*Aim*

Understanding the importance of environmental factors in the origin and maintenance of patterns of genetic variation among human-related plant species populations, in an explicit geographical context, is an important step in order to guarantee food security conditions for a growing human population in a current and future scenario of climatic uncertainty. Here I investigate the influence of soil water content, a potential factor for the development of drought tolerance in plant species, on the spatial genetic structure of Bambara groundnut, a crop mainly cultivated throughout sub-Saharan Africa.

*Location*

Sub-Saharan Africa.

*Methods*

I used Moran's I statistic and spatial correlograms to quantify spatial patterns in the genetic data (principal component analysis results based on landraces' allele frequency) from a previous study. Next, I regressed the genetic data (principal component scores) against soil water content using generalised least squares regression.

*Results*

The principal components and spatial correlograms showed a cline affecting the entire Sub-Saharan distribution.

*Main conclusions*

Previous studies indicated that distance is important to explain spatial patterns of genetic variation between the landraces, but correlograms and spatial regressions analyses in this study showed that there is also a significant correlation between soil water content and genetic variation. Reduced gene flow due to distinct farming practices and social networks between human populations, or local environmental conditions, could explain spatial variation mainly at smaller scales, while human migration can be related to the large-scale pattern.

## 5.1 Introduction

Food security is a core component of the United Nations (UN) Sustainable Development Goals (SDGs). The SDGs raise the urgency of reducing poverty and hunger, and increasing health and well-being in a world facing the challenge of feeding a population of billions of people and a climate under constant and increasing change (Griggs et al. 2013). Although there is still dispute about the extent of climate-changing effects on several human activities, and in particular on agriculture, there is concordance that poor communities are the most vulnerable to climate changes (Morton 2007; Wheeler and Braun 2013). These communities, often based on traditional and subsistence farming systems, rely on the use of plant landraces, which are a local adaptation of human-associated plant species (Dwivedi et al. 2016).

Plant landraces can help reduce the genetic erosion that dominates modern agriculture (Wouw et al. 2010, Bellon et al. 2015). Landraces are highly

heterogeneous, often considered to be well adapted to local environmental conditions (Villa et al. 2005), and likely to harbour genetic resources that can help meet the SDG challenges. Derived from the domestication of wild relatives, they are subjected to selective pressures over generations, at a particular geographic location, from the local environmental conditions (Peratoner et al. 2016). In parallel, artificial selection by farmers, whose criteria to select desirable agricultural traits typically include social factors, such as seed networks and ethno-linguistic membership (Orozco-Ramirez et al. 2014), may have also influenced landraces' genetic differentiation. Landraces are a key component in breeding programs due to their potentially useful genes for adaptation to abiotic stress, such as drought, salinity, and heat. Despite the landraces' importance, little is known about their patterns of spatial genetic variation and association with environmental factors.

Local adaptation is an important mechanism by which an individual population becomes, through phenotypic and genotypic change, better suited to a particular environment than other populations of the same species (Kawecki and D. Ebert 2004). While the relationship between local adaption processes and plants is relatively well explored in the literature, little is known about local adaptation related to farming activity (Corrado and Rao 2017). Farmer-mediated artificial selection may not be always aligned with the direction of selection by environmental factors in a particular region (Mercer and Perales 2010). Thus, establishing how populations respond to environmental conditions is not an easy challenge in agriculture (Savolainen et al. 2013) because desirable features of landraces are often

linked to specific farming practices and social networks. During climate change, remaining locally adapted can allow population persistence, and breeding for local adaptation has become an important goal for both crop and animal breeding in the face of climate change.

Soil water content (SWC) is a major limiting factor of plant growth, not least in sub-Saharan Africa, as a reduced amount of available water for physiological plant processes results in drought stress. Rainfall patterns and irrigation practices influence soil water content and are key determinants of plant distribution in arid and semi-arid regions (Wang et al. 2014). Drought-tolerant plants can respond to water stress by a number of distinct mechanisms, e.g.: completing the growth cycle before drought stress develops to damaging levels, improving root traits to increase extraction of soil moisture, or adapting plant capabilities to perform with minimal water (Khan et al. 2017). Traditional and subsistence farming practices are usually rainfall based, and drought tolerance generally plays a crucial role in the local adaptation of plants. Given that climate is spatially structured, drought tolerance is expected to affect the geographic patterns of species' genetic variation.

Although many patterns of plant species' genetic variation along environmental gradients have been documented, there remains a lack of understanding about the mechanisms associated with such spatial patterns (Engelbrecht et al. 2007), in particular with human-associated plant-species. According to Rao and Hodgkin (2002), differences in the geographic distribution of genetic diversity among species are extremely common, and

environmental differences (patches or gradients) may affect the genetic diversity in plant populations. Here, I use a human-associated crop species, Bambara groundnut, to test the hypothesis that there is correlation between genetic variation and soil water content (H4, presented in the introduction chapter).

## 5.2 Material and methods

### 5.2.1 Study species

Bambara groundnut (*Vigna subterranea* (L.) Verdc.) is an indigenous, drought-tolerant African grain legume, cultivated in a wide range of environmental conditions that vary from very arid (< 400mm annual rainfall) to very humid (> 2500mm annual rainfall). This crop, native to West Africa but cultivated throughout sub-Saharan Africa, plus Malaysia, Indonesia, Thailand and India, has an important socio-economic role, as a nutritious alternative to staple crops and as a source of income (Azam-Ali et al. 2001, Karunaratne et al. 2015, Massawe et al. 2015). Farmers cultivate Bambara groundnut from landraces, traditional varieties that may have developed specific agricultural traits to adapt to local environmental conditions. Previous research on genetic diversity of Bambara groundnut showed high levels of genetic heterogeneity between landraces (Aliyu et al. 2016, Khan et al. 2017, Jørgensen et al. 2010), indicating that evolutionary, ecological and human-driven processes may have influenced these populations leading to genetic divergence (see previous chapter's conclusion).

### 5.2.2 Genetic and environmental datasets

The genetic dataset (principal component analysis results based on landraces' allele frequency) from a previous study (see Chapter 3) was composed of genotyping information based on presence or absence of 19 microsatellite markers, also known as simple sequence repeats (SSRs). I collected count data for all 155 SSR alleles available in the genetic dataset, grouping them by landraces (33 in total), and dividing by the number of samples available in each landrace, generating an average allele frequency matrix. I then conducted a principal component analysis (PCA) on the allele frequency matrix to obtain a smaller set of independent variables that capture the overall genetic variation in the dataset.

I used the monthly Soil Water Content (SWC) variable of the Global High-Resolution Soil-Water Balance dataset (Trabucco and Zomer 2010) to estimate the water stress indexes around the landrace locations. SWC provides a coefficient as percentage of the soil water content available for evapotranspiration processes, with 0 meaning no water available and 100 meaning 100% of the soil water capacity is available. The dataset is available in ESRI Grid format with 30 arc second resolution and utilises the WorldClim[1] and Global-PET[2] databases as primary input. I extracted values of SWC at every landrace location using the package 'RASTER' (Hijmans et al. 2016b) and derived four indices associated with water stress from the Monthly Soil Water Content variable: the annual mean water content available (see Figure 5.1) and the water stress duration (the number of

[1] http://www.worldclim.org
[2] http://www.cgiar-csi.org/data/global-aridity-and-pet-database

months with SWC above a particular threshold - here 30%, 50% and 70%).

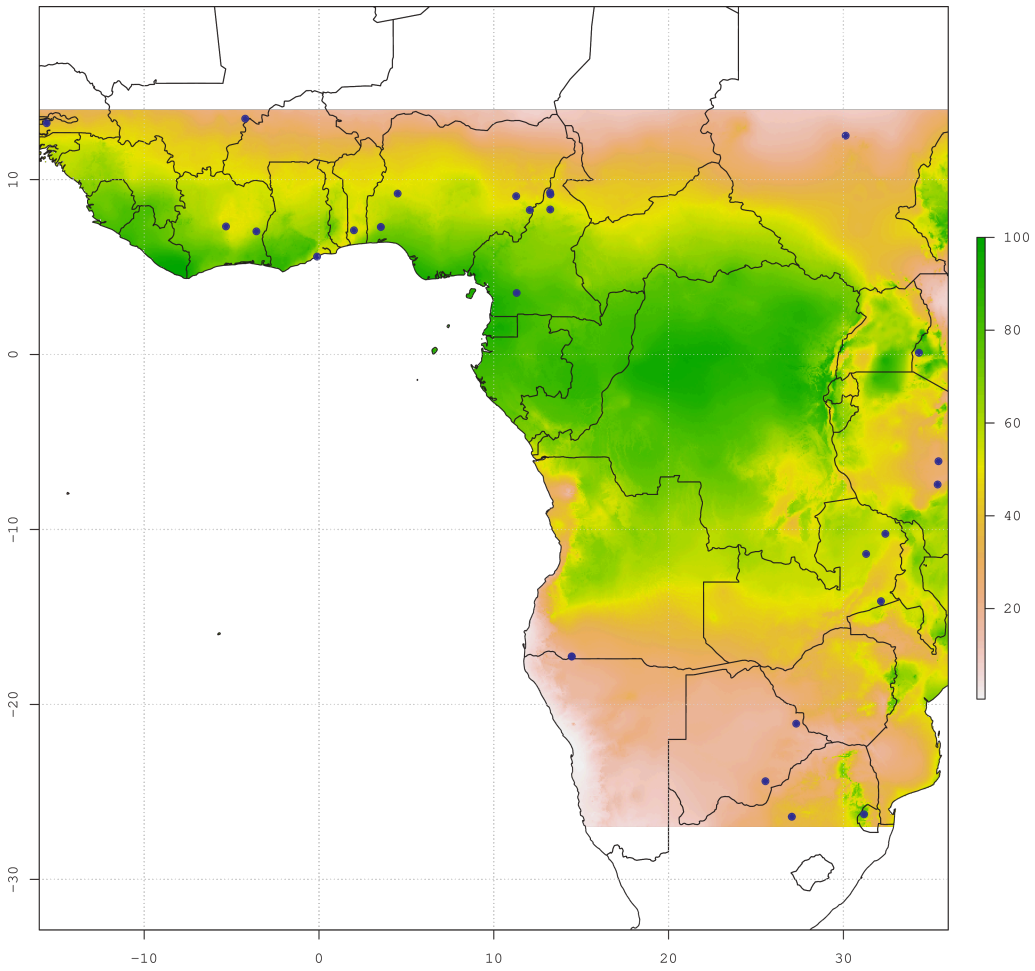



Figure 5.1: Location of the 33 landraces plotted over the distribution of soil water content monthly average (swc mean). Due to the broad range of environments where Bambara groundnut is cultivated, it is possible to identify landraces on environment conditions varying from very wet (swc mean above 80%) in the Central Africa next to the Atlantic coast, to very dry (swc mean below 40%) in the North Africa and border between Angola and Namibia.

### 5.2.3 Analysis

A core issue when analysing association between genetic and environmental data is the inherent spatial structure and spatial autocorrelation (Fortin and Dale 2005). Land use practices and physical processes impose spatial structure in environmental factors, to which organisms respond. Also, biotic

processes such as dispersal and spatially restricted mating create further spatial structure. Spatial autocorrelation, summarised in the First Law of Geography "everything is related to everything, but near things are more related than distant things" (Tobler 1970, p. 236), affects many traditional statistical tests and their use without proper control for spatial structure can increase Type 1 error rates.

I used Moran's I index and corresponding correlograms to estimate the spatial autocorrelation on the first (PC1) and second (PC2) components of the PCA results. I also calculated the Moran's I index on the residuals of an ordinary least squares regression (OLS) of PC1 and PC2 against the annual mean water content. Moran's I autocorrelation index is a correlation coefficient that aims to detect spatial structure in a dataset. Moran's I uses a spatial weights matrix to calculate the neighbours' influence on a particular observation (Fortin and Dale 2005). I calculate the great circle distance between the landrace locations using package 'GEOSPHERE' (Hijmans 2016a), Moran's I index using the function 'Moran.I' package 'APE' (Paradis et al. 2004), and correlograms using the function 'correlog' package 'NCF' (Bjornstad 2016).

I used generalised least squares (GLS) regression with R package 'NLME' version (Pinheiro et al. 2017) to model genetic variation among landraces as a function of the soil water capacity. With GLS regressions, it is possible to account for spatial autocorrelation in the error structure of the regression, effectively accounting for spatial autocorrelation between the samples in statistical tests of significance (Beale et al. 2010; Dormann et al. 2007). For

regressions of PC1 and PC2 against the annual mean water content, I tested five different error autocorrelation structures: exponential, gaussian, linear, quadratic ratio, and spherical. For each error structure, I calculated the Akaike information criterion (AIC) and retained the error structure with the lowest AIC as the final model. All calculations were performed using the R language for statistical computing (R Core Team 2017) version 3.3.0.

## 5.3 Results

The soil water content-based variables were highly correlated with each other (see Table 5.1). I used the monthly average value in the GLS regression analysis because although Bambara groundnut landraces are tolerant to water stress (Karunaratne 2009), each landrace has a distinct response to water stress duration.

Table 5.1: Pearson's correlation between four soil water content based variables. The water stress duration variables (swc duration 30, swc duration 50 and swc duration 70) showed an expected high positive (lower value was 0.84) correlation since their formulation only differ on the threshold definition of water stress (30, 50, and 70% of the soil capacity). They also showed a high negative correlation (higher value was -0.95), also expected since values of swc mean close to 1 would mean more water available in the soil, and a likely indication of shorter periods of water stress.

| | swcm | swcd30 | swcd50 | swcd70 |
|---|---|---|---|---|
| swc mean (swcm) | 1.00 | -0.95 | -0.97 | -0.95 |
| swc duration 30 (swcd30) | -0.95 | 1.00 | 0.90 | 0.85 |
| swc duration 50 (swcd50) | -0.97 | 0.90 | 1.00 | 0.95 |
| swc duration 70 (swcd70) | -0.95 | 0.85 | 0.95 | 1.00 |

The OLS regressions calculated to predict PC1 and PC2 based on the soil water content results were slope 0.05 and p-value 0.0696 for PC1 and slope -0.06 and p-value 0.0129 for PC2 (see Figure 5.2).

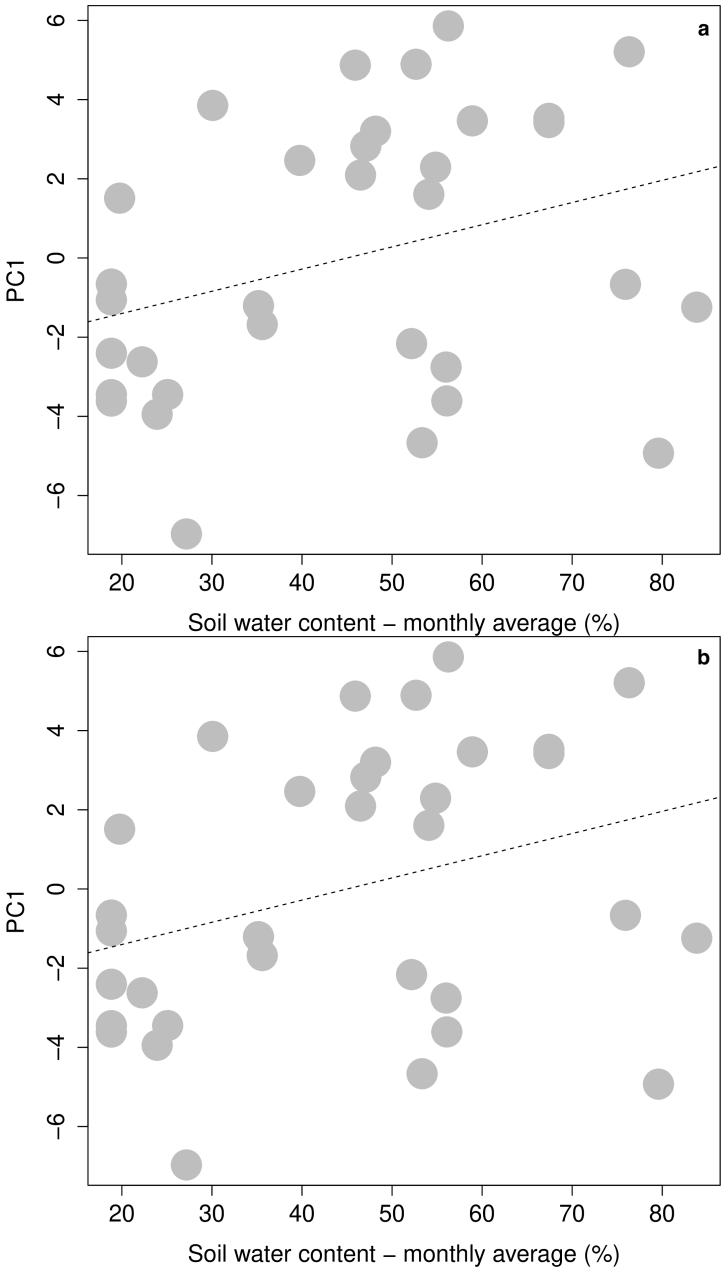



Figure 5.2: Two ordinary least squares regressions were calculate to predict the PCA component scores of the genetic data (based on the landraces allele frequency calculated in Chapter 3) based on the soil water content. Regression of PC1 (a) results were slope 0.05 and p-value 0.0696, and regression of PC2 (b) results were -0.06 and p-value 0.0129.

PC1 showed strong signal of spatial autocorrelation with global Moran's I index 0.44 and p-value $< 0.0001$, and significant spatial autocorrelation in the first 10 lags of the correlogram (see Figure 5.2). PC2 and the PC1 and PC2 residuals presented non significant spatial autocorrelation.

Table 5.2: Correlogram plot data source, containing n (the number of pairs within each distance class), mean of class (the actual average of the distances within each distance class), correlation (Moran's I), and p-value (the permutation two-sided p-value for each distance class).

| | | PC1 | | PC2 | |
|---|---|---|---|---|---|
| n | mean of class | correlation | p-value | correlation | p-value |
| 43 | 179.52 | 0.74 | 0.01 | 0.19 | 0.06 |
| 44 | 790.66 | 0.48 | 0.01 | 0.20 | 0.04 |
| 44 | 1270.75 | 0.58 | 0.01 | 0.03 | 0.30 |
| 58 | 1801.00 | 0.64 | 0.01 | 0.19 | 0.02 |
| 33 | 2214.64 | 0.39 | 0.03 | -0.41 | 0.02 |
| 87 | 2827.36 | -0.27 | 0.04 | -0.02 | 0.49 |
| 61 | 3183.77 | -0.43 | 0.02 | -0.32 | 0.02 |
| 52 | 3799.09 | -0.59 | 0.02 | -0.19 | 0.10 |
| 45 | 4254.94 | -0.69 | 0.02 | -0.05 | 0.44 |
| 39 | 4723.30 | -0.52 | 0.02 | 0.17 | 0.05 |
| 7 | 5207.91 | -0.52 | 0.09 | -0.68 | 0.04 |
| 7 | 5731.91 | -0.42 | 0.20 | -0.26 | 0.16 |
| 7 | 6154.37 | -0.73 | 0.08 | 0.13 | 0.19 |
| 1 | 6694.78 | -1.02 | 0.10 | -0.03 | 0.45 |

The GLS regressions assessed with Akaike information criterion (AIC) and Bayesian information criterion (BIC) presented similar results. The error correlation structure that best fit the data for PC1 was spherical with slope 0.04 with p-value 0.1502. The error correlation structure that best fit the data for PC2 was the exponential spatial structure with slope -0.06 and p-value 0.0376 (see Table 5.4 and Figure 5.3).

Table 5.3: Moran's I statistic results based on PC1, PC2, and the respective residuals of ordinary least squares regressions.

| | Moran's I index | Expected | Std. Dev. | p-value |
|---|---|---|---|---|
| PC1 | 0.44 | -0.03 | 0.12 | $< 0.0001$ |
| PC2 | 0.16 | -0.03 | 0.12 | 0.0518 |
| PC1 (residuals) | 0.27 | -0.03 | 0.12 | 0.0073 |
| PC2 (residuals) | 0.05 | -0.03 | 0.12 | 0.2521 |

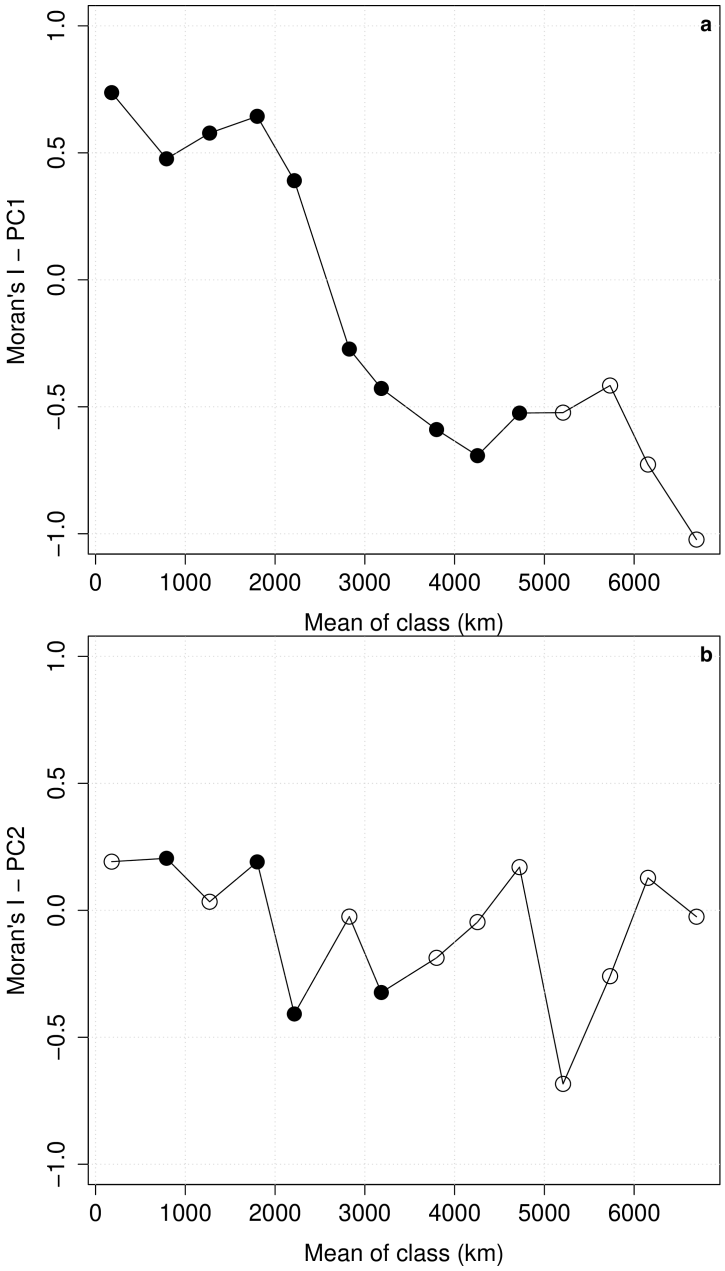



Figure 5.3: I tested the PCA result (based on landraces' allele frequency) for spatial autocorrelation using Moran's I. Figure (a) and (b) show correlogram plots of the coefficient Moran's I for PC1 and PC2 respectively, as a function of distance (km). PC1 correlogram indicates strong spatial clustering, while PC2 correlogram indicates week or absence of spatial clustering. I calculated the spatial autocorrelation coefficient Moran's I using package NCF - function 'correlog' with parameters increment = 500km and 999 permutations under the null hypothesis to assess the level of significance. Black filled circle represent correlations with p-value $< 0.05$.

## 5.4 Discussion

I used Bambara groundnut genetic dataset and the monthly Soil Water Content (SWC) variable of the Global High-Resolution Soil-Water Balance dataset, to test the hypothesis that there is a correlation between genetic variation and soil water content, hypothesis H4 (presented in Chapter 2). This hypothesis was particularly unusual given the weak relationship signal between genetic distance and environmental distance tested and discussed in the previous chapter. Landraces are usually the result of local adaptation and genotypes farmer-led selection to obtain or maximise desirable agricultural traits such as higher yield or drought/disease resistance. In this chapter, I tested for spatial autocorrelation in the genetic dataset and used generalised least squares regression to investigate the correlation between genetic variation and soil water content, a climatic variable particularly influent in crop yields.

The results of this study indicate that there is no significant relationship between genetic variation and soil water content influencing the genetic structure among landraces of Bambara groundnut. The distribution of landraces over the first principal component (see Figure 3.3) reflects a gradient from South/Southeast to North/Northwest and supports the existence of genetic differentiation between landrace populations that are geographically distant from each other. In contrast, no spatial pattern could be extracted from the distribution of landraces over the first and second principal components. The GLS regression analyses indicated no significant association between genetic variation in these components and soil water content. This

genetic structure showed in the results in Chapter 3 is consistent with local adaptation of Bambara groundnut. And If present, local adaptation may have been facilitated by a) the geographic distribution of landraces, covering very distinct habitats with a broad range of soil water content, b) limited seed dispersal, influenced by the distinct seed exchange practices and social networks in place in the area of the study, c) the main of objective of cultivating Bambara groundnut, mostly for subsistence, implying local consumption and restricted distribution of the genetic material. However, given the strong association between genetic and geographic distances shown in results of Chapter 4, and the absence of significant correlation between genetic variation and soil water content in this chapter, it could be concluded that at the larger scale, human movements may have played a role in distributing the seeds, but at the smaller scale, it is still unclear the effects of farmer-led selective pressure and adaptation to local environment affect the genetic structure of Bambara.

A possible explanation for the absence of correlation between genetic variation and soil water content is that given the limited number of microsatellites (19), it is unlikely that they could cover regions of the Bambara groundnut genome associated with drought tolerance. However, it is important to consider that drought tolerance is a quantitative trait, controlled by the combination of various genes, and the strong genetic differentiation shown in Chapter 3 results may represent a particular source of material, rather than loci linked directly to genes involved. It is important to consider that the absence of a reference genome sequence makes the identification

of loci under selection a challenging task (Kort et al. 2015). A promising avenue is the use of the well characterised genome of closely related species (Ho et al. 2017). The extent of the samples distribution and the potential uncertainty associated with the geographic coordinates of the samples, as well as the geographic resolution of the environmental dataset can also have contributed to mask the potential relationship between genetic variation and soil water stress.

## 5.5 Conclusion

Previous studies indicated that distance is important to explain spatial patterns of genetic variation between the landraces, but correlograms and spatial regressions analyses in this study showed that there is also a significant correlation between soil water content and genetic variation. Reduced gene flow due to distinct farming practices and social networks between human populations, or local environmental conditions, could explain spatial variation mainly at smaller scales, while human migration can be related to the large-scale pattern.

Table 5.4: Generalised least squares regression results. Exp. Var. - Explanatory variable, Struct - Spatial Structure, E - Exponential, G - Gaussian, L - Linear, R - Rational quadratic, S - Spherical. * The regression of PC1 with linear structure did not converge.

| Exp.Var. | Struct | AIC | BIC | logLik | Predictor | Slope | SE | t-value | p-value |
|---|---|---|---|---|---|---|---|---|---|
| PC1 | E | 182.45 | 189.62 | -86.22 | Intercept | -1.90 | 1.79 | -1.06 | 0.2973 |
| | | | | | swc mean | 0.04 | 0.03 | 1.25 | 0.2171 |
| | G | 181.18 | 186.91 | -86.59 | Intercept | -2.45 | 1.50 | -1.63 | 0.1124 |
| | | | | | swc mean | 0.05 | 0.03 | 1.81 | 0.0788 |
| | L | NA | NA | NA | Intercept | NA | NA | NA | NA |
| | | NA | NA | NA | swc mean | NA | NA | NA | NA |
| | R | 181.16 | 186.89 | -86.58 | Intercept | -2.41 | 1.52 | -1.57 | 0.1247 |
| | | | | | swc mean | 0.05 | 0.03 | 1.77 | 0.0863 |
| | S | 180.46 | 186.19 | -86.23 | Intercept | -2.08 | 1.68 | -1.23 | 0.2250 |
| | | | | | swc mean | 0.04 | 0.03 | 1.47 | 0.1502 |
| PC2 | E | 172.46 | 178.20 | -82.23 | Intercept | 2.77 | 1.28 | 2.16 | 0.0379 |
| | | | | | swc mean | -0.06 | 0.02 | -2.45 | 0.0197 |
| | G | 172.63 | 178.36 | -82.31 | Intercept | 2.88 | 1.26 | 2.28 | 0.0294 |
| | | | | | swc mean | -0.06 | 0.02 | -2.55 | 0.0158 |
| | L | 172.63 | 178.36 | -82.31 | Intercept | 2.88 | 1.26 | 2.28 | 0.0294 |
| | | | | | swc mean | -0.06 | 0.02 | -2.55 | 0.0158 |
| | R | 172.61 | 178.34 | -82.30 | Intercept | 2.86 | 1.26 | 2.26 | 0.0303 |
| | | | | | swc mean | -0.06 | 0.02 | -2.54 | 0.0163 |
| | S | 172.63 | 178.36 | -82.31 | Intercept | 2.88 | 1.26 | 2.28 | 0.0294 |
| | | | | | swc mean | -0.06 | 0.02 | -2.55 | 0.0158 |

# Chapter 6

# Conclusion

This thesis examined the geo-genetic structure of a human-associated plant species, Bambara groundnut, throughout sub-Saharan Africa. This chapter concludes the text, reviewing the research questions, aims, and objectives. Further, it discusses the limitations of this research, sets the future research agenda and closes this thesis.

The fundamental argument of my thesis is that genetic similarity among landraces of Bambara groundnut is related to the geographic proximity and, to a lesser extent, linguistic similarity between human communities. My research aimed to examine how geographic, environmental, and anthropogenic factors influence the genetic variation in human-associated plant species in human-associated landscapes. In this thesis, I set the detection of spatial patterns of genetic variation in the context of a biogeography framework and explored the potential to strengthen interdisciplinary research by providing repeatable, quantitative, and comparable correlation estimates

In Chapter 3, I explored the genetic dataset, investigating the genetic variation and genetic structure using ordination and clustering methods. I hypothesised that the current spatial distribution of the genetic variation of

Bambara groundnut could be explained by three processes acting over recent centuries: (H1) genetic similarity positively correlated with environmental similarity where this crop has been cultivated, (H2) isolation by distance among the populations that have genetically drifted apart, and (H3) human movement and cultural connections that have helped the dispersal of the seeds and increased gene flow of Bambara groundnut throughout sub-Saharan Africa. Likewise, I hypothesise that (H4) there is a correlation between genetic variation and soil water content, a significant environmental factor influencing crop yield.

In Chapter 4, I tested hypotheses H1, H2, and H3 using structural equation modelling, and the results indicate that genetic similarity among landraces of Bambara groundnut is related to the geographic proximity and, to a lesser extent, linguistic similarity between human communities. It shows that ecological and anthropogenic factors may have shaped the geo-genetic structure of Bambara groundnut, in large geographic scale by the diffusion of plants through human dispersal, and in finer grained scale through cultural and social practices related to seed exchange networks.

In Chapter 5, I tested the hypotheses (H4), that there is a correlation between genetic variation and soil water content. This hypothesis was particularly interesting given the weak relationship signal between genetic distance and environmental distance tested and discussed in Chapter 4. Landraces are usually the result of local adaptation and genotypes farmer-led selection to maximise yield or drought/disease resistance. In Chapter 5, I tested for the spatial autocorrelation in the genetic dataset and used generalised least

squares regression to investigate the correlation between genetic variation and soil water content, a climatic variation particularly influential in crop yields. The results did not indicate a significant correlation between these variables. A possible explanation for this is the microsatellite markers available in the dataset and used in the analysis did not cover regions of Bambara groundnut genome associated with water stress and drought tolerance. The extent of the samples distribution and the potential uncertainty associated with the geographic coordinates of the samples, as well as the geographic resolution of the environmental dataset can also have contributed to mask the potential relationship between genetic variation and soil water stress.

This work also highlights the challenges of integrating discipline-associated datasets, in particular, geographic and genetic data, and its effects on analyses and interpretation of the results. I identified some imprecision in the location data that did not affect the initial analyses (PCA, K-means, and mapping), but had an impact on methods that required different coordinates among the samples (usually due to requirements for geographic distance higher than 0). If, as a scientific community, we are serious about interdisciplinarity then we need a lot more work in coordinating data-collection activities, to guarantee the data acquired are useful for all disciplines involved. I propose that existing research on interoperability, an established concept in GIScience, be extended to other areas of science, and particularly the coordination of data collection. It holds potential for helping to address the challenges presented in the integration of multidisciplinary data.

Reproducibility in science has been receiving growing attention lately.

By using open source tools to perform the analysis, I had access to a set of libraries and packages created to address the particular research questions of the field of landscape genetics. Another positive aspect of the use of open source tools is the access to the source code and the possibility of reading and verifying the algorithms' implementation. At a point in the study, I was not certain of the Nei's genetic distance method version coded in the 'ADEGENET' package. To clarify this question, it was only necessary to read the source code of the function pop.dist. Access to the source code of scientific software empowers the users by giving them the opportunity to understand and verify how the tools used work.

Overall, the present study establishes a quantitative framework for detecting the relationship between geographic distance and environmental and linguistic similarities in the genetic diversity of Bambara groundnut. By investigating these relationships, this study highlights the role of language and cultural identity in the decision of which crops small-scale farmer grow and consume.

In conclusion, this work also contributes to the existing knowledge of food security by providing a reflection on a) underutilised crops as part the cultural identity of communities and their ways of dispersing seeds, and b) the importance of taking in account crops associated with cultural identity in the elaboration of plans to guarantee food security in present and future.

## 6.1 Suggested further research

This research highlighted the effect of geographic distance in the geo-genetic diversity of Bambara groundnut. It also identified signals of the relationship between genetic variation and linguistic groups, suggesting that social networks and cultural aspects may have a role in this intricate process.

However, to my knowledge, there is little scientific research on how cultural aspects and social networks affect the genetic variation in underutilised crops. Further investigation of this subject using Bambara groundnut genetic material from Malaysia and Indonesia or different underutilised crops in sub-Saharan Africa could help clarify their role in the selection of seeds, genetic material, by the farmers and the exchange of seeds among social groups.

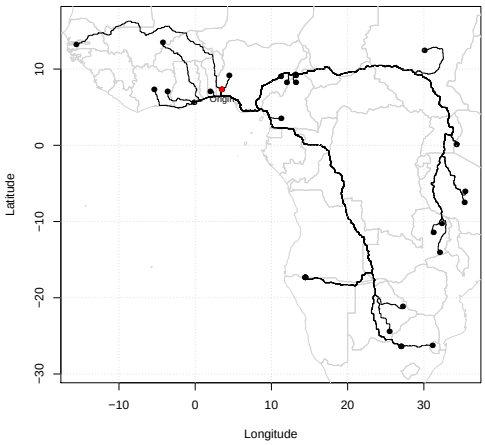
The geographic distribution of the genetic dataset, covering most of sub-Saharan Africa, presented challenges in itself. Also, as discussed in this thesis, the lack of georeferencing in many collections is currently a challenge when exploring the spatial dimension of biodiversity for many plant species. Further investigation could examine the impacts of adding noise in the dataset to avoid overlapping locations, as well as, research on methods to simulate artificial genetic data to cover potential geographic gaps and multiple geographic scales.

On the genetic side, recently released genetic datasets, based on DArT technology, which randomly extract information from the samples covering the complete genome, has potential to support investigating the relationship between genetic and environmental factors. In the same way, as presented

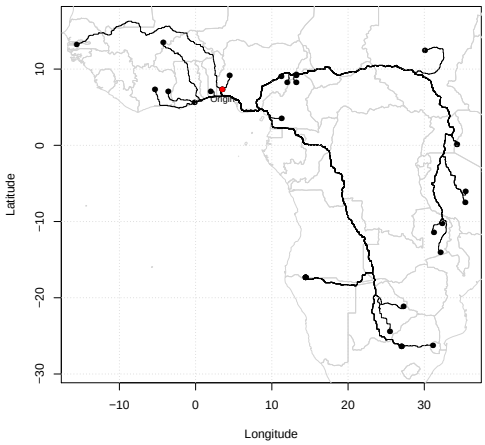
by Ho et al. (2017), the use of genetic data from similar major crops data has potential to add new dimensions to the dataset, and an example would be testing whether the geo-genetic patterns identified in this research are present in other crop species.

Also, testing the current framework to investigate the relationships in important crops such as Moringa, and indigenous varieties of Maize, Rice, and Wheat could help clarify the role of cultural identity in the spatial distribution of genetic diversity, as well as validate the application of the framework in other crops.

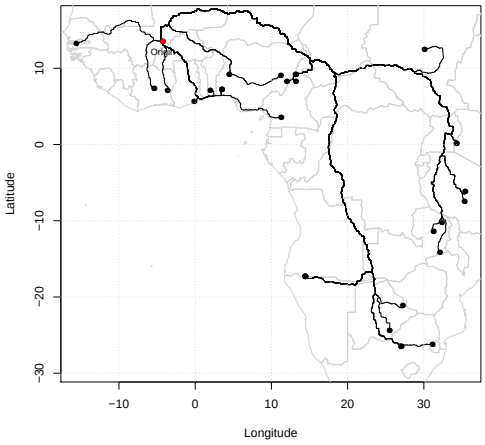
# Appendix 1 - Least cost path analyses for all the 33 landraces



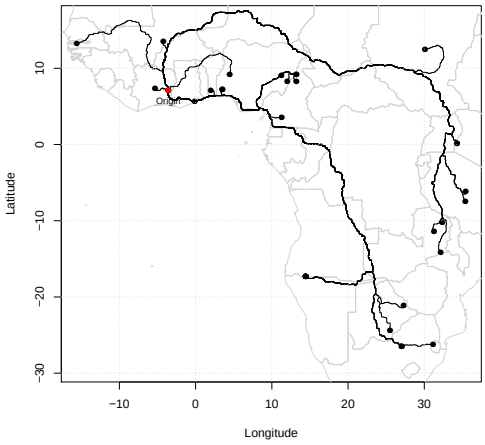

(a) Landrace 9



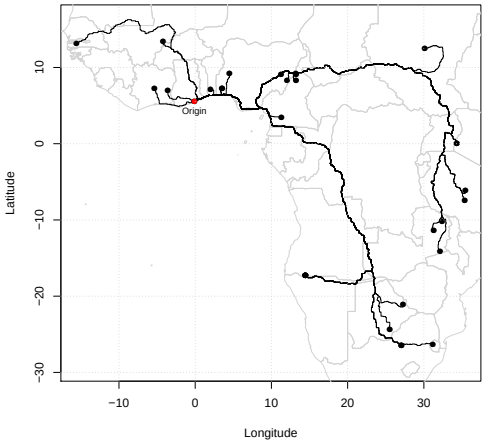

(b) Landrace 32



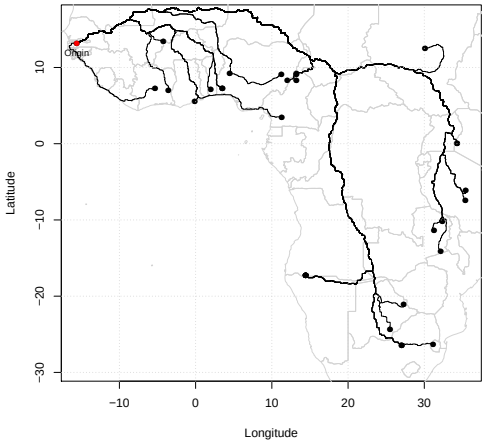

(c) Landrace 89




(d) Landrace 118

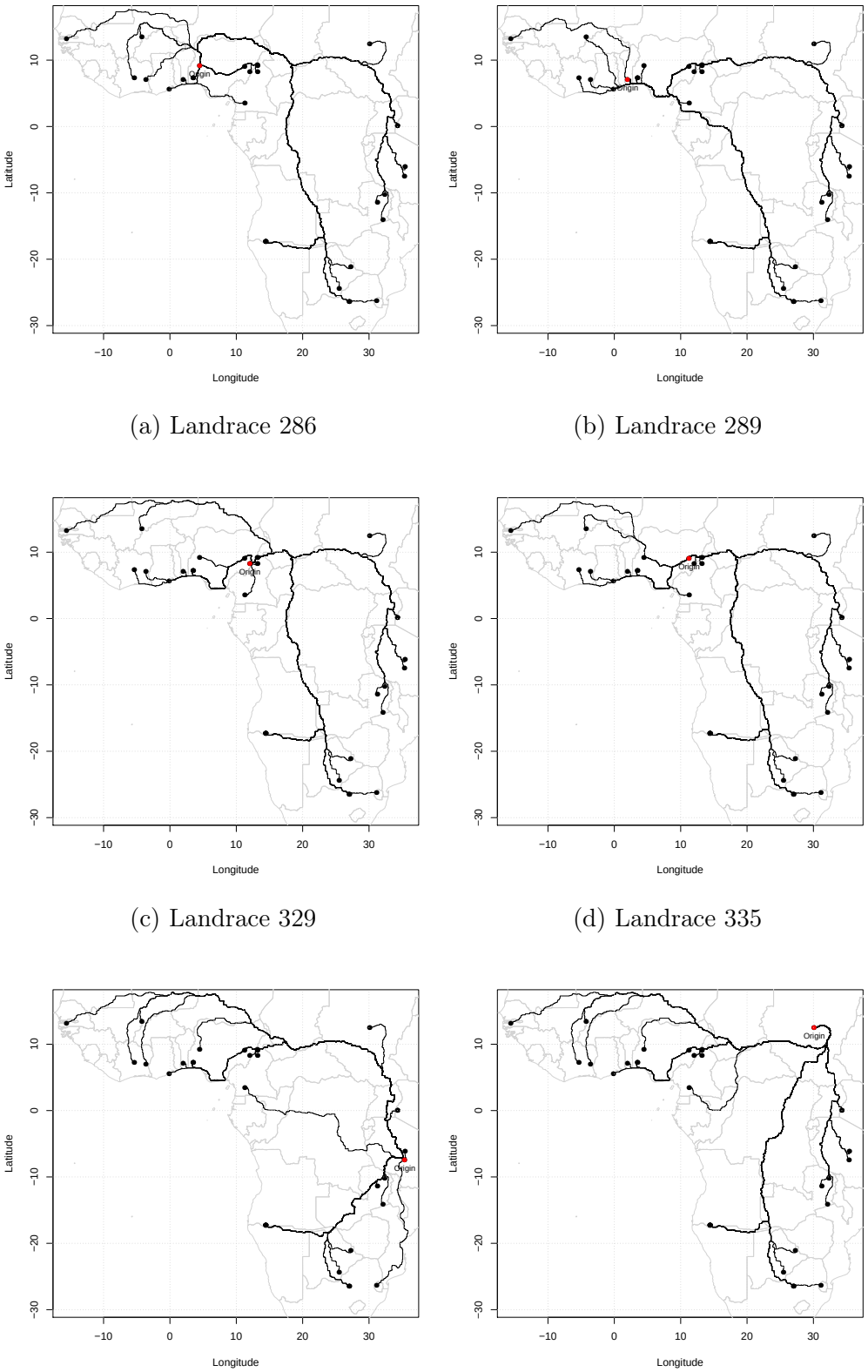



(e) Landrace 144




(f) Landrace 231

Figure 6.1: Least cost path analysis results for landraces 9, 32, 89, 118, 144, and 231.




(a) Landrace 286

(b) Landrace 289

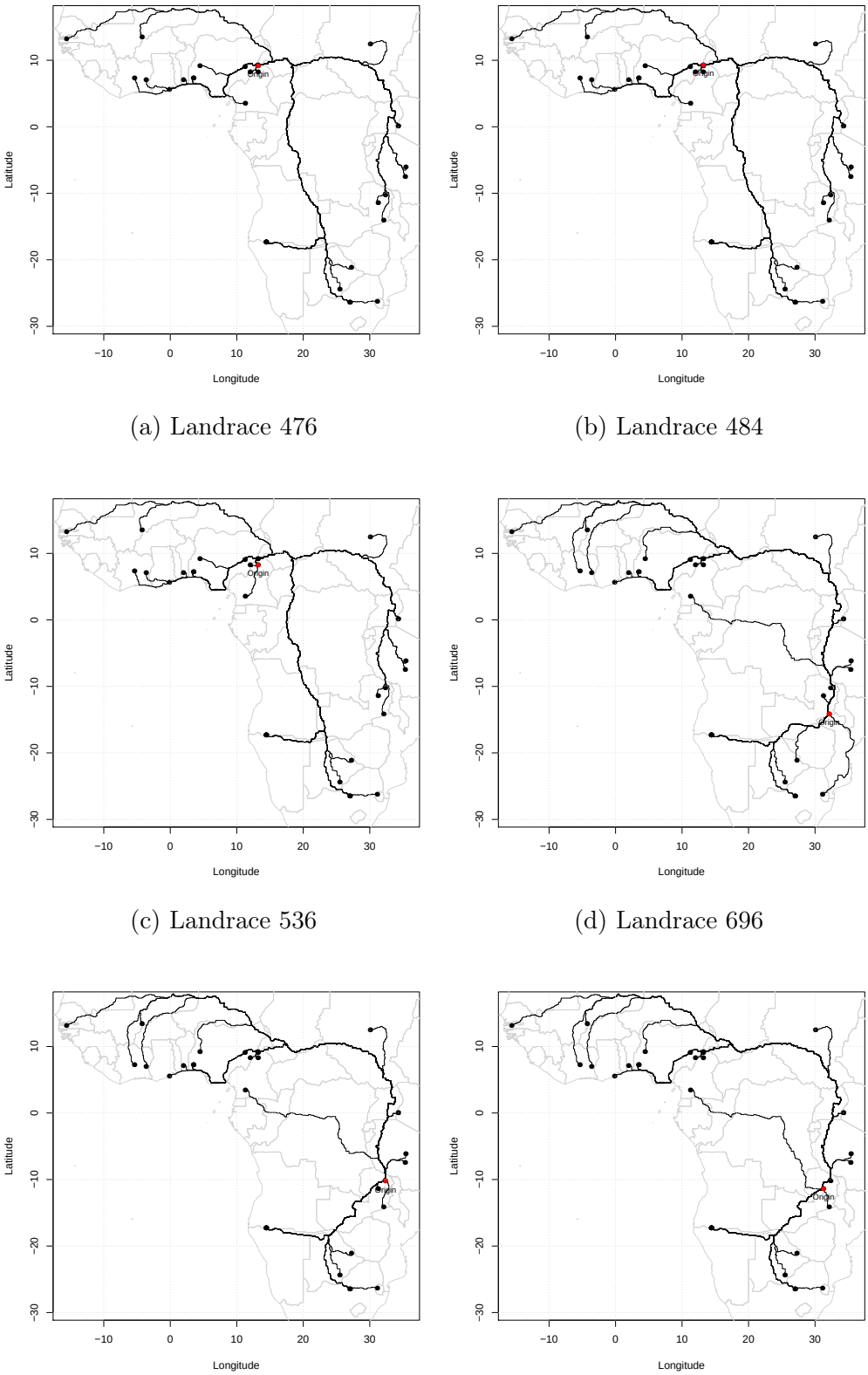
(c) Landrace 329

(d) Landrace 335

(e) Landrace 385

(f) Landrace 390

Figure 6.2: Least cost path analysis results for landraces 286, 289, 329, 335, 385, and 390.




(a) Landrace 476

(b) Landrace 484

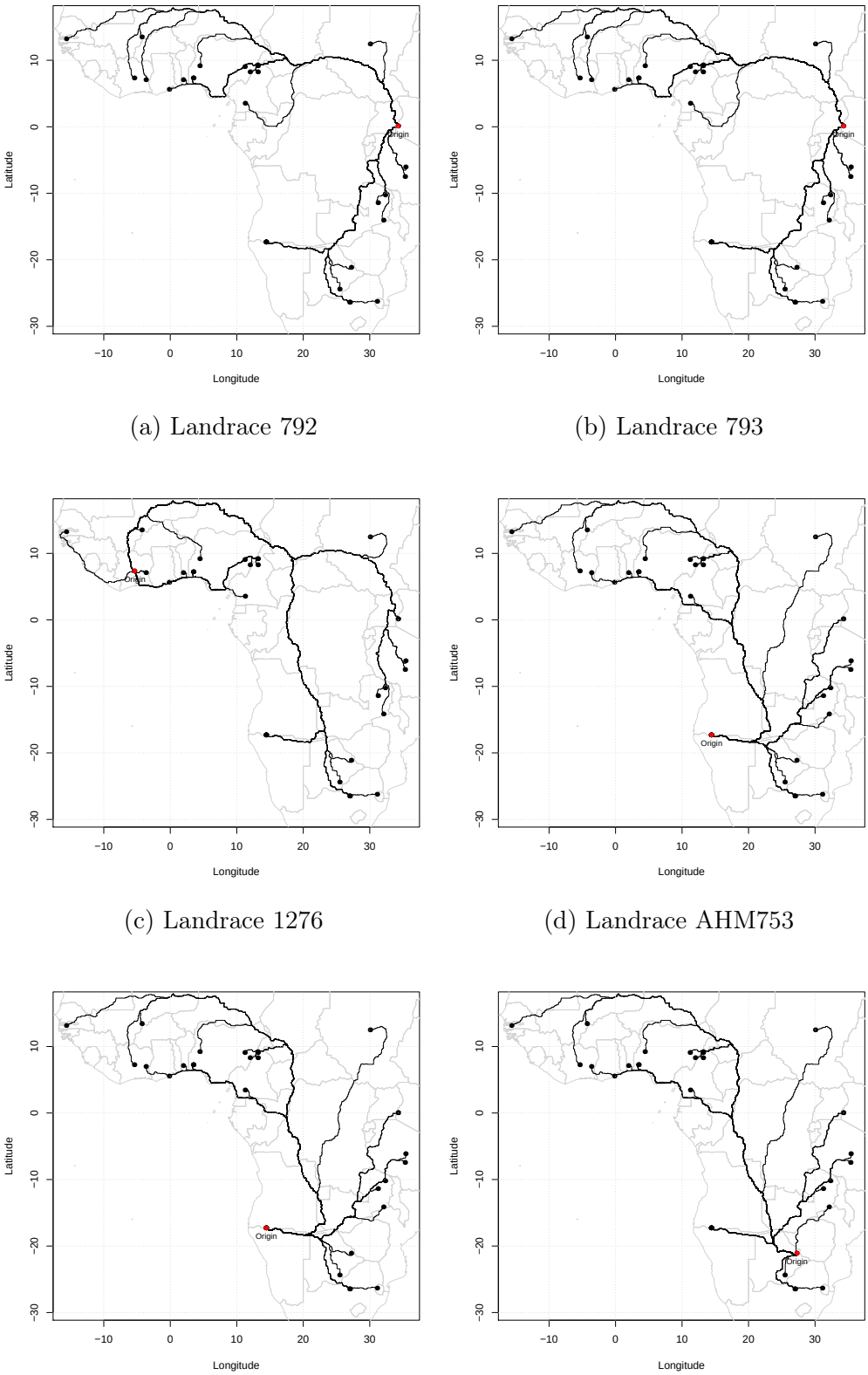
(c) Landrace 536

(d) Landrace 696

(e) Landrace 754

(f) Landrace 790

Figure 6.3: Least cost path analysis results for landraces 476, 484, 536, 696, 754, and 790.




(a) Landrace 792

(b) Landrace 793

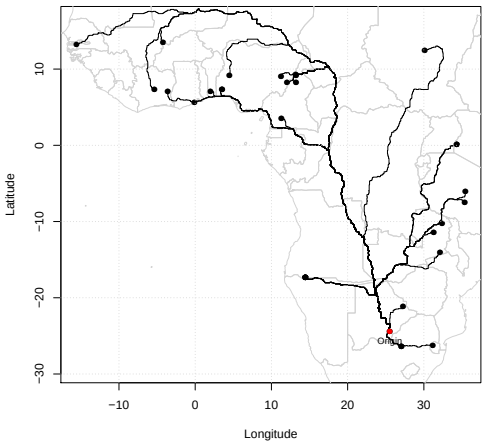
(c) Landrace 1276

(d) Landrace AHM753

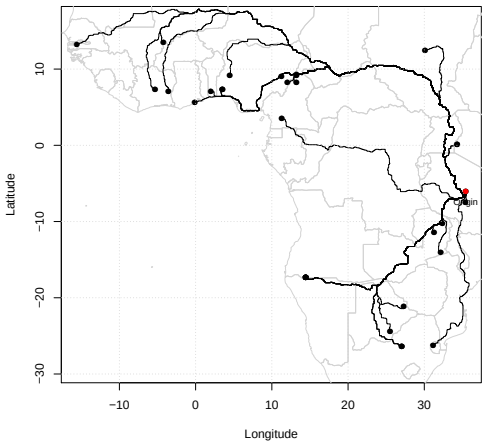
(e) Landrace AHM968

(f) Landrace BOTS 1

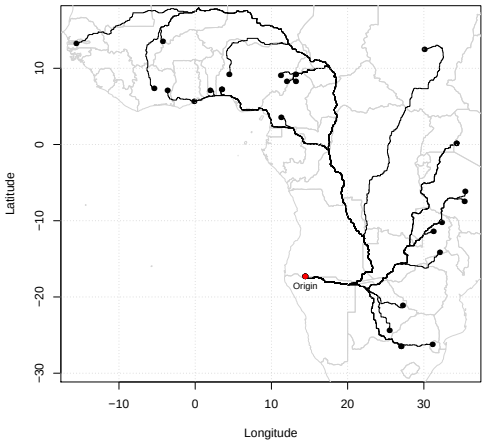
Figure 6.4: Least cost path analysis results for landraces 792, 793, 1276, AHM753, AHM968, and BOTS 1.



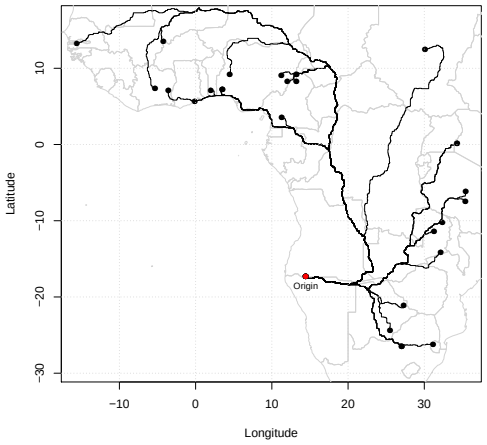

(a) Landrace BOTS 5



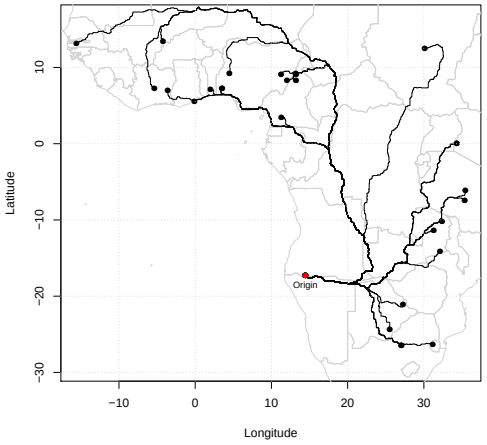

(b) Landrace DOD R



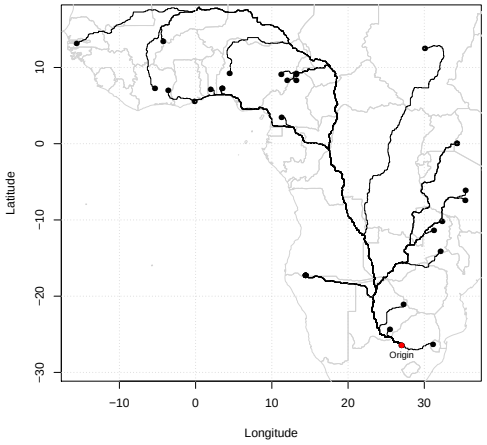

(c) Landrace Mahenene Black




(d) Landrace S-1913

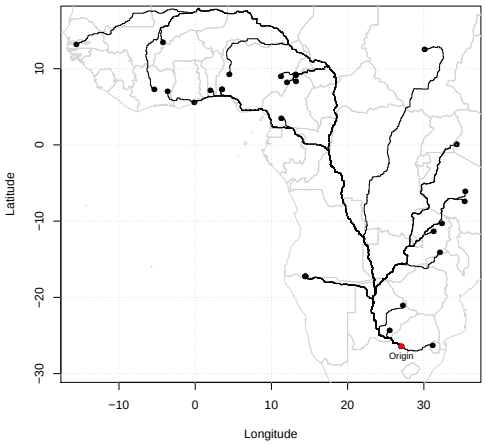



(e) Landrace S19-3

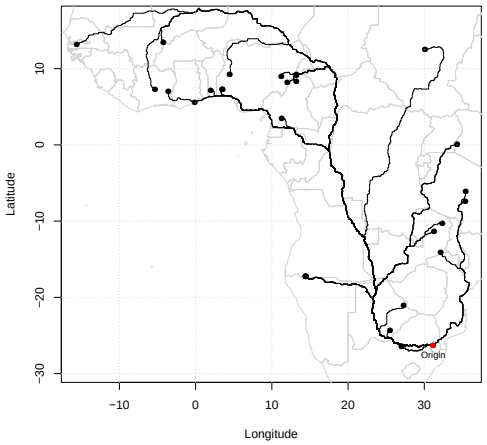



(f) Landrace SB4-2

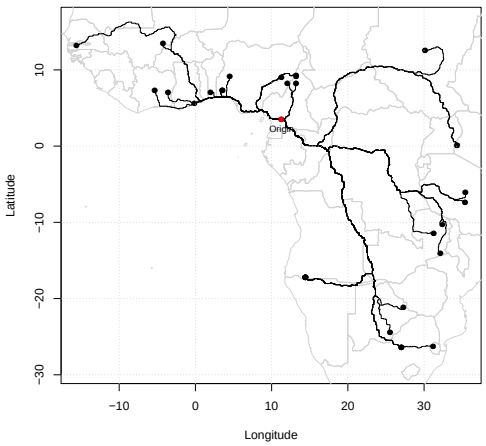
Figure 6.5: Least cost path analysis results for landraces BOTS 5, DOD R, Mahenene Black, S-1913, S19-3, and SB4-2.




(a) Landrace SB16 5A




(b) Landrace UNIS R




(c) Landrace VSSP 6

Figure 6.6: Least cost path analysis results for landraces SB16 5A, UNIS R, and VSSP 6.

# Bibliography


Abberton, Michael, Jacqueline Batley, Alison Bentley, John Bryant, Hongwei Cai, James Cockram, Antonio Costa de Oliveira, Leland J. Cseke, Hannes Dempewolf, Ciro De Pace, David Edwards, Paul Gepts, Andy Greenland, Anthony E. Hall, Robert Henry, Kiyosumi Hori, Glenn Thomas Howe, Stephen Hughes, Mike Humphreys, David Lightfoot, Athole Marshall, Sean Mayes, Henry T. Nguyen, Francis C. Ogbonnaya, Rodomiro Ortiz, Andrew H. Paterson, Roberto Tuberosa, Babu Valliyodan, Rajeev K. Varshney, and Masahiro Yano (2016). "Global agricultural intensification during climate change: A role for genomics". In: *Plant Biotechnology Journal* 14.4, pp. 1095–1098. ISSN: 14677652. DOI: 10.1111/pbi.12467.

Academies, National (2004). *Facilitating Interdisciplinary Research*. Washington, District of Columbia: National Academies Press, p. 332.

Alberts, B., A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter (2014). *Molecular Biology of the Cell, Sixth Edition*. Taylor & Francis Group. ISBN: 9781317563754. URL: https://books.google.co.uk/books?id=jK6UBQAAQBAJ.

Aliyu, Siise, Festo Massawe, and Sean Mayes (2016). "Genetic diversity and population structure of Bambara groundnut (Vigna subterranea (L.) Verdc.): synopsis of the past two decades of analysis and implications for crop improvement programmes". In: *Genetic Resources and Crop Evolution* 63.6, pp. 925–943. DOI: 10.1007/s10722-016-0406-z.

Anderson, Corey Devin, Bryan K. Epperson, Marie Josée Fortin, Rolf Holderegger, Patrick M A James, Michael S. Rosenberg, Kim T. Scribner, and Stephen Spear (2010). "Considering spatial and temporal scale in landscape-genetic studies of gene flow". In: *Molecular Ecology* 19.17, pp. 3565–3575. ISSN: 09621083. DOI: 10.1111/j.1365-294X.2010.04757.x.

Azam-Ali, S N, J Aguilar-Manjarrez, and M Bannayan-Avval (2001). *A global Mapping system for Bambara groundnut Production*, p. 47.

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman (2009). "Adding typology to lexicostatistics: A combined approach to language classification". In: *Linguistic Typology* 13.1, pp. 169–181. DOI: 10.1515/LITY.2009.009.

Balkenhol, Niko, Samuel A. Cushman, Andrew T. Storfer, and Lisette P. Waits, eds. (2015). *Landscape Genetics*. Chichester, UK: John Wiley & Sons, Ltd. DOI: 10.1002/9781118525258.

Barbieri, Rosa Lía, João Carlos Costa Gomes, Adriana Alercia, and Stefano Padulosi (2014). "Agricultural biodiversity in southern Brazil: Integrating efforts for conservation and use of neglected and underuti-

lized species". In: *Sustainability (Switzerland)* 6.2, pp. 741–757. ISSN: 20711050. DOI: 10.3390/su6020741.

Basu, S., ROBERTS, S. N. AZAM-ALI, and S MAYES (2007). "Development of microsatellite markers for bambara groundnut (Vigna subterranea L. Verdc.) an underutilized African legume crop species". In: *Molecular Ecology Notes* 7.6, pp. 1326–1328. ISSN: 14718278. DOI: 10.1111/j.1471-8286.2007.01870.x.

Beale, Colin M., Jack J. Lennon, Jon M. Yearsley, Mark J. Brewer, and David A. Elston (2010). "Regression analysis of spatial data". In: *Ecology Letters* 13.2, pp. 246–264. ISSN: 1461023X. DOI: 10.1111/j.1461-0248.2009.01422.x.

Beddington, J. (2010). "Food security: contributions from science to a new and greener revolution". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1537, pp. 61–71. ISSN: 0962-8436. DOI: 10.1098/rstb.2009.0201.

Bellon, Mauricio R., Elisabetta Gotor, and Francesco Caracciolo (2015). "Assessing the effectiveness of projects supporting on-farm conservation of native crops: Evidence from the high andes of South America". In: *World Development* 70, pp. 162–176. DOI: 10.1016/j.worlddev.2015.01.014.

Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio (2013). *Applied spatial data analysis with R*. Second. New York: Springer, p. 405. URL: www.asdar-book.org.

Bjornstad, Ottar N. (2016). *ncf: Spatial Nonparametric Covariance Functions*. R package version 1.1-7. URL: https://CRAN.R-project.org/package=ncf.

Bollen, Kenneth A. (2005). "Structural Equation Models". In: *Encyclopedia of Biostatistics*. Vol. 1459. March. Chichester, UK: John Wiley & Sons, Ltd. DOI: 10.1002/0470011815.b2a13089.

Bonthala, Venkata Suresh, Katie Mayes, Joanna Moreton, Martin Blythe, Victoria Wright, Sean Tobias May, Festo Massawe, Sean Mayes, and Jamie Twycross (2016). "Identification of gene modules associated with low temperatures response in bambara groundnut by network-based analysis". In: *PLoS ONE* 11.2, pp. 1–18. ISSN: 19326203. DOI: 10.1371/journal.pone.0148771.

Camacho-Henriquez, Alberto, Friederike Kraemer, Gea Galluzzi, Stef De Haan, Matthias Jäger, and Anja Christinck (2015). *Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools*. Ed. by Jameel M. Al-Khayri, Shri Mohan Jain, and Dennis V. Johnson. Vol. 1. Cham: Springer International Publishing, pp. 25–61. DOI: 10.1007/978-3-319-22521-0.

Carrasco, N. F., J. R. L. Oler, F. F. Marchetti, M. A. Carniello, M. C. M. Amorozo, T. L. Valle, and E. A. Veasey (2016). "Growing Cassava (Manihot esculenta) in Mato Grosso, Brazil: Genetic Diversity Conservation in SmallScale Agriculture". In: *Economic Botany* 70.1, pp. 15–28. DOI: 10.1007/s12231-016-9331-5.

Cheng, Acga, Sean Mayes, Gemedo Dalle, Sebsebe Demissew, and Festo Massawe (2017). "Diversifying crops for food and nutrition security - a

case of teff". In: *Biological Reviews* 92.1, pp. 188–198. ISSN: 1469185X. DOI: 10.1111/brv.12225.

Chivenge, Pauline, Tafadzwanashe Mabhaudhi, Albert T. Modi, and Paramu Mafongoya (2015). "The potential role of neglected and underutilised crop species as future crops under water scarce conditions in Sub-Saharan Africa". In: *International Journal of Environmental Research and Public Health* 12.6, pp. 5685–5711. ISSN: 16604601. DOI: 10.3390/ijerph120605685.

Corrado, Giandomenico and Rosa Rao (2017). "Towards the Genomic Basis of Local Adaptation in Landraces". In: *Diversity* 9.4, p. 51. DOI: 10.3390/d9040051.

Currie, Thomas E., Andrew Meade, Myrtille Guillon, and Ruth Mace (2013). "Cultural phylogeography of the Bantu Languages of sub-Saharan Africa". In: *Proceedings of the Royal Society of London B: Biological Sciences* 280.1762, p. 20130695. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2013.0695.

Deletre, M., D. B. McKey, and T. R. Hodkinson (2011). "Marriage exchanges, seed exchanges, and the dynamics of manioc diversity". In: *Proceedings of the National Academy of Sciences* 108.45, pp. 18249–18254. DOI: 10.1073/pnas.1106259108.

Diamond, Jared (2003). "Farmers and Their Languages: The First Expansions". In: *Science* 300.5619, pp. 597–603. ISSN: 00368075. DOI: 10.1126/science.1078208.

Dormann, Carsten F., Jana M. McPherson, Miguel B. Araújo, Roger Bivand, Janine Bolliger, Gudrun Carl, Richard G. Davies, Alexandre Hirzel, Walter Jetz, W. Daniel Kissling, Ingolf Kühn, Ralf Ohlemüller, Pedro R. Peres-Neto, Björn Reineking, Boris Schröder, Frank M. Schurr, and Robert Wilson (2007). "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review". In: *Ecography* 30.5, pp. 609–628. ISSN: 09067590. DOI: 10.1111/j.2007.0906-7590.05171.x.

Dwivedi, Sangam L., Salvatore Ceccarelli, Matthew W. Blair, Hari D. Upadhyaya, Ashok K. Are, and Rodomiro Ortiz (2016). "Landrace Germplasm for Improving Yield and Abiotic Stress Adaptation". In: *Trends in Plant Science* 21.1, pp. 31–42. ISSN: 13601385. DOI: 10.1016/j.tplants.2015.10.012.

Dyer, Rodney J (2015). "Is there such a thing as landscape genetics?" In: *Molecular Ecology* 24.14, pp. 3518–3528. ISSN: 09621083. DOI: 10.1111/mec.13249.

Ebert, Andreas W. (2014). "Potential of underutilized traditional vegetables and legume crops to contribute to food and nutritional security, income and more sustainable production systems". In: *Sustainability (Switzerland)* 6.1, pp. 319–335. ISSN: 20711050. DOI: 10.3390/su6010319.

Engelbrecht, Bettina M. J., Liza S. Comita, Richard Condit, Thomas A. Kursar, Melvin T. Tyree, Benjamin L. Turner, and Stephen P. Hubbell (2007). "Drought sensitivity shapes species distribution patterns in tropical forests". In: *Nature* 447.7140, pp. 80–82. DOI: 10.1038/nature05747.

Erp, Marieke van, Robert Hensel, Davide Ceolin, and Marian Van der Meij (2015). "Georeferencing Animal Specimen Datasets". In: *Transactions in GIS* 19.4, pp. 563–581. ISSN: 13611682. DOI: 10.1111/tgis.12110.

Etten, Jacob van (2017). "R Package gdistance : Distances and Routes on Geographical Grids". In: *Journal of Statistical Software* 76.13. ISSN: 1548-7660. DOI: 10.18637/jss.v076.i13.

FAO (2001). *The State of Food Insecurity in the World 2001*. Tech. rep. Rome, Italy.

— (2010). *The Second Report on the State of the World Plant Genetic Resources for Food and Agriculture*. Tech. rep. Rome, Italy.

Foley, Jonathan A, Navin Ramankutty, Kate A. Brauman, Emily S. Cassidy, James S. Gerber, Matt Johnston, Nathaniel D. Mueller, Christine O'Connell, Deepak K. Ray, Paul C. West, Christian Balzer, Elena M Bennett, Stephen R. Carpenter, Jason Hill, Chad Monfreda, Stephen Polasky, Johan Rockström, John Sheehan, Stefan Siebert, David Tilman, David P. M. Zaks, and Christine O'Connell (2011). "Solutions for a cultivated planet". In: *Nature* 478.7369, pp. 337–42. DOI: 10.1038/nature10452.

Food and Agriculture Organization of the United Nations (2017). *FAOSTAT Statistics Database*. URL: http://www.fao.org/faostat.

Fortin, M.J. and M.R.T. Dale (2005). *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press.

Galluzzi, Gea, Rigoberto Estrada, Vidal Apaza, Mirihan Gamarra, Ángel Pérez, Gilberto Gamarra, Ana Altamirano, Gladys Caceres, Victor Gonza, Ricardo Sevilla, Isabel Lopez Noriega, and Matthias Jager (2015). "Participatory breeding in the Peruvian highlands: Opportunities and challenges for promoting conservation and sustainable use of underutilized crops". In: *Renewable Agriculture and Food Systems* 30.05, pp. 408–417. DOI: 10.1017/S1742170514000179.

Galluzzi, Gea and Isabel López Noriega (2014). "Conservation and Use of Genetic Resources of Underutilized Crops in the Americas - A Continental Analysis". In: *Sustainability* 6.2, pp. 980–1017. ISSN: 2071-1050. DOI: 10.3390/su6020980.

Godfray, H Charles J and Tara Garnett (2014). "Food security and sustainable intensification". In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369.1639, pp. 1–13. DOI: 10.1098/rstb.2012.0273.

Goodchild, Michael F (2010). "Twenty years of progress: GIScience in 2010". In: *Journal of Spatial Information Science* 1. ISSN: 1948-660X. DOI: 10.5311/JOSIS.2010.1.2.

Grace, James B. (2006). *Structural Equation Modeling and Natural Systems*. 1st. Cambridge: Cambridge University Press, p. 365.

Griggs, David, Mark Stafford-Smith, Owen Gaffney, Johan Rockström, Marcus C. Öhman, Priya Shyamsundar, Will Steffen, Gisbert Glaser, Norichika Kanie, and Ian Noble (2013). "Sustainable development goals for people and planet". In: *Nature* 495.7441, pp. 305–307. DOI: 10.1038/495305a.

Hardy, Oliver J and Xavier Vekemans (1999). "Isolation by distance in a continuous population: reconciliation between spatial autocorrelation anal-

ysis and population genetics models". In: *Heredity* 83.2, pp. 145–154. ISSN: 0018067X. DOI: 10.1046/j.1365-2540.1999.00558.x.

Harlan, J R (1971). "Agricultural origins: centres and noncenters". In: *Science* 174.4008, pp. 468–474.

Hartl, D.L. and A.G. Clark (2007). *Principles of Population Genetics*. Sinauer. ISBN: 9780878933082. URL: https://books.google.co.uk/books?id=SB1vQgAACAAJ.

Hedrick, P. (2011). *Genetics of Populations*. Jones & Bartlett Learning.

Hijmans, Robert J. (2016a). *geosphere: Spherical Trigonometry*. R package version 1.5-5. URL: https://CRAN.R-project.org/package=geosphere.

— (2016b). *raster: Geographic Data Analysis and Modeling*. URL: https://cran.r-project.org/package=raster.

Hijmans, Robert J, Susan E Cameron, Juan L Parra, Peter G Jones, and Andy Jarvis (2005). "Very high resolution interpolated climate surfaces for global land areas". In: *International Journal of Climatology* 25.15, pp. 1965–1978. DOI: 10.1002/joc.1276.

Hillocks, R J, C Bennett, and O M Mponda (2012). "Bambara Nut: a Review of Utilisation, Market Potential and Crop Improvement". In: *African Crop Science Journal* 20.1, pp. 1–16. ISSN: 10219730.

Ho, Wai Kuan, Hui Hui Chai, Presidor Kendabie, Nariman Salih Ahmad, Jaeyres Jani, Festo Massawe, Andrzej Kilian, and Sean Mayes (2017). "Integrating genetic maps in bambara groundnut [Vigna subterranea (L) Verdc.] and their syntenic relationships among closely related legumes". In: *BMC Genomics* 18.1. DOI: 10.1186/s12864-016-3393-8.

Hubisz, Melissa J., Daniel Falush, Matthew Stephens, and Jonathan K. Pritchard (2009). "Inferring weak population structure with the assistance of sample group information". In: *Molecular Ecology Resources* 9.5, pp. 1322–1332. ISSN: 1755098X. DOI: 10.1111/j.1755-0998.2009.02591.x.

Jackson, M.J., G. Hobona, L. Bernard, J. Brauner, and C. Higgins (2011). "The role and development of a persistent interoperability test bed for geosciences research". In: *Geological Society of America Special Papers*. Geological Society of America, pp. 69–75. DOI: 10.1130/2011.2482(07).

Jacobsen, Sven Erik, Marten Sørensen, Søren Marcus Pedersen, and Jacob Weiner (2013). "Feeding the world: Genetically modified crops versus agricultural biodiversity". In: *Agronomy for Sustainable Development* 33.4, pp. 651–662. ISSN: 17740746. DOI: 10.1007/s13593-013-0138-9.

Jain, Anil K. (2010). "Data clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8, pp. 651–666. DOI: 10.1016/j.patrec.2009.09.011.

Jarvis, Andy, Sam Yeaman, Luigi Guarino, and Joe Tohme (2005). "The Role of Geographic Analysis in Locating, Understanding, and Using Plant Genetic Diversity". In: pp. 279–298. DOI: 10.1016/S0076-6879(05)95017-9.

Jombart, T. and I. Ahmed (2011). "adegenet 1.3-1: new tools for the analysis of genome-wide SNP data". In: *Bioinformatics* 27.21, pp. 3070–3071. DOI: 10.1093/bioinformatics/btr521.

Joost, Stéphane, Solange Duruz, Estelle Rochat, and Ivo Widmer (2016). "Open computational landscape genetics". In: *PeerJ Preprints* 4, e1721v3. DOI: 10.7287/peerj.preprints.1721v3.

Jørgensen, S. T., F. Liu, M. Ouédraogo, W. H. Ntundu, J. Sarrazin, and J. L. Christiansen (2010). "Drought Responses of Two Bambara Groundnut (Vigna subterranea L. Verdc.) Landraces Collected from a Dry and a Humid Area of Africa". In: *Journal of Agronomy and Crop Science* 196.6, pp. 412–422. ISSN: 09312250. DOI: 10.1111/j.1439-037X.2010.00435.x.

Kahane, Rémi, Toby Hodgkin, Hannah Jaenicke, Coosje Hoogendoorn, Michael Hermann, J. D H Dyno Keatinge, Jacqueline D'Arros Hughes, Stefano Padulosi, and Norman Looney (2013). "Agrobiodiversity for food security, health and income". In: *Agronomy for Sustainable Development* 33.4, pp. 671–693. ISSN: 17740746. DOI: 10.1007/s13593-013-0147-8.

Kalia, Rajwant K., Manoj K. Rai, Sanjay Kalia, Rohtas Singh, and A. K. Dhawan (2011). "Microsatellite markers: an overview of the recent progress in plants". In: *Euphytica* 177.3, pp. 309–334. ISSN: 0014-2336. DOI: 10.1007/s10681-010-0286-9.

Karp, Angela (1996). "Molecular Techniques in the Assessment of Botanical Diversity". In: *Annals of Botany* 78.2, pp. 143–149. DOI: 10.1006/anbo.1996.0106.

Karunaratne, A. S. (2009). "Modelling the response of Bambara groundnut (Vigna subterranea (L.) Verdc) for abiotic stress". PhD thesis. URL: http://etheses.nottingham.ac.uk/840/.

Karunaratne, A. S., S. N. Azam-Ali, G. Izzi, and P. Steduto (2011). "Calibration and validation of FAO-AQUACROP model for irrigated and water deficit Bambara groundnut". In: *Experimental Agriculture* 47.03, pp. 509–527. DOI: 10.1017/S0014479711000111.

Karunaratne, A. S., S. Walker, and S. N. Azam-Ali (2015). "Assessing the productivity and resource-use efficiency of underutilised crops: Towards an integrative system". In: *Agricultural Water Management* 147, pp. 129–134. ISSN: 18732283. DOI: 10.1016/j.agwat.2014.08.002.

Kawecki, Tadeusz J. and Dieter Ebert (2004). "Conceptual issues in local adaptation". In: *Ecology Letters* 7.12, pp. 1225–1241. DOI: 10.1111/j.1461-0248.2004.00684.x.

Kemp, Brian M, Angélica González-Oliver, Ripan S Malhi, Cara Monroe, Kari Britt Schroeder, John McDonough, Gillian Rhett, Andres Resendéz, Rosenda I Peñaloza-Espinosa, Leonor Buentello-Malo, Clara Gorodesky, David Glenn Smith, and Francisco Mauro Salzano (2010). "Evaluating the Farming/Language Dispersal Hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica". In: *Proceedings of the National Academy of Sciences* 107.15, pp. 6759–6764. DOI: 10.1073/pnas.0905753107.

Khan, Faraz, Hui Chai, Ishan Ajmera, Charlie Hodgman, Sean Mayes, and Chungui Lu (2017). "A Transcriptomic Comparison of Two Bambara Groundnut Landraces under Dehydration Stress". In: *Genes* 8.4, p. 121. DOI: 10.3390/genes8040121.

Kort, H De, K Vandepitte, J Mergeay, K V Mijnsbrugge, and O Honnay (2015). "The population genomic signature of environmental selection in the widespread insect-pollinated tree species Frangula alnus at different geographical scales". In: *Heredity* 115.5, pp. 415–425. DOI: 10.1038/hdy.2015.41.

Kunwar, Ripu M., Laxmi Mahat, Lila N. Sharma, Keshab P. Shrestha, Hiroo Kominee, and Rainer W. Bussmann (2012). "Underutilized plant species in Far West Nepal". In: *Journal of Mountain Science* 9.5, pp. 589–600. ISSN: 16726316. DOI: 10.1007/s11629-012-2315-8.

Labeyrie, Vanesse, Mathieu Thomas, Zachary K. Muthamia, and Christian Leclerc (2016). "Seed exchange networks, ethnicity, and sorghum diversity". In: *Proceedings of the National Academy of Sciences* 113.1, pp. 98–103. DOI: 10.1073/pnas.1513238112.

Lash, R, Darin S Carroll, Christine M Hughes, Yoshinori Nakazawa, Kevin Karem, Inger K Damon, and A Peterson (2012). "Effects of georeferencing effort on mapping monkeypox case distributions and transmission risk". In: *International Journal of Health Geographics* 11.1, p. 23. ISSN: 1476-072X. DOI: 10.1186/1476-072X-11-23.

Leibovici, Didier G., Suchith Anand, Roberto Santos, Sean Mayes, Rumiana Ray, Masoud Al-Azri, Abdul Baten, Graham King, Asha S. Karunaratne, Sayed Azam-Ali, and Mike J. Jackson (2017). "Geospatial binding for transdisciplinary research in crop science: the GRASPgfs initiative". In: *Open Geospatial Data, Software and Standards* 2.1, p. 20. ISSN: 2363-7501. DOI: 10.1186/s40965-017-0034-3.

Li, Yinghui, Rongxia Guan, Zhangxiong Liu, Yansong Ma, Lixia Wang, Linhai Li, Fanyun Lin, Weijiang Luan, Pengyin Chen, Zhe Yan, Yuan Guan, Li Zhu, Xuecheng Ning, Marinus J M Smulders, Wei Li, Rihua Piao, Yanhua Cui, Zhongmei Yu, Min Guan, Ruzhen Chang, Anfu Hou, Ainong Shi, Bo Zhang, Shenlong Zhu, and Lijuan Qiu (2008). "Genetic structure and diversity of cultivated soybean ( Glycine max ( L .) Merr .) landraces in China". In: pp. 857–871. DOI: 10.1007/s00122-008-0825-0.

Licker, Rachel, Matt Johnston, Jonathan A. Foley, Carol Barford, Christopher J. Kucharik, Chad Monfreda, and Navin Ramankutty (2010). "Mind the gap: How do climate and agricultural management explain the 'yield gap' of croplands around the world?" In: *Global Ecology and Biogeography* 19.6, pp. 769–782. ISSN: 1466822X. DOI: 10.1111/j.1466-8238.2010.00563.x.

Manel, Stéphanie, Michael K. Schwartz, Gordon Luikart, and Pierre Taberlet (2003). "Landscape genetics: Combining landscape ecology and population genetics". In: *Trends in Ecology and Evolution* 18.4, pp. 189–197. ISSN: 01695347. DOI: 10.1016/S0169-5347(03)00008-9.

Manni, Franz, Etienne Guerard, and Evelyne Heyer (2004). "Geographic Patterns of (Genetic, Morphologic, Linguistic) Variation: How Barriers Can Be Detected by Using Monmonier's Algorithm". In: *Human Biology* 76.2, pp. 173–190. DOI: 10.1353/hub.2004.0034.

Mantel, N. (1967). "The detection of disease clustering and generalised regression approach". In: *Cancer Research* 27, pp. 209–220.

Massawe, F J, S S Mwale, and J A Roberts (2005). "Breeding in bambara groundnut ( Vigna subterranea ( L .) Verdc .): strategic considerations". In: *African Journal of Biotechnology* 4.6, pp. 463–471. ISSN: 1684-5315. DOI: 10.5897/AJB2005.000-3084.

Massawe, Festo, Sean Mayes, and Acga Cheng (2016). "Crop Diversity: An Unexploited Treasure Trove for Food Security". In: *Trends in Plant Science* 21.5, pp. 365–368. ISSN: 13601385. DOI: 10.1016/j.tplants.2016.02.006.

Massawe, F.J., S. Mayes, A. Cheng, H.H. Chai, P. Cleasby, R. Symonds, W.K. Ho, A. Siise, Q.N. Wong, P. Kendabie, Y. Yanusa, N. Jamalluddin, A. Singh, R. Azman, and S.N. Azam-Ali (2015). "The Potential for Underutilised Crops to Improve Food Security in the Face of Climate Change". In: *Procedia Environmental Sciences* 29.Agri, pp. 140–141. DOI: 10.1016/j.proenv.2015.07.228.

Massawe, F.J., J.A. Roberts, S.N. Azam-Ali, and M.R. Davey (2003). "Genetic diversity in bambara groundnut (Vigna subterranea (L.) Verdc) landraces assessed by Random Amplified Polymorphic DNA (RAPD) markers". In: *Genetic Resources and Crop Evolution* 50.7, pp. 737–741. ISSN: 09259864. DOI: 10.1023/a:1025041301787.

Mayes, S., F. J. Massawe, P. G. Alderson, J. A. Roberts, S. N. Azam-Ali, and M. Hermann (2012). "The potential for underutilized crops to improve security of food production". In: *Journal of Experimental Botany* 63.3, pp. 1075–1079. ISSN: 0022-0957. DOI: 10.1093/jxb/err396.

Mercer, Kristin L. and Hugo R. Perales (2010). "Evolutionary response of landraces to climate change in centers of crop diversity". In: *Evolutionary Applications* 3.5-6, pp. 480–493. DOI: 10.1111/j.1752-4571.2010.00137.x.

Molosiwa, Odireleng (2012). "Genetic Diversity and Population structure analysis of bambara groundnut [Vigna subterranea (L.) Verdc.] Landraces using Morpho-agronomic Characters and SSR Markers". An optional note. PhD thesis. Loughborough, Leicestershire: University of Nottingham.

Molosiwa, Odireleng, Siise Aliyu, Florian Stadler, Katie Mayes, Festo Massawe, Andrzej Kilian, and Sean Mayes (2015). "SSR marker development, genetic diversity and population structure analysis of Bambara groundnut [Vigna subterranea (L.) Verdc.] landraces". In: *Genetic Resources and Crop Evolution* 62.8, pp. 1225–1243. ISSN: 0925-9864. DOI: 10.1007/s10722-015-0226-6.

Monmonier, Mark S. (1973). "Maximum-Difference Barriers: An Alternative Numerical Regionalization Method*". In: *Geographical Analysis* 5.3, pp. 245–261. DOI: 10.1111/j.1538-4632.1973.tb01011.x.

Morton, J. F. (2007). "The impact of climate change on smallholder and subsistence agriculture". In: *Proceedings of the National Academy of Sciences* 104.50, pp. 19680–19685. DOI: 10.1073/pnas.0701855104.

Mueller, Nathaniel D., James S. Gerber, Matt Johnston, Deepak K. Ray, Navin Ramankutty, and Jonathan A. Foley (2012). "Closing yield gaps through nutrient and water management". In: *Nature* 490.7419, pp. 254–257. DOI: 10.1038/nature11420.

Mwale, S. S., S. N. Azam-Ali, and F. J. Massawe (2007). "Growth and development of bambara groundnut (Vigna subterranea) in response to soil moisture. 1. Dry matter and yield". In: *European Journal of Agronomy* 26.4, pp. 345–353. ISSN: 11610301. DOI: 10.1016/j.eja.2006.09.007.

Nei, Masatoshi (1972). "Genetic Distance between Populations". In: *The American Naturalist* 106.949.

Ntundu, W. H., S. A. Shillah, W. Y F Marandu, and J. L. Christiansen (2006). "Morphological diversity of bambara groundnut [Vigna subterranea (L.) Verdc.] landraces in Tanzania". In: *Genetic Resources and Crop Evolution* 53.2, pp. 367–378. ISSN: 09259864. DOI: 10.1007/s10722-004-0580-2.

Ntundu, Wazael H., Inga C. Bach, Jørgen L. Christiansen, and Sven B. Andersen (2004). "Analysis of genetic diversity in bambara groundnut [Vigna subterranea (L.) Verdc] landraces using amplified fragment length polymorphism (AFLP) markers". In: *African Journal of Biotechnology* 3.4, pp. 220–225. ISSN: 16845315.

Ochatt, S. J., E. M. Patat-Ochatt, and A. Moessner (2011). "Ploidy level determination within the context of in vitro breeding". In: *Plant Cell, Tissue and Organ Culture (PCTOC)* 104.3, pp. 329–341. ISSN: 0167-6857. DOI: 10.1007/s11240-011-9918-6.

Olukolu, Bode A., Sean Mayes, Florian Stadler, N. Quat Ng, Iyiwole Fawole, Dumet Dominique, Sayed N. Azam-Ali, Albert G. Abbott, and Chittaranjan Kole (2012). "Genetic diversity in Bambara groundnut (Vigna subterranea (L.) Verdc.) as revealed by phenotypic descriptors and DArT marker analysis". In: *Genetic Resources and Crop Evolution* 59.3, pp. 347–358. ISSN: 09259864. DOI: 10.1007/s10722-011-9686-5.

OrozcoRamírez, Quetzalcóatl, Stephen B. Brush, Mark N. Grote, and Hugo Perales (2014). "A Minor Role for Environmental Adaptation in LocalScale Maize Landrace Distribution: Results from a Common Garden Experiment in Oaxaca, Mexico1". In: *Economic Botany* 68.4, pp. 383–396. ISSN: 00130001. DOI: 10.1007/s12231-014-9285-4.

Padulosi, S., N Bergamini, and T Lawrence (2011). *On-farm conservation of neglected and underutilized species : On-farm conservation of neglected and underutilized species : status, trends, and novel approaches to deal with climate change*, p. 307.

Padulosi, S., T. Hodgkin, J.T. Williams, and N. Haq (2002). "Underutilized Crops: Trends, Challenges and Opportunities in the 21st Century". In: *Managing Plant Genetic Diversity*, pp. 323–338.

Padulosi, Stefano, Karen Amaya, Matthias Jäger, Elisabetta Gotor, Wilfredo Rojas, and Roberto Valdivia (2014). "A holistic approach to enhance the use of neglected and underutilized species: The case of andean grains in Bolivia and Peru". In: *Sustainability (Switzerland)* 6.3, pp. 1283–1312. ISSN: 20711050. DOI: 10.3390/su6031283.

Paradis, E., J. Claude, and K. Strimmer (2004). "APE: analyses of phylogenetics and evolution in R language". In: *Bioinformatics* 20, pp. 289–290.

Pasquet, Rémy S., Sonya Schwedes, and Paul Gepts (1999). "Isozyme diversity in Bambara groundnut". In: *Crop Science* 39.4, pp. 1228–1236. ISSN: 0011183X. DOI: 10.2135/cropsci1999.0011183X003900040045x.

Perales, H. R., B. F. Benz, and S. B. Brush (2005). "Maize diversity and ethnolinguistic diversity in Chiapas, Mexico". In: *Proceedings of the National Academy of Sciences* 102.3, pp. 949–954. DOI: 10.1073/pnas.0408701102.

Peratoner, G., S. Seling, C. Klotz, C. Florian, U. Figl, and A. O. Schmitt (2016). "Variation of agronomic and qualitative traits and local adaptation of mountain landraces of winter rye (Secale cereale L.) from Val Venosta/Vinschgau (South Tyrol)". In: *Genetic Resources and Crop Evolution* 63.2, pp. 261–273. DOI: 10.1007/s10722-015-0245-3.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131. URL: https://CRAN.R-project.org/package=nlme.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org.

Ramanatha Rao, V. and Toby Hodgkin (2002). "Genetic diversity and conservation and utilization of plant genetic resources". In: *Plant Cell, Tissue and Organ Culture* 68.1, pp. 1–19. DOI: 10.1023/A:1013359015812.

Ray, Deepak K., Nathaniel D. Mueller, Paul C. West, and Jonathan A. Foley (2013). "Yield Trends Are Insufficient to Double Global Crop Production by 2050". In: *PLoS ONE* 8.6. ISSN: 19326203. DOI: 10.1371/journal.pone.0066428.

Richards, Gayle M Volk Christopher M (2011). "Integration of Georeferencing, Habitat, Sampling, and Genetic Data for Documentation of Wild Plant Genetic Resources". In: *HortScience* 46.11, pp. 1446–1449.

Rowe, G., M. Sweet, and T. Beebee (2017). *An Introduction to Molecular Ecology*. Oxford University Press.

Rurinda, J., P. Mapfumo, M. T. van Wijk, F. Mtambanengwe, M. C. Rufino, R. Chikowo, and K. E. Giller (2013). "Managing soil fertility to adapt to rainfall variability in smallholder cropping systems in Zimbabwe". In: *Field Crops Research* 154, pp. 211–225. ISSN: 03784290. DOI: 10.1016/j.fcr.2013.08.012.

Russell, Thembi, Fabio Silva, and James Steele (2014). "Modelling the spread of farming in the bantu-speaking regions of africa: An archaeology-based phylogeography". In: *PLoS ONE* 9.1. DOI: 10.1371/journal.pone.0087854.

Ruxton, Graeme D. and Markus Neuhäuser (2013). "Improving the reporting of P-values generated by randomization methods". In: *Methods in Ecology and Evolution* 4.11, pp. 1033–1036. DOI: 10.1111/2041-210X.12102.

Salmon, J.Meghan, Mark A. Friedl, Steve Frolking, Dominik Wisser, and Ellen M. Douglas (2015). "Global rain-fed, irrigated, and paddy croplands: A new high resolution map derived from remote sensing, crop inventories and climate data". In: *International Journal of Applied Earth*

*Observation and Geoinformation* 38, pp. 321–334. DOI: 10.1016/j.jag.2015.01.014.

Santos, Roberto Oliveira, Adam C. Algar, Richard Field, and Sean Mayes (2016). "Integrating GIScience and Crop Science datasets: a study involving genetic, geographic and environmental data". In: *PeerJ Preprints 4:e2248v3*. DOI: 10.7287/peerj.preprints.2248v3.

Savolainen, Outi, Martin Lascoux, and Juha Merilä (2013). "Ecological genomics of local adaptation". In: *Nature Reviews Genetics* 14.11, pp. 807–820. DOI: 10.1038/nrg3522.

Sentelhas, P. C., R. Battisti, G. M. S. Câmara, J. R. B. Farias, A. C. Hampf, and C. Nendel (2015). "The soybean yield gap in Brazil magnitude, causes and possible solutions for sustainable production". In: *The Journal of Agricultural Science* 153.8, pp. 1394–1411. DOI: 10.1017/S0021859615000313.

Siise, Aliyu and Festo J. Massawe (2012). "Microsatellites based marker molecular analysis of Ghanaian bambara groundnut (Vigna subterranea (L.) Verdc.) landraces alongside morphological characterization". In: *Genetic Resources and Crop Evolution* 60.2, pp. 777–787. DOI: 10.1007/s10722-012-9874-y.

Spear, Stephen F., Niko Balkenhol, Marie Josée Fortin, Brad H. McRae, and Kim Scribner (2010). "Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis". In: *Molecular Ecology* 19.17, pp. 3576–3591. DOI: 10.1111/j.1365-294X.2010.04657.x.

Stevens, Kim B. and Dirk U. Pfeiffer (2015). "Sources of spatial animal and human health data: Casting the net wide to deal more effectively with increasingly complex disease problems". In: *Spatial and Spatio-temporal Epidemiology* 13, pp. 15–29. ISSN: 18775845. DOI: 10.1016/j.sste.2015.04.003.

Storfer, a, M a Murphy, J S Evans, C S Goldberg, S Robinson, S F Spear, R Dezzani, E Delmelle, L Vierling, and L P Waits (2007). "Putting the "landscape" in landscape genetics." In: *Heredity* 98.3, pp. 128–42. ISSN: 0018-067X. DOI: 10.1038/sj.hdy.6800917.

Storfer, Andrew, Melanie A. Murphy, Stephen F. Spear, Rolf Holderegger, and Lisette P. Waits (2010). "Landscape genetics: Where are we now?" In: *Molecular Ecology* 19.17, pp. 3496–3514. ISSN: 09621083. DOI: 10.1111/j.1365-294X.2010.04691.x.

Thomas, Mathieu, Julie C. Dawson, Isabelle Goldringer, and Christophe Bonneuil (2011). "Seed exchanges, a key to analyze crop diversity dynamics in farmer-led on-farm conservation". In: *Genetic Resources and Crop Evolution* 58.3, pp. 321–338. DOI: 10.1007/s10722-011-9662-0.

Tobler, W. R. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region". In: *Economic Geography* 46, p. 234. DOI: 10.2307/143141.

Trabucco, A. and R. J. Zomer (2010). *Global Soil Water Balance Geospatial Database*. URL: http://www.cgiar-csi.org.

United Nations, Department of Economic and Social Affairs, Population Division (2015). *World Urbanization Prospects: The 2014 Revision*. Tech. rep. (ST/ESA/SER.A/366).

Van Ittersum, Martin K., Kenneth G. Cassman, Patricio Grassini, Joost Wolf, Pablo Tittonell, and Zvi Hochman (2013). "Yield gap analysis with local to global relevance-A review". In: *Field Crops Research* 143, pp. 4–17. ISSN: 03784290. DOI: 10.1016/j.fcr.2012.09.009.

Villa, Tania Carolina Camacho, Nigel Maxted, Maria Scholten, and Brian Ford-Lloyd (2005). "Defining and identifying crop landraces". In: *Plant Genetic Resources: Characterization and Utilization* 3.03, pp. 373–384. DOI: 10.1079/pgr200591.

Wang, Bing, Fenxiang Wen, Jiangtao Wu, Xiaojun Wang, and Yani Hu (2014). "Vertical Profiles of Soil Water Content as Influenced by Environmental Factors in a Small Catchment on the Hilly-Gully Loess Plateau". In: *PLoS ONE* 9.10. Ed. by Andrew C. Singer, e109546. DOI: 10.1371/journal.pone.0109546.

Wang, Ian J., Richard E. Glor, and Jonathan B. Losos (2013). "Quantifying the roles of ecology and geography in spatial genetic divergence". In: *Ecology Letters* 16.2, pp. 175–182. ISSN: 1461023X. DOI: 10.1111/ele.12025.

Westengen, Ola T., Mark Atam Okongo, Leo Onek, Trygve Berg, Hari Upadhyaya, Siri Birkeland, Siri Dharma Kaur Khalsa, Kristoffer H. Ring, Nils C. Stenseth, and Anne K. Brysting (2014). "Ethnolinguistic structuring of sorghum genetic diversity in Africa and the role of local seed systems". In: *Proceedings of the National Academy of Sciences* 111.39, pp. 14100–14105. DOI: 10.1073/pnas.1401646111.

Wheeler, T. and J. von Braun (2013). "Climate Change Impacts on Global Food Security". In: *Science* 341.6145, pp. 508–513. DOI: 10.1126/science.1239402.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown (2016). *The ASJP database (version 17)*.

Wouw, M van de, C Kik, T van Hintum, R van Treuren, and B Visser (2010). "Genetic erosion in crops: concept, research results and challenges". In: *Plant Genetic Resources-Characterization and Utilization* 8.1, pp. 1–15. ISSN: 1479-2621. DOI: 10.1017/S1479262109990062.

Wright, S. (1943). "Isolation by distance". In: *Genetics* 28, p. 114.

Zeven, A.C. (1998). "Landraces: A review of definitions and classifications". In: *Euphytica* 104.2, pp. 127–139. DOI: 10.1023/a:1018683119237.