Geostatistical Analysis of Genetic Diversity in the Present Male Argentine Population

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Dr. rer. agr. Amalia Nahír Díaz Lacava

aus

Buenos Aires

Bonn

2014

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter Prof. Dr. Wolfgang Alt

2. Gutachter Priv. Doz. Dr. Tim Becker

Tag der Promotion: 24.02.2015

Erscheinungsjahr: 2015

a Lucia y Aischa

a Markus

a mis padres

a Nico y Pablo

y a mis queridos que ellos trajeron

Summary

The geographical study of genetic diversity has a long history in fields related to population genetics. Its relevance has been already acknowledged in the 1940's and numerous tools have been developed since then. Most of these tools are based on principles and assumptions of classical population genetics and search for one or some particular spatial patterns fitting the observed data. The increasing availability of genetic markers and genotypes motivated the development of new tools in the last decades. Very powerful are the individual-based methods, which jointly analyze genetic and geographical information. Nevertheless, either methods are still based on quite simplistic assumptions or do not provide enough flexibility to analyze the complex spatial structure of a modern human group. Special challenges to the analysis are the presence of admixture –which implies the existence of genetically differentiated subgroups within geographically delimited groups- and guaranteeing individual anonymity –an utmost relevant reason why individual-based analyses in modern humans are almost always unfeasible.

This work presents a novel approach, a 'Genetic Geostatistical Framework', for the spatial analysis of genetic diversity with special focus on present-day populations, which are as a rule highly admixed. The proposed methodological approach is based on the integration of genetic statistics and spatial analysis, performed in the framework of Geographic Information Systems (GIS).

The underlying model defines the spatial pattern of genetic diversity of an admixed population as the spatial overlap of several groups of genetically similar individuals, each group accounting for an independent spatial frequency distribution. This is comparable to the idea of decomposing the genetic structure of an admixed population into genetic layers, each one represented by the spatial frequency distribution of a group.

This new Genetic Geostatistical Framework provides a flexible environment to precisely quantify and map inquires at population level related to the spatial distribution and frequency of alleles, haplotypes, and groups of closely related haplotypes. A novel feature is the identification, quantification, and mapping of the most frequent alleles or groups of closely related haplotypes per tract of land. The resulting maps partition the region into sub-regions, each one delimited by the area where one group of genetically closer individuals is the most predominant. This assessment

can be extended to map areas delimited by groups with the second highest spatial frequency and so forth.

In the case study the spatial diversity of the male population in central and northern Argentina was investigated. 145 samples were collected in 10 provinces and genotyped for highly polymorphic forensic Y-chromosomal haplotypes corresponding to seven short tandem repeat polymorphisms (Y-STR), also known as microsatellites –DYS19, DYS389I/II, DYS390, DYS391, DYS392, DYS393. Spatial analysis of the distribution of the most frequent alleles showed a clear differentiation among three regions. The northwest differed genetically in relation to central and littoral Argentina. The Argentinean Chaco presented a lower differentiation of the rest of the Argentine territory. The genetic geostatistical analysis identified four major groups of Y-chromosomal lineages, represented by groups of similar Y-STR haplotypes. Results showed that two genetically similar European lineages alternately dominate most of Argentina's territory. A third group was localized in the northwest, which included haplotypes representing South-American lineages. The fourth group was preponderantly localized in central Argentina, where it was the second in frequency. This last group included worldwide dispersed haplotypes. These results indicate a widespread European paternal ancestry throughout Argentina’s territory, substantial Amerindian proportion in the most northwestern area, and a noteworthy contribution of paternal lineages incoming from diverse worldwide origins. These findings are in agreement with ethno-historical, genetic, and demographic studies. It is noteworthy that a substantial genetic differentiation between Argentine territories, as it has been measured in this work, is of foremost relevance for statistical inferences in forensic studies based on Y-STR haplotypes. Additionally, a genetic geostatistical analysis was performed using published data of three populations inhabiting distant Argentine territories. The congruity between these and previously reported results, which were obtained with widely acknowledged population genetics methods, further demonstrates the reliability of this new method.

Findings of the spatial genetic diversity analysis of the georeferenced forensic Argentine data show that integrating genetic statistics and geostatistics within a GIS framework is a powerful instrument to address a wide range of spatial genetic inquiries at population level. These may include the

assessment of the spatial distribution and frequency of alleles, haplotypes, or groups of genetically similar individuals as well as the identification, quantification and mapping of the spatial coverage of the most frequent alleles or groups of related haplotypes per tract of land. In addition, this approach provides a flexible environment to adapt the analysis to various chromosomal levels and geographical scales and resolution. Results may be presented in form of summary statistics or charts. As well, patterns of genetic variation can be spatially quantified and precisely displayed in form of maps at the desired scale and resolution. The method has been primarily designed and tested on forensic Y-STR data. However, it will be straightforward to adapt it to other marker types or to increase substantially the number of loci. Since this new method aggregates genotypes spatially, violation of individual anonymity is also not at risk. In summary, it can be stated that this novel method offers an appropriate framework for detail investigation of spatial genetic diversity on the basis of genotypes accounting for a geographic reference.

Acknowledgments

This work is the result of my enthusiasm for GIS, genetics, and the diversity of life. It is the product of several years of research, conducted in parallel to my scientific activities at the Institute of Medical Biometrics, Informatics and Epidemiology (IMBIE) of the University of Bonn. At this institute and over many years I received full support from group leaders and colleagues. I am indebted to all these positive, creative and open-minded people. As well I am particularly appreciative to my mentors, collaborators and the dissertation committee.

Foremost, I want to express my special gratitude to Prof. Dr. Thomas Wienker, at that time head of the Research Group of Genetic Epidemiology. Thomas tutored and promoted this effort since its beginning. His encouragement, ideas and contacts allowed an initially explorative task to develop into a comprehensive thesis.

I am very grateful to Prof. Dr. Wolfgang Alt, head of the Interdisciplinary Group Theoretical Biology at the Faculty of Mathematics and Natural Sciences at the University of Bonn. Wolfgang encouraged me since the very beginning. I owe him deep gratefulness for his trust on my vision, for his dedicated mentorship, for his patient and careful guidance on mathematical modeling and for all the nice and fruitful talks.

I would like to express my heartfelt gratitude to Priv. Doz. Dr. Tim Becker, with whom I had the pleasure to cooperate in several projects and who took over the position of a supervisor of this thesis in its final stage. For this personal and professional valuable time, for our constructive discussions and for giving me the possibility to complete this thesis, I wish to dearly thank Tim.

My greatest thanks are also due to all my collaborators, who gave me access to real georeferenced genetic data used to challenge models and informatics systems presented in this thesis. I would like to give my special thanks to Dr. Gustavo Penacino, head of the DNA Analysis Unit, Official College of Pharmacists and Biochemists of Buenos Aires, Argentina, and coworkers, who provided

me the genetic data used in the case study. As well I wish to acknowledge Prof. Lutz Roewer, head of the Department of Forensic Genetics, Institute for Legal Medicine and Forensic Sciences, Charite-Universitätsmedizin Berlin, Germany, and coworkers, for giving me extended access to their valuable worldwide genotypic database. These data allowed me to conduct computations and tests in the initial stages of my research, to extend the scope of the findings in this thesis, and to accomplish further case studies using the proposed 'Genetic Geostatistical Framework' presented in this work.

My gratefulness goes as well to Prof. Dr. Andreas Hense –head of the Research Group for Climate Dynamics at the Meteorological Institute- and to Prof. Dr. Thomas Litt –head of the Research Group for Paleobotany at the Steinmann-Institute for Geology, Mineralogy and Paleontology-, both at the University of Bonn, for being members of my thesis committee. Moreover, this thesis greatly benefited from Prof. Hense's very helpful remarks. These motivated me to extend the analysis and to include a whole new chapter to this thesis in order to present a validation of the proposed methodology.

A large debt of gratitude is owed to my former colleague Maja Walier. Besides our regular exchange of ideas and her encouraging words, my thanks to Maja extend to our several scientific co-operations. In particular regarding this thesis, Maja developed the first version of the SAS code used to cluster haplotypes and to compute cluster frequencies. This code allowed me to perform basic explorative analysis. More specifically, it was very helpful to test the potentials and withdraws of several distances and clustering methods. For that delightful time and the efforts that we shared, I dearly want to thank her.

During the years I developed this thesis I received plenty of support from almost each one of my colleagues. This ranged from critical exchange of ideas to scientific collaborations. Great thanks go to Henning Henschke and Arnfried Schiller for their support and scientific contribution to the first conceptual framework which finally led to the methodology presented in this work to study the genetic diversity of modern humans in a continuous geographic space. My gratitude is also

extended to Dr. Manuel Mattheisen and Dr. Michael Steffens for their critical and constructive comments. Waldemar Spitz, Dr. Gustav Quade and Robert Fürst deserve as well my gratitude for their kindly IT support over many years. I would like to give special thanks to Waltraud Schillo for her standing support by literature search and Birgit Buchholz for her artwork assistance. And I wish to give thanks to Dr. Christine Herold for her friendly support.

To my dear friend Dr. Lodovica Borghese I am grateful for her helpful discussion and encouraging words.

And finally, to my family and friends I wish to express my dearest gratitude for their strong and lovely encouragement all the years I was dedicated to this scientific journey.

Contents

PART I - INTRODUCTION **1**

1 General Introduction to the Research Topic **1**

2 Aim and Objectives of the Study **5**

PART II - LITERATURE REVIEW **7**

3 Spatial Genetic Diversity in Modern Human Populations **7**

3.1. Elementary Definitions 7

3.2. Variability of the Human Genome 11

3.3. The Generation of Genetic Variation 12

3.4. 'Geo-'Genetics 14

3.5. Basic Measures of Genetic Differentiation 17

3.6. Statistical Toolbox in Spatial Genetics 19

3.7. Spatial Genetic Diversity in Humans 21

3.8. The Choice of the Molecular Marker 22

4 The Argentine Republic **27**

4.1. The Country 27

4.2. Geography 28

4.3. History 31

4.3.1. The Colonial Period 32

4.3.2. Mass Migration Period 37

4.3.3. Increasing Admixture 39

PART III - MATERIALS AND METHODS **43**

5 Geostatistical Analysis of an Admixed Human Population **43**

5.1. The Basic Scenario 43

5.2. Genetic Geostatistical Framework 44
5.2.1. The Basic Model 46
5.2.1.1. Parameters for the Delimitation of Groups of Genetically Similar Individuals 46
5.2.1.2. Spatial Probability 50
5.2.2. Computational Procedure for Determining Frequencies per Sampling Unit 52
5.2.3. Computational Procedure for Determining Spatial Probabilities 53
5.2.4. Spatial Overall Ranking 54
5.2.5. Composite Maps 54
5.2.6. Screening Algorithms 55
5.2.7. Assessment Routine 56
6 The Case Study 59
6.1. Urban Male Argentine Genetic Admixture 59
6.2. Study Region and Spatial Sampling Units 60
6.3. Subjects and Genotypes 62
6.4. Genetic Frequencies 64
6.4.1. Computation of Y-STR Allele Frequencies 64
6.4.2. Computation of Spatially Aggregated Frequencies 64
6.5. Geostatistical Analysis 66
6.5.1. Surface Interpolation 66
6.5.2. Creation of Composite Maps 68
6.6. Characterization of Y-Chromosome Ancestry 70

PART IV - RESULTS 73

7 The Genetic Heterogeneity of the Urban Argentine Population 73
7.1. Frequency Distribution of Y-STR Alleles 73
7.2. Frequency Distribution of Y-STR Haplotype Frequencies 74
7.3. Spatial Diversity at the Y-STR level 75
7.4. Spatial Diversity of Y-STR Haplotypes 78
8 Spatial Admixture of three Sub-Populations 87

8.1. Three Argentine Male Data Sets 87
8.2. Choice of Number of Y-STR Haplotype Clusters and Interpolation Parameters 89
8.3. Evaluation of the Spatial Structure of the Data 90
8.4. Haplotype Frequencies and Clustering Results 91
8.5. Geographical Patterns of the three Major Groups 92

PART V - DISCUSSION AND CONCLUSIONS 97

9 The Genetic Geostatistical Framework 97
10 Male Genetic Heterogeneity in Nowadays Argentina 103
11 The Geographical Study of Genetic Heterogeneity in Modern Humans 105
12 Conclusions 109

PART VI - REFERENCES 111

List of Tables

TABLE III-1	DESCRIPTION OF THE SAMPLING UNITS	62
TABLE IV-1	ABSOLUTE FREQUENCY AND CLUSTER ASSIGNMENT OF FREQUENT HAPLOTYPES	74
TABLE IV-2	GOODNESS OF FIT FOR SPATIAL INTERPOLATION FREQUENCIES OF Y-STR LOCI	75
TABLE IV-3	FREQUENCY DISTRIBUTION OF HAPLOTYPES AND SAMPLES PER CLUSTER	78
TABLE IV-4	GOODNESS OF FIT FOR SPATIAL INTERPOLATION OF Y-STR HAPLOTYPE CLUSTER FREQUENCIES	79
TABLE IV-5	FREQUENCY DISTRIBUTION OF SAMPLES AND HAPLOTYPES PER HAPLOGROUPS AND CLUSTERS	92
TABLE IV-6	GOODNESS OF FIT FOR SPATIAL INTERPOLATION OF Y-STR HAPLOTYPE CLUSTER FREQUENCIES	93

List of Figures

FIGURE II-1 SCHEMATIC REPRESENTATION OF A DNA MOLECULE. 8

FIGURE II-2 SCHEMATIC REPRESENTATION OF A CHROMOSOME AND THE LOCATION OF A GENE. 9

FIGURE II-3 HUMAN CHROMOSOMES. 9

FIGURE II-4 IDEOGRAPH OF THE Y CHROMOSOME. 10

FIGURE II-5 TOPOGRAPHIC MAP OF THE ARGENTINE REPUBLIC. 29

FIGURE III-1 SCHEMATIC REPRESENTATION OF A GEOSTATISTICAL APPROACH FOR THE DELIMITATION OF THE SPATIAL COVERAGE OF THE MOST FREQUENT GROUPS IN AN AREA. 45

FIGURE III-2 SCHEMATIC REPRESENTATION OF THE STUDY REGION. 51

FIGURE III-3 REPRESENTATION OF SINE AND ARCSINE TRANSFORMATION. 53

FIGURE III-4 WORKFLOW OF AN ASSESSMENT ROUTINE WITHIN THE GENETIC GEOSTATISTICAL FRAMEWORK. 57

FIGURE III-5 STUDY REGION. 61

FIGURE III-6 HISTOGRAM SHOWING THE NUMBER OF SAMPLES PER HAPLOTYPE AND PER CLUSTER. 65

FIGURE III-7 IMPACT OF SMOOTHING PARAMETER ON THE RESULTING SURFACE. 69

FIGURE IV-1 ALLELE FREQUENCY DISTRIBUTION OF SEVEN Y-STR LOCI. 74

FIGURE IV-2 SPATIAL DISTRIBUTION OF THE MOST FREQUENT ALLELES PER Y-STR LOCUS. 77

FIGURE IV-3 SPATIAL DISTRIBUTION OF THE MOST FREQUENT Y-STR HAPLOTYPE CLUSTERS. 80

FIGURE IV-4 SPATIAL DISTRIBUTION OF THE SECOND MOST FREQUENT Y-STR HAPLOTYPE CLUSTERS. 81

FIGURE IV-5 ISOLINE MAPS SHOWING THE SPATIAL DISTRIBUTION OF Y-STR HAPLOTYPE CLUSTERS. 82

FIGURE IV-6 PROFILE OF FREQUENCIES OF Y-STR HAPLOTYPE CLUSTERS ALONG TRANSECTS. 83

FIGURE IV-7 SAMPLING AREAS. 88

FIGURE IV-8 SPATIAL DISTRIBUTION OF Y-STR HAPLOTYPE CLUSTERS. 93

FIGURE IV-9 SPATIAL FREQUENCY DISTRIBUTION OF INTERPOLATION AND POINT DATA VALUES. 94

FIGURE IV-10 SPATIAL DISTRIBUTION OF Y-STR HAPLOTYPE CLUSTERS. 95

List of Abbreviations

ABO	ABO blood group system
DNA	deoxyribonucleic acid
GIS	geographic information system
GRASS GIS	Geographic Resources Analysis Support System; http://grass.itc.it/
IBD	identical by descent
IBS	identical by state
SAS	Statistical Analysis Software, SAS Institute Inc., Cary, NC, USA; http://www.sas.com/
SNP	single nucleotide polymorphism
STR	short tandem repeat

PART I - INTRODUCTION

1 General Introduction to the Research Topic

Good knowledge of the genetic structure of a population is relevant to plenty of fields related to genetic studies. This includes among others genetic epidemiology, pharmacogenetics, evolution, and forensic genetics. Numerous studies in population genetics provide substantial evidence of the geographical genetic structure of continuously populated territories at broader scales (Barbujani & Sokal 1990; Cavalli-Sforza et al. 1994; Roewer et al. 2005; Rosser et al. 2000; Zerjal et al. 2001). For instance, whole genome analysis clearly demonstrated spatial correspondence of major human groups to broad socio-demographic regions in Western Europe (Lao et al. 2008; Novembre et al. 2008).

Among the most applied methods to study geographic patterns of genetic structure are: assignment procedures, ordination methods, analysis of spatial variation in polymorphism frequency on the basis of visualization tools or population genetics summary statistics, synthetic maps, and the search of genetic boundaries –zones of increased differentiation of genetic frequencies, also denominated genetic discontinuities (reviewed in: Barbujani 2000; Epperson 2003; Manel et al. 2003; Storfer et al. 2007). Most of these methods search for one or some specific overall types of pattern fitting the data. This limitation is either related to method's assumptions or to methodological restrictions. Consequently, the detected spatial genetic structure of a population will depend on the chosen method.

Geographical genetic structures may follow complex patterns, including clines, discontinuities and even gaps (Fechner et al. 2008). For example, in regard to the population structure of Europe several studies demonstrated patterns of clinal variation (Gusmão et al. 2003; Rosser et al. 2000), while others identified spatial genetic discontinuities (Barbujani & Sokal 1990; Kayser et al. 2005; Zerjal et al. 2001). A more suitable approach to study complex patterns of genetic variation may: (a) increase the information content gained from the sample, and (b) avoid setting *a prior* on the overall type of spatial structure. Several methods proposed in the past recent years in the fields of molecular biology and ecology successfully increase the amount of information extracted from population genetic data (Guillot et al. 2005a; Hardy & Vekemans 2002; Manel et al. 2007). Such methods examine the geographical position of each sample individually and jointly analyze inter-individual genetic and spatial relationships.

The application of these methods to human data is close to unfeasible due to at least two main reasons: restrictions related to guaranteeing sample anonymity and admixture of modern societies.

Individual Anonymity: In the scope of population genetics studies of modern human populations, ethical restrictions usually require to guarantee individual anonymity (i.e. sample characterization must exclude the possibility of personal identification). To link an individual to a precise and unique geographical reference (e.g., postal address) is, in most cases, in conflict with such restrictions. Consequently, less precise geographical references are generally used. Samples may be georeferenced to sampling areas (e.g., state, country) or sites (e.g., sampling institution, hospital, urban settlement of recruitment). This procedure usually results in a geographical aggregation of samples. This situation arises because sampling design of genetic studies on modern humans usually involves a considerably low number of sampling locations in relation to the number of samples.

Admixture: To define each sampling location as a single population, i.e. a group of individuals who are likely to mate and reproduce, is a largely chosen approach in population genetics studies and it has been implemented for some of the individual-based methods cited above (Hardy & Vekemans 2002). The implicit assumption of such *modus operandi* is that each of those geographically defined populations can be considered genetically homogeneous. Since modern

societies may present a considerable degree of admixture, this assumption may not be quite realistic.
Geographical aggregation of samples may actually reduce the potentials of individual-based methods and it may hinder the search of complex spatial patterns of genetic heterogeneity.

An alternative approach of geographical data aggregation involves firstly grouping samples according to genetic similarities and then inspecting the spatial distribution of such groups. In the field of forensic genetics, studies applying such an approach observed clines of genetic variation and detected differences in frequencies among regions at continental and country scales. Spatial patterns were basically assessed on the basis of visualization methods. Commonly applied methods included: (a) comparison of groups' frequencies within a region, e.g., using pie or bar charts for the representation of group frequencies at each geographical location (Brion et al. 2004; Gusmão et al. 2003; Rosser et al. 2000), (b) evaluation of the spatial distribution of frequencies separately for each group, e.g., on the basis of visual examination of interpolated surfaces, which estimated the spatial frequency of each group (Lao et al. 2008; Roewer et al. 2005). In the first case, such type of visualization techniques facilitates the comparison of frequencies among groups at each sampling location; in addition this explorative approach does not set *a prior* on the overall type of spatial structure. However, characterizing the spatial pattern of genetic heterogeneity of the total sample on the basis of visual examination of pie or bar charts becomes more difficult with an increase in the number of compared groups, the number of observed locations, or pattern complexity. The second approach, based on the visual examination of one interpolation surface per group, largely facilitates the identification of patterns of genetic variation across space of each group. Interpolation surfaces indeed estimate the spatial distribution of values; in this specific case, the genetic frequencies. Several interpolation methods as well as stand-alone software are available to create interpolated surfaces. Nonetheless, in order to assess the spatial pattern of genetic heterogeneity jointly for all groups, the information contained in these surfaces must be somehow summarized (Barbujani 2000). The full use of spatial information content of the data requires to go beyond visual examination of partial spatial relationships and to perform deep geostatistical analysis of genetic relationships. A ***geographic information system*** (**GIS**) is a software package

specifically developed for analyzing geographic relationships. It is a very suitable platform for geostatistical analysis of complex spatial genetic structures. Once genetic data is framed within a GIS a wide range of geostatistical tests may be performed in order to generate and, eventually, to test hypothesis concerning the spatial genetic pattern of admixed groups populating a geographical space.

2 Aim and Objectives of the Study

The aim of this work is to present a new GIS-based approach for analyzing the spatial genetic diversity of admixed populations over spatially continuous regions. A '**Genetic Geostatistical Framework**' is set up, integrating genetic statistics and geostatistical analysis. Main components of this methodology are the assessment of the spatial distribution of groups of genetically similar individuals and the quantification of the spatial coverage and pattern of the most frequent groups.

This approach was primarily designed for and tested on forensic Y-chromosomal data.
Forensic Y-chromosomal short tandem repeat markers (Y-STR), also denominated Y-chromosomal microsatellites, are highly polymorphic. Non-recombinant Y-chromosomal loci are very likely selection-neutral and are jointly inherited. Because these loci are inherited as a block, unambiguous haplotypes can be defined. Y-STR haplotypes present consequently even much higher levels of polymorphism. Due to these characteristics and the standardized, high-quality genotyping procedure used to obtain this type of forensic markers, forensic Y-STR loci and Y-STR haplotypes are exceptionally informative genotypes and provide an outstanding basis for the analysis of population heterogeneity (Bosch et al. 2001; Brion et al. 2004; Diaz Lacava et al. 2011a; Diaz Lacava et al. 2011b; Gusmão et al. 2003; Kayser et al. 2005; Roewer et al. 2005; Zerjal et al. 2001). Forensic Y-STR loci and haplotypes have been proven to be suitable for genetic studies of spatially contiguous, admixed populations (Brion et al. 2004; Diaz Lacava et al. 2011a; Diaz Lacava et al. 2011b; Kayser et al. 2005).

This new methodology is introduced and discussed by analyzing Argentine forensic Y-chromosomal genotypes.
The Argentine population is of special interest due to two main factors:

(a) A complex, fine-scale spatial pattern of genetic variation is expected.
Argentina extends over vast and differentiated geographical regions. Immigrants, primarily

European lineages, and Amerindian populations mixed differentially in the distinct geophysical territories over the last centuries.

(b) There is abundant published data to verify result validity.
Plenty of published ethno-historical, genetic and census data is available; major, large-scale immigration occurred relative recently and the Argentine demographic history is consistently documented as well.

A major focus of this thesis is a case study. This case study evaluates the genetic diversity of the present urban male population in Argentine. Data was collected in the context of paternity tests in urban areas of 10 provinces of central and northern Argentina in 2007.

Further, the reliability of the presented methodology is demonstrated using published data of three geographically-distant Argentine populations, previously analyzed with well-established methods of population genetics.

Finally, potentials and limits of this Genetic Geostatistical Framework are inspected and discussed.

PART II - LITERATURE REVIEW

3 Spatial Genetic Diversity in Modern Human Populations

This chapter provides basic definitions as well as a concise overview of current knowledge and available methods and techniques for the analysis of genetic variation of modern human populations across the geographic space.

3.1. Elementary Definitions[1]

The hereditary material is denominated **genome**. The genome includes all required information to build and maintain an organism. This information is organized in **chromosomes** and is found practically in each human cell. The main constituent of the chromosome is the **deoxyribonucleic acid** (DNA). Most DNA is contained in the cell nucleus, denominated nuclear DNA. A nuclear DNA unit is basically an extremely large double strand of molecules that forms a spiral called **double helix**. These molecules are **nucleotides**, usually designated **bases**, which are joined to each other in a chain by covalent bonds, resulting in a **sugar-phosphate backbone**. There are in total four bases, which are usually represented by letters: adenine (**A**), guanine (**G**), cytosine (**C**), and thymine (**T**). Bases pair up with each other in a specific fashion according to their chemical properties, A with T and C with G, and form units called **base pairs** (Figure II-1).

1 reviewed in: Cavalli-Sforza et al. (1994) and in Hart & Clark (1997); alternative sources are indicated in text.

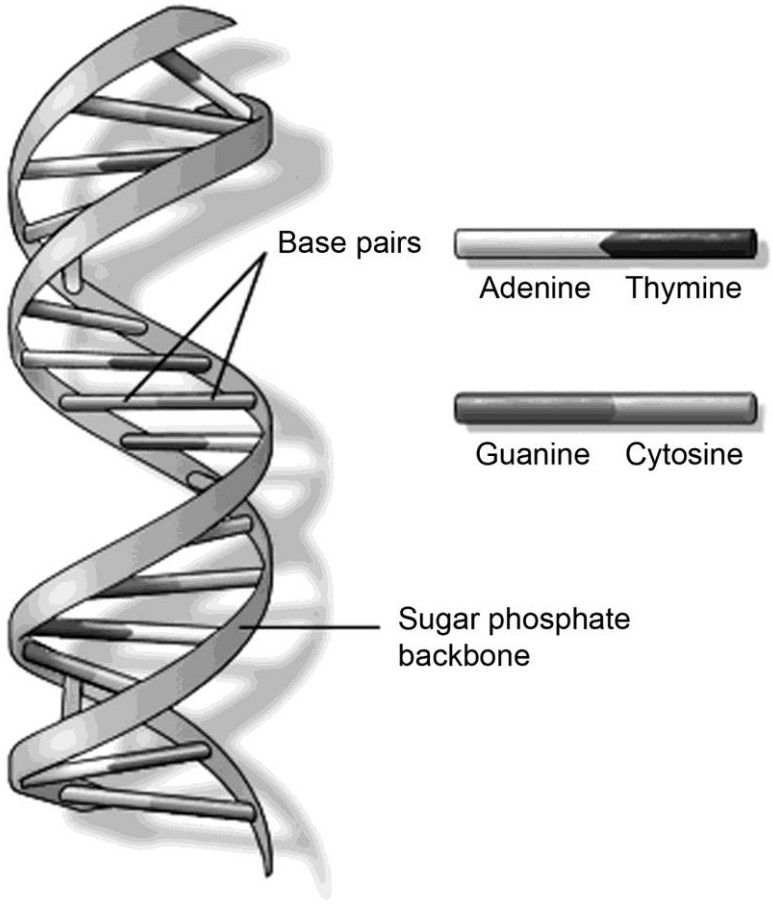

Figure II-1 Schematic representation of a DNA molecule.

Source: Genetics Home Reference [Internet]. [place unknown]: National Library of Medicine® 2013. [Figure], What is DNA? [cited 2013 Jan 21]. Available from: http://ghr.nlm.nih.gov/handbook/basics/how manychromosomes

The whole nuclear genome contains more than three billion base pairs. Their linear order is actually the storage of the information content. Humans share more than ninety nine percent of such linear order (Genetics Home Reference 2013; http://ghr.nlm.nih.gov/). Shorter segments of DNA with specific functions, **genes**, are spread along the chromosome (Figure II-2). On average many thousands of genes are present in one chromosome.

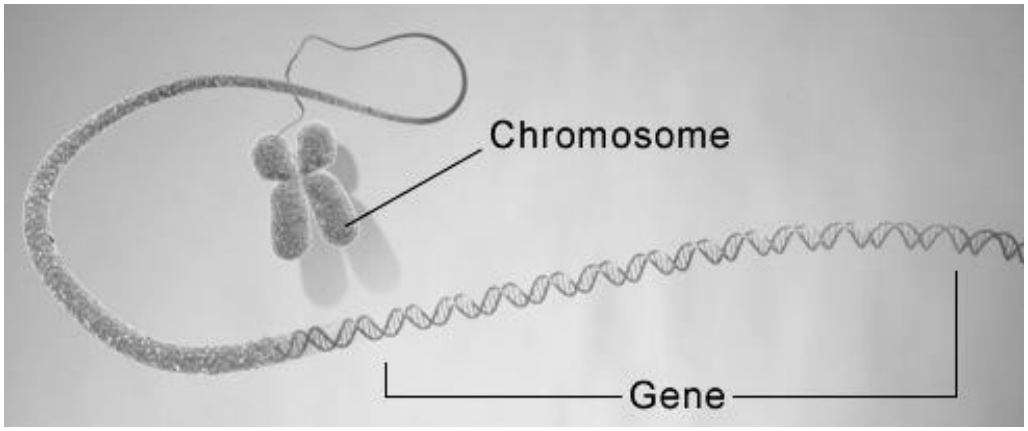

Figure II-2 Schematic representation of a chromosome and the location of a gene.

Source: Genetics Home Reference [Internet]. [place unknown]: National Library of Medicine® 2013. [Figure], What is a gene? [cited 2013 Jan 21]. Available from: http://ghr.nlm.nih.gov/handbook/basics/how manychromosomes

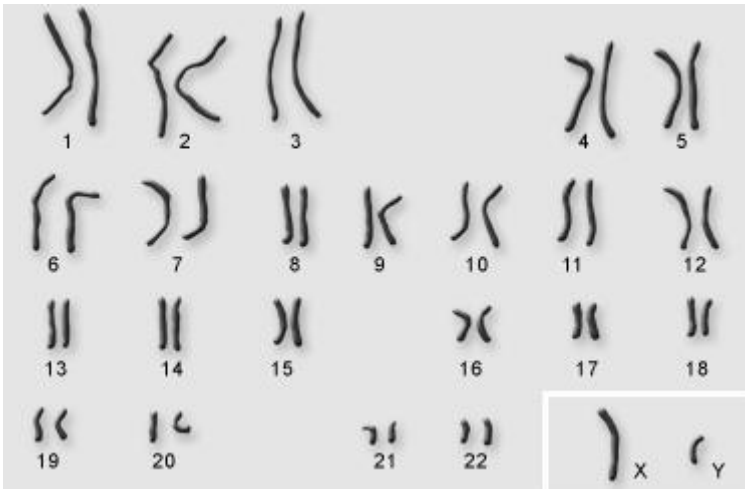

Figure II-3 Human chromosomes.

Number and appearance of nuclear chromosomes, arranged as pairs. Source: Genetics Home Reference [Internet]. [place unknown]: National Library of Medicine® 2013. [Figure], How many chromosomes do people have? [cited 2013 Jan 21]. Available from: http://ghr.nlm.nih.gov/handbook/basics/howmany chromosomes

The human nuclear genome comprises 23 pairs of chromosomes, which can be morphologically distinguished (Figure II-3). One pair determines the person's sex. These chromosomes are called **allosomes** or sex chromosomes. Females carry a pair of X chromosomes and males one X chromosome and one Y chromosome. The other 22 pairs are called **autosomal**. Both units of each pair of autosomal and X chromosomes are morphologically indistinguishable but account subtle differences detectable at the nucleotide level. The presence of two copies of each unique chromosome is referred to as **diploidy**.

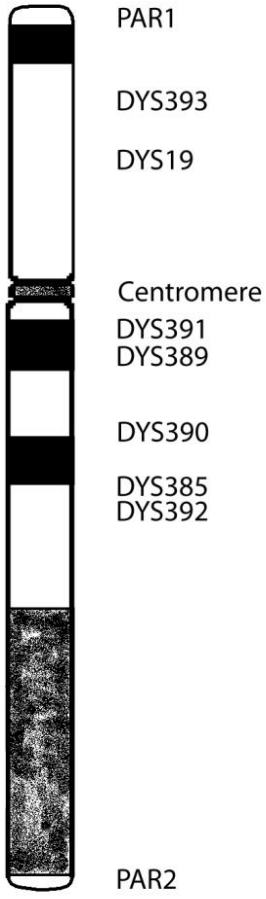

Figure II-4 Ideograph of the Y chromosome.

The ideograph illustrates the locations of the two pseudoautosomal regions (PAR1 and PAR2), the non-recombinant region, and the location of seven well-established forensic microsatellites (Y-STR). Source: Roewer (2001).

Considering a pair of chromosomes, one is contributed by the mother through the egg (ovum) and the other by the father through the sperm. Besides the special case of the Y chromosome, each one of all other chromosomes constitutes a **recombination** of the respective parental pair of chromosomes; the process by which the chromosomal material is exchanged is called **crossing over**. This can take place with a small probability anywhere along the chromosome and the probability of two crossing overs occurring on the same chromosome increases with the distances between their locations. Recombination takes place during the meiosis, i.e. a special type of cell division that produces the **gametes**, i.e. eggs or sperms.

The Y chromosome is a special case of nuclear genome. It is inherited from fathers to sons without recombination in most of its length (Figure II-4). To the largest extent the Y-chromosomal genetic material is not subjected to recombination and it is transmitted over generations almost in its ancestral form. Genetic variation in these regions of the human genome evolves through the accumulation of mutations. Because a male individual only carries one copy of this chromosome it is denominated **haploid.**

3.2. Variability of the Human Genome

Although geneticists have been aware of a certain amount of genetic variation among individuals of the same species for a long time, the full extent of individual variation was first appreciated in the 1980s, as the analysis at the level of the genetic material became possible (Cavalli-Sforza et al. 1994). Since then the techniques of DNA analysis have been developing rapidly (Schlötterer 2004) leading recently to sequencing of whole human genomes of globally representative populations (International Human Genome Sequencing Consortium 2004; Venter et al. 2001). An enormous wealth of genetic data of worldwide populations is now freely available (http://www.ncbi.nlm.nih.gov/; http://www.yhrd.org/).

The variation among individuals of the same species of one specific segment of the genome is usually indicated as **polymorphism**, **marker**, or **locus**; the different forms or attributes of one specific polymorphism is called **allele**. The proportion of a given allele, i.e. the **allele frequency**, varies considerably across space, but the greatest variation is observed at large distances (Cavalli-Sforza et al. 1994; Manica et al. 2005). A unique combination of alleles along a segment of a

chromosome is called **haplotype** (Hart & Clark 1997).

Current estimations indicate that 99.5 percent to 99.8 percent of human nuclear genome is essentially identical in every individual; this implies that although variants make up a small fraction of the total nuclear genome, in average 6 to 15 million nucleotides still vary between two randomly selected people (Kidd et al. 2004).

3.3. The Generation of Genetic Variation

The evolutionary process generating genetic variation is driven by **mutation**, **gene flow**, **natural selection** and **random genetic drift**.

Under **mutation** is meant the occurrence of errors in the DNA replication; it refers to replacement, addition, or deletion of one or several nucleotides. Mutations are rare events, but since the human genome is constituted by approximately three billion bases, some dozens of different mutations may be transmitted to the next born person. This process is considered basically random. Eventually a specific DNA segment may be hit by a second mutation, so that many alleles of one segment can arise and coexist in a population. In this case that segment or locus is called **polymorphic**.

Mutations affect the genome at the single locus level; they are the source of new genetic material. Therefore mutations are one of the key ingredients of evolution. But since the rate of recurrence of the same mutation is extremely low, the fate of one particular mutation depends on the other three elementary forces, i.e. natural selection, gene flow, and random genetic drift.

Natural selection is the automatic process of indirectly sorting out and favoring useful alleles while eliminating deleterious ones. It acts only on those mutations affecting the **fitness** of the individuals, i.e. the performance of the individual organism in relation to survival and fecundity. Alleles that are considered not to have a particular significance in the fitness of individuals, thereof 'invisible' to natural selection, are called **neutral** alleles. Neutral alleles are ideal for mapping the geographic distribution of average genetic variation of populations and for tracing ancestral lineages of genomic variants.

Gene flow is the result of migration processes and as a consequence new alleles are brought to a population. **Admixture** is the result of gene exchange between or among genetically different

populations. Gene flow affects the genome as a whole, all loci equally at the same time, regardless of the gene-flow amounts and the number of generations involved in the process (Wijsman 1984). It results in a new population. Allele frequencies of the admixed population are a constant linear combination of the allele frequencies in the parental ones (Cavalli-Sforza et al. 1994). Natural selection and gene flow drive allele frequencies in a specific, and to some extent predictable, direction.

Random genetic drift is the effect of random sampling of a gamete, the particular egg or sperm that gives raise a new individual, at each generation. Like mutations, drift is random.

Genetic drift affects all loci uniformly (Cavalli-Sforza et al. 1994). The effect on a given allele depends on the particular allele frequency and population size (Wijsman 1984). On average, greater changes in allele frequencies are expected when the population size is small (Cavalli-Sforza et al. 1994). Isolated populations tend to differ from other populations of the same geographic or ethnic origin. Since such populations are also less likely to present genetic similarities with other geographically more distant and genetically unrelated populations, they tend to behave as 'outliers' when compared all together (Cavalli-Sforza et al. 1994).

Leaving aside migration, natural selection and random drift are the cause of variation of allele frequencies in populations (Cavalli-Sforza & Feldman 2003). Estimating the effects of drift and selection on the generation of human genetic heterogeneity is considerably difficult (Wijsman 1984). The joint effect of these two factors must have differed in evolutionary times across regions and molecular evidence indicates that most probably different evolutionary models are required to explain genomic variation in different continents (Cavalli-Sforza & Feldman 2003). Present-day patterns of genetic variation could have been arisen due to different demographic, mutational as well as selective processes (Pannell & Charlesworth 2000). Assessments of evolutionary process and genetic composition of past populations remain inferential and require information from other fields, e.g., archaeology, palaeontology, or linguistics, which often gain information directly from ancient material (Goldstein & Chikhi 2002). Nevertheless, where admixture took place recently, involving large sources of populations, it is unlikely that drift may have had a considerable influence on the allele frequencies of the new admixed population and can be, to some extent, neglected (Wijsman 1984).

3.4. 'Geo-'Genetics

The relevance of including spatial components to the analysis of genetic variation was already acknowledged in the 1940's in theoretical studies of Dobzhansky and Wright (1941), Wright (1943) and Malécot (1948) (Guillot et al. 2009). Nevertheless, there is no spatial genetic theory available yet (Guillot et al. 2009).

Adequately summarizing spatial genetic variation is not straightforward. Genetic data merges information about historical and current processes (Balkenhol et al. 2009). Spatial patterns of genetic variation capture the cumulative effects of processes generating it over many generations (Epperson 2003). Contradictory results may be obtained analyzing the geographic structure of a population just by considering different assumptions of the mutation model, even when using the same data (Barbujani & Belle 2006). These 'incongruities' may not relate to data or methodological errors, but reflect a basic feature of human diversity: there is a heterogeneous distribution of genetic polymorphisms and in most cases these do not correlate (Barbujani & Belle 2006). This is so because, due to the complexity of the interrelated effects of selection, gene flow, mutations, and drift, every allele varies in frequency over evolutionary times in a relatively unpredictable manner, so that it can be considered, at least superficially, nearly random (Cavalli-Sforza et al. 1994). The interpretation of contemporary spatial genetic variation requires taking into account geographical constrains, demographic processes, and historical events (Balkenhol et al. 2009).

Classical population genetics considers populations to be spatially organized as discrete patches, i.e. corresponding to sub-populations, or to follow simple clinal patterns, e.g., isolation-by-distance (Balkenhol et al. 2009). Populations –and sub-populations- are defined *a priori*. In contrast, more recent individual-centered methods directly evaluate individual genotypes and their geographical co-ordinates using spatial methods (Manel et al. 2007). The latter approach is more appropriate when individuals are more uniformly distributed in space or are not arranged in a clustered distribution (Manel et al. 2003).

Spatial models in classical population genetics most generally assume a homogeneous geographical environment or, at least, these models assume that gene flow would not depend on the characteristics of the geographical region (Balkenhol et al. 2009). Models focus on **allopatric**

speciation, i.e. new species are created as consequence of restricted gene flow, followed by drift or/and local speciation (Balkenhol et al. 2009; Futuyma 1997). It is assumed that the system is in equilibrium and it models such processes independently of the heterogeneity of the geographical environment (Balkenhol et al. 2009). The inclusion of spatial components is constrained to test the effect of geographic distance (Guillot et al. 2009). Methods developed under this paradigm may calculate genetic distances among individuals or among populations. A battery of frequently applied tests is based on these assumptions. Most of these are based on global statistics, i.e. all spatial data points go jointly in the computation of model parameters and these mirror global properties of the sample over the total study region (Guillot et al. 2009). Considering that spatial environments are most usually not homogeneous, and these as well as the genetic processes occurring within them are rarely stable or at equilibrium, such models and the corresponding statistical tools are generally not suited to evaluate the influence of heterogeneous, evolving geographic environments, especially when the analysis involves individuals, i.e. continuous populations (Balkenhol et al. 2009).

Since one of the first studies of spatial genetic variation of Haldane (1940) –who reported clinal patterns of ABO*B frequencies in Europe- over largely acknowledged theoretical and empirical studies in more recent years (Barbujani & Sokal 1990; Barbujani 1987; Cavalli-Sforza et al. 1994; Lao et al. 2008; Piazza et al. 1995; Sokal 1988; Sokal & Oden 1978), all these works regarded the inclusion of geographic analysis as the examination of the effect of geographic distance. Epperson (2003) presented under the denomination of **Geographical Genetics** a deep analysis of the mathematical-statistics properties of spatial measures in relationship to measures of stochastic space-time processes. This work included a rigorous compendium of the most relevant statistical methods applied to model geographic patterns of genetic variation. Though the comprehensive treatment of this discipline, the cited methods still define the geographic region as a mere Euclidean distance dimension where genetic processes occurred.

The relatively new field **landscape genetics** integrates data and methods from geography, geostatitstics, landscape ecology and population genetics to analyze and understand the spatial distribution of genetic variation (Manel et al. 2003; Holderegger & Wagner 2006; Storfer et al.

2007). This field is explicitly concerned with interaction between the landscape –composition, configuration, and matrix quality (Turner et al. 2001)- and processes of gene flow and selection within the scope of microevolution, i.e. evolutionary processes within species (Holderegger & Wagner 2006; Storfer et al. 2007). Ideally, a spatial genetic pattern determined on the basis of genetic and spatial tools would correlate with landscape or environmental features that explain the observed spatial pattern (Manel et al. 2003).

So far, landscape genetics possess neither its own theory (Holderegger & Wagner 2006; Holderegger & Wagner 2008) nor its own methodological analytical framework (Holderegger & Wagner 2008). It differentiates from traditional population genetics by explicitly evaluating the effects of landscape composition, configuration and matrix quality, i.e. "the often-hostile space that separates the patches of a species' habitat in a given landscape", on gene flow and genetic variation within and among populations (Holderegger & Wagner 2008). Studies of these two disciplines differ also in relation to spatial and temporal scale (Holderegger et al. 2006). Population genetics studies are often carried out either at local scale, i.e. detailed studies usually involving intensive sampling within a narrow region and ignoring the landscape, or far beyond the extent of a single landscape. Usually the 'landscape' just refers to geographic distance. Population genetics investigations either referrer to very short periods or to extreme large, usually undetermined, time frames, i.e. several generations or evolutionary time frame. Time scale in landscape genetics is defined according to the organism perspective and the relevant landscape processes (Holderegger et al. 2006). While the relationship among groups or populations may be evaluated, in a landscape genetics approach the individual is the most appropriate operational unit of study due to two main reasons: it avoids potential bias in determining groups in advance; and it allows conducting studies at finer spatial scale (Manel et al. 2003).

Landscape genetics studies focus nowadays on four main research areas: barriers to gene flow; landscape variables that facilitate or hinder gene flow; effects of temporal scales; and species-specific hypothesis tests. A majority of these studies addresses questions related to the effect of barriers and connectivity on spatial genetic variation (Storfer et al. 2010). Most commonly applied analytical methods include a broad battery of spatial tests already largely applied in population genetics –e.g., Mantel test, isolation-by-distance test, linear regression, ordination, autocorrelation;

but also relatively new methods such as individual-based methods which estimate genetic connectivity on the basis of individual genotypes (Storfer et al. 2010).

Guillot et al. (2009) framed under the label of **Spatial Genetics** statistical methods applied to detect, quantify and test the spatial structure of genetic variation, i.e. "to analyze population genetic data in a spatially explicit framework". This area partially overlaps with landscape genetics. It focuses on modeling the spatial pattern with the aim of making "more accurate inferences by appropriately injecting *a priori* information and obtaining directly interpretable parameters". It uses information about genetic patterns "to gain deeper insights into the underlying ecological and evolutionary mechanisms", and it is mostly concerned with the study of neutral genetic variations (Guillot et al. 2009).

3.5. Basic Measures of Genetic Differentiation

Measures traditionally used to quantify and analyze the generation of genetic variation are intrinsically related to the concept of **population** (Epperson 2003). A population, i.e. **deme**, refers to a group of individuals who are likely to mate and reproduce (Dobzhansky 1970). In the field of population genetics the 'ideal' population is constant in size and follows **monoecious**, i.e. hermaphrodite, random mating (Epperson 2003).

Population substructure is practically universal among organisms. This is so because forces producing genetic variation within a population may lead through time to **genetic differentiation**, i.e. the acquisition of allele frequencies that differ among subpopulations of a previously homogeneous population (Hart & Clark 1997).

Traditionally, human genetic variation has been studied under the island-model framework and is still frequently implicitly assumed (Handley et al. 2007). The island model (Wright 1931) defines gene flow as the exchange of migrants among demes by a nonrandom fashion, in opposition to another widely applied model, the **stepping-stone** model, in which only neighboring groups exchange migrants (Handley et al. 2007). Both are actually special cases of the **infinite-lattice** model (Malécot 1950, reviewed in: Guillot et al. 2009), which assumes demes or individuals distributed on a lattice with homogenous demographic parameters of every lattice node, i.e. population size (or density) and dispersal rate (Guillot et al. 2009).

The phenomenon generating an increase of genetic differentiation of neutral loci with incremental geographical distance was first formally modeled by Sewall Wright (1943), who introduced it with the phrase **isolation by distance**. The theory of 'isolation by distance' was largely further developed by this and several other authors and these theoretical results were recurrently applied to compute different measures of genetic differentiation (Slatkin 1993). Most of the recently reported studies on this subject are based on the 'infinite lattice model' (Guillot et al. 2009).

For more than two decades Jorde (1985) classified most commonly applied measures of genetic distances into four major categories:

I. Chi-square distances: basically the square difference between gene frequencies in two populations is related to the standardization of this difference; a commonly applied standardization is the mean gene frequency of the total population, which is thought to represent the founding-population gene frequency. These measures are based on the assumption that differences in gene frequency between the two populations are not too large.

II. Angular transformation distances: an arcsine transformation is applied to gene frequencies in order to achieve independence between gene frequency variances and themselves. It assumes an evolution model based solely on drift process; it is less appropriate when widely divergent sample sizes are analyzed.

III. Gene substitution approaches: these include a set of genetic distances based on the number of codon substitution differences between two populations; it is based on the model of infinite alleles. In human studies some method assumptions may not be met and results must be regarded with caution.

IV. Information measures: these refer to indices of genetic diversity. Within- and between-groups diversity are computed and combined to estimate the overall genetic distance among groups.

When applied to the study of human populations, these major types of genetic distances produce highly correlated results (Jorde 1985).

Furthermore, the author pointed out a group of very useful display techniques (Jorde 1985):

I. Genetic maps: under this denomination is included the two-dimensional representation of distance matrices. The first proposed and widely used method to produce distance matrices is the principal coordinates analysis; another widely applied method is multidimensional scaling.
II. Evolutionary trees: these allow the illustration of evolutionary processes in form of a hierarchical structure.

3.6. Statistical Toolbox in Spatial Genetics

This section reviews spatially explicit genetic methods, in contrast to methods which include the geographic space or geographical references as a support for the visualization and interpretation of results.

A full battery of methods and tools are nowadays available for the spatial analysis of genetic data. Most commonly applied methods involve testing for isolation by distance, spatial autocorrelation, ordination, and spatial assignment of individuals (Storfer et al. 2010).

Populations usually exhibit spatial dependence of observation values. Testing for statistical dependence between geographic and genetic distances is usually carried out with Mantel test (Sokal & Rohlf 1995) or with the estimation of spatial autocorrelation (Guillot et al. 2009). Mantel test is based on the isolation-by-distance model. It involves the computation of geographic and genetic distance matrices for pairs of individuals or groups. These two matrices are compared on the bases of an empirical correlation coefficient; significant correlation is indicative of global spatial structure (Guillot et al. 2009). The empirical estimation of spatial autocorrelation is an exploratory tool and is usually performed computing variogram parameters and Moran's I index –a weighted correlation coefficient applied to test departures from complete spatial randomness (Moran 1950). Ordination methods, e.g., canonical correspondence analysis, principal component analysis or multidimensional scaling, can be applied as exploratory tools to identify continuous variables generating spatial arrangement of related individuals (Jongman et al. 1995). Such multivariate methods are used to order individuals or populations according to genetic characteristics with the aim of generating hypothesis of spatial continuous variables affecting gene flow. By using canonical correspondence analysis space can be explicitly incorporated as a covariate (Storfer et

al. 2007).

Spatial interpolation can be applied to predict values and the corresponding levels of uncertainty of continuous variables between observation sites. Although a quite large number of methods and tools are based on spatial interpolation, this methodology remains quite underutilized (Storfer et al. 2007). In the context of spatial genetics, it is particularly useful for analyzing spatial processes affecting gene flow and the genetic relationships of populations continuously distributed in space, i.e. populations that not have a clumped distribution, avoiding the necessity of *a priori* definition of groups and the difficulties related to the analysis of non-independent pair-wise data (Murphy et al. 2008).

A two-dimensional representation of the geographical distribution of genetic frequencies is designated with multiple terms, including 'geographic maps' (Cavalli-Sforza et al. 1994), 'landscape surface', 'genetic landscape', or 'genetic landscape surface' (Vandergast et al. 2011). Murphy et al. (2008) denominated **genetic surface** a representation of continuously, gradually changing spatial structures of genetic variation and **genetic surfacing** the process of generating genetic surfaces. These terms make direct reference neither to the interpolation method used to create them nor about potential further applications, aside from visual examination. A **synthetic map** represents a special case and designates the two-dimensional graphical representation of principal component values with geographical references (Manel et al. 2003). Synthetic maps have been widely used to represent spatial genetic variation in humans. For instance, Cavalli-Sforza et al. (1994) investigated in detail world-wide human genetic variation based on this method. While these graphical representations are extremely useful for visually identifying geographic patterns and for generating hypothesis about the origin of spatial diversity (Barbujani 2000), to model the geographical distribution of frequencies within the framework of a geographic information system (GIS) allows to implement deep geostatistical analysis to spatial frequency data, which generally includes a large spectrum of interpolation methods for data analysis (Vandergast et al. 2011).

A further very relevant set of techniques are related to the search of **genetic barriers**. This term, also denominated **genetic boundaries** (Barbujani 2000), refers to areas accounting for rapid change in allele frequencies. These usually implicate sharp genetic differences between populations

inhabiting the flanking areas and indicate any kind of obstacle constraining gene exchange. Local genetic structuring may be the result of barriers affecting gene flow (Barbujani 2000; Barbujani & Belle 2006). Two spatially explicit methods to infer genetic boundaries are **wombling** (Womble 1951; Barbujani et al. 1989) and **Monmonier's algorithm** (Monmonier 1973). The Wombling algorithm represents data in form of a continuous function defined over the entire study area –i.e. commonly represented by a two-dimensional spatial matrix or grid with (estimated or observed) values at the grid nodes- and is applied to detect areas where frequencies of biological measures (such as gene frequencies) change more rapidly than other areas (Barbujani et al. 1989). The method was originally proposed by Womble (1951) and implemented by Barbujani et al. (1989). Its application to spatial genetics is based on the idea that zones where allele frequencies show a high rate of change may correspond to areas of limited gene flow, i.e. genetic boundaries; this implementation searches for any type of gene-flow barrier with a geographical component. Wombling has been used to detect barriers between major populations at continental and national level (Barbujani et al. 1989; Barbujani et al. 1990). Monmonier's algorithm (Monmonier 1973) tries to detect pairs of neighboring predefined groups that account for relatively large genetic differentiation (Guillot et al. 2009). In recent years individual-based methods were proposed to detect barriers. For instance, GENELAND (Guillot et al. 2005b) applies a Bayesian spatial assignment of individuals combined with Voronoi tesselation. Further, Manel et al. (2007) proposed a moving-window technique to identify genetic spatial discontinuities which also applies an individual-centered approach.

3.7. Spatial Genetic Diversity in Humans

Global human genetic diversity is extremely low. Humans account for the lowest species diversity among primates (Kaessmann et al. 2001). Because in evolutionary terms our own species is relatively young, human genetic diversity is reduced even compared to that of our closest relatives, the great apes (Gagneux et al. 1999). For instance, genetic differences between major socio-demographic groups are less pronounced than those between chimp populations, one of the primates most genetically similar to humans (Fischer et al. 2006; Hurles & Jobling 2001). Though such low global differentiation, human-genome diversity presents a complex geographical structure

(Alves et al. 2012). Geographic distances account for three quarters of the genetic variance between populations worldwide (Ramachandran et al. 2005). This strongly indicates that phenomena occurring in geographic space, i.e. demographic expansions, have primarily shaped geographical patterns of genetic variation (Barbujani & Colonna 2010). While global patterns of genetic variation are greatly influenced by geography, at regional scales culture and language also play an important role (Handley et al. 2007).

At global scales the proportion of neutral alleles shared among populations decays smoothly with increasing geographical distance (Manica et al. 2005). Such a structure of continuous, gradual changes is designated **clinal pattern** (Handley et al. 2007). Clinal patterns of genetic variation presumably result from migratory contact between populations through most of their history (Barbujani & Belle 2006). At these scales, broad clinal patterns are interrupted by genetic barriers (Handley et al. 2007; Ramachandran et al. 2005). A reduction of gene flow, e.g., as a consequence of geographical and/or cultural barriers, may most probably lead to a certain degree of isolation between those groups (Barbujani & Belle 2006).

On average, the largest proportion of human genetic diversity (ca. 85 percent) is represented by differences within populations (Barbujani & Belle 2006). Differences between populations are extreme subtle. Precise, absolute boundaries, i.e. clear-cut genetic differences among groups, cannot be delineated (Barbujani & Belle 2006). Nevertheless, if the number of genetic markers is large enough genetic differences may be observed between any pair of populations or groups thereof (Bamshad et al. 2003). This holds true even for groups separated by very few kilometers (Manni et al. 2004; Rosser et al. 2000; Wooding et al. 2004).

Complex patterns may be found as well in narrow regions. In areas recently populated by groups from different geographical origins, which did not mix much, specific genetic characteristics of the ancestral populations may be observed in a single location (Shriver et al. 1997; Wooding et al. 2004).

3.8. The Choice of the Molecular Marker

Since the discovery of the first molecular markers of genetic variation among human populations at the beginning of the 19th century, i.e. ABO blood groups, plenty of types of markers have been

discovered and validated (Cavalli-Sforza & Feldman 2003). Molecular properties and information provided by each type of marker vary considerably (Jombart et al. 2009). When the spatial pattern under study is assumed to be related to gene-flow the analysis must be performed with neutral markers (Jombart et al. 2009).

In studies of human populations' diversity single nucleotide polymorphisms (SNP) and short tandem repeats (STR) –also denominated microsatellites- (Schlötterer 2004) are among the most commonly used these days (Barbujani & Colonna 2010); nowadays enormous amounts of genetic data of selected populations are freely available (http://www.ncbi.nlm.nih.gov/; http://www.yhrd.org/).

Studies of present and past population genetic diversity and assessment of past demographic processes rely on the estimation of allele distribution. The substantial difference in mutation rates and genomic abundance between SNPs (lower rate; higher abundance) and STRs (higher rate; lower abundance) frames the temporary scale of study and affects the number of loci required to quantify genetic heterogeneity among and within populations. Due to their low mutation rates SNPs are useful to trace demographic events in evolutionary scales. Nevertheless, because they are not very informative, a large number of loci are required for studies of genetic diversity; this is especially important in studies of human populations due to the low amount of within-species variability. STRs are highly informative and a smaller amount of loci are required to perform accurate comparison between groups and even between individuals (Schlötterer 2004). Plenty of studies demonstrated their value for tracing human demographic processes; conclusions were verified by findings of various related disciplines such as anthropology, archaeology, demography, or linguistics (Cavalli-Sforza & Feldman 2003). So far STRs are the most widely used type of marker in studies of spatial genetic diversity in any animal taxonomic group (Storfer et al. 2007).

Mechanisms of hereditary transmission differing among the components of the genomic material affect the type of information provided by molecular markers. Haploid loci are inherited uni-parentally as a block, without recombination, i.e. most of the Y chromosome is passed unchanged from fathers to sons (Jobling & Tyler-Smith 2003). These types of markers are very informative on gene flow and are useful to detect and to trace paternal lineages. Population frequency of haploid loci is not affected by recombination and these types of loci are better suited to infer more recent

demographic and evolutionary events than autosomal loci (Barbujani & Colonna 2010). Specifically, Y-chromosomal markers supply substantially finer details about population differentiation than the recombining counterparts of our genome (Underhill & Kivisild 2007). Because the genetic information of Y chromosomes is only modified by mutation, rather than suffering reshuffling due to recombination as the autosomal genome does, it preserves a simpler register of human history (Jobling & Tyler-Smith 2003). These exceptional features most probably accredit the Y chromosome as a genetic tool with the highest sensitiveness of the human nuclear genome for detecting admixture (Hurles & Jobling 2001). By convention, classification of Y chromosomes using binary SNPs refers to **haplogroups** or **clades**; Y chromosomes which are differentiated on the basis of STRs are denominated **haplotypes**; finally, the term **lineage** is used to designate Y-chromosome classifications including both types of markers (de Knijff 2000).

The identification and validation of several Y-chromosomal STR loci (Y-STR) provided new markers with higher resolution of paternal lineage differentiation (Jobling & Tyler-Smith 2003). Due to lack of recombination in Y chromosome, genealogical information can be estimated at maximum molecular resolution as a function only of STR sequence length, a quantity, which depending on the STR marker, may be highly variable (Underhill & Kivisild 2007). Additionally, more reliable measures of genetic diversity within and among groups may be obtained with Y-chromosomal STRs than in case of Y-chromosomal SNPs, because these latter are more prone to marker-ascertainment bias, i.e. systematic distortion in the data generated by the way markers are chosen (Jobling & Tyler-Smith 2003).

A further special feature of the Y chromosome is the high geographical specificity of its variants determined by two main factors: drift and patrilocality. Allele frequencies of the Y chromosome are more susceptible to be affected by drift, i.e. the random sampling of Y chromosomes from one generation to the next one, than allele frequencies of the other nuclear chromosomes. This higher susceptibility to drift accelerates the differentiation of groups of Y chromosomes between geographical regions. Patrilocality, i.e. the case in which it is the woman the one that moves when heterosexual individuals marry and these do not belong to the same place, is the most general practice in most societies worldwide. Men, carriers of the Y chromosome, live generally closer

than women to their birthplace. This behavior further enhances local differentiation of the Y chromosome. Because of these particular features, the Y chromosome is characterized by a rapid evolutionary change and geographically structure drift (reviewed in: Jobling & Tyler-Smith 2003). All in all, the non-recombinant part of the Y chromosome may be considered as a single extremely polymorphic genetic locus. Since evolutionary forces acting on this segment of the Y chromosome, such as drift and mutations, affect it as a unit, it evolves along lineages (Jobling & Tyler-Smith 2000). The present distribution of haplotype frequencies is the result of plenty of past events. The use of Y-chromosomal genetic data can provide unique insights of past events relevant to anthropologists, paleontologists, historians, and linguists. Nevertheless, by interpreting past events simplistic interpretations of Y-STR information content must be avoided, e.g., equating a lineage with a human socio-demographic group or a migration event. The combined effect of recent events, such as modern intercontinental travel or migrations during the last centuries, profoundly affects present frequency distribution. It is therefore of great importance to evaluate the effect of more recent demographic events before considering potential explanations regarding ancient ones (reviewed in: Jobling & Tyler-Smith 2003).

4 The Argentine Republic

4.1. The Country[2]

Argentina, located at the southern region of South America, is the world's eighth largest country. Its official name is **República Argentina** (Argentine Republic) and it constitutes a federal republic with two legislative houses: Senate and Chamber of Deputies. Argentina's direct neighbors are Bolivia and Paraguay to the north, Brazil and Uruguay to the east, and Chile to the south and west (Figure II-5). Its Atlantic coastline stretches some 2,900 miles (4,700 km) to the east.

The capital city is Buenos Aires. It is located on the western shore of the La Plata River (Río de La Plata). Buenos Aires, together with its conurban ring **Greater Buenos Aires** (Gran Buenos Aires), achieves a population of around 13 million inhabitants, almost a third of the total population (INDEC 2010). Argentina's population is mostly urban (92.2 percent). It is mainly concentrated in the central and northern territory. Around 96 percent of the total population lives in this region (INDEC 2010). 62 percent of the total population resides in the triangle formed by the territory embraced by the three provinces Buenos Aires (including Buenos Aires city), Santa Fe, and Cordoba (INDEC 2010). Argentina's economy is dependent on services and manufacturing, although production of cereals and livestock together with mining and tourism are important income sources. Major cities are La Plata, Mar del Plata, and Bahía Blanca on the Atlantic coast, Rosario in the littoral, and Córdoba, San Miguel de Tucumán, and Neuquén in the interior.

Argentina's most pregnant natural landscapes include vast plains, large mountain chains, various forest, deserts, and tundra ecotypes, as well as rivers, lakes, and thousands of miles of ocean shoreline.

2 reviewed in: Encyclopædia Britannica (2012a).

4.2. Geography[3]

Argentina's territory is shaped like an inverted triangle. It covers some 1,073,00 square miles (2,780,000 km^2), 880 miles (1,420 km) from east to west at its widest length and stretching 2,360 miles (3,800 km) from the subtropical north at the border with Brazil, Paraguay and Bolivia to the subantarctic south, neighboring Chile and the Atlantic Ocean. Drainage follows a northwestern-southeastern direction. Major exceptions are, among others, various large rivers of the Paraguay–Paraná–La Plata River system Basin (Cuenca del Plata), including Iguazú, Uruguay, and Paraná River. Other main Argentine rivers are La Plata River, holding major Argentine port, and Colorado River, which defines the boundary between the central plains and Patagonia.

Argentina's natural landscapes may be divided into five distinct zones: the Andean region, the Gran Chaco, Mesopotamia, the Pampas, and Patagonia.

The most characteristic feature of the broad **Andean region**, the Andes, is a system of north-south-trending mountains, which extends along the total western boundary. The northern region includes elevations from 16,000 to 22,000 feet (4,900 to 6,700 meters), high plateaus (*punas*) and basins (10,000 to 13,400 feet; 3,000 to 4,080 meters). In the Mendoza Province lies the Aconcagua, South America's highest mountain (22,831 feet; 6,959 meters). Distinct features of the southern Andean region, the Patagonian Andes, are series of basins –called Lake District- and several glacier formations, being Perito Moreno the most well-known.

Annual average temperatures range from more than 36 °F (20 °C) in some parts of the Andean northwest, where continental climatic conditions occasionally occur, to below freezing towards the south and at higher elevations. Tundra climate (average annual temperatures below 50° F; 10°C) occurs above 11,500 feet (3,500 meters) in the north and at sea level in southern Tierra del Fuego. In the south the vegetation is austral. Mid-latitude rain forests are abundant. In the north there is poor and desert-like vegetation, although at higher altitudes steppe vegetation is found.

3 reviewed in: Encyclopædia Britannica (2012a).

Figure II-5 Topographic map of the Argentine Republic.

Provinces (federal states) are delimited with a black line. Note that this map shows only large urban centers (black dot). Capital cities with relatively low number of inhabitants are not represented in this map. The location of the capital cities used as geographical reference of point data (see Figure III-5; Table III-1) was added to the original figure (red dot). Source: Encyclopædia Britannica Online Academic Edition [Internet]. [place unknown]: Encyclopædia Britannica ©2013. [Figure], Patagonia [cited 2013 May 7]. Available from: http://www.britannica.com/EBchecked/topic/446174/Patagonia

Argentina's northern central region, denominated The **Gran Chaco** or **Chaco**, comprises dry lowlands between the Andes and the Paraná River. It contains many wide rivers of shallow nature, which do not permit regular navigation. Most of them cause summer floods and dry up in winter. Only the Pilcomayo, Bermejo, and Salado –three of the numerous rivers watering this region- flow from the Andes to Paraguay–Paraná–La Plata River Basin system without evaporating en route and forming salt pans (*salinas*).

The Gran Chaco is principally a sub-tropical zone, with severe climatic conditions. It has continental climate, with hot wet summers and dry extreme cold winters. While rainfall decreases to the west, reaching semi-arid conditions, temperatures decrease from north to south.

The Gran Chaco vegetation is highly varied and exceedingly complex. In the western region it is dominated by thorn forests. Vegetation is increasingly abundant towards the east; the eastern Chaco presents park-like landscapes of tall, herbaceous savannas alternated with clustered trees and shrubs.

The **Mesopotamia** is situated in eastern Argentina, between the Paraná, Iguazú, and Uruguay rivers. It is a narrow depression of 60 to 180 miles (100 to 300 km) wide. It is bounded on the west by the Gran Chaco, on the north by Paraguay, on the northeast by Brazil, and on the southeast by Uruguay. The highest region of Mesopotamia is located at the north, bordered by the highlands of southern Brazil, by Paraguay and Chaco to the west, and Brazil and Uruguay to the east. It stretches for 1,000 miles (1,600 km) southward, where merges with the Pampas south of the La Plata River. The north has a sub-tropical weather, with long and humid summers and mostly mild winters. Towards the south the climate turns milder, four seasons can be delimited. Occasional cold fronts from Patagonia, especially in July, bring frosts. Rainfall decreases southwards, precipitation ranges with precipitation average up to 2,300 mm in the northern areas and 1,300 mm in the southern part (Servicio Meteorológico Argentino; http://www.smn.gov.ar/serviciosclimaticos).

Subtropical evergreen rain forest occurs in the northeast. Tall wax palms grow in the flood zones. Groups of trees patching grassy areas form park-like landscapes. Along the rivers grow gallery forests, which become denser and taller towards the north.

Argentina's rich grasslands, called **The Pampas**, are centrally located. They extend between the

Atlantic shore and Mesopotamia, on the east, and the Andean Region, on the west. These endless stretching plains are subdivided into a more humid eastern and an arid western region, denominated, respectively, Humid Pampa and Dry Pampa.

The Humid Pampa has hot, humid summers and cool, mild winters. Average temperatures are about 72–75 °F (22–24 °C) in summer and about 46–55 °F (8–13 °C) in winter. The Dry Pampa presents more clearly differentiated temperatures between the four seasons. Average rainfall varies from 39 inches (990 mm) in the east to 20 inches (500 mm) in areas near the Andean region (Servicio Meteorológico Argentino; http://www.smn.gov.ar/serviciosclimaticos). In winter cold fronts that move northward from Patagonia bring occasional frosts to the Pampas.

Natural vegetation changes from knee-high grasses in the most humid areas to *monte* forest, where precipitation decreases. Much of the original flora has been replaced by agricultural crops and forestal trees.

The southernmost portion of Argentina, **Patagonia**, is a cold, windy plateau. It extends some 1,200 miles (1,900 km) south of the Colorado River to the tip of South America. Patagonian plateau descends sequentially east of the Andes in form of broad, flat steps. Coastal terraces extend along the Atlantic coastline, with higher cliffs towards the south (of more than 150 feet; 45 meters).

Rainfall decreases westward. Semiarid, or steppe, conditions in the Atlantic region rim the arid (desert) core of Patagonia. In both zones evaporation exceeds precipitation and the Patagonian plateau remains treeless. Strong winds carry abrasive sand and dust, which markedly reduce visibility.

Southern of the Colorado River shrub vegetation occurs, which further south gives way to low scrub vegetation alternated with green grass steppe.

4.3. History

Strong admixture characterizes nowadays Argentine population (Alfaro et al. 2005; Avena et al. 2001; Corach et al. 2010; Diaz Lacava et al. 2011a; Marino et al. 2007; Toscanini et al. 2007). This population does not resemble the land where admixture between Spaniards and natives took place in the sixteenth century (Levene 2002). Three major patterns of admixture processes may be

differentiated according to the geographical origin and dimension of immigration: admixture during the colonial period, mostly between Spaniards, native people, and Africans; admixture as result of mass immigration, mainly from Italy and Spain; and admixture between modern Argentine inhabitants and migrants from neighboring countries (Levene 2002; Romero 2011).

4.3.1. The Colonial Period

At the time of the Spanish arrival in the XVI century, followed by colonial settlement, various ethnic groups populated this territory. The most developed tribes inhabited the northwest, the Diaguita tribe, and the northeast, the Guaraní tribe. Diaguitas settled in small hamlets with dwellings made of stone. They practiced irrigated agriculture in terraces, grew llamas and vicuñas, and were capable of producing tools, clothes, and ornaments. The Guaraní tribe inhabited the Mesopotamia and was semi-agricultural. Both, Diaguitas and Guaraníes, were forced into labor by the conquerors (Levene 2002).

Most of the Argentine territory was inhabited by hunters and gathers. Tribes populating the flatland between La Plata River and the Andes received the denomination of Pampas –pampa is a native word and means flat land without trees (Levene 2002). This denomination included, among others, the Querandí tribe, populating the region of Buenos Aires, and the Araucarians, who traveled over the Andes. Matacos and Guaycurúes inhabited the Chaco forest; these tribes combined hunting and gathering with rudimentary agriculture. Tehuelches inhabited Patagonia; Tierra del Fuego was inhabited by Onas and Yaganes. All in all, the vast Argentine lands were low dense populated, with considerable less native population than other South-American territories (Levene 1992; Levene 2002; Romero 2011). For instance, while the native Mexican population at the time of Spanish conquest has been estimated as 25 million, only 300,000 to 750,000 natives may have lived at that time in the territory of future Argentina; roughly two-third of the native population comprised maize-based cultures concentrated in the northwest, in the rectangle embracing the modern provinces of Jujuy, Salta, La Rioja, Catamarca, Cordoba, and Santiago del Estero (Rock 1987).

In the sixteenth century a quite small number of Spaniards founded twenty-five cities, among which fifteen survived. Settlement pattern of the sixteenth and seventeenth centuries mirrored the geography of pre-conquest native culture. Except for Buenos Aires and Santa Fe, founded to secure

navigable routes to the La Plata River estuary, settlements were founded in agriculturally fertile areas inhabited by sedentary, agrarian tribes. These were forced to provide tribute and labor by a relatively small number of Spaniards set themselves as overlords of the native people (Rock 1987). Settlement founders came from neighboring Spanish colonies –Asunción of Paraguay, Peru, and Chile. These attempted to capture and to exploit natives as well as to establish direct routes to Europe over the main rivers flowing to the Atlantic (Romero 2011). Through Brazil arrived a large number of Jewish Portuguese. These were in 1622 a quarter of Buenos Aires population and they were also very populous in other inland cities (Levene 2002). Until the creation of the Viceroyalty of the La Plata River in 1776, Argentina was divided in *gobernaciones* and was under the political jurisdiction of the Spanish authorities in Peru and Chile. During this period only Spanish ships were allowed to approach its coasts. Overseas migration included at most Spaniards, Africans, and a few possible smugglers (Romero 2011). The largest number of incoming Africans was sent to the modern territory of Bolivia and northern regions (Rock 1987). Since almost no Spanish women arrived to this territory in this period –although Spain encouraged the arrival of Spanish married couples-, patrician families were usually created by the union of Spaniard men and chief-tribe daughters (Levene 1951). Beyond these special cases, the offspring of a Spaniard and a native was called mestizo; mestizos just incremented the number of subdued labor (Levene 2002).

In the sixteenth and seventeenth centuries the largest colonial settlements laid along an arc toward the northeast of about thousand miles in length, connecting Buenos Aires on the La Plata River estuary and the silver mining city of Potosí in Upper Peru. The largest settlements were Santa Fe, Cordoba, Santiago del Estero, San Miguel de Tucumán, Salta, and Jujuy. The arc had two branches, one to the west, connecting La Rioja and Catamarca, and another to the north, connecting Asunción of Paraguay and Corrientes with Buenos Aires. East of the Andes was Mendoza, San Juan and San Luis. The rest was practically unsettled and some areas remained practically unexplored until the twentieth century (Rock 1987). Villages were practically isolated (Levene 2002). Main population elements of these settlements were Indians, mestizos, slaved Africans, who have been smuggled into the country through Buenos Aires, their slave descendants, and a minority of *Creoles*, i.e. Argentine-born individuals of European origin (Levene 2002; Romero 2011). Throughout the

territory lived in 1570 less than 2,000 Spaniards and 4,000 mestizos; the largest settlement of that time, Cordoba, had only 250 Spaniards (Rock 1987). At the beginning of the seventeenth century, in the *Gobernación* of Tucuman, a territory that double the size of modern Italy, there were no more than 700 Spaniards and twenty-four thousand natives (Levene 2002). The region where Buenos Aires was founded counted originally with scarce native population. Spaniards farmed with help of relatively few African slaves. At the beginning of the seventeenth century Buenos Aires had a population of around 1,000 Spaniards and a horde of African slaves (Romero 2011).

In 1776 Spain created the Viceroyalty of the La Plata River, including the territories of modern Argentina, Paraguay, southern Bolivia and Uruguay. Buenos Aires was its capital; this city reached at that time a population of twenty thousand inhabitants. Cordoba, the largest settlement of the viceroyalty, had around a million of inhabitants (Romero 2011).

During the first two centuries the socioeconomic growth of the grassland regions was based on livestock production, supported by overwhelming abundance of livestock, high price of leather and Spanish restrictions to agriculture production. In the eighteenth century this situation changed. Since 1791 Spain allowed Spanish ships to trade through Buenos Aires port, facilitating trade of cattle products and triggering enormous impulse to ranching activities. The Creole population had grown considerably. Buenos Aires, for instance, had already a population of twenty-two thousand inhabitants. Livestock was not any more overabundant. Nomadic natives did not find easily cattle outside the farms and began to raid farms' herds and women. Marginal areas began to be unsafe. Many Creoles neither got own farmland nor could earn their living as pawns. Livestock production demanded very low number of labor forces (usually performed by Africans slaves and their slave descendent). In those times appeared the typical figure of the open flatland, the *gaucho*. This term refers to horseback men, who lived outside the settlements, mostly homeless, malnourished and poorly fed (reviewed in: Levene 2002).

The Spanish approval to transatlantic trade through Buenos Aires happened at the time of the Britain industrial revolution. This resulted in a markedly increase in trade with Britain accompanied by British settlement, who soon assimilated the Creole culture. Nevertheless, population composition and culture remained substantially unaltered until 1880 (Romero 2011).

In 1810 an open *cabildo* –municipal council- established Creole administration of the Viceroyalty of the La Plata River autonomous from Spain. The surface of the viceroyalty was equivalent to half of the European continent and had a total population of only one million people. In that time Buenos Aires had a population of forty five thousand inhabitants; one third of them were slaved Africans and slaved *mulatos*, i.e. person of African and European mixed origin. Regional economies had already developed important socioeconomic and ethnic differences, due to differences in natural conditions and unequal distribution of natives, Africans, and Spaniards. Consequently, marked conflict of interest among the regions characterized the beginnings of the Argentine society, which nevertheless remained relatively united towards the purpose of establishing and structuring the independent country (Levene 2002). The new local authorities introduced forthwith several progressive measures. Among others, titles of nobility were removed, freedom was grant to those who were born of slave parents, inquisition was abolished, and its torture tools were publicly burned. Few years later the introduction of African slaves was prohibited and official promotion of immigration was regulated. In 1813 it was enacted, that any slave became free just by setting foot in this territory; as well, all natives were relieved of any further personal services towards Spaniards and church members (Levene 2002).
In 1816 the independence of Spain was declared and the new country received the name of United Provinces of the La Plata River. The modern territories of Paraguay, Bolivia and Uruguay soon separated from the new country; the territory of modern Argentina was first united in 1860 after years of internal wars (Levene 2002; Rock 1987; Romero 2011).

At the time of separation from Spain, in 1810, native tribes had established themselves no more than one hundred kilometers from Buenos Aires and lived relatively peacefully (Levene 2002). Since the abolition of the Spanish monopoly exports from cattle production provided enormous benefits. Ranching and cattle manufacturing expanded rapidly and required more land and labor (Levene 2002; Romero 2011). In the eighteenth thirties the rural population, predominantly composed of Creoles, mestizos, and natives, was forced to hard labor by few ranchers with extensive political power (Romero 2011). All this expansion did not go along with a population of

the vast flatland, since ranching demanded relatively little manpower (Levene 2002).
In 1833 a large military expedition succeeded in pushing back, devastating, or subduing the flatland native population up to the Negro River, which until then kept attacking frontier ranchers and appropriating cattle. The ranching area was enormously enlarged to the south and distributed among few hands, victors, friends, and regime supporters. Gauchos and *arrieros* (herder), solitary men of mixed origin, were the main labor force of the vast ranches, pursuing for wild horses and Creole cattle (Romero 2011). In the next two decades, until 1852, a federation led by Buenos Aires province ruled the country and restrained provinces' development. Buenos Aires province kept growing thanks to leather and *tasajo* production (beef dried with salt) and the revenues of its port, opened to any type of European product. Economy of the grass lands remained based on open-land ranching of lean livestock destined to the *saladeros*, i.e. manufacturing plants of *tasajo* (Romero 2011).

In 1853 a Constitution went into effect. After some modifications, this remains in essence the basis of the current legal system in Argentina. The new constitution adopted a political system based on the division of power, significantly reducing catholic-church influence in political affairs, and it incorporated several progressive measures of civil rights previously enacted by Creole authorities, e.g., it abolished slavery and nobility titles. A Constitutional law explicitly promoted European immigration. The law of freedom of worship was included, a step also already undertaken by some provinces, aiming to encourage and to support immigration. Argentina counted only one million inhabitants (reviewed in: Levene 2002; Rock 1987; Romero 2011).

In the eighteen fifties immigration and agriculture settlement was promoted along with improved sheep production (Romero 2011). Railways construction began –the first line connecting the littoral, Cordoba, and Chile- and followed in the next decades. Migrants' towns arose along these railways (Levene 2002). Crop production started to grow and littoral provinces, with very rich soils and adequate climatic conditions, slowly began to develop. Nevertheless, provinces lived isolated from each other, flatland natives had resettled themselves two hundred kilometers from Buenos Aires city, and marginal areas, especially in the Buenos Aires province, were still assaulted by

nomadic tribes (Levene 2002; Romero 2011).

4.3.2. Mass Migration Period

In the eighteen sixties 76,000 migrants settled in Argentina and in the next decade 85,000. There was rapid but extremely unequally distributed growth. Most migrants settled in the littoral and in the largest cities (Romero 2011). The first census (1869) showed a population of 1,830,000 inhabitants; of them 90,000 were natives and 200,000 were migrants; 80 percent were illiterate and 80 percent lived in small mud huts (Levene 2002). Buenos Aires grew from 150,000 inhabitants in 1865 to 230,000 ten years later (Romero 2011). Sheep production was intensified. Railways construction throughout the country employed large number of immigrants and their descendants (Romero 2011).

In 1876 a new immigration law was passed; this new law promoted immigrant settlement but did not guarantee land tenure. As a consequence, migrants were recruited from regions of low living standards –especially from Spain and Italy- and of low technical level; these performed as rural or industrial labors. Some of these migrants moved seasonally between Argentina and either the land of origin or other countries with high demand for labor, e.g., Brazil or United States. Those who settled chose the littoral regions or the largest urban centers (Devoto 2009; Romero 2011).

At the end of the 19th century native population was still dominating northern Patagonia, southern, and western flat lands. In 1879 the largest expedition against these groups began. This military campaign is known as the *conquista del desierto* –conquest of the wilderness. As its result, 32,800 square miles (85,000 km^2), a land comparable in size to the Austrian territory, passed into the hands of 381 people. The surviving natives were pushed aside, constrained to reservations (Rock 1987).

After years of disputes among provinces about customs revenue of Buenos Aires' port, a law passed in 1880 established Buenos Aires as the capital city of the Republic Federal State of Argentina. Buenos Aires city became itself a federal state. At this time Argentina was still divided in three regions according to economic and political interests: Buenos Aires region, the littoral provinces, and the inland provinces (Romero 2011). While sheep production had expanded sharply and brought important revenues, crop production covered just self-supply needs. Urban centers had

grown thanks to immigration (Romero 2011). The 2,500 km of railways completed in 1880 grew to 10,000 km in 1890, centralized in Buenos Aires port (Levene 2002). Buenos Aires got the largest profit of this economic development (Romero 2011).
In the eighteenth-eighties laws were passed guaranteeing secular, free and mandatory education, and civil marriage; these measures separated considerably creed from civic live and were in accordance with the requirements of a pluralistic society of admixed origins and creeds (Levene 2002; Romero 2011).

Mass migratory movements kept affecting Argentina until 1914. Since no concrete governmental migratory settlement plan was carried out, migrants settled according to their preferences. Not all Argentina was affected similarly. The northwestern region was more related to the actual Bolivia and it almost did not participate from the mass migratory movements. Seventy percent of migrants settled in the fertile littoral and in harbor cities. There was an increase in socio-economic differences among regions (Levene 2002; Romero 2011). In 1895 a quarter of the population was foreigner. Almost one third (30 percent) of all migrants lived in Buenos Aires city; in total, 80 percent were concentrated either in this city or in the provinces of Buenos Aires or Santa Fe (Devoto 2009). Until 1914 the total population grew from about four million to eight million inhabitants. In 1914 thirty percent were foreign-born and 35 percent were illiterate (Levene 2002; Romero 2011). Of the 30 percent foreign-born that resided in Argentina in 1914, 51 percent were in Buenos Aires city, 43 percent were in Santa Fe and Corrientes provinces, and only 2 percent in the northwestern provinces of Catamarca and La Rioja. Migrants of this period were predominantly men with an age between 15 and 35 years (Devoto 2009).
Since the end of the nineteenth century there were also important migratory movements of rural Creole population, who moved to the littoral seeking work and better wages. Cattle production grew in importance thanks to the installation of freezing establishments, which allowed adding frozen beef to livestock exports. Sheep production was shifted to southern areas, to the new available territories of La Pampa and Rio Negro provinces. Both were carried out mainly by large estates. Agriculture production had grown considerably as well. Its export volumes took similar dimensions of those of cattle exports. Agriculture was substantially performed in littoral regions

by tenant smallholders (Levene 2002; Romero 2011).

At the end of the nineteenth century factory workers were subject to too many labor hours and were poorly rewarded. This situation triggered large social conflicts; in the next decades workers' movements were repeatedly brutally repressed (Romero 2011).

In 1914 railways were already 30,000 km long and functioned as export routes of wood, crops, and beef from the Buenos Aires port; railways as well as most freezing plants belonged to foreign companies (Levene 2002).

4.3.3. Increasing Admixture

Immigration stopped during the First World War and restarted afterward (Devoto 2009; Levene 2002). After 1920 immigration origin changed. Since its beginnings in mid-nineteenth century 80 percent came from Italy or Spain. The new movements incorporated a large proportion of central and eastern Europeans: Germans, Poles, Czechoslovakians, etc. (Levene 2002). Education allowed migrant descendants to grow socially. Rural population kept reducing. While in 1914 there was 42 percent of rural population, in 1930 there was only 30 percent. Cattle and crop production kept growing in importance and agriculture was diversified; sheep production decreased in volume and was further displaced to the southwestern steppes, to the regions covered from western Buenos Aires to Patagonia (Romero 2011).

The international crisis of 1929 affected principally cattle producers. Conservatory sectors overtook the government and political opposition was persecuted (Romero 2011). Immigration was restrained as well (Devoto 2009; Romero 2011).

Industrial production grew during the period between 1935 and 1941 and a new worker sector gathered in the urban areas (Romero 2011). At the end of nineteenth-forties only one third of all farmers were owners (Levene 2002). Over three million people, one fifth of the total population, have moved from their birth place in Argentina’s interior regions to urban centers seeking for labor sources. Of these, half moved to Buenos Aires and almost one third to the littoral areas (Romero 2011).

In the period between 1944 and 1955 Argentina underwent new social transformations: extensive social legislation was passed, protecting industrial and rural labors; as well, labor unions were

created and organized at governmental level (Levene 2002; Romero 2011).
Overseas migration, which had almost stopped during the Second World War, arrived intensely to Argentina in the period between the years 1947 and 1951; it practically finished in 1960. In 1947 overseas migrants were only 13 percent of the total population. 90 percent of these foreign-born resided in the littoral provinces and in Buenos Aires city. This metropolis had 35 percent of all overseas migrants; these were 26 percent of Buenos Aires population. In the nineteenth fifties overseas migration declined and in the final phase family reunification was the main migratory force (Devoto 2009).

In 1960 only 10 percent of the population was born in Europe. Almost half were Italians, followed by a third of Spaniards, and 5 percent of Poles. In the region of Buenos Aires city and Buenos Aires province, which together accounted for 37 percent of the total population, resided 64 percent of these migrants; this region together with the provinces of Santa Fe and Cordoba added 90 percent of all Europeans, and 66 percent of the total population (Devoto 2009).
Overseas migrants were historically predominantly male, but men-women ratios showed a declining tendency. At the beginnings of the nineteenth century there were 610 men for 100 women. In the period between 1881 and 1914 there were 300 men for 100 women, in 1895 there were 177, in 1947 there were 142, and in 1960 it had decreased to 110 men to 100 women. In those two hundred years of overseas migration persisted a tendency to marry someone of the same origin. This behavior was more accentuated in women than in men and in people who arrived as adults. Migrants' children were more open for partners of other origins (Devoto 2009).
Migration from neighboring countries oscillated historically between two and three percent. There was though variation in the geographical distribution of this migration. Migrants were initially more attracted to areas close to their original countries; in the period between 1930 to 1970 migrants distributed mostly along the border zones, where they worked as rural labors. Since the 1960s, a time when neighboring migrants were 50 percent of the total immigration, these migrants concentrated principally in Buenos Aires, where they took positions in the construction sector – mainly as industry workers- or as domestic service. In this last period the largest numbers came from Paraguay, followed by Chile and Bolivia (Benecia 2009).

Several military coups followed since 1955. Unions' activities were recurrently banned. Nevertheless Argentina's social structure had already changed. Factory workers had achieved the sense of union and political experience (Romero 2011). Armed underground groups took action since 1969. Violence and the economic crisis started decades ago and persisted long time after. Democracy was restored in 1983 but Argentina required many decades to start to recover from the extreme social, industrial and economic losses resulted from the military periods (Romero 2011). Migration from neighboring countries still arrived in similar proportions to previous decades (INDEC 1997; INDEC 2001). Demographic movements of domestic origin and from neighboring countries towards main metropolitan centers, primarily to Buenos Aires, Santa Fe, Cordoba, and the triangle among these cities (Rock 1987), may have reintroduced a considerable genetic Amerindian component in the littoral and central urban Argentina (Avena et al. 2001).

Nowadays Argentine population counts with a low proportion of clearly defined native population. Self-identifying Amerindian groups (thirty two in total) or individuals identifying themselves as Amerindians or descendants contribute only to one and half percent of the total population (INDEC 2004-2005; INDEC 2010). As groups, these are practically restricted to scanty, marginal areas (Bartolome 1976; INDEC 2004-2005; Sanchez-Albornoz 1994). The largest groups are the Mapuches, who reside in the southern-central Andean region, the Kollas, located in the northwest, and the Tobas, in the central-northern area (INDEC 2004-2005). The Kollas, basically agriculturalists and herders originals from the Bolivian high plateaus, still maintain substantial heritage of their pre-Columbian culture (Encyclopædia Britannica 2012b). The Mesopotamia, including the northeastern provinces of Misiones, Corrientes, and Entre Rios, presents a special case. While the local Amerindian language Guarani survived, the Guaranies did not. Rural population, predominantly small holders, as well as urban population consists overwhelmingly of European descendants. The left Amerindian communities are very small scattered, isolated groups, who live in precarious and impoverished conditions. Their ancestors migrated to these areas for some generations, presumably from Paraguay and Brazil (Bartolome 1976).

Part III - Materials and Methods

5 Geostatistical Analysis of an Admixed Human Population

5.1. The Basic Scenario[4]

The core concept behind the geostatistical analysis of the spatial structure of an admixed human population presented in this work is the evaluation of the spatial distribution of the most frequent groups of genetically similar individuals within an area. A scenario of a continuously populated, admixed region is presumed. In such a region a complex spatial pattern of genetic composition is expected. This region includes individuals of several genetic backgrounds (Figure III-1 a). A main assumption of this scenario is that individuals do not populate an area randomly; rather, genetically similar individuals, related to each other by familiar, cultural or just lingual bounds, tend to reside in geographically adjacent areas (Figure III-1 b). Considering a group of genetically similar individuals, this assumption implies **spatial dependence**[5] of the probability to find one of its individuals at a certain location. This probability is conditional on the occurrence of this group at that location or in its neighborhood, i.e. **within-group spatial dependence**. It is further assumed that the spatial distribution of a group of genetically similar individuals is independent of the location of other groups, i.e. **no spatial interference among groups**. This implies that the

4 The introductory passages of this chapter and section 5.2.6 are partially identical to Diaz Lacava & Walier (2012).

5 Spatial dependence occurs when observations sampled in spatially closer sites present more similar attribute values than observations sampled at more distant sites (Anselin & Bera 1998).

geographical coverage of genetically distinct groups overlaps. The spatial frequency of each group varies across the region, presenting areas with higher and lower frequencies (Figure III-1 b). Based on these two assumptions and from a methodological point of view it follows that, as first, different groups must be identified and then the geographical distribution of all groups must be explored and quantified. The regional distribution of the most frequent groups per tract of land can thus be assessed by jointly analyzing the distribution of all groups (Figure III-1 c).

5.2. Genetic Geostatistical Framework

This section presents the Genetic Geostatistical Framework proposed to evaluate the spatial diversity of an admixed population in a continuous region. To begin with, the methodology is schematically described.

This assessment begins with the identification of groups of similar individuals. Individuals are grouped according to a single attribute (e.g., STR allele) or according to a set of attributes (e.g., genetically similar haplotypes). Afterwards, the spatial frequency distribution per spatial unit of each group is estimated by using interpolation surface methods, and stored in a **raster layer**, i.e. georeferenced regular grid representing a spatially continuous surface (Figure III-1 d-i). The number of computed raster layers, i.e. interpolated surfaces, equals the number of groups. The spatial distribution per group is regarded as one of the total number of **genetic layers** into which the genetic admixture could be decomposed. Based on these genetic layers the most frequent groups per tract of land can be identified. Such a query is performed on the basis of a pixel-wise screening through all layers. Results are stored in two new layers, also called **composite raster map layers** (Figure III-1 d-ii). At each pixel, the first layer stores the maximum frequency value of all frequency layers, whereas the second layer stores the label of the group accounting for the maximum frequency. Composite maps may be used for further spatial analysis. For example, coverage and frequency of the most frequent groups per tract of land may be jointly examined by overlapping data of both composite maps (Figure III-1 d-iii). A pixel-wise screening of the second highest frequency values may be performed in order to assess the spatial pattern of genetic heterogeneity remaining after excluding the groups with the highest spatial frequency. For an example see Figure

IV-5.

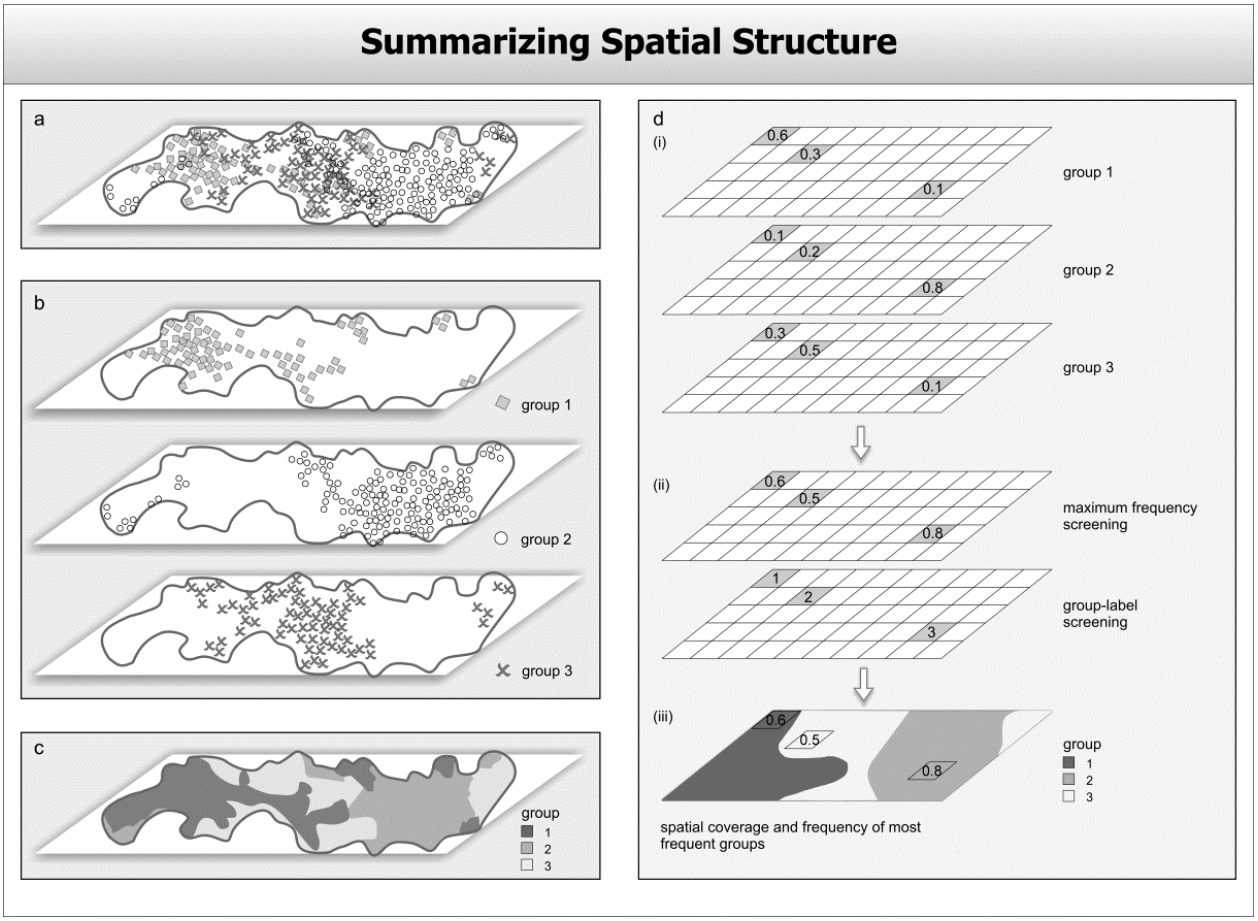

Figure III-1 Schematic representation of a geostatistical approach for the delimitation of the spatial coverage of the most frequent groups in an area.

(a) The basic scenario: an admixed area is assumed, in this scenario individuals with a certain degree of similarity are clustered into 3 groups indicated by squares (group 1), dots (group 2), and crosses (group 3); (b) spatial distribution of groups of similar individuals; (c) schematic view of the spatial coverage of the most frequent groups; (d) schematic example of the geostatistical assessment of the spatial coverage of the most frequent groups, numerically demonstrated for 3 pixels: (i) for each group, relative spatial frequency per spatial unit is estimated and represented in a georeferenced raster layer; each pixel stores the estimated spatial frequency of the group at that geographical position; (ii) for each pixel, the maximum frequency value of all layers is screened and two composite maps (raster layer) are created; maximum frequency values are stored in one layer, label of the group accounting for the maximum frequency value at each pixel is stored in a second layer; (iii) juxtaposing these two composite maps a new composite map is created, which describes the spatial coverage and frequency of the most frequent groups in an area.

Below are presented the basic model, its key components and a schematic illustration of the assessment routine of the Genetic Geostatistical Framework.

A list of starting variables follows:

$n = 1, \ldots N$: number of samples (subjects), for each of which genotypic characteristics, symbolized by α_n , and geographical coordinates of the sampling site $s_n = (x_n , y_n)$ is given as data; in this case study, genotypic data refer exclusively to Y-chromosomal data; since the human Y chromosome is constitutively haploid, the number of subjects equals the number of the genotyped Y chromosomes

Q = number of parameters for the analysis of genetic variation

$k = 1, ... K$: number of groups (C), or clusters, of genetically similar individuals

$m = 1, \ldots M$: number of Y-STR loci

$a = 1, \ldots A$: in case of single-locus analysis, number of alleles identified in the sample for each Y-STR locus, with possibly different $A = A^m$, $m = 1, \ldots M$; or in case of multi-marker analysis, number of distinct haplotypes identified in the sample, determining a unique $A = A_H$; in both cases $A \geq K$ will be satisfied

$l = 1, \ldots L$: number of spatial sampling units; in this case study a spatial sampling unit is a geographical space (region) defined by one or more contiguous provinces including a certain number of sampled urban areas

x, y : any chosen geographic coordinate within the area delimited by the overall study region (W)

s = any geographical site within the study region

U = geographical space covered by an urban area and, if present, its spatially contiguous periphery

5.2.1. The Basic Model

The following description of the basic model includes details of the most relevant parameter values used in the case study. This model may be suited for other parameter specifications, which are not specifically presented or discussed in this section.

5.2.1.1. Parameters for the Delimitation of Groups of Genetically Similar Individuals

In this work, parameters for the analysis of genetic variation (Q), used for the delimitation of groups of genetically similar individuals, were defined at two chromosomal levels: the single

marker level (G), and the multi-marker level (H).
At the single-marker level, genetic similarity of individuals was evaluated on the basis of single Y-STRs; each Y-STR defined a parameter of genetic variation. The number of evaluated parameters at this level equals the number of analyzed loci.

$$Q_G = M$$

The genetic variance of each Y-STR m is represented by the frequency distribution of the alleles (α) identified in the sample for this locus, with A^m denoting the number of observed alleles.
For each $m = 1, \ldots, M$

$$\alpha \stackrel{\text{def}}{=} \alpha_n{}^m \in [\ 1, ..., A^m\ ]$$

$$A_G = A^m$$

$$\mathcal{A}_G = [\ 1, ..., A^m\ ]$$

where $\mathcal{A}_G$ is the total genetic variance at the single-marker level, i.e. the total number of observed alleles in the M loci.

At the multi-marker level, Y-STR haplotypes were used to determine groups of genetically similar individuals. In the case study only one marker sequence was considered to construct the Y-STR haplotypes used for the geostatistical analysis; the marker sequence used to construct the haplotypes included all analyzed M Y-STRs. At this level of analysis only one parameter of analysis of genetic variation was evaluated.

$$Q_H = 1$$

In this case the genetic variance is represented by the frequency distribution of all haplotypes (α) observed in the sample. This variance arises from the combinations of alleles present in the sample of the M Y-STRs used to create the Y-STR haplotypes (H).

$$\stackrel{\text{def}}{=} \alpha_n = (\alpha_n{}^1 , \ldots, \alpha_n{}^M)$$

for $n = 1, \ldots, N$

$$A_H = \# \mathcal{A}_H$$

and

$$\mathcal{A}_H = \{ \alpha \in A^1 \times \ldots \times A^M \}$$

where $\mathcal{A}_H$ is the set of all different Y-STR haplotypes observed in the sample.

Within one genetic category, an individual carrying

- an allele $\alpha \in [1, ..., A^m]$ found in the sample for one Y-STR, or
- a haplotype $\alpha \in \mathcal{A}_H \subset A^1 \times \ldots \times A^M$ as combination of Y-STR alleles conforming the haplotypes,

is assumed to be genetically closer to the members of one and only one group k than to the rest of the K–1 groups in that genetic category.

$$k (\alpha) \in [1, \ldots, K]$$

For each type of parameter of genetic variation (Q_G, Q_H), specific rules were applied to define a group a genetically similar individuals (C) and the total number of groups (K) in each genetic category.

For the case of single Y-STRs (G) a straightforward rule was chosen: each Y-STR allele defined a group of genetically similar individuals. Considering one Y-STR, the number of groups corresponded to the total number of alleles identified in the sample for that marker.

For each $\alpha \in [\ 1, ..., A^m\ ]$ trivially:

$$k = k(\alpha) = \alpha \in [\ 1, ..., A^m\ ]$$

with

$$K_G{}^m = A^m$$

so that for each $k = 1, \ldots, K_G$

$$C_k = \{k\}$$

represents the k^{th} group of genetically similar individuals.

The delimitation of groups of genetic similarity on the bases of Y-STR haplotypes (H) involves an evaluation of the multidimensional space of genetic variants; these variants are the Y-STR loci used to create the haplotypes. In the case study, groups were determined performing cluster analysis. Genetic similarity of haplotypes, the metric used to create the distance matrix required for the clustering procedure, was evaluated according to the single-step mutation model (Gusmão et al. 2003). The final number of clusters (K_H) was specified in an iterative fashion. For this process it was taken into account: (a) *a priori* information including, for instance, historical and demographic references related to the number of main ethnic groups who gave origin to the present-day population of analysis; and (b) the consistency of the resulting composite maps, evaluated from a geostatistical point of view.

For each $\alpha \in \mathcal{A}_H$

$$k(\alpha) \in [\ 1, \ldots, K_H\ ]$$

with

$$K_H \leq A_H$$

so that for each $k = 1, \ldots, K_H$

$$C_k = \{\ \alpha \in \mathcal{A}_H \mid k(\alpha) = k\ \}$$

represents the k^{th} group of genetically similar individuals.

In summary, the number of groups of genetically similar individuals at the two proposed

chromosomal levels is given by

$$K = K_G \text{ or } K_H$$

where:

$K_G{}^m = A^m$ is the number of alleles identified in the sample for the Y-STR locus m

and

$K_H \leq A_H$ is the number of haplotype clusters delimited for the specific set of M Y-STR loci used in the case study.

5.2.1.2. Spatial Probability

The geographical space of analysis is defined by the study region (W).

In the case study, samples corresponded to individuals who reside in urban centers or, if present, in their spatially contiguous urban or semi-urban peripheral zones.

Samples were geographically assigned to the corresponding urban centers (U). Sampled urban centers were considered representative of larger areas, denoted spatial sampling units (w_l), $l = 1, \dots, L$.

Spatial computations were performed for the total area, that is for all points $(x, y) \in W$, i.e. represented in the case study by pixels, the minimal spatial unit of grid or raster layer. Since the data refer only to urban samples, strictly speaking, in the case study computed values are interpretable only for the areas covered by the urban centers (U), that is $(x^U, y^U) \in U$ (Figure III-2).

A spatial sampling unit l may include several urban centers and it covers a larger area than the total areas covered by the urban centers included in it. For each spatial sampling unit one representative geographical point of the urban centers included in it (x_c^l, y_c^l) was specified (Figure III-2).

$$(x_c^l, y_c^l) \in U_c \subset W_l$$

for each $l = 1, \dots, L$.

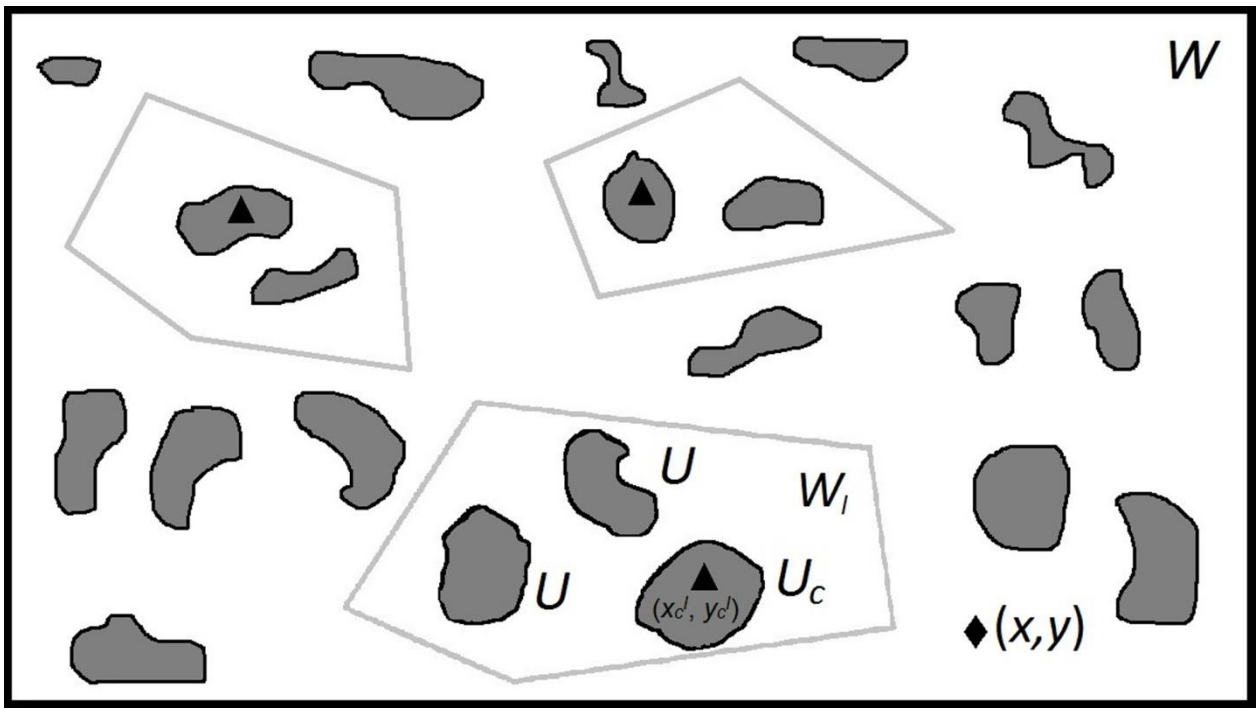

Figure III-2 Schematic representation of the study region.

In the case study, following geographical elements of the study region (*W*) were specified: geographical location (*x, y*) of a certain site (*s*) indicated by a diamond; spatial sampling units (W_l), delimited by a light grey polygon; urban centers (*U*), represented by grey areas; and location of the representative geographic points of the spatial sampling units (x_c^l, y_c^l), signaled by a triangle.

Each group *k* is assumed to be distributed in the geographical space of the study region (*W*) according to a conditional probability $f = f_k (x, y)$, the probability that the allele α, sampled at a certain site $s = (x, y)$, belongs to the group *k*, conditioned to the geographical position (*x, y*) of the sampling site.

For each (*x, y*)

$$f_k (x, y) = P (\alpha \in C_k \mid s = x, y) =$$

$$= P (\alpha \in C_k \wedge s = x, y) / P (s = x, y)$$

In the present case study, only the urban population of the study region is inspected.

5.2.2. Computational Procedure for Determining Frequencies per Sampling Unit

Data is given by (α_n, s_n) with $\alpha_n \in \mathcal{A}$ and $s_n \in U \subset W_l$ for some urban areas within some sample units W_l ($l = l_n$).

Each sample *n* (corresponding to one and only one individual) assigned to a group *k*, geographically linked to a spatial sampling unit *l*, is counted to yield the total sum per spatial sampling unit.

$$N_{k,l} = \#\{ n : \alpha_n \in C_k \wedge s_n \in W_l \}$$

$$N_l = \sum_{k=1}^{K} N_{k,l} \quad > \quad 0$$

The proportion of individuals (p) belonging to the k^{th} groups at the spatial sampling unit *l* is calculated as the number of individuals in the group ($N_{k,l}$) in relation to the total number of individuals (N_l) at that location.

$$p_{k,l} = N_{k,l} \ / \ N_l$$

$$p \in [\, 0 \, , \, 1 \,]$$

The estimated frequency (q), expressed in percentage, of individuals belonging to one group *k* at a spatial sampling unit *l* is obtained applying an *arc sine* transformation (Barbujani 1985).

$$q_{k,l} = 100 \cdot \arcsin\left(\sqrt{p_{k,l}}\right) / \arcsin\left(1\right)$$

$$q \in [\, 0, 100 \,]$$

This *arc sine* transformation is introduced in order to spread frequency values in the spectrum of small and large values (Figure III-3).

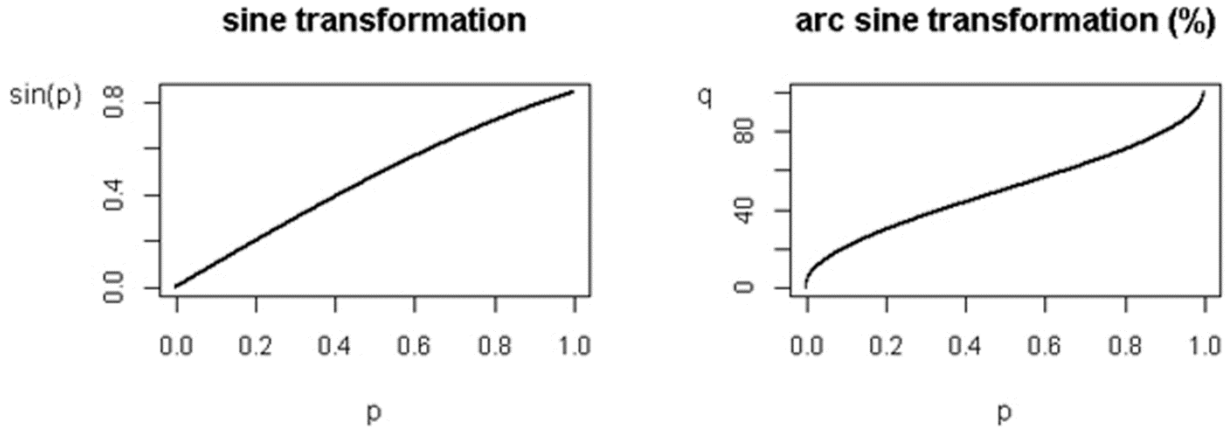

Figure III-3 Representation of sine and arcsine transformation.

5.2.3. Computational Procedure for Determining Spatial Probabilities

The conditional probability $f : W \rightarrow \mathbb{R}$ (f expressed in percentage) is estimated from q by applying a three-dimensional interpolation function (Mitasova & Mitas 1993) with the property:

For each l and k the conditional probability f at the geographical location of the site (x_c^l, y_c^l) representing the sampling unit l equals the estimated frequency (q) of individuals belonging to the k^{th} group at the spatial sampling unit l plus a deviation of the predicted values from the measured values expressed as the root mean square deviation (rms) of f in relation to q

$$f_k (x_c^l, y_c^l) = q_{k,l} \pm rms \qquad \text{with } (x_c^l, y_c^l) \in W_l$$

so that for each raster pixel (x, y) of the study area it holds that:

$$0 \leq f_k (x, y) \leq 100 \pm rms$$

and

$$\sum_{k=1}^{K} f_k (x, y) = 100 \pm rms$$

The spatial distribution of f may be represented on a 'continuous' three-dimensional raster surface $(r_{x,y,z})$, where x and y indicate the geographical location of the r^{th} pixel, $r = 1, \ldots, R$ pixels defining

the study region (W), and z indicates the magnitude of f.

While p and q refer to aggregated, spatially discontinuous data (point or raster), solely characterizing L spatial sampling units, f corresponds to spatially 'continuous' (raster) data and it characterizes the total study area (W).

5.2.4. Spatial Overall Ranking

Following a screening procedure (see s. *5.2.6 Screening Algorithms*) it is possible to rank the K groups of one parameter at each geographical location according to f:
For each $k = 1, \ldots, K$

$$f_k(x, y) \rightarrow b_k(x, y)$$

where $b_k \in \{ I, II, III, \ldots \}$ is the rank assigned to the group k at each geographical location (x, y) so that:

$$f_k^I \geq f_k^{II} \geq f_k^{III} \ldots$$

5.2.5. Composite Maps

The screening results may be summarized in form of composite maps (see Figure III-1). A composite map, which displays the results for a given rank b, $b \in \{ I, II, III, \ldots \}$, presents the spatial coverage of the groups included in the rank and their estimated conditional probability f , expressed in percentage (%), at each geographical location (x, y) of the study area.
For each given $b \in \{ I, II, III, \ldots \}$ the new composite map, created on the basis of two overlapping composite raster map layers, may display following data:

(a) estimated conditional probability f , expressed in percentage (%)

$$f = f_k^b(x, y)$$

and

(b) label of the corresponding group at each geographical location of the study area

$$k = k_b(x, y)$$

defined by the condition $b_k(x, y) = b$.

These two types of information may be graphically displayed in two-dimensional overlapped composite map layers (raster or vector) as follows:

1) $f(x, y) \leftarrow$ shading or isolines

2) $k(x, y) \leftarrow$ colours or gray tones

Applications of these techniques can be seen for instance in Figure IV-8 or Figure IV-10. The same procedure may be applied to three-dimensional graphical representation of the composite maps (see Figure IV-3).

5.2.6. Screening Algorithms

Separately for each parameter of genetic variation at each one of the two chromosomal levels of analysis (Q_G, Q_H), a pixel-wise search of the k^{th} group with the highest f is performed. For each pixel of the R pixels (x, y) representing the study region W, group label (k) and frequency per spatial unit (f) are stored in two separated raster layers: MAX_1_freq and MAX_1_group.
This procedure generates a number of maps equivalent to twice the number of parameter of genetic variation: $2 \cdot Q$ maps (Figure III-1).

Following screening algorithms is proposed:

i = 1 to Q ## independent run for each parameter of genetic variation

  r = 1 to R ## pixel-wise screening across the study region

    $\text{max\_1\_freq}_{i,r} = \max(f_{i,1,r}, \ldots f_{i,k,r})$

    $\text{max\_1\_group}_{i,r} = K_{i,r}(\max(f_{i,1,r}, \ldots f_{i,k,r}))$

  done

  >> array: MAX_1_freq_i

  >> array: MAX_1_group_i

done

This procedure is repeated searching for the second (third, …) maximum value at each r^{th} pixel. Results are stored in two further sets of layers: MAX_2_freq$_i$ and MAX_2_group$_i$.

5.2.7. Assessment Routine

This proposed Genetic Geostatistical Framework was applied in this thesis to the study of spatial diversity in Argentina using forensic Y-chromosomal genotypes. Figure III-4 exemplifies an application of this methodology to the analysis of the spatial diversity of Y-chromosomal haplotype data. This assessment includes two major procedures:

(a) Clustering Procedure: non-spatial delimitation of groups, or clusters, which involves in case of haplotypes analysis a clustering analysis; this step was followed by
(b) Geostatistical Analysis: this step includes surface interpolation of point frequency data and detection of spatial patterns of genetic variation.

These two major analytical procedures include following analytical steps:

(a) Clustering Procedure:
    i. all distinct haplotypes identified in the total sample are clustered according to genetic similarity;
    ii. cluster frequency per location (i.e. spatial sampling unit) is computed.

(b) Geostatistical Analysis:
    i. cluster frequencies are georeferenced using the geographic coordinates of the geographic locations (sites) representing the spatial sampling unit; one point data layer is created for each cluster;
    ii. spatial frequency distribution of each cluster in the total study region is estimated conducting three-dimensional spatial interpolation of the point data layers, on the basis of the function Regular Tension with Spline (Mitasova & Mitas 1993), and it is stored in form of raster data layer;

iii. frequency data per pixel of all clusters are ranked performing a pixel-wise screening of all raster data layers;

iv. spatial pattern of genetic variation is quantified and represented, for instance, in form of composite maps or three-dimensional topographies.

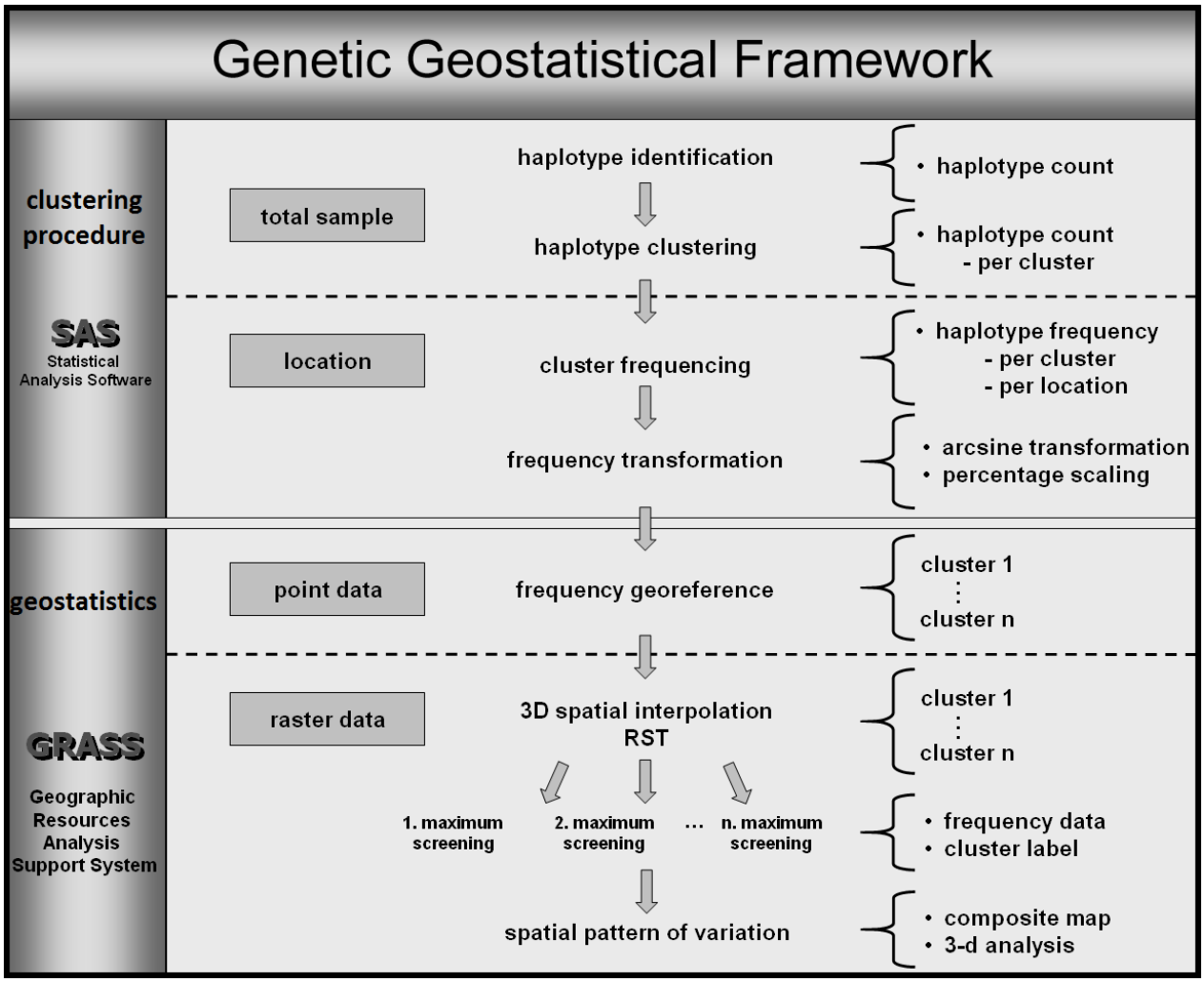

Figure III-4 Workflow of an assessment routine within the Genetic Geostatistical Framework.

Main analysis steps of an assessment routine to evaluate a spatial pattern of genetic variation are shown.

In the present thesis non-spatial descriptive statistics and clustering procedure were performed with the software package **SAS** (Statistical Analysis Software, SAS Institute Inc., Cary, NC, USA; http://www.sas.com/) and geostatistical analysis with the open-source geographic information system **GRASS GIS** (Geographic Resources Analysis Support System; http://grass.itc.it/).

6 The Case Study

6.1. Urban Male Argentine Genetic Admixture[6]

The approach described in the previous chapter was applied to the analysis of forensic Y-STR Argentine data.

In recent years there has been much attention to Argentine population stratification (Alfaro et al. 2005; Corach et al. 2010; Diaz Lacava et al. 2011a; Dipierri et al. 2005; Marino et al. 2008; Toscanini et al. 2007). In a rather short historical period the Argentine population went through massive demographic changes. Overwhelmingly large immigration waves populating Argentina since the 1850s introduced a strong male European component (Avena et al. 2001; Corach et al. 2010; Levene 1992; Rock 1987).

This relatively recent immigration largely diluted the previous admixture and played a major role in modulating the present Argentine genetic background (Avena et al. 2001; Corach et al. 2010). As result of almost two hundred years of admixture among previously settled populations and worldwide incoming lineages, a complex spatial pattern of genetic variation is expected. Questions related to the history, extent, and geographical structure of admixture and the characteristics of nowadays genetic composition were addressed under several perspectives (Alfaro et al. 2005; Avena et al. 2001; Corach et al. 2010; Diaz Lacava et al. 2011a; Dipierri et al. 2005; Marino et al. 2008; Sala et al. 1998; Toscanini et al. 2007).

From a methodological point of view, two main factors contribute to make the study of this population interesting. On one hand, the expected complex admixture poses a methodological challenge. On the other hand, since most influencing socio-demographic processes affecting contemporary Argentine genetic background took place in a relatively recent and short historical period, there are enough historical, ethnological, and census data available to analyze and validate

6 Contents of this chapter are partially identical to Diaz Lacava & Walier (2012).

the results.

Two evaluations regarding the spatial pattern of genetic admixture of the contemporary Argentine population were specifically addressed in this work:

(a) regional distribution of the most frequent alleles per Y-STR locus, and

(b) regional distribution of the most frequent groups of genetically similar Y-STR haplotypes.

In each case, composite maps were created showing the estimated spatial coverage of the most frequent groups. Transects were constructed displaying the spatial distribution of group frequencies across the study region. The worldwide provenance of most frequent haplotypes was inspected.

6.2. Study Region and Spatial Sampling Units

The study region covers central and northern Argentina. It includes 10 sampled provinces and 10 further provinces within the area circumscribed by the sampled provinces (Figure III-5). These 20 provinces (out of a total of 24 Argentine provinces) represent 80 percent of the total Argentine area. It is worth noting that the sampled area contains the absolute majority of the total population. Argentina is an extremely centralized and highly urbanized country. While the 10 sampled provinces include 75 percent of the total population, the study region includes 98 percent (INDEC 2010).

Spatial sampling units were primarily defined by sampled provinces. Sampled provinces with small number of samples were aggregated to neighboring sampled provinces. As a result, out of 10 sampled provinces 6 spatial sampling units were defined (Figure III-5; Table III-1).

Geographical coordinates of provincial capital cities were used to georeference genetic frequencies *q* to the spatial sampling units (see s. *5.2.1 The Basic Model*). In case of spatial sampling units containing more than one provincial capital city (i.e. aggregated sampled provinces) the capital city of the province with the largest number of samples was used as georeferenced point.

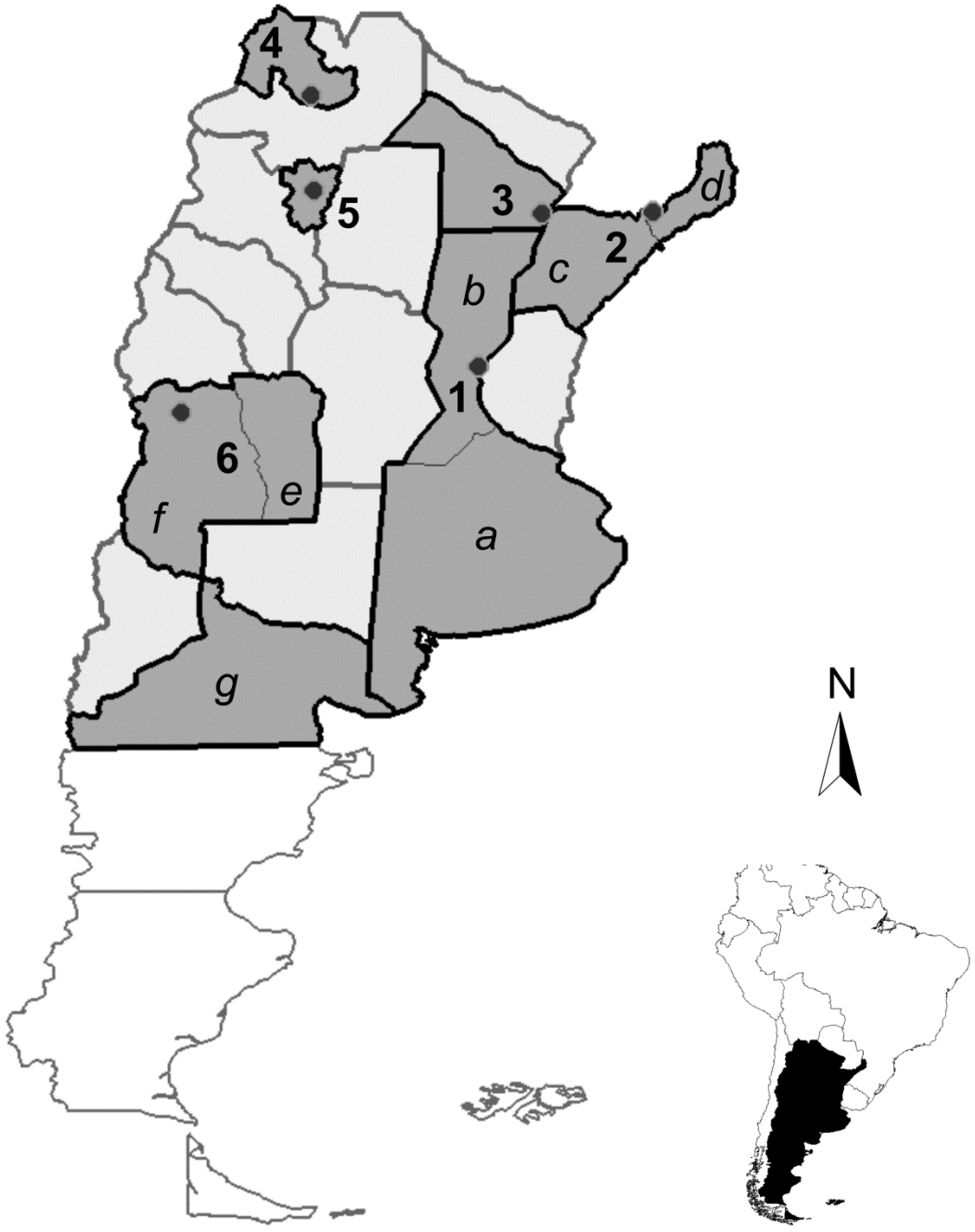

Figure III-5 Study region.

The study region (thicker contour) included 20 of the 24 Argentine provinces. Six spatial sampling units were delimited out of 10 sampled provinces (1a: Buenos Aires; 1b: Santa Fe; 2c: Corrientes; 2d: Misiones; 3: Chaco; 4: Jujuy; 5: Tucuman; 6e: San Luis; 6f: Mendoza; 6g: Rio Negro). Grey circles show capital cities used as geographical reference of point data. The inset shows Argentina's location in South America.

Table III-1 Description of the sampling units

SID	*Sampling Unit*	*n*	*Cov*	*Georeference*	*Province*
1	**Santa Fe**	**20**	**15.9**	**Santa Fe**	**Santa Fe**
					Buenos Aires
2	**Misiones**	**8**	**4.3**	**Posadas**	**Misiones**
					Corrientes
3	**Chaco**	**7**	**3.5**	**Resistencia**	**Chaco**
4	**Jujuy**	**18**	**2.0**	**Jujuy**	**Jujuy**
5	**Tucuman**	**33**	**0.8**	**Tucuman**	**Tucuman**
6	**Mendoza**	**59**	**15.3**	**Mendoza**	**Mendoza**
					San Luis
					Rio Negro

SID: sampling-unit identification (see Figure III-5); *n*: number of samples; *Cov*: spatial coverage in percentage in relation to the study region; *Georeference*: capital city used as geographical reference of point data; *Province*: provinces (federal states) included in the spatial sampling unit.

6.3. Subjects and Genotypes

DNA material was obtained from 145 unrelated male Argentine citizens in year 2007. Donors were recruited in the framework of legal paternity testing. No sampling bias may be assumed due to socio-economic condition of donors; depending on the socio-economic situation of the donor paternity-testing costs were either privately or publicly financed. Sampling did not include any restriction related to the donor's ethnic background. Care was taken that neither closely related individuals nor non-Argentine citizens were included in the sample. Since sampling did not include any further restrictions, this data set may be regarded as a random sample of the present male urban Argentine population in the study region. Samples and genotypes were treated anonymously, following local ethical restrictions. Samples were solely referenced to the sampled province (federal state), thus guaranteeing donor anonymity.

DNA was extracted from blood stains or buccal swab specimens by conventional organic extraction methods. DNA was prepared and genotyped at the DNA Analysis Unit, Official College of Pharmacists and Biochemists, Buenos Aires, using the commercial typing kit PowerPlex Y System (Promega, Madison, Wisconsin; http://www.promega.com/) –DYS19, DYS385a/b, DYS389I/II,

DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439- in a 25 µl reaction volume, as specified by the manufacturer. Detection of the amplified fragments was done using the ABI Prism 377 (Applied Biosystems, Foster City, California; http://www.appliedbiosystems.com/). PowerTyper Y Macro (Promega, Madison, Wisconsin; http://www.promega.com/) was used to assign the alleles. DYS385 was excluded from the analysis since it is not possible to unequivocally assign one allele to a specific locus (Gusmão et al. 2003).

Alleles were coded in terms of the number of variable repeats, in agreement with the freely accessible worldwide YHRD database (http://www.yhrd.org/). Y-STR loci –DYS19, DYS389I/II, DYS390, DYS391, DYS392, and DYS393- were combined to define haplotypes. Haplotypes were denoted by listing the seven alleles in the same order as in YHRD, concatenated by a dash (i.e. DYS19 _ 389I _ 389II _ 390 _ 391 _ 392 _ 393).

Most probable worldwide provenance of frequent haplotypes was investigated. Historical and ethnological data (Bartolome 1976; Rock 1987; Sanchez-Albornoz 1994) indicate that the contemporary Argentine population to the greatest extent is the result of modern admixture arisen from major immigration processes, officially promoted since the 1850s. This relatively new and long-lasting immigration arrived from all over the world. It largely outnumbered the colonial population and most probably diluted a previous admixture among a reduced group of sedentary Amerindian populations, Spaniards, and some other minor ethnicities (Levene 1951; Levene 1992; Rock 1987; Sanchez-Albornoz 1994). Since Y-STR haplotypes are paternally inherited as a block, without recombination, it was assumed that the worldwide distribution of a haplotype would provide a good clue about the region where that haplotype came from before it integrated into the Argentine population. All different haplotypes identified in the sample were ranked according to their absolute frequency. The worldwide geographical distribution of the most frequent haplotypes was inspected. Relative frequency and worldwide geographical distribution were searched in the YHRD database (http://www.yhrd.org/).

6.4. Genetic Frequencies

6.4.1. Computation of Y-STR Allele Frequencies

Allele frequencies of each of the seven Y-STR loci were scored by single-gene counting procedures. Bar plots of each Y microsatellite showing the allele frequencies were created with the basic module `barplot` of the R software package, version 2.11.1 (R : A Language and Environment for Statistical Computing; http://www.r-project.org/; R Development Core Team, 2010-05-31).

6.4.2. Computation of Spatially Aggregated Frequencies

At the single locus level, groups of genetic similarity were defined based on the Y-STR alleles identified in the sample.

At the multi-locus level, first the Y-STR sequence used to create haplotypes was specified. Two Y-STR sequences were inspected:

a) 5 STRs: DYS19, DYS390, DYS391, DYS392, DYS393 (i.e. Corach et al. 2001).
b) 7 STRs: DYS19, DYS389I/II, DYS390, DYS391, DYS392, DYS393 (standard 'core' of the minimal YHRD 7-Y-STR haplotype; http://www.yhrd.org/; Roewer et al. 2001).

For this purpose and separately for each of both sequences, all different haplotypes were identified and then grouped in 3 clusters according to the clustering procedure described below in this section. Subsequently, one histogram per cluster and per sequence was created. Haplotypes were sorted alphanumeric. Only haplotypes with an absolute frequency larger than two were included. Of both sets, the YHRD 7 Y-STR sequence was selected for the construction of the haplotypes. This set showed a better differentiation of the sample indicated in the histogram by a more balanced distribution shape among clusters (Figure III-6).

All different haplotypes (minimal YHRD 7-Y-STR haplotype) were identified and grouped according to molecular distance following the widely accepted stepwise mutation model (Gusmão et al. 2003; Gusmão et al. 2005). A total of 4 clusters were delimited using the statistical software package SAS v. 9.1. (Statistical Analysis Software, SAS Institute Inc., Cary, NC, USA; http://www.sas.com/; clustering method: Ward; Euclidean distance).

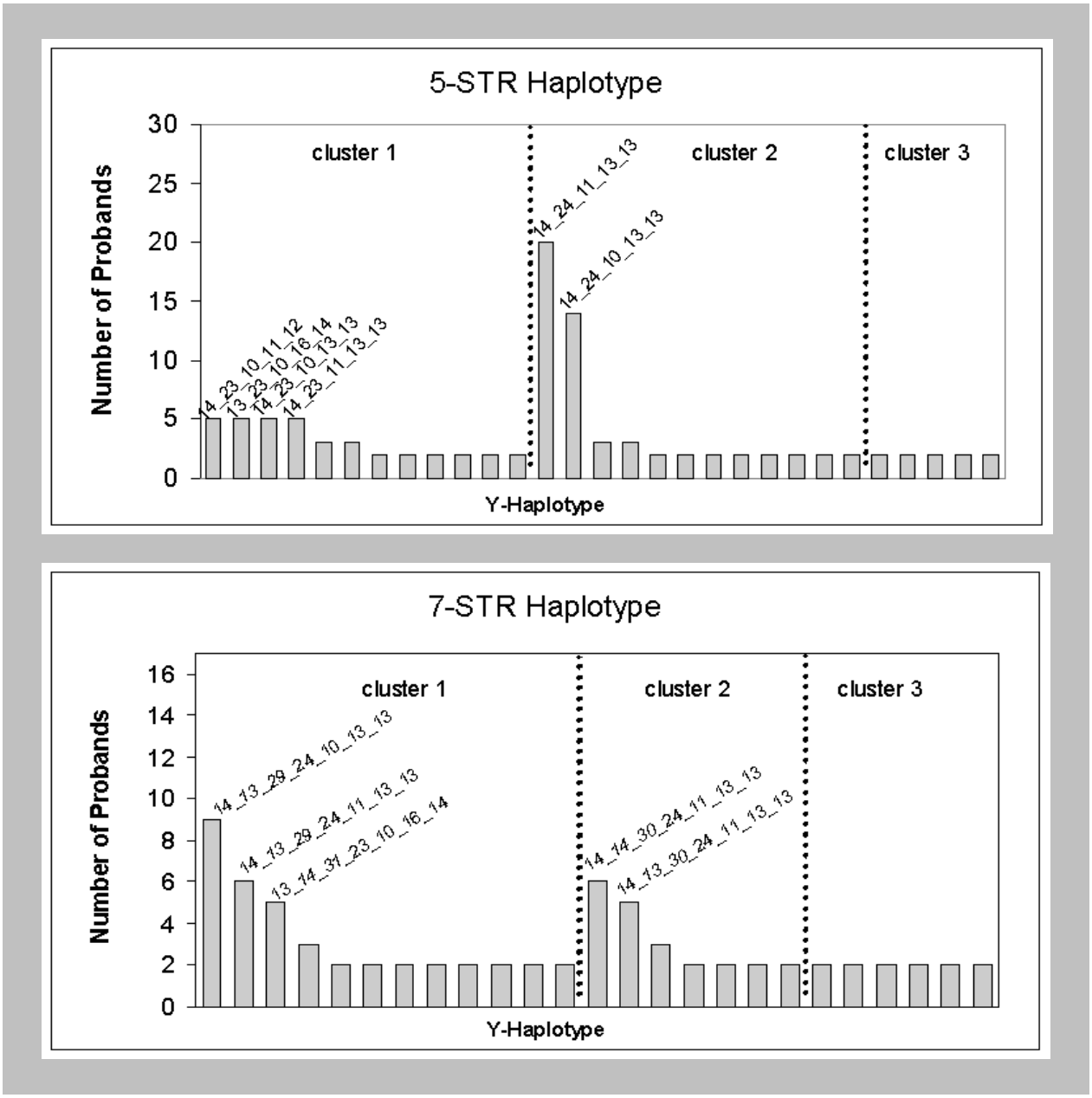

Figure III-6 Histogram showing the number of samples per haplotype and per cluster.

Most frequent haplotypes per cluster are displayed and correspond to following two Y-STR set: 5 Y-STR set (DYS19, DYS390, DYS391, DYS392, DYS393; i.e. Corach et al. 2001); 7 Y-STR set (DYS19, DYS389I/II, DYS390, DYS391, DYS392, DYS393; http://www.yhrd.org/; Roewer et al. 2001); italics highlight additional STRs to the 5-Y-STR set[7].

[7] This figure is partially identical to Diaz Lacava et al. (2007).

The most appropriate number of clusters fitting the data was chosen on the basis of geostatistical results, the composite maps (see s. *5.2.4 Spatial Overall Ranking*). This step pursued to maximize the regionalization and to minimize the number of regions not supported by the data, i.e. maximization of the number of differentiated areas including spatial sampling units. This explorative procedure was performed inspecting the spatial results based on 3, 4, 5, and 7 clusters. The proportion of individuals per spatial sampling unit (p) was computed with the SAS procedure **`proc freq`** according to the specifications presented in s. *5.2.2 Computational Procedure for Determining Frequencies per Sampling Unit*. The arc sine transformation $p \rightarrow q$ (see s. *5.2.2 Computational Procedure for Determining Frequencies per Sampling Unit*) was performed with SAS basic modules.

6.5. Geostatistical Analysis

Geostatistical analysis was performed with the open-source geographic information system GRASS GIS v. 6.4. (Geographic Resources Analysis Support System; http://grass.itc.it/).

6.5.1. Surface Interpolation

Separately for each *k* group, i.e. delimited by each allele of the seven Y-STR loci and by each one of the four haplotype clusters, genetic frequencies per spatial sampling unit (q_k) (see s. *5.2.2 Computational Procedure for Determining Frequencies per Sampling Unit*), were referenced to the spatial sampling units (see s. *6.2 Study Region and Spatial Sampling Units*) and imported into GRASS GIS as point data layers. Based on each point data layer one interpolated surface (i.e. raster layer) was created. Surface interpolation was performed using the function 'regularized spline with tension', implemented in the module `v.surf.rst` (Mitasova & Mitas 1993). This method computes the values of the interpolated surface using a function which simulates a thin flexible plate passing through or close to the points (Mitasova & Mitas 1993). Splines are flexible to model differential local patterns based on change of elastic properties of the interpolation function. Splines proved to be rather successful for cases where the phenomena have less random components (uncertainty) and are more driven by processes which minimize energy (Neteler & Mitasova 2004). Since it seemed intuitive to assume that the genetic admixture of a region may be more influenced

by local socio-demographic processes than by random processes (uncertainty), the spline function was selected as the best choice to model genetic layers representing the spatial distribution of groups of genetically similar individuals by means of surface interpolation. The GRASS GIS implementation of the spline function offers the possibility to optimize final surfaces by parameter tuning. Explorative runs were performed varying the tuning parameters **smooth** and **tension**. While the tension parameter adjusts the character of the interpolated surface (peak/pit ('crater') vs. stiff-plate landscape), the smoothing parameter regulates the deviation between point data and interpolated data at the interpolation site, i.e. smooth=0 implies no deviation between both values at the interpolation sites (Neteler & Mitasova 2004). Tuning of these two parameters was performed pursuing to achieve minimal statistical error (also called predictive error) defined by root mean squared deviation (*rms*) (Neteler & Mitasova 2004), to reduce interpolation artifacts (e.g., several neighboring stripes, extremely small or unexpected patches), and to increase the number of regions detected with the final composite maps (see s. *5.2.5 Composite Maps*). It could be verified that the effect on the resulting surface of varying one parameter showed a strong dependence on the value of the other parameter (probably attributable to the relatively low number of interpolation sites). Therefore, in order to simplify interpretation of geostatistical results, while the value of tension parameter was fixed to the default value (tension=40), smoothing values were optimized. The effect of solely varying smoothing values is presented in Figure III-7. As the smoothing parameter increases, the landscape shape becomes more even and the differentiation between the site values vanishes. Note that the interpolation function obtained with smooth=0 passes exactly through the data points 'Jujuy' and 'Mendoza', but it presents a valley, which is not supported by the data (i.e. interpolation artifact). A lower smoothing value may be appropriate for cases in which interpolation-site data may represent the proximal surroundings. A higher smoothing value may be suited when the interpolation-site data may represent a broader region. Since only one geostatistical parameter was tuned (smooth), to model the genetic landscape best-fitting the data was equivalent to select the most appropriate smoothing value. Applying this concept to the modeling of the geographical distribution of Y-chromosome frequencies, principally affected by migration in the time scale of the present study, varying the smoothing value may be considered equivalent to tuning the degree of migration of each specific group (f_k) among regions. An interpolation surface

obtained with lower smoothing values may best represent a scenario of lower migrant rates among regions; higher values may fit better a scenario of higher migration rates among regions.

6.5.2. Creation of Composite Maps

Composite map layers were created with the basic GRASS GIS module `r.mapcalc` using the procedure described in s. *5.2.6 Screening Algorithms*. One layer type displayed the spatial distribution of the maximum (and second maximum) frequency values among groups and the second showed the spatial distribution of groups accounting for the maximum (and second maximum) frequency (where groups were defined by Y-STR alleles or by haplotype-cluster membership).

In case of the analysis at the locus level (Y-STR), the composite maps storing frequency values, MAX_1_freq and MAX_2_freq (see s. *5.2.6 Screening Algorithms*), were used for tuning and visual evaluation of result consistency. The composite maps MAX_1_group and MAX_2_group were exported and saved as graphics for result presentation. The exported graphics separately showed for each Y-STR the regions where one allele presented the maximum (as well as the second maximum) frequency per tract of land in relation to the other alleles.

At the haplotype level, composite maps storing frequency values, MAX_1_freq and MAX_2_freq (see s. *5.2.6 Screening Algorithms*), were used for tuning and visual evaluation of result consistency as well as for further analysis. Coverage and frequency of the most frequent and the second most frequent clusters of Y-STR haplotypes (MAX_1_group/MAX_1_freq and MAX_2_group/MAX_2_freq respectively) were displayed in form of composite maps and three-dimensional views (see s. *5.2.5 Composite Maps*).

Composite maps were constructed using two approaches. First, colors were used to display group (cluster) coverage and shades to display frequency values. Second, tones were used to display group (cluster) coverage and contour lines were used to display frequency values.

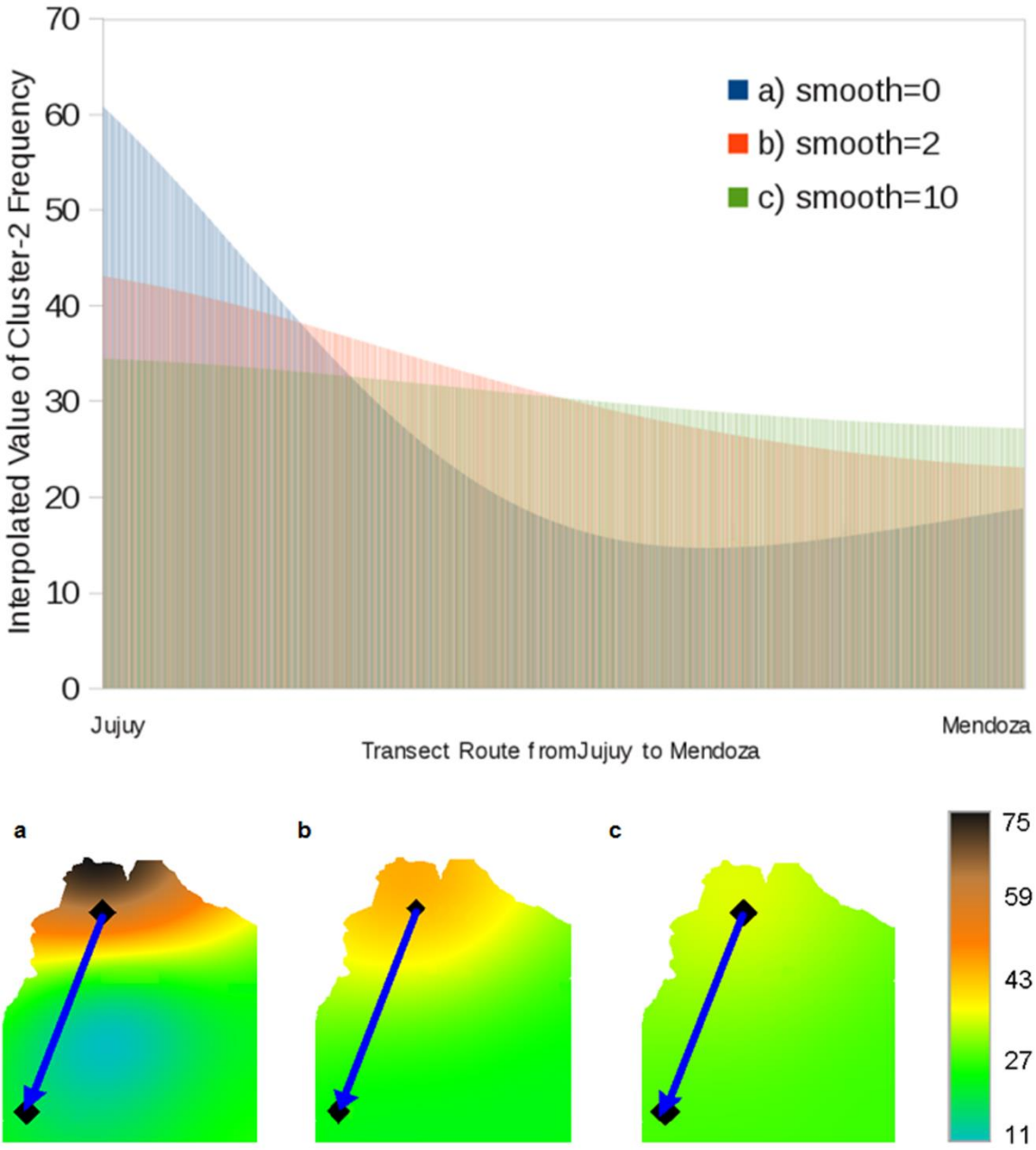

Figure III-7 Impact of smoothing parameter on the resulting surface.

Three interpolated surfaces were created from the cluster-2 frequency data with default tension values (tension=40) and varied smoothing values: a) smooth=0; b) smooth=2; c) smooth=10. A profile transect was drawn between the interpolation sites Jujuy (starting point) and Mendoza (end point). The chart on the left shows the raster values of the interpolated surfaces along the transect. Note that the interpolation function obtained with smooth=0 passes exactly through the data points (*q*) measured in Jujuy and Mendoza (see s. *5.2.2 Computational Procedure for Determining Frequencies per Sampling Unit*).

Contour-line layers were created based on these frequency maps using the GRASS GIS module `r.contour`. Contour lines were set along points of equal *f*, where *f* is the estimated frequency per pixel (in percent) of a certain group (see s. *5.2.3 Computational Procedure for Determining Spatial Probabilities*), with a step of 0.5 percent difference. A second set of contour-line layers was created with larger steps. On the basis of the map MAX_1_freq, contour lines were delineated along the values: 35, 37.5, 40, 42.5, and 45 percent. For the map MAX_2_freq contour lines were constructed along the values: 30, 32.5, 35, and 36.5 percent. Final graphics were exported showing the regional distribution of the clusters with the maximum (as well as the second maximum) frequency per tract of land and the respective spatial distribution of frequencies. The spatial distribution of frequencies was displayed juxtaposing contour lines to the regional maps, MAX_1_group and MAX_2_group.

Three dimensional views (3d) were created to complement this latter type of representation of the data. In these 3d views coverage of groups were represented with the same colors used in the composite maps; surface elevation displayed spatial frequencies. 3d views were created with the GRASS GIS module `NVIZ visualization suite`.

The spatial pattern of frequencies was further inspected on the basis of transects. For each frequency map, MAX_1_freq and MAX_2_freq, one transect was set up between the local maximum value in the central region and the local maximum value in the northwest region. Frequencies (in percent) were measured along transects and profile charts were created with the interactive GRASS GIS module **`profile`**. Maps, 3d views and charts were exported and saved as graphics for result presentation.

6.6. Characterization of Y-Chromosome Ancestry

Geographical ancestry of haplotypes was investigated using two approaches. First, it was assumed that the most probable geographical origin of a haplotype is the region where that haplotype presents the highest frequency (i.e. Salas et al. 2008; Gusmão et al. 2003). Second, the geographical ancestry of the Y-chromosome haplogroup most likely corresponding to a haplotype was analyzed. In this latter case, each haplotype was searched in YHRD (http://www.yhrd.org/). For each haplotype, the distribution of worldwide frequencies as well as the number of haplogroup matches

was inspected. The number of occurrences per geographical region of a haplotype in this worldwide database was assumed to provide a strong evidence of its most probable geographical origin. The worldwide geographical distribution of each haplogroup of interest was extracted from the literature.

PART IV -RESULTS

7 The Genetic Heterogeneity of the Urban Argentine Population

7.1. Frequency Distribution of Y-STR Alleles[8]

All seven Y-STR loci –DYS19, DYS389I/II, DYS390, DYS391, DYS392, and DYS393- were polymorphic and presented from 3 to 8 alleles. Allele frequency distributions of the seven loci are shown in Figure IV-1. All loci presented a unimodal distribution except for DYS392, clearly showing a bimodal allele-frequency distribution. With the exception of DYS389II, unimodal loci presented one frequent allele and less-frequent alleles differentiated from the next most-frequent allele by a single repeat unit. By DYS389II, allele 27 was not detected in the sample. Allele frequency distribution of all seven loci is in good accordance with the results of a global survey of 986 males including a large number of European samples (n = 470) (Kayser et al. 2001). In every case, the most-frequent allele as well as the shape of the allele frequency distribution in relation to the less-frequent alleles matched global allele-frequency distribution, as described by Kayser et al. (2001). A less-frequent allele of DYS392 (allele 17) was found in the Argentine sample and not in the global survey (Kayser et al. 2001). This result is consistent with a higher genetic variance of DYS392 in the Argentine population in relation to globally distributed populations, as reported by Kayser et al. (2001). On the other hand, a less-frequent allele of DYS389II (allele 27) detected in

8 Contents of this chapter are partially identical to Diaz Lacava & Walier (2012).

the global survey (Kayser et al. 2001) was absent in this sample (see above).

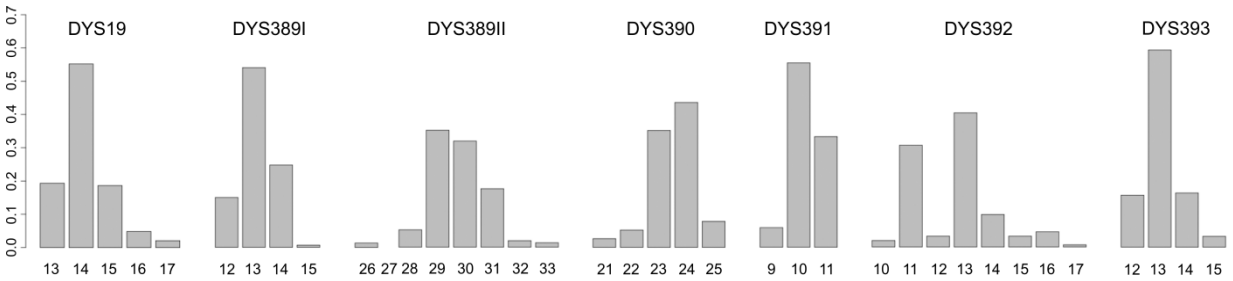

Figure IV-1 Allele frequency distribution of seven Y-STR loci.

For each locus, allelic notation (designated in number of repeats) is indicated on the x-axis; observed frequencies (*n* = 145) are shown on the y-axis.

7.2. Frequency Distribution of Y-STR Haplotype Frequencies

Out of a total of 145 males 97 different haplotypes were identified. Seven different haplotypes were found in three or more individuals (Table IV-1). 18 haplotypes were detected twice and 72 haplotypes only once.

Table IV-1 Absolute frequency and cluster assignment of frequent haplotypes

Haplotype	***n***	**(%)**	**Cluster**	**(%)**
14_13_29_24_10_13_13	**9**	**(6.2)**	**1**	**(20.9)**
14_13_29_24_11_13_13	**6**	**(4.1)**	**1**	**(14.0)**
14_14_30_24_11_13_13	**6**	**(4.1)**	**3**	**(15.8)**
14_13_30_24_11_13_13	**5**	**(3.4)**	**3**	**(13.2)**
13_14_31_23_10_16_14	**5**	**(3.4)**	**2**	**(18.5)**
14_13_29_23_11_13_13	**3**	**(2.1)**	**1**	**(7.0)**
14_13_30_23_10_11_12	**3**	**(2.1)**	**3**	**(7.9)**
Total	**37**	**(25.5)**		

This table lists haplotypes found in three or more individuals; the percentage (%) of samples carrying the haplotype in relation to the total sample (*n* = 145) and to the number of samples per cluster are indicated in parentheses.

About 25 percent of the sample (*n* = 37) carried one of seven different haplotypes (Table IV-1). The rest 75 percent of the sample (*n* = 108) corresponded to the haplotypes with an absolute frequency equal or smaller than two. These figures are in agreement with the expected strong admixture of the Argentine population (Alfaro et al. 2005; Avena et al. 2001; Corach et al. 2010; Diaz Lacava et al. 2011a; Marino et al. 2007; Toscanini et al. 2007).

According to the worldwide YHRD database (http://www.yhrd.org/) the most frequent haplotype (14_13_29_24_10_13_13; *n* = 9; Table IV-1) shows higher frequencies in Western Europe, and only one frequent haplotype of this sample (13_14_31_23_10_16_14; *n* = 5) shows higher frequencies in South America.

7.3. Spatial Diversity at the Y-STR level

One interpolated surface was created for each allele of each one of the seven Y-STRs. In total, 36 interpolated surfaces were obtained. Table IV-2 lists the selected smooth-parameter value and minimum and maximum root mean square deviation (*rms*) of all interpolated surfaces per Y-STR locus (estimated spatial distribution of allele frequencies).

Table IV-2 Goodness of fit for spatial interpolation frequencies of Y-STR loci

			rms	
Y-STR	***a***	***smooth***	***min***	***max***
DYS19	**5**	**0.5**	**1.01**	**6.88**
DYS389I	**4**	**0.2**	**0.63**	**2.56**
DYS389II	**7**	**1.0**	**1.32**	**6.35**
DYS390	**5**	**0.5**	**1.04**	**4.88**
DYS391	**3**	**0.2**	**1.78**	**2.17**
DYS392	**8**	**0.5**	**1.00**	**5.43**
DYS393	**4**	**0.5**	**1.82**	**3.86**

This table shows the goodness of fit for the interpolated data per Y-STR locus on the basis of root mean square deviation values (*rms*); *min* and *max*: minimum and maximum *rms* values of all *a* interpolated surfaces, obtained by surface interpolation of *q*, where *q* is the estimated Y-STR allele frequency (in percent) at one sampling location (see s. *5.2.2 Computational Procedure for Determining Frequencies per Sampling Unit*); *a*: number of alleles per locus, corresponding to the total number of interpolated surfaces per locus; *smooth*: smooth-parameter value used for the interpolation procedure.

For each Y-STR two composite maps were created by jointly screening the spatial frequency distribution of the alleles. Regional-specific allelic differences were indicated by the composite maps (Figure IV-2). Allelic differentiation of the northwest region was observed for the loci DSY19, DSY389I, DSY389II, and DSY390. In every case, one allele presented the highest frequency almost everywhere in the study region and the second highest frequency in the northwest, where a second allele showed the highest frequency. DSY19 and DSY389I presented a third allele, which had the second highest frequency in central Argentina. DSY389I presented as well an allelic differentiation in the Misiones Province, located in northeastern Argentina. DSY389II was the only locus which presented three areas of allelic differentiation in the MAX_1_group map: central, northwest, and northeast. The MAX_2_group map showed in the northwest a transition zone, evidenced by a stripe pattern. This stripe pattern arose due to very similar frequencies of all three alleles in the same region. This result is in good agreement with the expected high genetic variance of DSY389II (Kayser et al. 2001). DSY389II composite maps indicate that (a) this microsatellite may allow a differentiation of the male Argentine population among three main socio-demographic regions, and (b) the highest variance of this locus may be primarily centered in the northwest of Argentina. Further deeper analysis will be necessary to confirm these observations.

DSY391 and DSY392 presented a quite similar spatial pattern to the previously described loci but differed in the geographical area where a second allele showed the highest frequency. While one allele had the highest frequency over the largest portion of the study region, another had the highest frequency in the central northern area, a region known as the ‘Argentine Chaco’ (Figure IV-2).

A single allele of DSY393 had the highest frequency all over the study region. The two next most-frequent alleles (Figure IV-1) had the second highest frequency, one in the northern and the other in the southern portion of the study region (Figure IV-2).

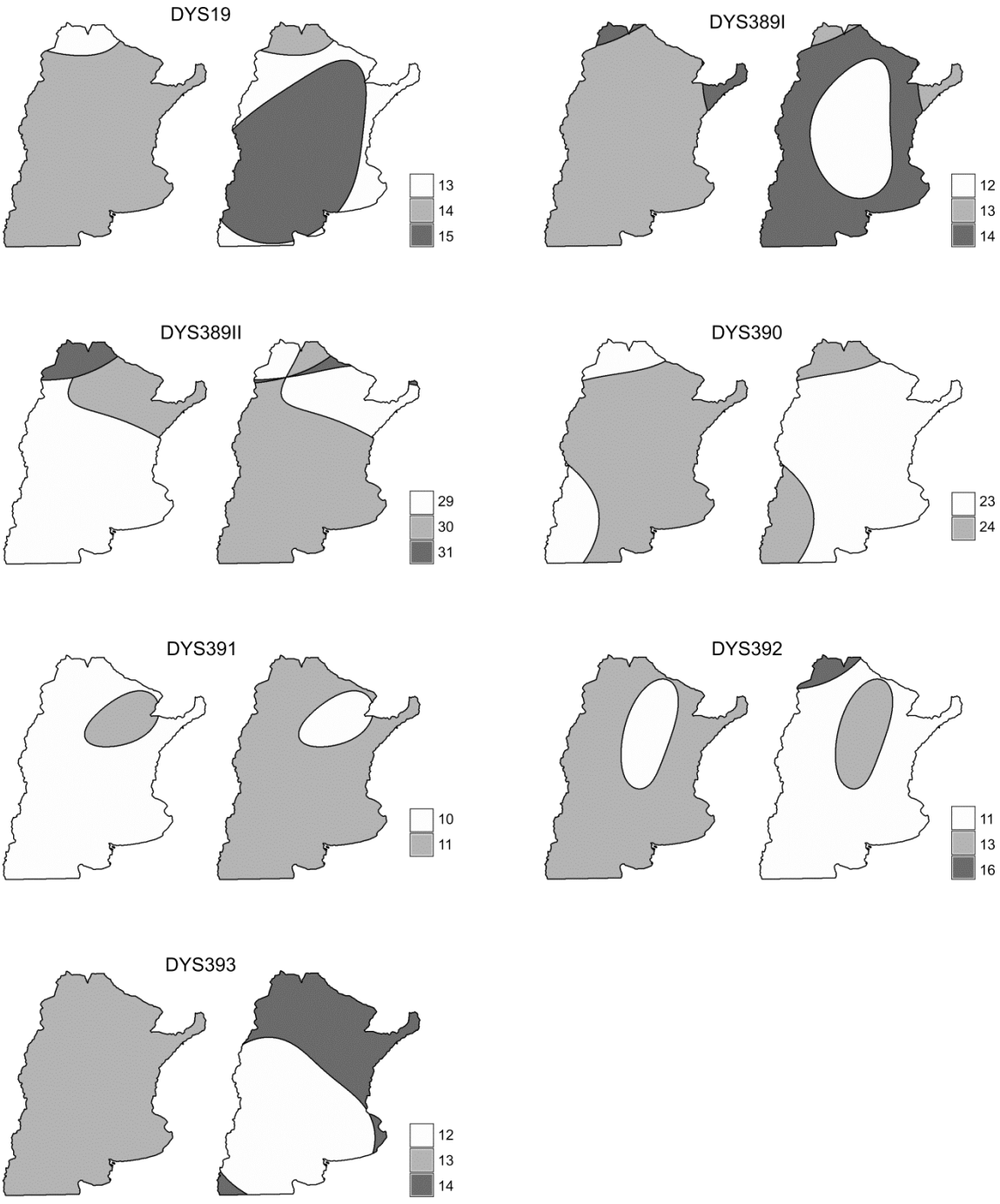

Figure IV-2 Spatial distribution of the most frequent alleles per Y-STR locus.

Gray coding indicates the allelic notation (in number of repeats). Each region in the composite maps delimits an area where one allele presented: (a) the highest frequency (left map), and (b) the second highest frequency (right map).

The observed spatial patterns of allelic differentiation of all seven Y-STR loci considered as a whole indicate strong regional differentiation of the northwest in relation to the rest of the Argentine territory as well as lower differentiation of two other regions: the Argentine Chaco and the central and littoral Argentina. The spatial patterns observed at the Y-STR level corroborate prior expectations of regional genetic differentiation of the contemporary male Argentine population (Alfaro et al. 2005; Corach et al. 2010; Marino et al. 2007; Marino et al. 2008; Toscanini et al. 2007).

7.4. Spatial Diversity of Y-STR Haplotypes

The YHRD data base (http://www.yhrd.org/) was searched to investigate the most probable worldwide origin of frequent haplotypes according to the haplotype frequency distribution and the inferred haplogroup status.

The clustering procedure sorted different haplotypes into relatively similar clusters with respect to both total count of different haplotypes and total count of samples per cluster (Table IV-3). Haplotypes with an absolute frequency larger than two were assigned to the first three clusters (Table IV-1).

Table IV-3 Frequency distribution of haplotypes and samples per cluster

Cluster	*hap*	(%)	*n*	(%)
1	**24**	**(24.7)**	**43**	**(29.7)**
2	**19**	**(19.6)**	**27**	**(18.6)**
3	**23**	**(23.7)**	**38**	**(26.2)**
4	**31**	**(32.0)**	**37**	**(25.5)**
Total	**97**	**(100.0)**	**145**	**(100.0)**

The percentage (%) of different haplotypes (*hap*) and samples (*n*) in relation to total figures are indicated in parentheses.

Smooth-parameter value, minimum and maximum values of the interpolated surfaces, minimum and maximum values of the sampled data, and the root mean square deviation (*rms*) per cluster layer are shown in Table IV-4.

Table IV-4 Goodness of fit for spatial interpolation of Y-STR haplotype cluster frequencies

Cluster	*smooth*	*rms*	*min* [*q*]	*max* [*q*]
1	**0.5**	**1.78**	**33.6 [32.8]**	**44.9 [46.8]**
2	**1.5**	**7.45**	**21.3 [18.8]**	**47.1 [60.8]**
3	**2.0**	**16.71**	**15.7 [0.0]**	**39.8 [58.0]**
4	**2.0**	**14.58**	**13.3 [0.0]**	**37.3 [54.6]**

The goodness of the interpolated data is shown on the basis of the root mean square deviation (*rms*) of the interpolation surfaces and the minimum (*min*) and maximum (*max*) values obtained by surface interpolation of *q*, where *q* is the estimated cluster frequency (in percent) at a sampling location (see s. *5.2.2 Computational Procedure for Determining Frequencies per Sampling Unit*); *smooth*: smooth-parameter value used to perform surface interpolation.

According to the YHRD database (http://www.yhrd.org/) the most frequent haplotypes ($n \geq 3$) of cluster 1 and of cluster 3 (Table IV-1) are most frequently registered throughout the world in Western Europe and with considerable frequency in North and South America. One of these haplotypes (14_13_30_23_10_11_12; $n = 3$; cluster 3; Table IV-1) also shows higher frequencies in the Middle East and in Southern Asia.

The most frequent haplotype of cluster 2 (13_14_31_23_10_16_14; $n = 5$; Table IV-1) presents higher frequencies of matches in Latin America and in North Africa. Non-unique cluster-4 haplotypes ($n = 2$) are scarcely distributed throughout the world, with higher frequency of matches either in South America, Africa or southeastern Asia (http://www.yhrd.org/). These figures indicate that haplotypes grouped into cluster 1 and cluster 3 were introduced most probably by immigrants or immigrant descendants from Western Europe. Haplotypes grouped into cluster 2 correspond most probably to descendants from local Amerindian populations and those grouped into cluster 4 are of admixed origin.

Cluster 1 was found to be present with the highest frequency in 86 percent of the study region (Figure IV-3) and second highest frequency in the rest (Figure IV-4). In the northwest, cluster 2 accounted for the highest frequencies (Figure IV-3). This region covered 13 percent of the study region and included the provinces of Jujuy and Salta. In the northeast, in the region corresponding to the Misiones Province, cluster 3 was the highest in frequency (one percent of the study region) (Figure IV-3). In the littoral area cluster 4 had the second highest frequency after cluster 1 and it covered 52 percent of the total study region (Figure IV-4).

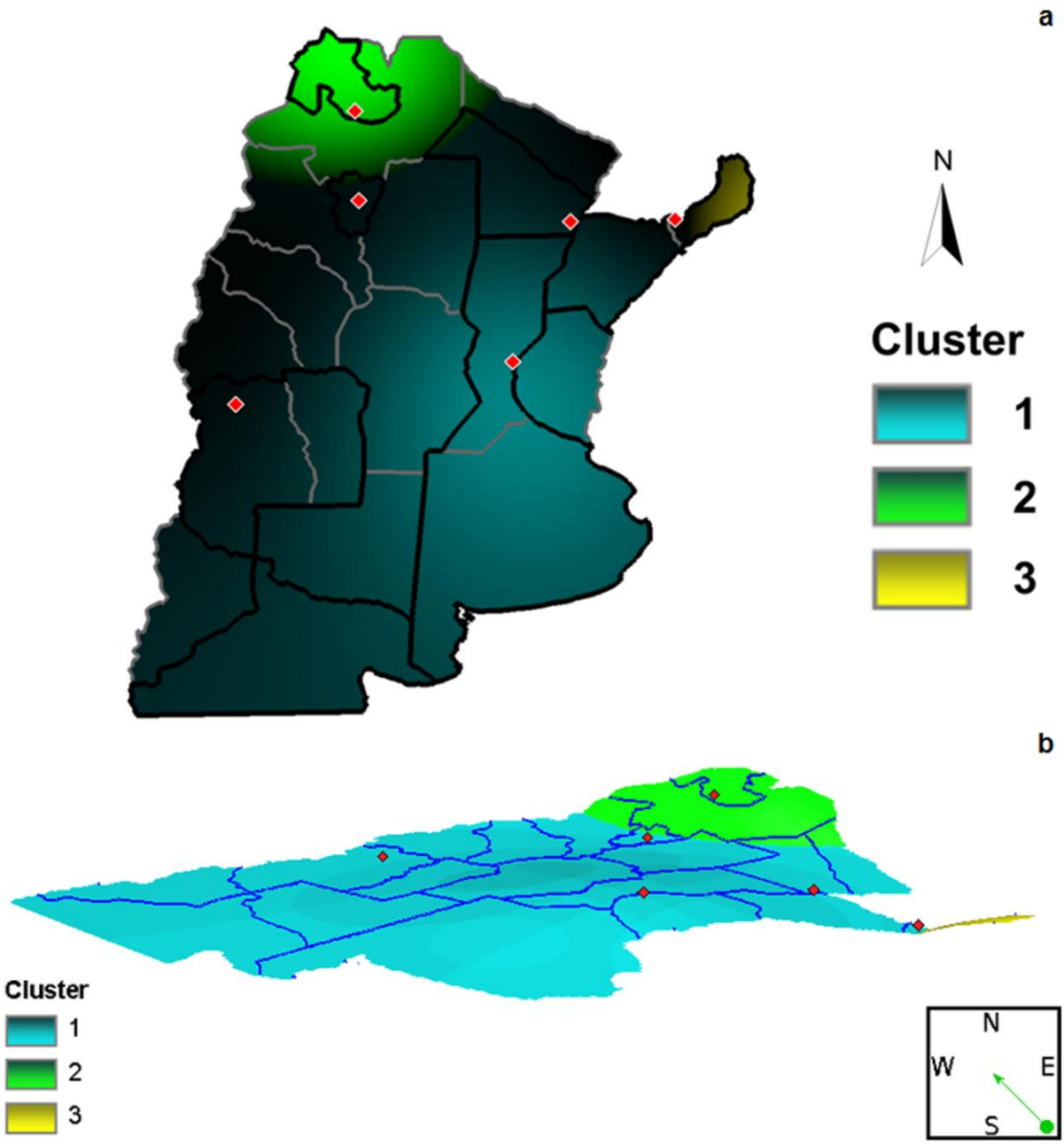

Figure IV-3 Spatial distribution of the most frequent Y-STR haplotype clusters.

(a) Composite map showing the geographical distribution of the most frequent clusters of Y-STR haplotypes; (b) 3d view of the spatial distribution of the most frequent clusters. Lighter shading indicates higher spatial frequency values. Diamonds show capital cities used as georeference of point data.

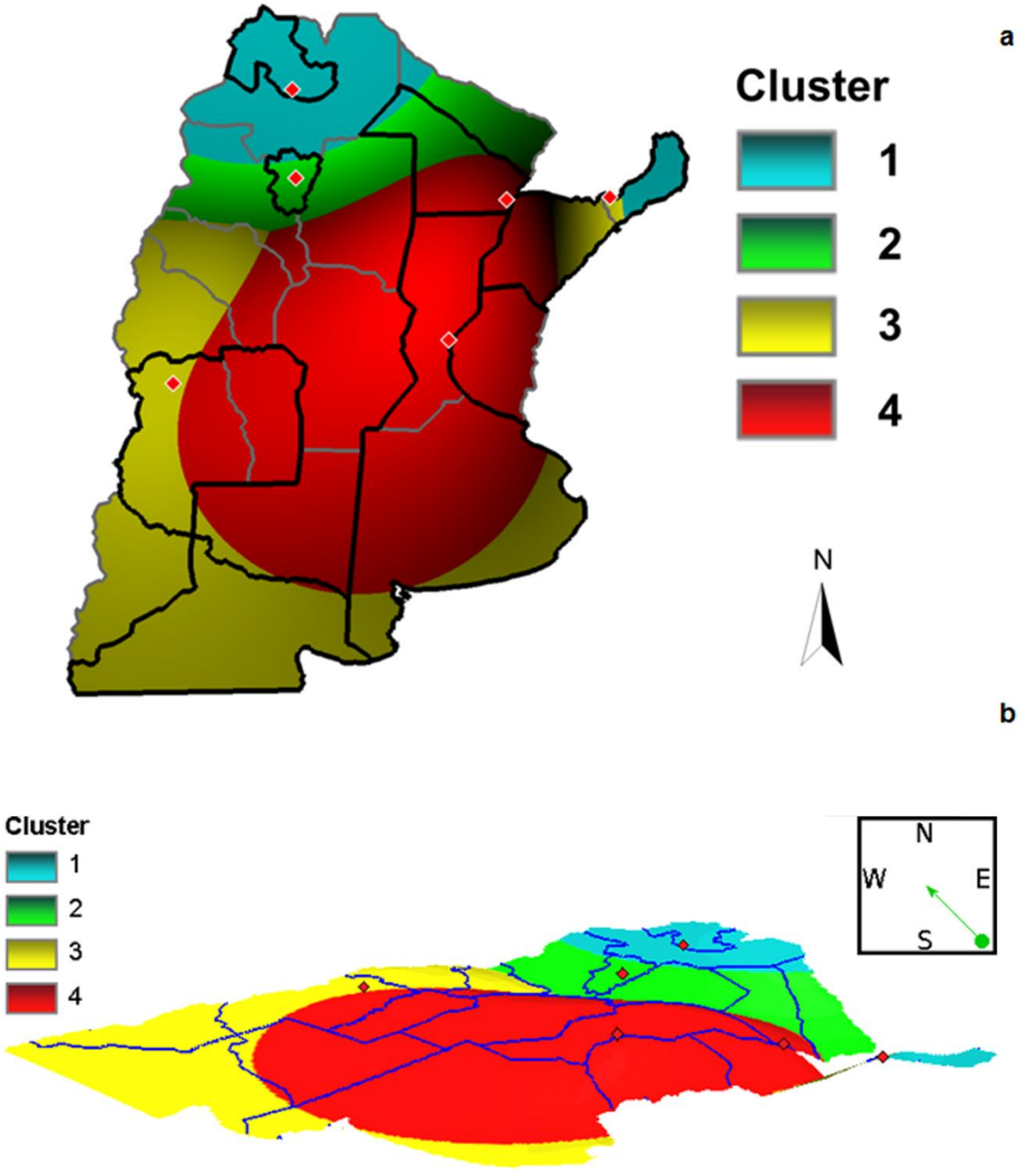

Figure IV-4 Spatial distribution of the second most frequent Y-STR haplotype clusters.

(a) Composite map showing the geographical distribution of the second most frequent clusters of Y-STR haplotypes; (b) 3d view of the spatial distribution of the second most frequent clusters. Lighter shading indicates higher spatial frequency values. Diamonds show capital cities used as georeference of point data.

The highest frequencies of cluster 1 were measured in the Argentine littoral, centered in Santa Fe and surroundings (Figure IV-5 a). Although cluster 1 presented the highest frequencies over the largest area, its maximum frequency values were lower than maximum cluster-2 frequency values, measured in the northwestern region (Figure IV-3 b).

Evaluating the spatial distribution of frequencies in the study region, it was observed that the highest values were registered in the northwest, in the Jujuy Province, where cluster 2 was the most frequent cluster (Figure IV-5). These highest global frequencies are an indication that the northwest, where cluster 2 is the most frequent cluster, is less admixed than the rest of the territory.

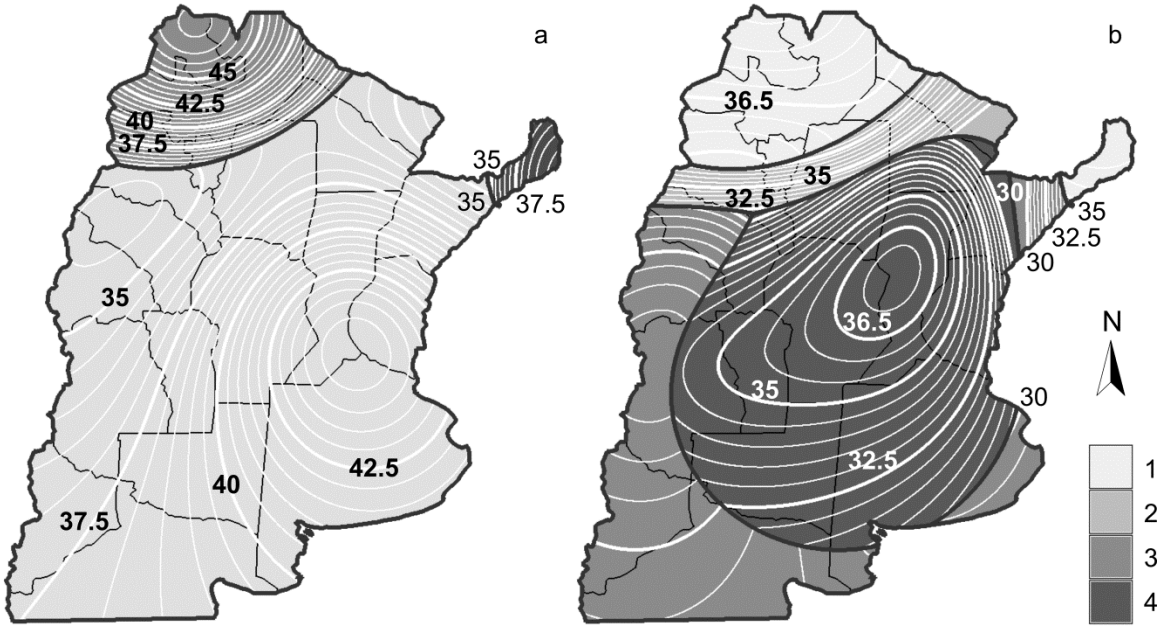

Figure IV-5 Isoline maps showing the spatial distribution of Y-STR haplotype clusters.

(a) Spatial distribution of clusters accounting for the highest frequencies; (b) spatial distribution of clusters accounting for the second highest frequencies. Contour lines (white lines) of cluster frequencies (in percent) have been drawn (thinner line: step=0.5 percent; frequency values corresponding to thicker lines are shown in the map).

Figure IV-6 a shows a profile of frequencies (in percent) along a transect running between the geographical positions accounting for the maximum frequency values of cluster 1 and cluster 2. While cluster-1 frequencies presented a smooth variation pattern, cluster-2 frequencies showed an abrupt decline towards the south. The abrupt decline of cluster 2 towards the south is an indication that the lineages included in this cluster have a relative narrow spatial distribution, centered in the extreme northwest of Argentina.

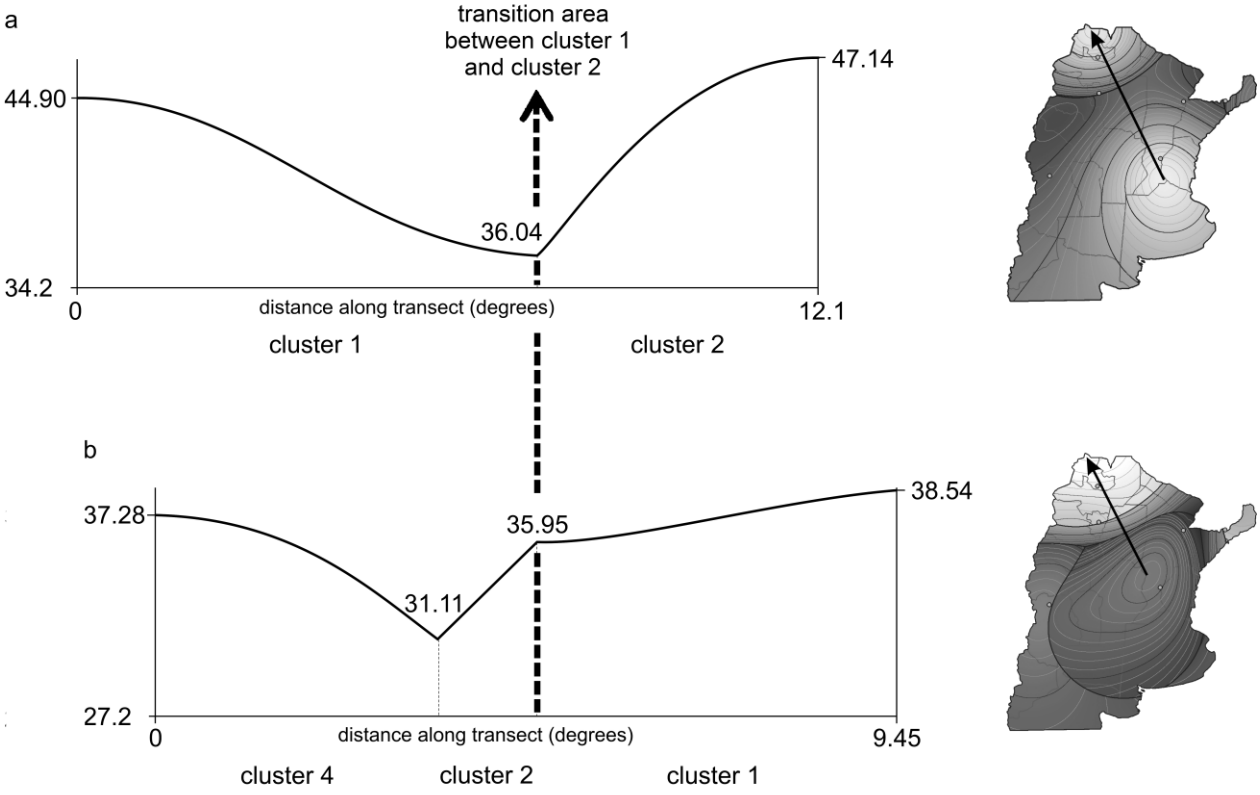

Figure IV-6 Profile of frequencies of Y-STR haplotype clusters along transects.

(a) Transect running between the geographical sites accounting for the maximum spatial frequency values of cluster 1 and cluster 2 (see Figure IV-3); (b) transect running between the geographical sites accounting for the maximum spatial frequency of cluster 4 and the second maximum spatial frequency of cluster 1 (see Figure IV-4). Transects are shown with an arrow in the composite map on the right side (see Figure IV-5). Frequencies (in percent) are indicated on the y-axis. The relative spatial portion of the transect corresponding to each cluster and transect length (in degrees) are indicated on the x-axis.

Figure IV-4 shows on the basis of color shading and in form of a 3d view the spatial distribution of frequencies of the clusters second in frequency in the study region. Similar maximum values were registered in all four regions. Frequencies (in percent), measured along a transect set up between the littoral and the northwestern area, showed a similar smooth pattern of spatial variation of frequencies of cluster 1 and cluster 4 in the corresponding regions (Figure IV-6 b). A smooth spatial variation of frequency values is an indication of the widespread distribution of lineages included in these clusters within Argentina.

As mentioned above, the worldwide distribution of the most frequent haplotypes of cluster 1 (http://www.yhrd.org/) supports an expected European origin. The regional coverage of cluster 1 is in good agreement with previous studies and literature, indicating a major male European component (Avena et al. 2001; Corach et al. 2010; Levene 1992; Rock 1987). Higher frequencies of cluster 1 in the area of Santa Fe (Figure IV-3) correspond well with historical data indicating the Argentine littoral as one of the most important centers of long-lasting European immigration (Rock 1987). On the other hand, lower frequencies of cluster 1 in relation to cluster 2 indicate that the central-littoral region is more admixed (Figure IV-3). This implicates that besides a widespread European heritage, represented in this sample by cluster 1, other male lineages constitute a considerable fraction of the population in this area, grouped primarily into cluster 4 (Figure IV-4). Cluster 4, second in frequency in central and littoral Argentina, included haplotypes incoming from all over the world (http://www.yhrd.org/). These results support an observed multi-ethnic genetic admixture (Avena et al. 2001) in an area which has been a recursive destination of most recent immigration (Rock 1987).

The differentiation of cluster 3 in the northeast, grouping most probably males of European ancestry (http://www.yhrd.org/), is consistent with historical data as well, indicating that a large proportion of European peasants settled in this region (Rock 1987). As soon as the officially promoted immigration decreased, the northeast practically ceased to attract new immigration due to a relative geographical isolation of this region in relation to main industrial and metropolitan centers, located primarily in central and littoral Argentina (Rock 1987). Further studies might more precisely indicate the differential origins of cluster-1 and cluster-3 haplotypes.

The Argentine northwest differentiated strongly from the rest of the study region. As mentioned above, the overall highest cluster frequencies (cluster 2) were measured in the region of Jujuy and Salta (Figure IV-3; Figure IV-5). Such difference in frequencies indicates that the northwest is, on an average, less admixed than other regions. Cluster 2 retained the second highest frequency in the surrounding area to the south including the provinces of La Rioja, Tucuman, northern Chaco and Formosa. As already mentioned, cluster 2 gathered haplotypes frequently registered in South America (http://www.yhrd.org/), indicating a presumable Amerindian origin (Alfaro et al. 2005; Marino et al. 2007). These haplotypes belong most probably to the Amerindian populations inhabiting the northwest. According to census data, Jujuy Province counts Argentina's highest percentage of population (10 percent) identifying him- or herself as an indigenous person or descendant (INDEC 2004-2005). These are predominantly represented by the Kollas (INDEC 2004-2005).

The Kolla community, one of the last Argentine Amerindian populations, survivor of the European conquest, resides mainly in the Jujuy Province and extends with lower frequencies to the south (INDEC 2004-2005). This ethnic group still conserves traditions and habits corresponding to the most advanced culture in the pre-Columbian Argentine territory. The spatial distribution of the Kolla community coincides with the geographical regional coverage where cluster 2 showed the maximum and second maximum frequency values.

Copious studies evaluated the Amerindian contribution to the Argentine genetic composition (Alfaro et al. 2005; Avena et al. 2001; Corach et al. 2010; Marino et al. 2007). While Argentina registers a high proportion of Amerindian maternal heritage (Corach et al. 2010), the paternal Amerindian contribution to the whole population is much lower (Corach et al. 2010; Marino et al. 2007). Specifically concerning the northwest, previous studies support a larger Amerindian contribution (Alfaro et al. 2005; Marino et al. 2007). Our findings further reinforce a predominant indigenous paternal heritage in the northwest, much less diluted by either colonial or modern immigration than in other Argentine regions. The spatial coverage of cluster 2, most probably gathering individuals of Amerindian ancestry, strongly represented in the northwest, is in good agreement with prior expectations.

All in all, analysis conducted at both single-locus and haplotype level, reinforce the notion of Argentina as an admixed country, with a widespread predominance of male European lineages, a strong component of prevalent Amerindian heritage in the northwest and a strongly admixed fraction in central and littoral Argentina.

8 Spatial Admixture of three Sub-Populations

This section tests the performance of the Genetic Geostatistical Framework using Argentine Y-chromosomal STR haplotypes taken from published literature. The selected data sets were previously analyzed using wide-established genetic population methods. This new analysis is based on methods described in s. *5 Geostatistical Analysis of an Admixed Human Population*. On the one side this section illustrates several options of the proposed genetic geostatistical analysis. On the other, findings were contrasted with the previously published results and conclusions. Finally, new insights gained with the application of the method presented in this thesis were inspected.

8.1. Three Argentine Male Data Sets

Individuals were sampled in three locations in central-northern Argentina, located in the Cordoba, Tucuman and Chaco Provinces (Figure IV-7). While Cordoba samples (n = 102) corresponded to healthy unrelated individuals from the general population (Salas et al. 2008), Tucuman and Chaco samples were recruited from two differentiated indigenous communities (Toscanini et al. 2008). Tucuman samples (n = 29) corresponded to a small Kolla community, composed of about 50 middle-size families, widespread in the mountains 50 km away from the Tucuman capital city (San Miguel de Tucuman); these are descendant of the Calchaqui people (originals from the northwestern territories of modern Argentina) admixed with individuals of European origin (Toscanini et al. 2008). The third sample set (n = 49) included individuals recruited from a Toba community, residing in a region called *Chaco impenetrable* (impenetrable Chaco), located between the Teuco and Bermejito rivers (Toscanini et al. 2008).

Each individual was genotyped at seven microsatellite loci (DYS19, DYS389I/II, DYS390, DYS391, DYS392, and DYS39) as reported elsewhere (Salas et al. 2008; Toscanini et al. 2008). Y-STRs were combined into haplotypes according to the YHRD minimal set (http://www.yhrd.org/). Y-chromosome haplogroup inference was taken from the literature (Salas et al. 2008; Toscanini et

al. 2011).

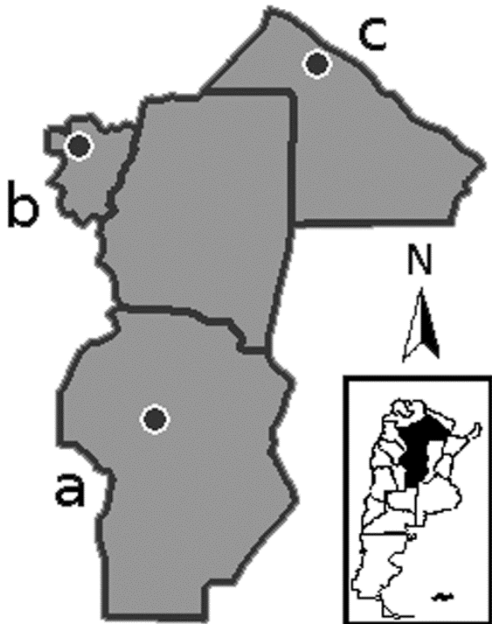

Figure IV-7 Sampling areas.

Map showing the geographical location of the sampling areas (black dot) within the respective province in the Argentine territory: (a) Cordoba Province; (b) Tucuman Province; (c) Chaco Province.

These data sets were considered useful for challenging the Genetic Geostatistical Framework due to two main reasons. First, all three sets have been previously analyzed in full detail with well-established statistical population genetics methods (i.e. Salas et al. 2008; Toscanini et al. 2008; Toscanini et al. 2011). For instance, Toscanini et al. 2008 computed genetic distances among the three sets; plotted the joint genetic relationship of these samples on the basis of widespread genetic distances; and inspected the ancestry of the two indigenous communities with Y-chromosome haplogroups. Ancestry analysis of the urban sample using haplogroups was also performed by Salas et al. (2008). This meticulous in-deep genetic knowledge of the samples, including precise measures of within- and between-group genetic relationships, provides a robust platform for evaluating the accuracy of results gained with a new type of methodology. Second, a relatively simple but differentiated spatial pattern of genetic diversity was expected. According to published

results, the total sample is composed of at least three main genetically differentiated groups or lineages: NW Argentine indigenous lineages, central-north Argentine indigenous lineages, and mixed overseas immigrant lineages (i.e. Salas et al. 2008; Toscanini et al. 2008). Relevant to this experiment is that each lineage is represented with different proportions in the three samples (Salas et al. 2008; Toscanini et al. 2008). Consequently, each one of the three main genetic groups must distribute differentially in the geographical space. Because only three locations were available to create the genetic layers, each layer could only present a spatial pattern of low complexity. Accordingly, the pattern of spatial genetic diversity of the joint sample was expected to remain relatively simple. A simple spatial pattern is of benefit for this experiment because the comparisons between the two types of resutls are straightforward. Noteworthy is that this experiment was not intended to fully describe the current genetic diversity in the geographic area, on the contrary it was attempted to capture the spatial relationship of three groups independently of other actually coexisting groups for purposes of testing the proposed methodology.

8.2. Choice of Number of Y-STR Haplotype Clusters and Interpolation Parameters

Conditioned by the low number of geographical sampling areas (Figure IV-7) and the *a priori* expected minimal number of differentiated groups of lineages (three groups as well; Salas et al. 2008; Toscanini et al. 2008) the appropriate number of clusters for modeling the data was expected to range between three or four clusters. For the purpose of testing the presented framework, to delimit a fourth cluster of haplotypes could be considered well-grounded if the additional cluster presented a differentiated spatial pattern in respect to the first three other haplotype clusters. This situation may arise if for example the overseas immigrant lineages could be split into two largely represented groups, each of both accounting for different frequencies in each one of the three sampling areas. According to the previously published studies related to these samples (Salas et al. 2008; Toscanini et al. 2008; Toscanini et al. 2011) the latter situation was not expected. Nevertheless several runs with three and four final number of clusters, varying the interpolation parameters, were conducted. The explorative runs indicated that delimiting a fourth cluster would only split the group most frequently present in Cordoba, without adding any additional geographic information.

Once the number of final clusters was set to three, further explorative runs were conducted in order to select the appropriate interpolation parameters. Finally, for all three clusters the default tension value (tension=40) and smooth=1 were considered to generate the best spatial pattern fitting the data. The final values were chosen according to the decision criteria detailed in s. *6.5.1 Surface Interpolation*. It should be noted that the chosen smooth parameter value is relatively large in comparison with the default value (smooth=0.1). As it was explained in s. *6.5.1 Surface Interpolation*, setting a larger smooth value is a more appropriate choice when it can be assumed that the truth value at one site is considered to account a large dependence from the truth values of other sites (see Figure III-7). In the spatial genetic context this situation may arise for example when a certain degree of migration is expected as it should be the case of this data set.

8.3. Evaluation of the Spatial Structure of the Data

Due to the low number of clusters and the low number of sampling areas (three in both cases), a relatively simple spatial pattern of genetic differentiation was expected. Simple spatial patterns are of advantage for the purpose of demonstrating methodological potentials on the side of geostatistical analysis and visualization options. Therefore, to the already presented tools in previous sections further techniques were implemented.

The joint spatial distribution of the three clusters was assessed in two and three dimensions (3d). In two dimensions the pattern was analyzed using isolines of cluster frequency. This type of presentation technique may be helpful to obtain a comprehensive understanding of complexity of overlapping spatial distributions. Contour lines were created along isolines of equal frequency values separately for each cluster layer using the GRASS GIS module `r.contour`. In each case a step of 2.5 percent between isolines was used. Isolines of all three clusters were displayed jointly in one map. Cluster membership was distinguished by color. Isoline values were differentiated by tone shading and thickness of the line.

A three-dimensional view of the joint distribution of the three clusters was created with the GRASS module `NVIZ visualization suite`. First, surface layers of all clusters were displayed. Second, cluster frequency values corresponding to the georeferenced point data used for the interpolation procedure were added. Point data were displayed in form of 3d points (icon type:

sphere), where the elevation indicates the frequency values at the site for the corresponding cluster. This combined view may be supportive for visualizing the relative difference of estimated frequencies of the clusters across the study area and observing the deviation at each site between point data and interpolated values. Cluster membership of both surface layers and 3d points were displayed with colors.

Finally, one composite map was created showing the coverage of the most frequent clusters per area and a second composite map displaying the distribution of the second most frequent clusters per area. Values underlying the final composite maps were computed according to the procedure detailed in s. *5.2.6 Screening Algorithms*. Cluster membership was indicated with colors and the spatial gradient of cluster-frequency values was indicated with tone shading.

8.4. Haplotype Frequencies and Clustering Results

The total sample (n = 506) included 142 different haplotypes. The clustering procedure distributed the total amount of different haplotypes unequally among the three clusters, resulting in an even more unequal distribution of samples among the three clusters (Table IV-5). Cluster 1 grouped the large majority of different haplotypes and samples and cluster 2 the smallest proportion of both, different haplotypes and samples.

Haplotypes corresponding to the two major European Y-chromosome haplogroups, clade I and clade R (Karafet et al. 2008), were included to the largest extent in cluster 1 and cluster 2 (Table IV-5). None of the 60 haplotypes corresponding to clade R was grouped under cluster 3. Cluster 3 included over 90 percent of the haplotypes assigned to clade Q, a major lineage among the Native Americans (Karafet et al. 2008). About half of the different haplotypes grouped under cluster 2 corresponded to the clades E, G and J. Clade E can be found at high frequencies in Africa, clade G is present mostly in the Middle East, Mediterranean, and the Caucasus, and clade J was observed at high frequencies in Middle East, North Africa, Europe, Central Asia, Pakistan, and India (Karafet et al. 2008). According to these figures cluster 1 and cluster 2 grouped mainly overseas immigrant lineages while cluster 3 comprised mostly native lineages.

Table IV-5 Frequency distribution of samples and haplotypes per haplogroups and clusters

cluster	E*	G*	I*	J*	Q*	R*	T*	*n* / hap
1	1 / 1	5 / 5	9 / 4	26 / 9	8 / 3	249 / 44	2 / 1	300 / 67
	[0.3]	[1.7]	[3.0]	[8.7]	[2.7]	[83.0]	[0.7]	
2	4 / 4	7 / 5	2 / 2	4 / 4	1 / 1	18 / 16	0	36 / 32
	[11.1]	[19.4]	[5.6]	[11.1]	[2.8]	[50.0]	[0]	
3	7 / 4	0	2 / 2	3 / 2	158 / 35	0	0	170 / 43
	[4.1]	[0]	[1.2]	[1.8]	[92.9]	[0]	[0]	
Total	**12 / 9**	**12 / 10**	**13 / 8**	**33 / 15**	**167 / 39**	**267 / 60**	**2 / 1**	**506 / 142**

This table shows the absolute frequency of samples (*n*) and different haplotypes (*hap*) per haplogroup and cluster; relative sample frequencies (in percent) are indicated in square brackets.

8.5. Geographical Patterns of the three Major Groups

The surface-interpolation step exhibited differentiated spatial patterns for each one of the three clusters. Based on the spatial distribution of isolines of cluster frequencies it was observed that cluster 2 and cluster 3 distributed along a north-south gradient, albeit they followed opposite directions and presented large differences in frequency values (Figure IV-8). Figure IV-9 shows a three dimensional representation of the joint spatial distribution of the interpolated surface of all three clusters and point data values. Minimum and maximum values of the interpolated surfaces, minimum and maximum values of the point sampled data, and the root mean square deviation (*rms*) per cluster layer are shown in Table IV-6.

Cluster 2 had its maximum in Cordoba and cluster 3 in Chaco. While both clusters showed a restricted area of higher frequencies and decreasing frequencies everywhere else, cluster 3 presented a steeper decline (Figure IV-9). Cluster 3 was found to have the highest global frequencies, indicating that Chaco samples were genetically more homogeneous than the rest, and it contributed to a large portion of the Tucuman samples, between 40 and 50 percent.

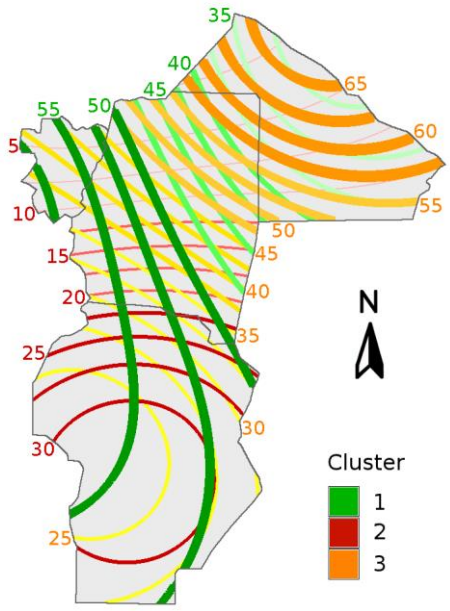

Figure IV-8 Spatial distribution of Y-STR haplotype clusters.

Isolines represent cluster frequencies (in percent).

Table IV-6 Goodness of fit for spatial interpolation of Y-STR haplotype cluster frequencies

Cluster	*rms*	*min [q]*	*max [q]*
1	**5.6**	**30.7 [24.7]**	**59.2 [62.4]**
2	**5.5**	**1.3 [0.0]**	**31.9 [38.3]**
3	**7.0**	**23.1 [17.0]**	**67.1 [75.3]**

This table shows fit goodness for interpolated data on the basis of the root mean square deviation (*rms*) of the interpolation surfaces and the minimum (*min*) and maximum (*max*) values obtained by surface interpolation of *q*, where *q* is the estimated cluster frequency (in percent) at a sampling location (see s. *5.2.2 Computational Procedure for Determining Frequencies per Sampling Unit*).

Consequently, Tucuman samples were found to be composed primarily of cluster 1 and cluster 3 samples. Cluster 1 differed in the pattern distribution of frequencies (Figure IV-8). It showed a relatively flat surface, with moderate decrease in the eastern direction. Similar frequencies were found in Tucuman and Cordoba, where cluster 1, as it was mentioned above, showed the largest

frequencies. In Cordoba, cluster 2 and cluster 3 presented relatively similar frequencies, showing that Cordoba samples were genetically more admixed than the rest (Figure IV-9).

The composite maps created with the screening procedure showed that cluster 1 and cluster 3 are the predominant groups in the study area (Figure IV-10). Cluster 1 was the most predominant group in the largest portion of the area, only excluding the Chaco Province, where was second in frequency. Cluster 3 showed the highest frequencies in the Chaco Province. Cluster 2 was second in frequency in the area of Cordoba.

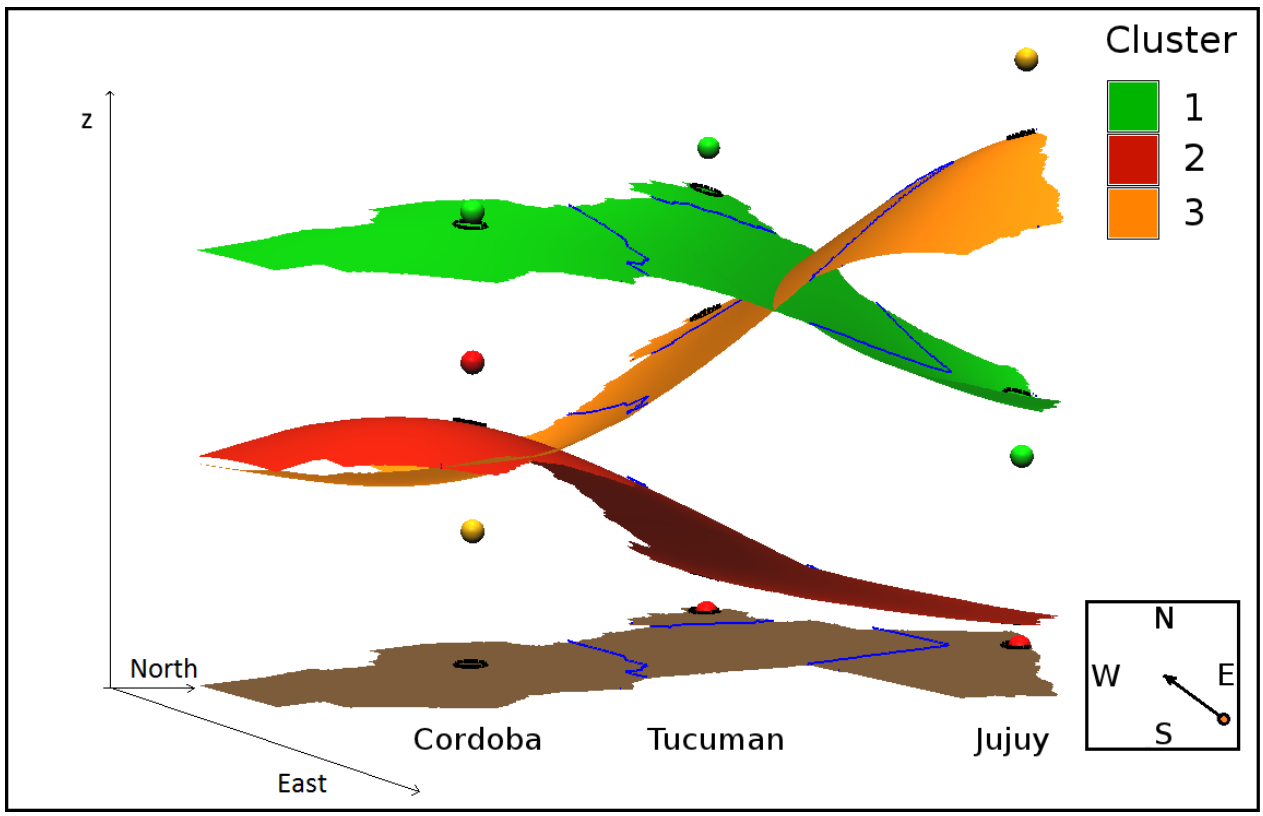

Figure IV-9 Spatial frequency distribution of interpolation and point data values.

Multiple surface 3d view showing the estimated spatial distribution of cluster frequencies (raster layer) in comparison with point data values (dot). Frecuency values are represented as elevation (z) in a 3d space. The ground level (brown raster layer) represents the zero frequency values (z=0). North and east directions are represented in the three-dimensional coordinate system. Provincial borders are indicated with a blue line. The geographical position of sampling sites is indicated with a black circle. The inset with the compass rose shows the cardinal position used to create this 3d perspective.

In the three dimensional representation shown in Figure IV-9 can be observed, that all three interpolation surfaces achieved a relatively good agreement between interpolated and point values. The better agreement was achieved by cluster 1, with more similar point values among the three sites. In this figure it can be verified as well, that the presented method properly identified the spatial ranks of cluster frequencies. Especially remarkable is the case of cluster 2, which presented a positive frequency value only in the sampling area of Cordoba and zero otherwise. The spatial screening of the most frequent groups per area correctly identified this cluster as the one with the second largest frequency within a very restricted area circumjacent to Cordoba interpolation site.

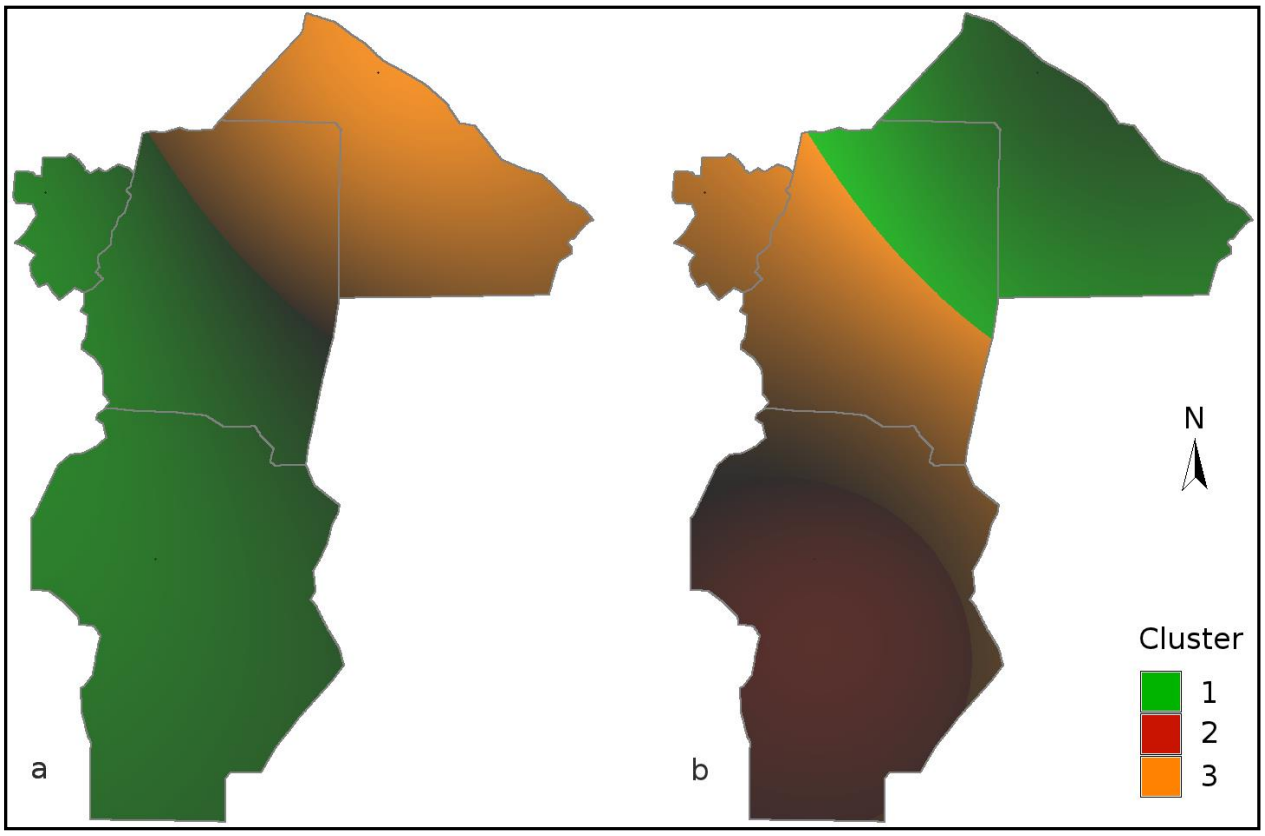

Figure IV-10 Spatial distribution of Y-STR haplotype clusters.

(a) Most frequent clusters; (b) second most frequent clusters. Lighter shading indicates higher spatial frequency values.

A noteworthy result of this data re-analysis is the good agreement between the detected spatial genetic pattern using the Genetic Geostatistical Framework and previously published results obtained with traditional genetic statistics. Salas et al. (2008) demonstrated that the urban population of Cordoba includes several paternal lineages, comprising a majority of admixed overseas lineages and a small proportion of indigenous lineages. According to Toscanini et al. (2011), the two indigenous populations inhabiting Tucuman (Kolla) and Chaco (Toba) differ in the degree of European ancestry. Of both, the Tucuman sample is considerably closer to European populations than to other indigenous populations. Toscanini et al. (2008) compared these three populations along with other published data using neighboring tree and multidimensional scaling plots based on R_{ST} genetic distances. The authors observed that while the urban population of Cordoba was very close to European samples and the Toba population of Chaco to several Amerindian populations, the Kolla population of Tucuman fell in between. All in all, these previous findings are in agreement with the observed spatial genetic pattern presented in this study. The small genetic distance between the Cordoba and Tucuman populations –lead back in the cited studies to the European paternal background- is in agreement with the widespread distribution of cluster 1, including mostly chromosomes of European origin (Table IV-5). The previously reported low proportion of Amerindian chromosomes in the urban population of Cordoba, a higher fraction in the Kolla community, and in high frequency in the Toba population correspond well with the distribution of cluster 3. Cluster 3 included most chromosomes of Amerindian origin (Table IV-5). The higher frequencies of cluster 3 were measured in Chaco, decreasing values in Tucuman, and consistently lower in Cordoba. A rest overseas component, in addition to the European lineages detected by Salas et al. (2008) in the urban sample of Cordoba and in much lower proportion in the two autochthonous communities (Toscanini et al. 2011), is in correspondence with the spatial distribution of cluster 2, which was mainly restricted to the area of Cordoba and included chromosomes of several overseas lineages.

PART V - DISCUSSION AND CONCLUSIONS

9 The Genetic Geostatistical Framework

This work presents an integration of population genetics and geostatistics framed within the powerful geographical information system GRASS GIS. The suitability and potentials of this methodology was investigated on the basis of a case study: the geographical assessment of the genetic diversity in the central-northern urban Argentina using seven forensic Y-chromosomal STRs. Further, an analysis of the genetic diversity of Y-STR haplotypes gained from individuals sampled in three geographically and culturally separated populations, previously characterized on the basis of standardized genetic statistics, was performed. Aim of this second evaluation was to demonstrate the capability of the Genetic Geostatistical Framework to detect spatial genetic heterogeneity and to increase the information content extracted from the data.

A central element of this methodology is to model and to summarize spatial patterns of genetic heterogeneity in form of composite maps. The composite maps are created on the basis of regional frequencies. Regional frequencies are calculated based on several computational steps, each one involving case-specific assumptions. The main steps include: (a) designing groups (clusters) of genetically similar individuals; (b) defining the spatial parameters of analysis and specifying the spatial units required for spatial data aggregation; (c) computing genetic frequencies per group and per spatial unit; (d) selecting spatial interpolation procedures and tuning parameters; (e) creating

composite maps. In each step population genetics as well as social-demographic assumptions are needed. Consequently, a composite map created with this framework is one possible regionalization, an outcome which requires an interpretation in the context of case-specific assumptions and criteria.

The first step of the analysis constitutes the grouping or 'clustering' of genetically similar individuals. In relation to this step several aspects must be observed. Spatial genetic heterogeneity of the male urban Argentine population was investigated at two levels: (a) locus level, i.e. separately for each one of the seven Y-chromosomal STRs; (b) multilocus level, i.e. combing all seven Y-chromosomal STRs. The parameters used for detecting groups of genetic similarity were either the allele, in the first case, or the haplotype, in the latter one. In either case interpretation of results must take into account marker-specific characteristics. Due to their mutation rates, STR alleles are prone to homoplasy in evolutionary times (Forster et al. 2000; Gusmão et al. 2003). This marker-specific feature raises the question if the assumed genetic similarity, e.g., the shared allele or haplotype, is the result of **identical by state** (IBS), i.e. two individuals share an allele or haplotype by chance, or **identical by descent** (IBD), i.e. there is a common source for the shared feature, meaning that the two individuals are biologically related.

Taken into account the high Y-STR mutation rates –0.0051-0.0011 in average for the seven Y-STR loci analyzed in this work (Goedbloed et al. 2009)-, it has to be assumed that there must be a certain degree of IBS at the locus level in the studied sample. This means that each one of the seven Y-STR composite maps by itself may be strictly interpreted as a characterization of present spatial heterogeneity at that locus. Further information and analysis is necessary to raise conclusions related to other topics, for example, to inspect recent or past migration of Y lineages based on single STR.

Low probability of Y-STR haplotype IBS is expected for the study population. Pereira et al. (2003) found that the IBS proportion of pairs of haplotypes differing in either zero or one repeat unit is $\leq$ 0.02 for European populations and $\leq$ 0.01 for an Argentine population. Based on these previous results, it might be reasonable to assume that the studied sample would include a proportion of Y-STR haplotype IBS tending to zero. It must be pointed out that even if there were a certain (small)

proportion of IBS in the sample, excluding those parallel-developed haplotypes leading to IBS would not modify the pattern of spatial heterogeneity presented in the composite maps. The exclusion of few individuals would only affect the pattern of spatial frequency distribution in the areas where the cluster including those haplotypes presents low spatial frequencies. Since the final composite maps show at each spatial location the cluster that accounts for the highest (second highest) frequency, the inclusion or removal of the individuals carrying such haplotypes cannot affect the final composite maps. Consequently, this method is very robust against low proportion of haplotype IBS. Nevertheless, using this method with other data for evolutionary inquiries may require to test the assumption of IBD and, if necessary, to correct for IBD deviation.

The 'clustering' step, i.e. grouping individuals according to genetic similarity, was implemented differently depending on the level of analysis. At the locus level a straightforward and directly interpretable grouping criterion was used: each Y-STR allele detected in the sample determined one group of genetically similar individuals. The final composite maps show the spatial distribution pattern of the most frequent alleles per Y-STR. It must be indicated that the number of groups per locus depends on the number of alleles present in the sample. Examining another sample of the same population may result in a modification of the number of delimited groups per locus since the presence and number of rare alleles would vary from sample to sample. Reducing or enlarging the sample would cause the same effect. The spatial pattern shown by the genetic layers of each allele would vary in areas of low frequency. As explained above, pattern variation in areas of low frequency will not impact the spatial pattern detected in the final composite maps of genetic heterogeneity.

Comparison across loci should take into account marker-specific features, e.g., mutation mechanisms, mutation rate, number of alleles per locus. Joint analysis of different loci may reveal diverse evolutionary developments (Hurles & Jobling 2001). Once again, integrating results referring to several loci may require additional information and deeper spatial analysis. Setting this further spatial analysis within the Genetic Geostatistical Framework will offer the broad spectrum of flexible geostatistic tools of GRASS GIS.

At the haplotype level a more complex grouping strategy was necessary. Grouping genetically

similar individuals based on their haplotypes requires including further genetic assumptions. Y-chromosomal STR haplotypes were grouped according to molecular distance accepting the single-step mutation model (Gusmão et al. 2003). This step involved in the case study seven Y-STR loci. Genetic similarity was determined comparing the number of repeated tandems of nucleotides in a seven dimensional space. An implicit assumption behind this implementation is that a difference of one tandem at one locus is comparable to the same difference at the other loci. Abundant literature confirms that to each Y-STR corresponds an independent, different mutation probability (Gusmão et al. 2005; Goedbloed et al. 2009). Nevertheless and according to previous studies, the variation in genetic differences in respect to a one-tandem-step difference among these loci could be neglected in this case. For instance Gusmão et al. (2003) succeeded in describing the micro-regional variation of this same Y-STR haplotype in a northern region of Spain just pooling one-step neighboring haplotypes. Moreover, the authors concluded that a sound relationship analysis of Y-STR haplotypes requires molecular-distance analysis between haplotypes. Accordingly, it can be fairly hypothesized in reference to the case study that frequency analysis of Y-STR haplotype clusters, grouped according to the single-step mutation model, provides detailed insights about nowadays male lineages cohabiting Argentina.

Data analysis recurrently involved selection of adequate geostatistical tools and parameter tuning. The specific decision-making process of tool selection and parameter specification applied to the case study was explained in detail in *Part III -Materials and Methods*. It must be noted that a common line characterized the whole procedure. On one hand, each step required further information of related fields in order to define case-specific assumptions. This included assumptions regarding main source of genetic variation, spatial pattern of migration and number of genetically similar groups measurable at the defined geographic scale of analysis. On the other hand, parameters were fine tuned in an iterative process in order to maximize regionalization, i.e. to increase the number of regions identified in the composite maps while reducing artifacts (small patches or patterns of narrow, multiple stripes not explainable by the data at the geographical scale of analysis). At last, outcome plausibility was verified in a deductive manner. On the one side results were compared with findings of genetic studies conducted with the same or analogous set

of genetic markers, at similar geographical and time scales. On the other, outcome plausibility was analyzed on the light of demographic, historical, and ethnographic information. After proceeding in such a deductive fashion it could be confirmed that results of both applications to real data presented in this thesis are in clear congruence with genetic, historical and demographic findings.

10 Male Genetic Heterogeneity in Nowadays Argentina[9]

The pattern of genetic admixture of the urban male population in central and northern Argentina was summarized on the basis of composite maps. The most probable origin of the most frequent haplotypes was assessed inspecting worldwide Y-STR haplotype distribution and lineage origin (i.e. search of the estimated origin of the corresponding SNP-haplogroups). In agreement with previous studies (Corach et al. 2010; Marino et al. 2007) findings of this work further support a majority of males with European lineage in the present Argentine population. This heritage is represented in this sample by two groups of closely related haplotypes, cluster 1 and cluster 3, showing the highest frequencies across the largest extension of the study region. A larger sample or different sets of genetic markers would be necessary to more precisely evaluate the difference in origin between these two groups.

Historical, ethnological and census data confirm, explain, and validate the results indicating an exiguous portion of Amerindian component in the largest territories of Argentina and a major component of Amerindian population in the northwest. One of the largest groups of individuals, who identify themselves as indigenous people or descendants, resides in this region (INDEC 2004-2005). This expected spatial structure was detected and quantified on the basis of composite maps at two levels: (a) at the single-locus level, by the repeatedly observed pattern of allelic differentiation in the northwest and, (b) at the haplotype level, by the strong representation of cluster 2 in the northwest, which included haplotypes of presumable Amerindian origin.

It is through the composite maps presented in this work that a geographical estimation of spatial coverage and geographical pattern of frequency variation of one of the largest extant Argentine Amerindian groups, the Kollas, was obtained. The inspection of the three dimensional representation of cluster frequencies and transect charts largely facilitates comparing the degree of

9 This section is partially identical to Diaz Lacava & Walier (2012).

admixture of the Amerindian group in relation to others. Specifically for this group three regions were identified: (a) the northwest, including the provinces of Jujuy and Salta, where this group presents the highest frequency above all groups; (b) a surrounding transition zone, where European lineages present similar frequency; and (c) the territory where its frequency abruptly declines. According to these results, the group represented by this cluster, the Kollas, is the one with less admixture (since it accounts for the overall highest frequency) and it is mostly restricted to the northwest. Consequently, the northwest is the area with lower gene flow of overseas lineages. This finding goes in line with historical sources indicating that European immigration did not affect the Argentine territory equally. According to these sources the northwest was the territory less affected by the large immigration movements populating Argentina (Devoto 2009).

New immigration waves after the 1950's, predominantly affecting central and littoral Argentina (Rock 1987), might have reintroduced Amerindian lineages from neighboring countries and might have also introduced further lineages arriving from all over the world (Avena et al. 2001). In this work such expected strongly admixed group was led back to cluster 4, which gathered haplotypes registered worldwide. With this analysis it was possible to determine that lineages included in this cluster are the second group in frequency in the most populated centers, which are located in central and littoral Argentina. Moreover, on the basis of transects the pattern of spatial variation in frequency of cluster 4 was compared to others. The broad regional coverage and the substantial contribution of this group to the overall genetic background further support observed changes in the genetic composition of littoral Argentina towards a multi-ethnic population (Avena et al. 2001). Taking into account that the sampled provinces gathered 75 percent of the total Argentine population (INDEC 2010) and that the study region extended over 80 percent of the total country, this analysis may be considered representative of the total contemporary male Argentine population.

11 The Geographical Study of Genetic Heterogeneity in Modern Humans

Human genotypes are sampled in the most general case together with certain type of geographical references. Due to ethical constrains precise references, e.g., postal address, may not be available for population genetics studies of modern humans. Therefore, although individual- or allele-based methods may be very useful for analyzing spatial genetic stratification of natural populations (Kelly et al. 2010), are less appropriated for human-population studies. On the other hand, the traditional population-based methods suffer loss of resolution due to two primary simplifications: (a) defining biological populations on the base of sampling locations; (b) averaging individuals or alleles at population level (Kelly et al. 2010). The presented Genetic Geostatistical Framework increases the information gained from the data avoiding the drawbacks of traditional population-based statistics. This methodology neither requires user-defined populations nor demands for parameters specifying the underlying theoretical model of gene flow or drift (Kelly et al. 2010). It is based on the intuitive idea that groups of genetically similar individuals must be first delimited, and afterwards their geographic distribution must be analyzed (Gusmão et al. 2003).

The geographical distribution of groups of genetically similar individuals is modeled in form of 3d surfaces. This procedure is comparable to decompose the total genetic diversity of a region into spatial genetic layers. The overall geostatistical analysis of these genetic layers allows examining main features of the genetic composition of a region and detecting spatial genetic structures. The application to real data showed that these spatial structures could be revealed even in highly admixed populations with considerable gene flow. Juxtaposing the genetic layers in a 3d space substantially facilitates the geographical analysis of group frequencies in a comparative fashion.

Case study results demonstrate that the proposed Genetic Geostatistical Framework successfully exploits the geographic information content of the data and it consistently reveals the spatial genetic structure of an even highly admixed population, i.e. the methodology succeeds in detecting spatial patterns of genetic diversity at relatively high levels of migration. It is specifically convenient for data with low geographic resolution (e.g., geographic references of samples available at the

resolution of sampling locations), and it is sensitive enough to detect spatial genetic structures based on very low number of sampling sites. It is not directly restricted by the type of genetic markers and it can provide solid results even by using a small marker set.
A substantial contribution of this work is to provide tools to search for spatial genetic structures within the frame of GIS. Framing the analysis within a powerful GIS turns the implementation of spatial screening algorithms straightforward. It is self-evident that the numerous spatial tools (e.g., from univariate spatial queries to multivariate geostatistics and complex spatial modeling) considerably increase potentials and flexibility of geographical analysis of genetic diversity. As well as relevant are the manifold GIS reporting tools, which include among others extracting and summarizing spatial data in tables and graphics and creating a large variety of map types.

Thoroughly discussed in population genetics literature are the needs and potentials of applying spatial analysis to statistical genetics (Epperson 2003). With this work it was shown that the Genetic Geostatistical Framework succeeds in merging geostatistics and genetics to meaningfully analyze the genetic diversity of an admixed population. Nonetheless limitations and scope of the proposed Genetic Geostatistical Framework must be considered.
The recently fast growing field of landscape genetics "integrates data and analysis methods from landscape ecology, spatial statistics, geography and population genetics to understand the spatial distribution of genetic variation" (Storfer et al. 2010). A key element of landscape genetics is the incorporation of the landscape matrix as an essential element being considered by the analysis of gene flow and genetic heterogeneity within and among populations (Holderegger & Wagner 2008). For instance, Storfer et al. (2010) classified empirical studies into the field of landscape genetics if they include at least one geographical landscape variable in addition to Euclidean distances. The methodology proposed in this work could be easily extended to fulfill this criterion, or even to include landscape elements. But at the present stage and according to these definitions, it may be more properly assigned to the field of geographical genetics, i.e. a field which emphasizes the use of spatial statistics to study spatial or geographical patterns of genetic variation and the space-time process that produce them (Epperson 2003). Consequently, the presented Genetic Geostatistical Framework and its outcome must be confined to the category of explorative analysis. A relevant

extension of this methodology may involve the computation of statistics to measure the degree of significance of the detected patterns of genetic variation. Furthermore, since this methodology was primarily proposed for population studies of modern humans, a sound analysis extension towards the field of landscape genetics may require to include on the one side data related to spatial socio-demographic variation –for instance, spatial distribution of present and past infrastructure, political, economic, and socio-cultural elements, e.g., language, religion- and data describing natural landscape features on the other. Upon them hypothesis driven analysis could be performed to understand how natural and socio-demographic landscapes affect observed patterns of genetic variation.

12 Conclusions

This work shows that integrating statistical genetics and geostatistics in the framework of GRASS GIS is a successful approach for precisely examining fine-scale spatial patterns of genetic diversity. This task becomes even more challenging in case of high degree of admixture. Most studies using samples collected within modern societies are confronted with strong admixture, as it was presented in the case study. The complex nature of the Argentine genetic structure was revealed at high degree of spatial resolution applying geostatistical interpolation methods (Manel et al. 2003; Storfer et al. 2007).

The strength of this new Genetic Geostatistical Framework is the comparative quantification of spatial coverage of groups of genetically similar individuals in relation to one another. On the basis of only seven Y-STR loci admixture was spatially modeled as the overlapping distribution of coexistent groups. The urban male population of central and northern Argentina was decomposed into four major groups: two groups including European lineages, one including a strong component of Amerindian heritage and a rest one, including lineages of worldwide origin. The overlapping geographical coverage and differential spatial frequency of these groups was summarized in the final composite maps. All in all this study provides a further confirmation of the ethnic pluralistic composition of the urban male Argentine population as well as the differentiated regional impact of historical and modern migration.

On the basis of a case study and data re-analysis it has been demonstrated that the systematic, comprehensive search for spatially overlapping structures of genetic variation allows the identification of spatially coherent regions populated by genetically closely related individuals as well as the comparative quantification of differentiated degree of admixture among groups and regions. Further advantages of this kind of analysis are twofold. First, as it was discussed in Barbujani (2000), spatial findings reported in form of maps are intuitive to understand and provide the most direct way to evaluate geographical relationships. Second, accurate geostatistical characterization of the genetic diversity of a region provides an optimal basis for further

evaluations. These may include: within-group heterogeneity, quantification of the degree of genetic relationship among detected groups, connectivity analysis (detection of barriers, pathways, and corridors), spatial and temporal lag as well as performing explicit tests of the effects of geographical variables on gene flow and genetic composition of populations. In addition, results may be straightforwardly summarized in form of statistics or charts.

It is the hope of the author that the presented Genetic Geostatistical Framework as well as the spatial characterization of the male Argentine population, further validated by historical, ethnic, and census data, will provide a basis for future works ranging from investigations in forensic genetics, population genetics, genetic epidemiology, to studies in closely related areas such as ethnology, history, or demographic surveys.

PART VI -REFERENCES

Alfaro EL, Dipierri JE, Gutiérrez NI, Vullo CM (2005) Genetic structure and admixture in urban populations of the Argentine North-West. Ann Hum Biol 32(6) 724-737.

Alves I, Sˇra´mkova´ Hanulova´ A, Foll M, Excoffier L (2012) Genomic data reveal a complex making of humans. PLoS Genet 8(7) e1002837.

Avena S, Goycoechea A, Dugoujon J, Slepoy M, Slepoy A, Carnese FR (2001) Análisis antropogenético de los aportes indígena y africano en muestras hospitalarias de la ciudad de Buenos Aires [Anthropogenic analysis of the indigenous and African contribution in hospital samples of Buenos Aires cities]. Revista Argentina de Antropología Biol 3(1) 79-99.

Balkenhol N, Gugerli F, Cushman S, Waits L, Coulon A, Arntzen JW, Holderegger R, Wagner H (2009) Identifying future research needs in landscape genetics: where to from here? Landsc Ecol 24 455-463.

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72 578-589.

Barbujani G (1985) A two-step test for the heterogeneity of F_{ST} values at different loci. Hum Hereditas 35(5) 292-295.

Barbujani G (1987) Autocorrelation of gene frequencies under isolation by distance. Genetics 117(4) 777-82.

Barbujani G (2000) Geographic patterns: how to identify them and why. Hum Biol 72 133-153.

Barbujani G, Belle EM (2006) Genomic boundaries between human populations. Hum Hered 61

15-21.

Barbujani G, Colonna V (2010) Human genome diversity: frequently asked questions. Trends Genet 26(7) 285-95.

Barbujani G, Jacquez GM, Ligi L (1990) Diversity of some gene frequencies in European and Asian populations V. Steep multilocus clines. Am J Hum Genet 47 867-875.

Barbujani G, Oden NL, Sokal RR (1989) Detecting regions of abrupt change in maps of biological variables. Syst Zool 38 376-389.

Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 87(5) 1816-1819.

Bartolome MB (1976) Argentinien. In: Walter D (ed) Die Situation der Indios in Südamerika. Grundlagen der interethnischen Konflikte der nichtandinen Indianer [The situation of the Indians in South America. Fundamentals of inter-ethnic conflicts of the non-Andean Indians] (pp 352-398). Wuppertal: Hammer Verlag.

Benecia R (2009) Apéndice: La Inmigración limítrofe [Appendix: Immigration from Neuighbouring Countries]. In: Devoto F (ed) Historia de la Inmigración en la Argentina [History of Immigration in Argentina] (third edition) (pp 433-484). Buenos Aires: Editorial Sudamericana.

Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68(4) 1019-1029.

Brion M, Quintans B, Zarrabeitia M, Gonzalez-Neira A, Salas A, Lareu V, Tyler-Smith C, Carracedo A (2004) Micro-geographical differentiation in Northern Iberia revealed by Y-chromosomal DNA analysis. Gene 329 17-25.

Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet Suppl 266-75.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton NJ: Princeton University Press.

Corach D, Filgueira Risso L, Marino M, Penacino G, Sala A (2001) Routine Y-STR typing in forensic casework. Forensic Sci Int 118(2-3) 131-5.

Corach D, Lao O, Bobillo C, van Der Gaag K, Zuniga S, Vermeulen M, van Duijn K, Goedbloed M, Vallone PM, Parson W, de Knijff P, Kayser M (2010) Inferring continental ancestry of Argentineans from autosomal, Y-chromosomal and mitochondrial DNA. Ann Hum Genet 74(1) 65-76.

de Knijff P (2000) Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am J Hum Genet 67 1055-61.

Devoto F (2009) Historia de la Inmigración en la Argentina [History of Immigration in Argentina] (third edition). Buenos Aires: Editorial Sudamericana.

Diaz Lacava A, Walier M (2012) Forensic genetics from a geographic perspective: integrating forensic genetics and geostatistics. In: Yacine N, Fellag R (eds) Forensic Science (pp 59-83). New York: Nova Science Publishers.

Diaz Lacava A, Walier M, Penacino G, Wienker TF, Baur MP (2007) Forensic Genotypes and Genetic Landscapes of Extant Argentinean Population. Poster presented at the 22nd Congress of the International Society for Forensic Genetics (ISFG), Tivoli Garden, 2007 21–25 Sep, Copenhagen.

Diaz Lacava A, Walier M, Penacino G, Wienker TF, Baur MP (2011a) Spatial assessment of Argentinean genetic admixture with geographical information systems. Forensic Sci Int Genet 5 297-302.

Diaz Lacava A, Walier M, Willuweit S, Wienker TF, Fimmers R, Baur MP, Roewer L (2011b) Geostatistical inference of main Y-STR-haplotype groups in Europe. Forensic Sci Int Genet 5 91-94.

Dipierri JE, Alfaro EL, Scapoli C, Mamolini E, Rodriguez-Larralde A, Barrai I (2005) Surnames in Argentina: a population study through isonymy. Am J Phys Anthropol 128(1) 199-209.

Dobzhansky T (1970) Genetics of the Evolutionary Process. New York: Columbia University Press.

Dobzhansky T, Wright S (1941) Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of Drosophila pseudoobscura. Genetics 26 23-51.

Encyclopædia Britannica (2012a) Argentina. Encyclopædia Britannica Online. Retrieved from

http://www.britannica.com/EBchecked/topic/33657/Argentina

Encyclopædia Britannica (2012b) Aymara. Encyclopædia Britannica Online. Retrieved from http://www.britannica.com/EBchecked/topic/46515/Aymara

Epperson BK (2003) Geographical Genetics. Princeton NJ: Princeton University Press.

Fechner A, Quinque D, Rychkov S, Morozowa I, Naumova O, Schneider Y, Willuweit S, Zhukova O, Roewer L, Stoneking M, Nasidze I (2008) Boundaries and clines in the West Eurasian Y-chromosome landscape: insights from the European part of Russia. Am J Phys Anthropol 137(1) 41-47.

Fischer A, Pollack J, Thalmann O, Nickel B, Pääbo S (2006) Demographic history and genetic differentiation in apes. Curr Biol 16(11) 1133-8.

Forster P, Röhl A, Lünnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B (2000) A short tandem repeat-based phylogeny for the human Y chromosome. Am J Hum Genet 67(1) 182-96.

Futuyma DJ (1997) Evolutionary biology. Sunderland: Sinauer.

Gagneux P, Wills C, Gerloff U, Tautz D, Morin PA, Boesch C, Fruth B, Hohmann G, Ryder OA, Woodruff DS (1999) Mitochondrial sequences show diverse evolutionary histories of African hominoids. Proc Natl Acad Sci USA 96 5077-5082.

Goedbloed M, Vermeulen M, Fang RN, Lembring M, Wollstein A, Ballantyne K, Lao O, Brauer S, Krüger C, Roewer L, Lessig R, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M (2009) Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR® Yfiler® PCR amplification kit. Int J Legal Med 123(6) 471-482.

Goldstein DB, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. Annu Rev Genomics Hum Genet 3 129-52.

Guillot G, Estoup A, Mortier F, Cosson JF (2005a) A spatial statistical model for landscape genetics. Genetics 170 1261-1280.

Guillot G, Leblois R, Coulon A, Frantz AC (2009) Statistical methods in spatial genetics. Mol Ecol 18 4734-4756.

Guillot G, Mortier F, Estoup A (2005b) Geneland: A program for landscape genetics. Mol Ecol

Notes 5 712-715.

Gusmão L, Sánchez-Diz P, Alves C, Beleza S, Lopes A, Carracedo A, Amorim A (2003) Grouping of Y-STR haplotypes discloses European geographic clines. Forensic Sci Int 134(2-3) 172-179.

Gusmão L, Sánchez-Diz P, Calafell F, Martín P, Alonso CA, Alvarez-Fernández F, Alves C, Borjas-Fajardo L, Bozzo WR, Bravo ML, Builes JJ, Capilla J, Carvalho M, Castillo C, Catanesi CI, Corach D, Di Lonardo AM, Espinheira R, Fagundes de Carvalho E, Farfán MJ, Figueiredo HP, Gomes I, Lojo MM, Marino M, Pinheiro MF, Pontes ML, Prieto V, Ramos-Luis E, Riancho JA, Souza Góes AC, Santapa OA, Sumita DR, Vallejo G, Vidal Rioja L, Vide MC, Vieira da Silva CI, Whittle MR, Zabala W, Zarrabeitia MT, Alonso A, Carracedo A, Amorim A (2005) Mutation rates at Y chromosome specific microsatellites. Hum Mutat 26(6) 520-528.

Haldane JB (1940) The blood-group frequencies of the European peoples and racial origins. Hum Biol 12(4) 457.

Handley LJ, Manica A, Goudet J, Balloux F (2007) Going the distance: human population genetics in a clinal world. Trends Genet 23(9) 432-9.

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2 618-620.

Hart DL, Clark AG (1997) Principles of Population Genetics. Sunderland MA: Sinauer Associates Sinauer.

Holderegger R, Kamm U, Gugerli F (2006) Adaptive vs. neutral genetic diversity:

Implications for landscape genetics. Landsc Ecol 21 797–807.

Holderegger, Wagner (2006) A brief guide to landscape genetics. Land Eco 21 793-796.

Holderegger R, Wagner HH (2008) Landscape genetics. BioScience 58 199–208.

Hurles ME, Jobling MA (2001) Haploid chromosomes in molecular ecology: lessons from the human Y. Mol Ecol 10(7) 1599-613.

INDEC Argentina (1997) La Migración Internacional en la Argentina: sus Características e Impacto [The International Migration in Argentina: Characteristics and Impact]. Estudios INDEC 29. Buenos Aires: Instituto Nacional de Estadistíca y Censo.

INDEC Argentina (2001) Censo Nacional de Población, Hogares y Viviendas 2001 [National Population, Household and Housing Census of 2001]. Instituto Nacional de Estadistíca y

Censo. Retrieved from http://www.indec.mecon.gov.ar

INDEC Argentina (2004-2005) Resultados de la Encuesta Complementaria de Pueblos Indígenas (ECPI) 2004-2005 - Complementaria del Censo Nacional de Población, Hogares y Viviendas 2001 [Results of the Ssupplementary Survey of Indigenous Peoples (ECPI) 2004-2005 - Complement of the National Census of Population, Household and Housing 2001]. Instituto Nacional de Estadistíca y Censo. Retrieved from http://www.indec.gov.ar

INDEC Argentina (2010) Censo Nacional de Población, Hogares y Viviendas 2010 [National Population, Household and Housing Census of 2010]. Instituto Nacional de Estadistíca y Censo. Retrieved from http://www.censo2010.indec.gov.ar

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011) 931-45.

Jobling MA, Tyler-Smith C (2000) New uses for new haplotypes the human Y chromosome, disease and selection. Trends Genet 16(8) 356-62.

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4(8) 598-612.

Jombart T, Pontier D, Dufour AB (2009) Genetic markers in the playground of multivariate analysis. Hered 102(4) 330-41.

Jongman RHG, Braak CJFT, Tongeren OFRV (eds) (1995) Data Analysis in Community and Landscape Ecology. Cambridge: Cambridge University Press.

Jorde LB (1985) Human genetic distance studies: present status and future prospects. Annu Rev Anthropol 14 343-373.

Kaessmann H, Wiebe V, Weiss G, Pääbo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27(2) 155-6.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18(5) 830-8.

Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V, Gehrig C, Bernini LF, Jespersen J, Bakker E, Roewer L, de Knijff P (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet 68(4)

990-1018.

Kayser M, Lao O, Anslinger K, Augustin C, Bargel G, Edelmann J, Elias S, Heinrich M, Henke J, Henke L, Hohoff C, Illing A, Jonkisz A, Kuzniar P, Lebioda A, Lessig R, Lewicki S, Maciejewska A, Monies DM, Pawłowski R, Poetsch M, Schmid D, Schmidt U, Schneider PM, Stradmann-Bellinghausen B, Szibor R, Wegener R, Wozniak M, Zoledziewska M, Roewer L, Dobosz T, Ploski R (2005) Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. Hum Genet 117(5) 428-443.

Kelly RP, Oliver TA, Sivasundar A, Palumbi SR (2010) A method for detecting population genetic structure in diverse, high gene-flow species. J Hered 101(4) 423-36.

Kidd KK, Pakstis AJ, Speed WC, Kidd JR (2004) Understanding human DNA sequence variation. J Hered 95(5) 406-20.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18(16) 1241-8.

Levene R (1951) Las Indias no eran Colonias [The Indies were not Colonies]. Buenos Aires: Espasa Calpe.

Levene R (1992) Lecciones de Historia Argentina [Lessons of Argentine History] (25th edition, vol I). Buenos Aires: Ediciones Corregidor.

Levene GG (2002) Argentina se hizo así [So arose Argentina]. Buenos Aires: Distal.

Malécot G (1948) Les Mathématiques de l'Hérédité [Mathematics of Inheritance]. Paris: Masson.

Malécot G (1950) Quelques Schémas Probabilistes sur la Variabilité des Populations Naturelles [Some Probabilistic Patterns on the Variability of Natural Populations]. Annales de l'Université de Lyon A13 37-60.

Manel S, Berthoud F, Bellemain E, Gaudeul M, Luikart G, Swenson JE, Waits LP, Taberlet P, Intrabiodiv Consortium (2007) A new individual-based spatial approach for identifying genetic

discontinuities in natural populations. Mol Ecol 16(10) 2031-2043.

Manel S, Schwartz M, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol 18 189-197.

Manica M, Prugnolle F, Balloux F (2005) Geography is a better determinant of human genetic differentiation than ethnicity. Hum Genet 118 366-371.

Manni F, Guerard E, Heyer E (2004) Geographic patterns of (genetic, morphologic, linguistic) variation: How barriers can be detected by using Monmonier's algorithm. Hum Biol 76 173-190.

Marino M, Sala A, Corach D (2007) Genetic attributes of the YHRD minimal haplotype in 10 provinces of Argentina. Forensic Sci Int Genet 1(2) 129-133.

Marino M, Sala A, Bobillo C, Corach D (2008) Inferring genetic sub-structure in the population of Argentina using fifteen microsatellite loci. Forensic Sci Int Genet 1(1) 350-352.

Mitasova H, Mitas L (1993) Interpolation by regularized spline with tension: I. Theory and implementation. Math Geol 25 641-655.

Monmonier M (1973) Maximum-difference barriers: an alternative numerical regionalization method. Geog Anal 3 245-261.

Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37(1) 17-23.

Murphy MA, Evans JS, Cushman S, Storfer A (2008) Representing genetic variation as continuous surfaces: An approach for identifying spatial dependency in landscape genetic studies. Ecography 31 685-697.

Neteler M, Mitasova H (2004) Open Source GIS: A GRASS GIS Approach (second edition). Boston: Kluwer Academic Publishers/Springer.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456(7218) 98-101.

Pannell JR, Charlesworth B (2000) Effects of metapopulation processes on measures of genetic diversity. Philos Trans R Soc London Ser B 355 1851-64.

Pereira L, Prata MJ, Amorim A (2003) An evaluation of the proportion of identical Y-STR haplotypes due to recurrent mutation. International Congress Series 1239 57-60.

Piazza A, Rendine S, Minch E, Menozzi P, Mountain J, Cavalli-Sforza LL (1995) Genetics and the origin of European languages. Proc Natl Acad Sci USA 92(13) 5836-40.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102 15942-15947.

Rock D (1987) Argentina, 1516-1987: from Spanish Colonization to Alfonsín. Los Angeles: University of California Press.

Roewer L (2001) Die Haplotypisierung des Y-Chromosoms - Grundlagen und Anwendungen einer neuen molekulargenetischen Identifizierungsmethode [The Haplotyping of the Y chromosome - Fundamentals and Applications of a New Molecular Genetic Identification Method] (habilitation). Humboldt-Universität zu Berlin. Retrieved from http://edoc.hu-berlin.de/habilitationen/roewer-lutz-2001-05-29/HTML

Roewer L, Croucher PJ, Willuweit S, Lu TT, Kayser M, Lessig R, de Knijff P, Jobling MA, Tyler-Smith C, Krawczak M (2005) Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum Genet 116(4) 279-291.

Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A, Anslinger K, Augustin C, Betz A, Bosch E, Cagliá A, Carracedo A, Corach D, Dekairelle AF, Dobosz T, Dupuy BM, Füredi S, Gehrig C, Gusmaõ L, Henke J, Henke L, Hidding M, Hohoff C, Hoste B, Jobling MA, Kärgel HJ, de Knijff P, Lessig R, Liebeherr E, Lorente M, Martínez-Jarreta B, Nievas P, Nowak M, Parson W, Pascali VL, Penacino G, Ploski R, Rolf B, Sala A, Schmidt U, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Kayser M (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int 118(2-3) 106-13.

Romero JL (2011) Breve historia de la Argentina [Short History of Argentina] (fifth edition, 10th reprinting). Buenos Aires: Fondo de Cultura Económica.

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, Beckman G, Beckman L, Bertranpetit J, Bosch E, Bradley DG, Brede G, Cooper G, Corte-Real HB, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Golge M, Hill EW, Jeziorowska A, Kalaydjieva L, Kayser M, Kivisild T, Kravchenko

SA, Krumina A, Kucinskas V, Lavinha J, Livshits LA, Malaspina P, Maria S, McElreavey K, Meitinger TA, Mikelsaar AV, Mitchell RJ, Nafa K, Nicholson J, Norby S, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, Previdere C, Roewer L, Rootsi S, Rubinsztein DC, Saillard J, Santos FR, Stefanescu G, Sykes BC, Tolun A, Villems R, Tyler-Smith C, Jobling MA (2000) Y chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67 1526-1543.

Sala A, Penacino G, Corach D (1998) Comparison of allele frequencies of eight STR loci from Argentinian Amerindian and European populations. Hum Biol 70(5) 937-947.

Salas A, Jaime JC, Alvarez-Iglesias V, Carracedo A (2008) Gender bias in the multiethnic genetic composition of central Argentina. J Hum Genet 53(7) 662-74.

Sanchez-Albornoz N (1994) La Poblacion de America latina: desde los tiempos precolombinos al año 2005 [The Population of Latin America: from pre-Columbian times to 2005]. Madrid: Editorial Alianza.

Schlötterer C (2004) The evolution of molecular markers - just a matter of fashion? Nat Rev Genet 5(1) 63-9.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60 957-964.

Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium. Evolution 47 264-279.

Sokal RR (1988) Genetic, geographic, and linguistic distances in Europe. Proc Natl Acad Sci USA 85(5) 1722-6.

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology 2. Some implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society 10 229-249.

Sokal RR, Rohlf FJ (1995) Biometry. San Francisco CA: Freeman.

Storfer A, Murphy MA, Evans JS, Goldberg CS, Robinson S, Spear SF, Dezzani R, Delmelle E, Vierling L, Waits LP (2007) Putting the "landscape" in landscape genetics. Heredity 98(3) 128-142.

Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP (2010) Landscape genetics: where are we now? Mol Ecol 19(17) 3496-514.

Toscanini U, Gusmão L, Berardi G, Amorim A, Carracedo A, Salas A, Raimondi E (2008) Y chromosome micro-satellite genetic variation in two Native American populations from Argentina: Population stratification and mutation data. Forensic Sci Int Genet 2 274-280.

Toscanini U, Gusmão L, Berardi G, Amorim A, Carracedo A, Salas A, Raimondi E (2007) Testing for genetic structure in different urban Argentinian populations. Forensic Sci Int 165(1) 35-40.

Toscanini U, Gusmão L, Berardi G, Gomes V, Amorim A, Salas A, Raimondi E (2011) Male lineages in South American native groups: evidence of M19 traveling south. Am J Phys Anthropol 146(2) 188-196.

Turner MG, Gardner RH, O'Neill RV (2001) Landscape Ecology in Theory and Practice: Pattern and Process. New York: Springer.

Vandergast AG, Perry WM, Lugo RV, Hathaway SA (2011) Genetic Landscapes GIS Toolbox: tools to map patterns of genetic divergence and diversity. Mol Ecol Resour 11(1) 158-61.

Venter JC, Adams MD, Myers EW, et al. (2001) The sequence of the human genome. Science 291 1304–1351.

Wijsman EM (1984) Techniques for estimating genetic admixture and applications to the problem of the origin of the Icelanders and the Ashkenazi Jews. Hum Genet 67 441-448.

Womble WH (1951) Differential systematics. Science 114(2961) 315-22.

Wooding S, Ostler C, Prasad BV, Watkins WS, Sung S, Bamshad M, Jorde LB (2004) Directional migration in the Hindu castes: Inferences from mitochondrial, autosomal and Y-chromosomal data. Hum Genet 115 221-229.

Wright S (1931) Evolution in Mendelian populations. Genetics 16 97-159.

Wright S (1943) Isolation by distance. Genetics 28 114–138.

Zerjal T, Beckman L, Beckman G, Mikelsaar AV, Krumina A, Kucinskas V, Hurles ME, Tyler-Smith C (2001) Geographical, linguistic, and cultural influences on genetic diversity: Y-chromosomal distribution in Northern European populations. Mol Biol Evol 18(6) 1077-1087.

EIDESSTATTLICHE ERKLÄRUNG

An Eides statt versichere ich, dass ich die Dissertation „Geostatistical Analysis of Genetic Diversity in the Present Male Argentine Population“ selbst und ohne jede unerlaubte Hilfe angefertigt habe. Des Weiteren erkläre ich, dass diese oder eine ähnliche Arbeit noch keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht wurde.
Teile dieser Arbeit wurden in den unten aufgelisteten Originalpublikationen veröffentlicht.

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.
Some sections of this thesis were published in the original publications listed below.

Diaz Lacava A, Walier M (2012) Forensic genetics from a geographic perspective: integrating forensic genetics and geostatistics. In: Yacine N, Fellag R (eds) Forensic Science (pp 59-83). New York: Nova Science Publishers.

Diaz Lacava A, Walier M, Penacino G, Wienker TF, Baur MP (2007) Forensic Genotypes and Genetic Landscapes of Extant Argentinean Population. Poster presented at the 22nd Congress of the International Society for Forensic Genetics (ISFG), Tivoli Garden, 2007 21–25 Sep, Copenhagen.

Diaz Lacava A, Walier M, Penacino G, Wienker TF, Baur MP (2008) Forensic genotypes and genetic landscapes of extant Argentinean population. 22nd Congress of the International Society for Forensic Genetics Copenhagen 2007. Forensic Science International: Genetics Supplement Series 1(1): 326-328.

Diaz Lacava A, Walier M, Penacino G, Wienker TF, Baur MP (2011a) Spatial assessment of

Argentinean genetic admixture with geographical information systems. Forensic Sci Int Genet 5 297-302.

Diaz Lacava A, Walier M, Willuweit S, Wienker TF, Fimmers R, Baur MP, Roewer L (2011b) Geostatistical inference of main Y-STR-haplotype groups in Europe. Forensic Sci Int Genet 5 91-94.