Renan Barbosa Lemes

INBREEDING STUDIES IN A *QUILOMBO* ISOLATE FROM THE STATE OF SÃO PAULO

ESTUDOS SOBRE ENDOCRUZAMENTO EM UM ISOLADO QUILOMBOLA DO ESTADO DE SÃO PAULO

São Paulo

2017

Renan Barbosa Lemes

INBREEDING STUDIES IN A *QUILOMBO* ISOLATE FROM THE STATE OF SÃO PAULO

ESTUDOS SOBRE ENDOCRUZAMENTO EM UM ISOLADO QUILOMBOLA DO ESTADO DE SÃO PAULO

Dissertation Presented in a Partial Fulfillment of the Requirements for the Degree of Doctor of Science (Biology/Genetics) in the Institute of Biosciences of the University of São Paulo.

Advisor: Prof. Dr. Paulo A. Otto

São Paulo

2017

Catalog Record

Lemes, Renan Barbosa

Inbreeding Studies in a Quilombo Isolate from the State of São Paulo
78 pages

PhD Dissertation - Institute of Biosciences of the University of São Paulo. Department of Genetics and Evolutionary Biology.

Palavras-chave: Endocruzamento; Coeficiente de endocruzamento; Índice de fixação de Wright; *Runs of Homozygosity*; Isolado populacional.
Keywords: Inbreeding; Inbreeding coefficient; Wright's fixation index; Runs of homozygosity; Population isolate.
I. University of São Paulo. Institute of Biosciences. Department of Genetics and Evolutionary Biology.

Examination board:

____
Prof. Dr.

____
Prof. Dr.

____
Prof. Dr.

____
Prof. Dr.

____
Prof. Dr.
Paulo Alberto Otto
Advisor

To Juliana Carnavalli and
to my family for their
support.

"All we have to decide is what to do with the time that is given to us."

J. R. R. Tolkien

Acknowledgements

To my advisor Prof. Dr. Paulo A. Otto for guiding me safely through the path of science during the last eight years.

To Professors Diogo Meyer and Regina Célia Mingroni Netto for their valuable collaboration.

To Juliana Carnavalli, Kelly Nunes, and Lilian Kimura for the assistance in the sample preparation and for their friendship.

To Professor Fabio M. do Nascimento for his trust and encouragement.

To my laboratory colleagues and friends Adriano, Alex, Allysson, Ana Carol, André, Darine, Dayane, Esther, Gustavo, Leandro, Luis Gustavo, Maria, Rodrigo Barbosa, Rodrigo Salazar, Talita, Thaise, Uirá, and Vinicius; and to Professors Ana Krepischi, Angela Morgante, Carla Rosenberg, Tábita Hünemeier for the many stimulating discussions.

To the head of the Department and direction of the Institute for the infrastructure facilities provided.

To CAPES, FAPESP, and CNPq for financial support.

To friends Fátima, Israel, Mara, Maraisa, Paulo Rogério, and Silvia for technical support.

To City Hall members of the municipalities of Eldorado, Iporanga, and Barra do Turvo for their support, sisters Angela Biagioni and Maria Sueli Berlanga and Antônio Carlos Nicomedes for their assistance.

The inhabitants of quilombo communities.

Finally, to my friends and members of my family.

Summary

Summary of Figures

CHAPTER 1

CHAPTER 2

CHAPTER 3

Summary of Tables

CHAPTER 1

CHAPTER 3

I. GENERAL INTRODUCTION

This dissertation deals with issues related to the estimation of inbreeding levels and substructure levels, as well as with demographic inferences from a Brazilian population *quilombo* isolate. The document is structured in five sections: (1) this general introduction, where basic concepts related to inbreeding are reviewed; (2) chapter 1, dealing with the estimation of inbreeding and substructure levels in a *quilombo* population; (3) chapter 2, in which a simplified method is presented to estimate the variance of inbreeding coefficient; (4) chapter 3, containing results from inbreeding and demographic analyses performed in the *quilombo* isolate by means of the information of hundreds of thousands of biallelic markers; and (5) a final section with general conclusions. Demographic, historical, and geographical details about the quilombo studied here are exhaustively presented on pages 276-277 of the published article attached to Chapter 1.

I.1. Inbreeding coefficient (Wright's fixation index)

Inbreeding is a non-random mating system in which the choice of mate is influenced or directed by the degree of biological relationship between individuals that mate (Crow and Felsenstein, 1968; Lewontin *et al.*, 1968). Since relatives have one or more ancestors in common, the proportion of alleles identical by descent (IBD) in the genome of their offspring is associated to the amount of ancestry that is shared by their parents.

Endogamy levels are usually estimated by the inbreeding coefficient, which can be defined in terms of correlation as well as of probability (Templeton, 2006; Hartl and Clark, 2007).

Inbreeding coefficient **f** can first be understood as the population correlation coefficient between gametes that come together to generate a zygote (Wright, 1922) and that estimates the deviation λ (covariance among uniting gametes) from genotype frequencies in Hardy-Weinberg (HW) proportions. Considering this parameter, the expected genotype proportions from a biallelic locus (**A**, **a**) can be written down as

$\{\mathbf{d} = \mathbf{P}(\mathbf{AA}) = \mathbf{p}^2 + \lambda,\ \mathbf{h} = \mathbf{P}(\mathbf{Aa}) = \mathbf{2pq} - 2\lambda,$ and $\mathbf{r} = \mathbf{P}(\mathbf{aa}) = \mathbf{q}^2 + \lambda\}$,

where $\mathbf{p} = \mathbf{P(A)}$ and $\mathbf{q} = \mathbf{1\text{-}p} = \mathbf{P(a)}$.

The inbreeding coefficient **f** is then defined as the correlation coefficient $\mathbf{f} = \rho_{\mathbf{x},\mathbf{y}} = \dfrac{\sigma^2_{\mathbf{x},\mathbf{y}}}{\sqrt{\sigma^2_{\mathbf{x}}\sigma^2_{\mathbf{y}}}} = \dfrac{\lambda}{\mathbf{pq}}$, from where we obtain $\lambda = \mathbf{fpq}$. In the equation for **f**, **x** and **y** are dummy variables that take the value **1** when the gamete is **A** and **0** otherwise, when the gamete is **a**.

Replacing λ by **fpq** in the genotype proportions **{d, h, r}** above we immediately obtain the usual formulation for genotype frequencies under inbreeding:

$\{\mathbf{d} = \mathbf{p}^2 + \mathbf{fpq},\ \mathbf{h} = \mathbf{2pq}(\mathbf{1} - \mathbf{f}),\ \mathbf{r} = \mathbf{q}^2 + \mathbf{fpq}\}$,

from which the value of the inbreeding coefficient can be directly estimated: $\mathbf{f} = \mathbf{1} - \dfrac{\mathbf{h}}{\mathbf{2pq}}$.

An alternative approach to estimate the inbreeding coefficient, referred here as **F**, takes into account the probability that two alleles segregating at an autosomal locus are IBD. **F** is usually estimated from genealogies and can be interpreted as the genomic proportion of an individual that is IBD (Haldane and Moshinsky, 1939; Cotterman, 1940; Malécot, 1948). Since the inbreeding coefficient of an individual is the probability that any pair of his homologous genes are identical by descent, its value coincides with the probability (coefficient of consanguinity) that two homologous genes drawn randomly, one from each individual, are identical. Thus, the inbreeding coefficient $\mathbf{F_k}$ of the individual **k** is also the coefficient of consanguinity $\mathbf{F_{ij}}$ of his/her parents **i** and **j**.

The correct estimation of **F** depends however on the existence of an arbitrary founder population completely unrelated. It is, therefore, very difficult or even impossible to trace back the reliable ancestry information from more ancient generations, which rarely includes relationships more remote than third cousins (Cavalli-Sforza and Bodmer, 1971; Speed and Balding, 2015).

Conceptually, **f** and **F** as defined above are different both biologically as well as mathematically, since **F** is a probability (belonging to the domain $0 \leq \mathbf{F} \leq 1$) that estimates the amount of identity by descent for an individual, while **f** is a coefficient of correlation (belonging to the domain $-1 \leq \mathbf{f} \leq 1$) that measures the population proportions of genotypes above or below the ones randomly expected.

I.2. Hierarchical structure of a population

Natural populations frequently are aggregates formed by partially isolated subpopulations within which mating preferentially occurs. Given the reduced subpopulation sizes, the consequence of substructure is an increase of homozygous levels within the population considered as a whole even if mating within subpopulations takes place randomly, due to changes in allelic frequencies secondarily to genetic drift within subpopulations (Crow and Kimura, 1970).

Hierarchically structured populations were first considered by Wright (1951), who defined three different types of fixation indices: $\mathbf{f_{IS}}$ (fixation index due to inbreeding within each subpopulation), $\mathbf{f_{ST}}$ (fixation index due to genetic drift responsible for differences in allele frequencies among subpopulations), and $\mathbf{f_{IT}}$ (fixation index due to the combined effects of inbreeding and genetic drift), related by the following equations:

$$\mathbf{f_{IT} = f_{ST} + f_{IS} - f_{IS}f_{ST} = 1 - \frac{P(Aa)}{2pq}} \quad ;$$

$$\mathbf{f_{ST} = \frac{f_{IT} - f_{IF}}{1 - f_{IS}} = \frac{var(p)}{pq}} \quad ;$$

$$\mathbf{f_{IS} = \frac{f_{IT} - f_{ST}}{1 - f_{ST}}} \quad ,$$

where **p**, **q**, **P(Aa)**, and **2pq** are respectively the estimated allelic frequencies of alleles **A** and **a**, and the directly observed frequency and the expected panmictic proportion of heterozygous individuals in the whole population. As Chakraborty (2016) noticed, these indices have been conceptually defined in several ways: Wright (1943, 1951) defined them in terms of correlations between uniting gametes, Nei (1973, 1977) and Nei and Chesser (1986) defined them as functions of heterozygotes and differences from their respective expectations under HW equilibrium proportions, while Cockerham (1969, 1973), Weir and Cockerham (1984) and Long (1986) formulated them in terms of functions of parameters of components of nested analysis of variance.

I.3. Consequences of inbreeding

The immediate consequence of inbreeding (**f** > 0) is the increase in the frequency of homozygotes in the population, which favors the expression of deleterious recessive alleles previously hidden in heterozygous state. Inbreeding usually leads also to other harmful effects (inbreeding depression), such as the decrease in size, fertility, vigor, yield and fitness, as described for the first time with experimental accuracy by Darwin, who observed its effects in cultivated plants (Fisher, 1949; Crow and Kimura, 1970; Hartl and Clark, 2007).

The effects of endogamy in humans are, in general, more difficult to detect when compared to other species, since the inbreeding levels are usually low and methods that can be developed easily in experimental populations cannot be applied to humans. Empirical studies as well as theoretical risks based on realistic population genetic models show that the chances of affected progeny are largely increased in the offspring of consanguineous marriages (Otto *et al.*, 2007). Strategies based on homozygous mapping (Lander and Botstein, 1987) were developed recently to detect deleterious variants and have been successfully used (1) in identifying new variants related to many disorders of Mendelian recessive inheritance (Lander and Botstein, 1987; Sheffield *et al.*, 1994; Christodoulou *et al.*, 1997; Parvari *et al.*, 1998; Winick *et al.*, 1999; Abou Jamra *et al.*, 2011; Alkuraya, 2013; Ghadami *et al.*, 2015); and (2) in determining susceptibility genes associated with polygenic or complex diseases (Lencz *et al.*, 2007; Nalls *et al.*, 2009; Yang *et al.*, 2012).

I.4. Consanguinity in humans

In humans, consanguineous marriages are still today a relatively common practice, being regarded as customary in many countries throughout the world, because of its traditional status in some cultures. The highest inbreeding levels are found in populations of the Middle East, Central South Asia and the Americas (Leutenegger *et al.*, 2011).

During the last decades, empirical estimates of consanguinity levels were grossly obtained for many populations over the world, by censoring the frequencies of marriages between second cousins and more closely related pairs of individuals; the information was assembled into a database (consang.net) by Bittles and Black (2015). Despite the very low values (much less than 1% on average) observed for most urbanized populations, the prevalence of consanguineous marriages for the global human population was estimated in about 10%, reaching values above 50% in some extremely inbred populations (Bittles, 2002; Bittles and Black, 2010; Hina and Malik, 2015; Ahmad *et al.*, 2016; Riaz *et al.*, 2016).

I.5. Population isolates

Among humans (and other organisms as well), individuals are, in general, heterogeneously distributed in the population territory, tending to form clusters called population isolates, that can be defined as sets of individuals with imprecise boundaries of different natures: geographical, religious, social, ethnic, political, and so on. (Salzano and Freire-Maia, 1967).

Population isolates offer many advantages to medical and evolutionary studies, mainly when isolates have well documented pedigrees, high prevalence of individuals affected by rare genetic conditions, a high degree of inbreeding due to cultural practices or limited population size, and demographic history of foundation consisting in a bottleneck followed by a founder effect (Arcos-Burgos and Muenke, 2002).

Inbreeding and demographic analyses have been the focus of many studies developed in isolates with different ancestries, with the aim (1) to establish relationships among socio-cultural factors and individual homozygous proportions, (2) to provide demographic information for complementing historical records, and (3) to explain in some extent differences in the prevalence of diseases among different populations (Carothers *et al.*, 2006; McQuillan *et al.*, 2008; Lemes *et al.*, 2014; Abdellaoui *et al.*, 2015; Ben Halim *et al.*, 2015; Jalkh *et al.*, 2015; Karafet *et al.*, 2015).

I.6. Runs of Homozygosity

As known from basic population genetic theory, when two individuals are related in some degree, they share segments that are identical by descent (IBD), that is, autozygous. The offspring of biologically related individuals inherit these segments from both parents, which explains the presence, in them, of long stretches of consecutive homozygosity, called runs of homozygosity (ROH). Broman and Weber (1999) were the first to point out the obvious fact that ROH could be identified by means of the occurrence in homozygous state of a large number of contiguous markers detected by molecular analysis.

Individuals may inherit identical chromosomal segments even when the biological relationship between their parents is very distant. Since elapsed time is positively correlated with the event of recombination occurrence responsible for the breaking up of previously existing segments, ROH from more ancient origin tend to be shorter, while those from recent origin tend to be longer (Kirin *et al.*, 2010).

Recently more precise identification of ROH has been greatly enhanced by the use of genomic data. The inbreeding coefficient, referred here as $\mathbf{F_{ROH}}$, can be directly estimated from the proportion of the genome composed of these long tracts in homozygous state (McQuillan *et al.*, 2008). $\mathbf{F_{ROH}}$ is very similar to that directly obtained from pedigree analyses, but much more conservative, since it also takes reliable information from ancient and cryptic inbreeding.

Recent studies of ROH data performed in the worldwide human population detected high levels of autozigosity even in cosmopolitan non-inbred populations. It revealed an increment of endogamy levels and a reduction of genetic diversity according to the population distance from African ones, as expected by the out-of-Africa model of modern human migration. The differences have been explained by the occurrence of small and medium ROH resulting from background relatedness, which also enables the use of ROH to obtain reliable information about demographic and evolutionary events (Kirin *et al.*, 2010; Pemberton *et al.*, 2012).

I.7. General Objective

The aim of this work is to obtain reliable estimates of the average inbreeding coefficient using data obtained from a traditional Brazilian tri-hybrid quilombo population. To achieve this, we used different alternative methods, some of them adapted by us for the specific task of dealing with such a genetically complex population aggregate.

We also tried to establish demographic inferences about the foundation of this population isolate.

The specific objectives are presented in the sections labeled as chapters 1 to 3.

1. CHAPTER 1

We present here a study dealing with the estimation of inbreeding and substructure levels in an African-derived Brazilian *quilombo* isolate. The analyses were partially performed during my Master's project in which: (1) all available genealogies of ten *quilombo* communities were used to estimate the inbreeding coefficient and the consanguinity rates; and (2) data from 30 autosomal loci (14 SNPs, and 16 microsatellites) were used to estimate inbreeding and substructure levels. During my PhD project we concluded the study considering for the genealogical analyses only the more reliable information obtained from individuals with full ascendant records over at least two generations; for the analysis of molecular markers, in order to take into account errors in the process of genotype determination, we used data obtained from two subsets of individuals, one considering those genotyped for at least 27 of the 30 markers, and another containing the original data presented on the MSc thesis (results of all genotyped individuals). The inbreeding coefficients identified in the introduction as **f** and **F** are referred to in the article representing this chapter as **F** and $\mathbf{F_G}$, respectively; the article was published in the specialized journal Human Biology (Lemes, RB, Nunes, K, Meyer, D, Mingroni-Netto, RC, Otto, PA. Estimation of inbreeding and substructure levels in African-derived Brazilian quilombo populations. **Hum. Biol. 86: 276-88.** 2014).

Estimation of Inbreeding and Substructure Levels in African-Derived Brazilian Quilombo Populations

Renan B. Lemes,[1] Kelly Nunes,[1] Diogo Meyer,[1] Regina Célia Mingroni-Netto,[1] and Paulo A. Otto[1]*

ABSTRACT

This article deals with the estimation of inbreeding and substructure levels in a set of 10 (later regrouped as eight) African-derived quilombo communities from the Ribeira River Valley in the southern portion of the state of São Paulo, Brazil. Inbreeding levels were assessed through *F*-values estimated from the direct analysis of genealogical data and from the statistical analysis of a large set of 30 molecular markers. The levels of population substructure found were modest, as was the degree of inbreeding: in the set of all communities considered together, *F*-values were 0.00136 and 0.00248 when using raw and corrected data from their complete genealogical structures, respectively, and 0.022 and 0.036 when using the information taken from the statistical analysis of all 30 loci and of 14 single-nucleotide polymorphic loci, respectively. The overall frequency of consanguineous marriages in the set of all communities considered together was ~2%. Although modest, the values of the estimated parameters are much larger than those obtained for the overall Brazilian population and in general much smaller than the ones recorded for other Brazilian isolates. To circumvent problems related to heterogeneous sampling and virtual absence of reliable records of biological relationships, we had to develop or adapt several methods for making valid estimates of the prescribed parameters.

Over three million Africans were brought to Brazil as slaves over a period of three hundred years. Runaway, abandoned, and freed slaves created small communities known as *quilombos*, the remnants of which in the state of São Paulo are confined to its southern border along the Ribeira River Valley (Figure 1). The region's geography afforded these communities a certain degree of isolation. These settlements became traditional rural communities surviving on subsistence agriculture for many decades. Some drastic recent changes have taken place in the lifestyle of their inhabitants, with traditional agriculture replaced by the cultivation of more commercially valuable products (Santos and Tatto 2008; Pasinato and Rettl 2009). This nutritional transition process has resulted in the high rates, among its inhabitants, of multifactorial (complex) diseases, such as essential hypertension and obesity (Angeli et al. 2011; Kimura et al. 2012).

Quilombos have long been the subject of interest for population and evolutionary geneticists. They usually originate from a relatively small number of individuals (founder effect) and remain isolated over several generations, thus being subjected to the classical process of micro-differentiation due mainly to random genetic drift. Many (but not all) isolates studied in Brazil and elsewhere (see Table

[1]Department of Genetics and Evolutionary Biology, Institute of Biosciences, Universidade de São Paulo, São Paulo, Brazil.

*Correspondence to: Paulo A. Otto, Department of Genetics and Evolutionary Biology, Institute of Biosciences, Universidade de São Paulo, Caixa Postal (PO Box) 11.461, 05422-970 São Paulo, SP, Brazil. E-mail: otto@usp.br.

KEY WORDS: INBREEDING, POPULATION ISOLATES, QUILOMBO REMNANTS, SUBSTRUCTURE ANALYSIS.

Human Biology, Fall 2014, v. 86, no. 4, pp. 276–288. Copyright © 2015 Wayne State University Press, Detroit, Michigan 48201

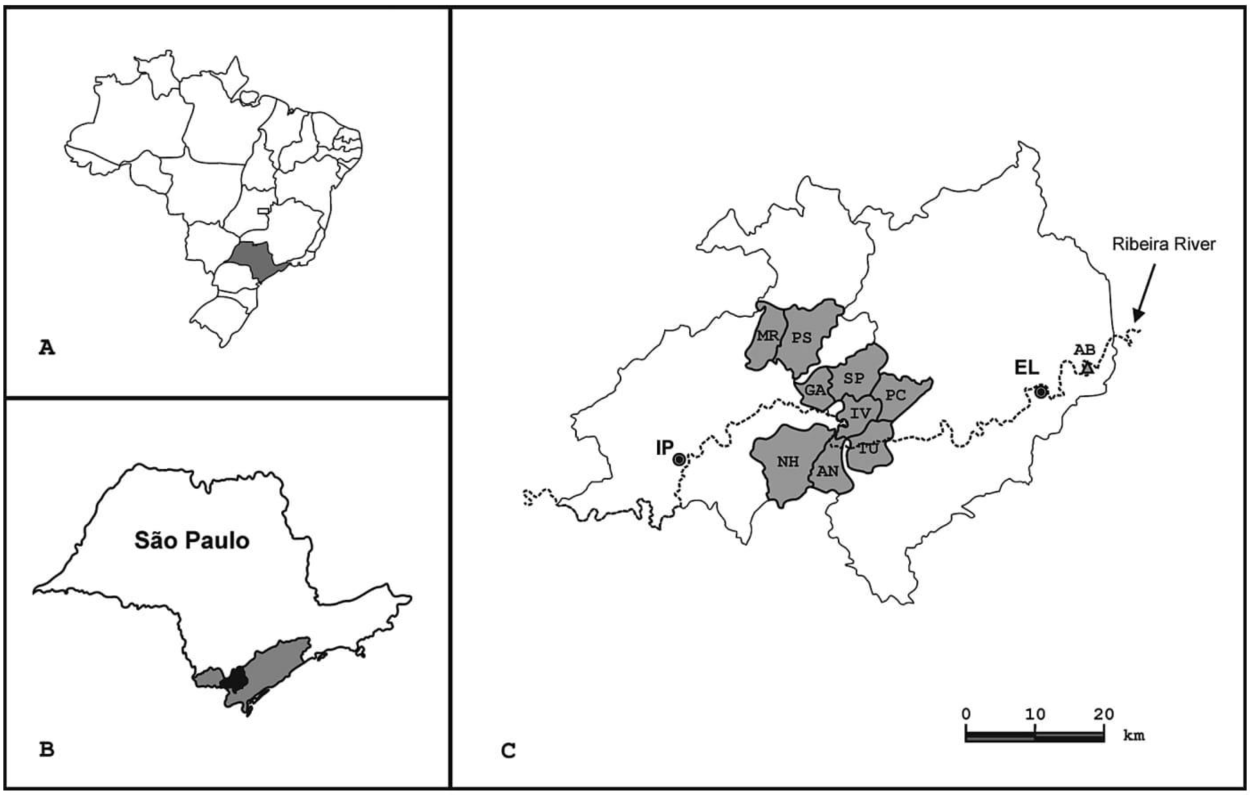


FIGURE 1. (A) State of São Paulo highlighted within the Brazilian territory. (B) Location of both Ribeira Valley region in São Paulo (gray) and the municipalities of Eldorado (EL) and Iporanga (IP) (black). (C) Territory in which the 10 quilombo communities are located (from Kimura et al. 2013): AB, Abobral; AN, André Lopes; GA, Galvão; IV, Ivaporanduva; MR, Maria Rosa; NH, Nhunguara; PC, Pedro Cubas; PS, Pilões; SP, São Pedro; TU, Sapatu.

4 below) show detectable levels of inbreeding. This is measured by the average inbreeding coefficient *F* of its individuals or, as usually happens, using simplified methods that weigh the various inbreeding coefficients of the progenies corresponding to the different types of marriages occurring in the population. As Cavalli-Sforza and Bodmer (1971: 352) point out, "these inbreeding estimates take into account only easily detectable consanguinity, which rarely includes relationships more remote than third cousins." Therefore, genealogical estimates of the mean inbreeding coefficient, in spite of being able to demonstrate the presence of consanguinity even at very modest rates, clearly constitute an underestimate of the real parameter value. More realistic estimates of consanguinity rates can be inferred from the population analysis of genetic markers (classical or molecular). The main problem with this strategy is that incredibly large samples are required in order to reveal statistically significant departures from $p^2{:}2pq{:}q^2$ Hardy-Weinberg equilibrium rates, as Figure 2 clearly shows. For instance, a sample size of about 1,500 individuals is necessary to detect a significant value of the inbreeding coefficient in an inbred population having a parameter value of $F = 0.05$. Another problem with *F*-coefficients so estimated is that they should be differentiated from similar coefficients that might be spuriously interpreted as indicative of inbreeding and that commonly arise when the populations under study are hierarchically stratified (Wahlund's effect).

The primary objective of this study was to provide estimates of inbreeding and of substructure levels from a set of 10 quilombo communities. In order to circumvent problems related to the paucity of written and oral historical records and those related to heterogeneous molecular sampling (detailed in the sections below), we had to develop or adapt several methods for obtaining reliable

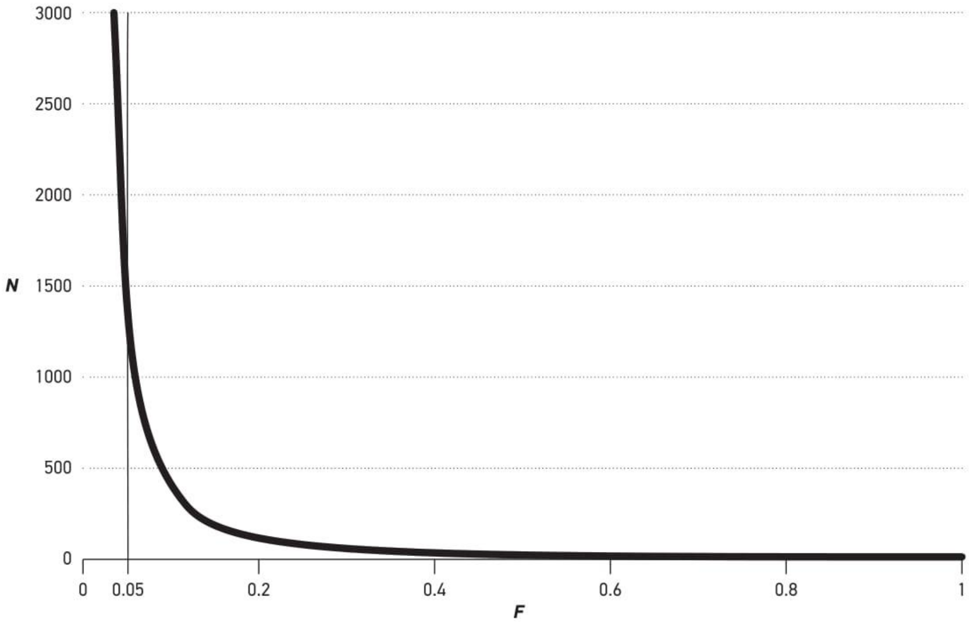


FIGURE 2. Sample sizes (y-axis) required for obtaining statistical significance of F-values (x-axis) at the rejection level of 5%. The vertical line corresponds to an F-value of 0.05.

estimates of the prescribed parameters of inbreeding and population substructure. The presentation of these methodological variations is an important contribution of this report.

Subjects and Methods

Populations and Subjects

Like other quilombos in Brazil, the communities here presented were founded, in the last decades of the 19th century, by a relatively small number of runaway, abandoned, and freed African-derived slaves. Over the years the communities grew to include individuals from different ancestries (most of them African derived, but also some Amerindians and admixed individuals with African and European ancestry). Given their proximity (most communities of the Ribeira River Valley are contiguous and within walking distance), relatively high levels of gene flow are expected to have occurred among the communities over the next five or six generations that have elapsed since their founding. Taking all this into account, a relatively high degree of homogeneity is expected to be found among them, as well as a relatively low inbreeding level within them. Table 1 lists the present number of living individuals in each community and the corresponding numbers of individuals interviewed for assessing genealogical data (per community) and of individuals molecularly genotyped (per locus and community). The data from two pairs of communities (Galvão + São Pedro and Maria Rosa + Pilões) were grouped and analyzed together since they occupy adjacent territories, being basically formed by the same family groups.

This study was approved by the ethics committee of the Instituto de Ciências Biomédicas, Universidade de São Paulo. Informed consent was obtained from all participants in the study.

Genotype Determination

Molecular (DNA markers) and genealogical data from the eight communities were obtained in different surveys organized and performed by members of the Laboratory of Human Genetics of our Department and partly reported in Mingroni-Netto et al. (2009a, 2009b), Cotrim et al. (2004), Angeli et al. (2005, 2011), Auricchio et al. (2007), Yeh et al. (2008), and Kimura et al. (2012, 2013). Our analyses used data from 14 autosomal single-nucleotide polymorphisms (SNPs) previously genotyped in our laboratory (for details on methodology, see Angeli et al. 2011; Kimura et al. 2012): *ACE* (rs1799752), *NOS3* (rs1799983), *GNB3* (rs5443), *GNB3* (rs5441), *AGT* (rs669), *ADD2* (rs3755351), *GRK4* (rs1801058), *PLIN1* (rs2289487), *INSIG2* (rs7566605), *LEP*

Table 1. Numbers of Genotyped Individuals for Each Molecular Marker at a Given Community

Measure	Community								Total
	AB	*AN*	*GA/SP*	*IV*	*MR/PS*	*NH*	*PC*	*TU*	
N	573	320	266	270	184	447	286	295	2,641
N_G	364	247	224	217	148	237	263	179	1,879
SNP markers									
ACE (rs1799752)	96	86	99	77	55	89	78	56	636
NOS3 (rs1799983)	59	79	92	76	30	67	78	56	537
GNB3 (rs5443)	95	78	98	77	39	67	76	56	586
GNB3 (rs5441)	93	65	94	62	54	84	77	66	595
AGT (rs669)	58	48	92	76	30	63	78	56	501
ADD2 (rs3755351)	92	75	90	76	45	48	73	59	558
GRK4 (rs1801058)	91	85	97	75	52	86	77	72	635
PLIN1 (rs2289487)	93	108	115	128	64	109	93	78	788
INSIG2 (rs7566605)	93	103	112	125	65	102	93	79	772
LEP (rs2167270)	94	106	114	116	61	109	92	80	772
LEPR (rs1137101)	94	107	115	116	60	109	91	79	771
ADRB2 (rs1042713)	95	102	111	110	61	104	91	78	752
PPARG (rs1801282)	93	103	115	102	61	106	93	80	753
RETN (rs1862513)	89	105	113	126	65	104	91	76	769
Microsatellite markers									
D1S551	36	24	34	51	37	41	39	28	290
D4S3248	36	24	34	50	37	41	39	28	289
D5S816	36	25	34	51	37	41	39	28	291
D6S1040	35	22	33	52	37	43	39	31	292
D7S821	36	24	34	51	37	41	39	28	290
D7S3061	36	24	34	51	37	41	39	28	290
D8S2324	36	22	34	52	37	43	39	31	294
D9S301	37	23	34	52	37	43	39	31	296
D9S922	36	24	34	51	36	41	39	28	289
D10S1426	29	18	30	49	34	39	38	25	262
D13S317	37	22	34	52	37	43	39	31	295
D16S539	36	24	34	51	37	41	39	28	290
D18S535	37	23	34	52	37	43	39	31	296
D19S559	36	22	34	52	37	43	39	31	294
D20S482	37	23	34	52	37	43	39	31	296
D21S1437	32	19	22	38	26	34	25	11	207

N, estimated number of adult individuals (Auricchio et al. 2007); N_G, number of individuals interviewed for gathering genealogical data. Communities are as defined in Figure 1.

(rs2167270), *LEPR* (rs1137101), *ADRB2* (rs1042713), *PPARG* (rs1801282), and *RETN* (rs1862513).

Using DNA samples from some 300 individuals of the communities, we determined the genotypes of the following 16 autosomal microsatellite loci: D1S551, D4S3248, D5S816, D6S1040, D7S821, D7S3061, D8S2324, D9S301, D9S922, D10S1426, D13S317, D16S539, D18S535, D19S559, D20S482, and D21S1437. The primer sequences were generated using software Primer3 (Rozen and Skaletsky 2000), and the forward sequences were marked with fluorescence (Supplementary Table S1). Microsatellite genotypes were determined by polymerase chain reaction in four multiplex systems submitted to capillary electrophoresis on ABI 3730 DNA analyzer (Applied Biosystems, Foster City, CA). All analyses were carried out using the Peak Scanner software, version 1.0 (Applied Biosystems).

Different groups of individuals were selected for determination of molecular markers on different occasions with distinct purposes: the first set of seven SNP markers of the 14 listed above were used primarily in association studies with arterial hypertension, and the last seven, in association studies with obesity. As a result, data for each set of marker only partially overlap, introducing an additional source of variation, leading us to expect to find a significant degree of heterogeneity among loci and populations.

Genealogical Data

Genealogical analysis of data based on detailed interviews provided information for about 2,000 individuals, which allowed us to estimate a mean inbreeding coefficient or fixation index (F_G) for each community and in the set of all communities. Our analysis included all living individuals who were born in a given community. We also considered as belonging to a given community migrant individuals who had offspring with native quilombo inhabitants from that community. Information from deceased individuals was used only to assess biological relationships among individuals within communities. The total number of inhabitants and individuals interviewed for genealogical data (2,641 and 1,879, respectively) varied from 573 and 364 to 184 and 148 per community, respectively; the total number of genotype determinations varied from 788 to 207 in relation to different loci in the total population (see Table 1).

The quilombo communities here studied were isolated for a long period of time, with few historical records (written or oral) of biological relationships. In order to correct or decrease this bias, average inbreeding coefficients (per community and for the set of all communities grouped together), in addition to being estimated using all available information, were assessed just from individuals that possessed double-checked information on their ascendants over at least two generations. From the total of 3,959 individuals represented in the genealogies, 2,171 provided complete information on their ascendants over at least two generations; just 794 among them had reliable information (in order to establish the presence of eventual biological relationships) for at least half of their great-grandparents; and fewer than 100 individuals had reliable information for all their great-grandparents.

Quantitative Analyses

Genealogical Analysis

Genealogical estimates of the mean inbreeding coefficient (fixation index F_G) for each community and in the set of all communities were obtained by averaging the individual inbreeding coefficients (f_G) from all individuals represented in the genealogies and from a subsample of individuals that possessed information on their ascendants over at least two generations. The values of each f_G were obtained by the usual Wright's (1922) formula $f_G = \Sigma[1/2^n \times (1 + f_A)]$, in which n is the number of individuals between the parental pair and the common ancestor, including these three individuals, and f_A is the inbreeding coefficient of the common ancestor of the parental pair.

Molecular Markers Data Analysis

Reliable estimates of genotype and allele frequencies and of the average inbreeding coefficient (Wright's fixation index) $F = 1 - \Sigma\Sigma P(a_i a_j)/(2\Sigma\Sigma p_i p_j)$, which reduces to $F = 1 - P(Aa)/(2pq)$ in the two-allele case, were obtained through programs developed in a Windows-based structured BASIC dialect (Liberty BASIC, version 4.04; Shoptalk Systems, Framingham, MA) and using the package of mathematical routines Mathematica, version 8.0.4.0 (Wolfram Research, Champaign, IL). By means of chi-squared tests and bootstrap simulation techniques, these programs test the samples for departures of Hardy-Weinberg ratios, estimate their corresponding fixation index values, construct "exact" confidence intervals for them, and perform appropriate substructure analyses.

Mean values of F for the whole population in relation to each locus were obtained by adding the corresponding data of all communities. In the case of the set of all loci per population or in the set of all populations, average F-values were estimated by the usual method of combining them by the reciprocal values of their corresponding variances:

$$F = \Sigma[F_i/\text{var}(F_i)]/\Sigma[1/\text{var}(F_i)],$$

with i varying from 1 to the number of different loci. The appropriate estimation of the variance of the inbreeding coefficient, var(F), is a complicated issue, and the formula derived by Fyfe and Bailey (1951) for the case of two autosomal alleles is generally used:

$$\mathrm{var}(F) = \frac{(1-F)^2(1-2F)}{N} + \frac{F(1-F)(2-F)}{2Np(1-p)},$$

in which $p = P(A) = [2N(AA) + N(Aa)]/2N$, $F = 1 - [N(Aa)/N]/[2p(1-p)]$, $N = N(AA) + N(Aa) + N(aa)$, and A and a are a pair of alleles segregating in an autosomal locus.

We were able to derive a different formula for the variance of F whose numerical values for the two-allele case are virtually the same as those obtained using either the formula proposed by Fyfe and Bailey (1951) or the average population values estimated by simulations using bootstrapping techniques. Our formula is expressed in the two-allele case by the equation

$$\begin{aligned}\mathrm{var}(F) &= N_1{\cdot}N_2{\cdot}N_3/[(Npq)^2{\cdot}(N_2{\cdot}N_3 + 4{\cdot}N_1{\cdot}N_3 + N_1{\cdot}N_2)] \\ &= (1-F)(p+qF)(q+pF)/[Npq(1+F)],\end{aligned}$$

where $N_1 = N(AA)$, $N_2 = N(Aa)$, $N_3 = N(aa)$, $N = N_1 + N_2 + N_3$, $p = 1 - q = (2N_1 + N_2)/2N$, and $F = 1 - N_2/2pq$. Unlike Fyfe and Bailey's formula, it is possible to adapt this formula to the generalized case of any number of alleles segregating at an autosomal locus. The subject has theoretical interest; mathematical details about its derivation and properties will be published and discussed elsewhere.

To determine which values of F could be considered as outliers and should be excluded from a global analysis, we proceeded as follows: in the long run the various per locus estimates of F inside the same community are expected to be normally distributed around the average F value for that community, so the outlier values should be outside the usual 95% range $F \pm 1.96[\mathrm{var}(F)]^{1/2}$, where $F = \Sigma x_i F_i$, $\mathrm{var}(F) = \Sigma x_i F_i^2 - F^2$, and $x_i = \mathrm{var}^{-1}(F_i)/\Sigma_{(j=1,n)} \mathrm{var}^{-1}(F_j)$.

"Exact" 95% confidence intervals for the estimated values of the mean inbreeding coefficient (fixation index) F were obtained for each combination locus/community through 1,000 computer-assisted bootstrap simulations of samples, each having the same size and genotypic proportions observed in the actual one. A similar approach with variations was used to construct the confidence intervals of Wright's substructure indexes F_{ST}, F_{IT}, and F_{IS}.

For the substructure analysis, we recoded the microsatellite markers as biallelic, with the first allele corresponding to the allele with the highest frequency in the population and the second allele being equivalent to the total of the remaining alleles.

To circumvent problems related to heterogeneous sampling of loci and communities, besides performing the analyses detailed above in the whole data set (considering all genotyped individuals), we repeated the procedures using a subsample containing only individuals genotyped for all loci. Since with this strategy the sample size dropped to only 87 individuals (Supplementary Table S2), we also used a subsample containing all individuals who were genotyped for at least 27 of the 30 marker systems, resulting in a sample of 207 individuals (Supplementary Table S3). To take into account the different nature of the sets of molecular markers used, we estimated all parameters in relation to SNPs and microsatellites separately.

Results and Discussion

Genealogical Analysis

Table 2 lists the estimated values of the inbreeding coefficient (F_G) from the genealogical analysis of the eight communities considered separately and together, taking into account the data from all 3,959 individuals with genealogical information. Table 3 lists the same values estimated from the set of 2,171 individuals who had complete information about his ascendants over at least two generations. Unlike other estimates derived from genealogical analysis, which calculate the population F value weighing the different F-values by the mean sizes of the sibships from which they were estimated, our F estimate is the average value of the parameter estimated for each living individual of the population.

Before applying our methodology to the quilombos reported here, we tested its performance by applying it to the published genealogical structure of the quilombo isolate of Valongo (Souza and Culpi 1992) in the southern state of Santa Catarina (Supplementary Figure S1), founded by just four couples, where the frequency of consanguineous

Table 2. Estimates of *F* Obtained by Genealogical Analysis: All Individuals

Community	*N*	F_G	% CM
AB	773	0.00344	3.63
AN	567	0.00245	2.31
GA/SP	446	0.00070	1.72
IV	575	0.00033	0.63
MR/PS	324	0.00024	0.88
NH	434	0.00176	5.26
PC	368	0	0
TU	472	0	0
Total	*3,959*	*0.00136*	*1.87*

Communities are as defined in Figure 1; *N*, number of individuals included in the analyses; F_G, estimated value of the inbreeding coefficient; % CM, observed frequencies of consanguineous marriages.

Table 3. Estimates of *F* Obtained by Genealogical Analysis: Individuals with Complete Information for Their Ascendants over at Least Two Generations

Community	*N*	F_G	% CM
AB	380	0.00699	8.18
AN	383	0.00363	5.68
GA/SP	235	0.00133	4.76
IV	288	0.00065	1.47
MR/PS	152	0.00052	2.22
NH	221	0.00346	13.95
PC	368	0	0
TU	472	0	0
Total	*2,171*	*0.00248*	*4.58*

Communities are as defined in Figure 1; *N*, number of individuals included in the analyses; F_G, estimated value of the inbreeding coefficient; % CM, observed frequencies of consanguineous marriages.

Table 4. Estimates of *F* and Percent Consanguineous Marriages (% CM) from Several Isolates Reported in the Literature

Population	*F*	% CM	Reference
Jewish isolate from Curitiba (Brazil)	0.0013	4.0	Freire-Maia and Krieger 1963
Amish of Adams county (USA)	0.0195	66.5	Jackson et al. 1968
Törbel (Switzerland)	0.0058	—	Ellis and Starmer 1978
Quilombo of Valongo (Brazil)	0.0477	85.0	Souza and Culpi 1992
Amish of Lancaster (USA)	0.0166	—	Dorsten et al. 1999
Hutterites of South Dakota (USA)	0.0340	—	Abney et al. 2000
India	0.0075	11.9	Bittles 2002
Southern India	0.0212	31.0	Bittles 2002
Amman (Jordan)	0.0142	28.4	Hamamy et al. 2005
Quilombo of Ribeira River Valley (Brazil)	0.0025	4.6	Present study

unions is 85%. We obtained the estimate F_G = 0.0457 for the whole community, a value that is not significantly different from the estimate of 0.0477 obtained by Souza and Culpi (1992) using the formula $F = 2(N_r - 1)/[2N_e - (2N_e - 1)(1 - m_e)^2]$, where N_r is the breeding population size, $N_e = 2(N_r - 1)/(k - 1 + \sigma_k^2/k)$ is the effective population size, m_e is the effective migration rate, and k is the average offspring size in the breeding population.

The estimated values of *F* for the set of all communities grouped together range from 0.00136 (considering all individuals) to 0.00248 (considering only the subset of 2,171 individuals with more reliable information). These values are approximately 1.5–3 times higher than the corresponding estimate for the total Brazilian population (F = 0.00088) and about 2–4 times higher than the estimate for the population of the state of São Paulo (F = 0.00067) (Freire-Maia 1957, 1990). The community values of *F* ranged from zero in two aggregates to 0.00344 (Table 2) or 0.00699 (Table 3) in the population of Abobral (AB).

As already commented, the values of F_G in the quilombos reported here surely are underestimates of the true values due to many factors, such as lack of information on many branches of the genealogies and generalized absence of reliable records as to the origin of the populations, as well as to biological relationships among their members. In any case, the strategy of reassessing the parameter in the subsample containing only individuals with more reliable information was able to partially eliminate this bias.

Table 4 compares our estimates of both inbreeding coefficient and the frequency of consanguineous marriages with the results from isolate surveys in the literature. With the exception of the Brazilian Jewish isolate studied by Freire-Maia and Krieger (1963), all other communities listed in Table 4 show relatively large *F*-values, almost always associated with substantial levels of consanguineous unions, unlike our results shown in Tables 2 and 3.

The strikingly high inbreeding levels of Valongo quilombo are perfectly compatible with the fact that the community presently comprises fewer than 100 individuals, all originated from only four founding couples. Unlike this community, the whole isolate of the Ribeira River Valley has more than 2,500 adult individuals. Its size, together with other factors (see Subjects and Methods), probably

Table 5. Average *F* (95% Confidence Intervals) in Relation to Microsatellites, SNPs, and All Markers Together

Community	Microsatellites	SNPs	All Markers
AB	−0.010 (−0.104, 0.085)	0.020 (−0.151, 0.192)	0.011 (−0.149, 0.171)
AN	−0.042 (−0.244, 0.160)	0.003 (−0.113, 0.119)	−0.002 (−0.132, 0.129)
GA/SP	−0.138 (−0.225, −0.052)	0.045 (−0.145, 0.235)	−0.057 (−0.226, 0.112)
IV	−0.051 (−0.176, 0.074)	−0.006 (−0.249, 0.236)	−0.014 (−0.239, 0.211)
MR/PS	−0.036 (−0.157, 0.086)	0.060 (−0.247, 0.366)	0.031 (−0.246, 0.309)
NH	−0.064 (−0.117, −0.010)	−0.051 (−0.206, 0.105)	−0.059 (−0.169, 0.052)
PC	−0.041 (−0.060, −0.021)	−0.037 (−0.180, 0.106)	−0.035 (−0.117, 0.047)
TU	−0.028 (−0.149, 0.094)	0.001 (−0.231, 0.232)	−0.002 (−0.223, 0.218)
Total	*−0.002 (−0.064, 0.060)*	*0.036 (−0.049, 0.121)*	*0.022 (−0.050, 0.093)*

Communities are as defined in Figure 1.

Table 6. Average *F* (95% Confidence Intervals): Only Individuals Genotyped for at Least 27 of 30 Markers

Community	Microsatellites	SNPs	All Markers
AB	−0.071 (−0.101, −0.042)	−0.013 (−0.166, 0.140)	−0.057 (−0.140, 0.026)
AN	−0.049 (−0.272, 0.175)	−0.035 (−0.323, 0.253)	−0.039 (−0.309, 0.230)
GA/SP	−0.065 (−0.138, 0.009)	0.017 (−0.183, 0.216)	−0.078 (−0.249, 0.093)
IV	−0.031 (−0.105, 0.043)	−0.045 (−0.288, 0.198)	−0.013 (−0.195, 0.170)
MR/PS	−0.057 (−0.151, 0.038)	−0.069 (−0.348, 0.209)	−0.038 (−0.273, 0.197)
NH	−0.089 (−0.227, 0.050)	0.059 (−0.286, 0.404)	−0.053 (−0.238, 0.133)
PC	−0.104 (−0.204, −0.005)	0.011 (−0.298, 0.321)	−0.065 (−0.242, 0.111)
TU	−0.049 (−0.224, 0.127)	0.005 (−0.322, 0.332)	0.001 (−0.277, 0.280)
Total	*−0.024 (−0.467, 0.419)*	*0.055 (−0.464, 0.575)*	*0.013 (−0.167, 0.192)*

Communities are as defined in Figure 1.

account for the unusually low inbreeding levels detected in the isolate here reported.

Molecular Marker Analysis

Our analysis of a set of independent autosomal loci provided us with estimates of mean *F*-values both for the individual quilombo communities and for all of them together, in relation to each locus and for the set of all loci considered together. Outlier values, determined using the method described in Subjects and Methods, were not considered for any calculations.

Considering the frequency of *P*-values < 0.05, only in six of a total of 239 combinations (~2.5%) of locus/community was the hypothesis of $p^2{:}2pq{:}q^2$ ratios of Hardy-Weinberg equilibrium rejected, which is slightly less than the expected proportion by chance in the long run. When all quilombo communities were considered together, the genotype frequencies at 2 of 30 loci (~6.7%) deviated significantly from Hardy-Weinberg ratios at the same rejection level of 5%, which clearly indicates just a nonsignificant excess of positive results. Including the data obtained from pooling, per locus, all communities together, a total of approximately 250 tests for verifying the hypothesis $F = 0$ were performed. A Bonferroni-type correction of our data will show that none of the tests produced a significant *P* value.

Table 5 summarizes the results for each isolate and for the set of all communities considered together, in relation to (a) the set of 16 microsatellite markers, (b) the set of 14 SNPs, and (c) all loci considered together. Table 6 shows the results for the analysis of a data set containing all individuals that were genotyped for at least 27 of the 30 markers.

Table 7. Estimates of Fixation Indexes (95% Confidence Intervals) by Marker

Marker	F_{IT}		F_{ST}		F_{IS}	
ACE (rs1799752)	**0.097**	(0.014, 0.179)	**0.045**	(0.029, 0.076)	**0.054**	(−0.032, 0.128)
NOS3 (rs1799983)	0.054	(−0.048, 0.163)	0.021	(0.011, 0.051)	0.033	(−0.067, 0.132)
GNB3 (rs5443)	0.030	(−0.058, 0.110)	0.037	(0.022, 0.067)	−0.007	(−0.096, 0.063)
GNB3 (rs5441)	0.085	(−0.013, 0.175)	0.025	(0.011, 0.057)	0.062	(−0.046, 0.151)
AGT (rs669)	−0.028	(−0.118, 0.069)	0.013	(0.005, 0.039)	−0.041	(−0.137, 0.052)
ADD2 (rs3755351)	0.062	(−0.027, 0.147)	0.020	(0.011, 0.047)	0.043	(−0.053, 0.118)
GRK4 (rs1801058)	0.018	(−0.061, 0.102)	0.015	(0.008, 0.038)	0.003	(−0.082, 0.083)
PLIN1 (rs2289487)	**0.104**	(0.026, 0.172)	**0.031**	(0.018, 0.056)	**0.075**	(−0.006, 0.139)
INSIG2 (rs7566605)	0.002	(−0.077, 0.076)	0.153	(0.008, 0.036)	−0.014	(−0.099, 0.058)
LEP (rs2167270)	0.017	(−0.058, 0.089)	0.023	(0.012, 0.045)	− 0.006	(−0.082, 0.064)
LEPR (rs1137101)	0.001	(−0.063, 0.068)	0.032	(0.021, 0.055)	−0.033	(−0.103, 0.031)
ADRB2 (rs1042713)	−0.034	(−0.113, 0.046)	0.027	(0.014, 0.053)	−0.063	(−0.152, 0.014)
PPARG (rs1801282)	0.056	(−0.013, 0.140)	0.061	(0.037, 0.103)	−0.002	(−0.074, 0.065)
RETN (rs1862513)	−0.004	(−0.071, 0.065)	0.015	(0.009, 0.034)	−0.019	(−0.092, 0.046)
D5S816	**−0.122**	(−0.219, −0.029)	**0.001**	(0.003, 0.028)	**−0.123**	(−0.231, −0.041)
D1S551	0.097	(−0.014, 0.207)	0.024	(0.014, 0.068)	0.075	(−0.049, 0.174)
D7S3061	0.092	(−0.030, 0.209)	0.007	(0.005, 0.045)	0.086	(−0.046, 0.190)
D4S3248	0.067	(−0.056, 0.186)	0.012	(0.007, 0.049)	0.056	(−0.081, 0.160)
D16S539	−0.015	(−0.122, 0.098)	0.011	(0.006, 0.047)	−0.026	(−0.149, 0.073)
D9S922	−0.062	(−0.182, 0.045)	0.018	(0.010, 0.057)	−0.082	(−0.215, 0.013)
D10S1426	0.047	(−0.102, 0.180)	0.054	(0.030, 0.115)	−0.007	(−0.168, 0.118)
D7S821	−0.087	(−0.195, 0.023)	0.011	(0.006, 0.046)	−0.099	(−0.220, −0.009)
D13S317	0.017	(−0.089, 0.131)	0.033	(0.021, 0.078)	−0.016	(−0.140, 0.089)
D8S2324	0.106	(−0.032, 0.251)	0.013	(0.006, 0.054)	0.095	(−0.058, 0.230)
D19S559	−0.007	(−0.131, 0.112)	0.018	(0.009, 0.057)	−0.026	(−0.164, 0.083)
D6S1040	−0.077	(−0.202, 0.039)	0.006	(0.004, 0.036)	−0.084	(−0.218, 0.018)
D20S482	0.111	(−0.012, 0.229)	0.022	(0.010, 0.074)	0.090	(−0.048, 0.195)
D21S1437	0.197	(0.015, 0.347)	0.026	(0.010, 0.097)	0.175	(−0.017, 0.324)
D9S301	−0.023	(−0.139, 0.080)	0.035	(0.021, 0.081)	−0.061	(−0.188, 0.035)
D18S535	−0.021	(−0.140, 0.092)	0.007	(0.005, 0.038)	−0.028	(−0.158, 0.072)

Values in boldface indicate cases in which we can assume unambiguously that the F_{ST} index is different from zero.

Unlike what happens when only the SNPs are used, the average *F* estimates using microsatellite data have negative values for practically all communities. This is especially noted when the sample sizes are drastically reduced in order to minimize data heterogeneity (Table 6), and it is known from sampling theory that small-sized samples favor the occurrence of heterozygous individuals (see Cannings and Edwards 1969). This should be critical when the number of segregating alleles is high, a situation in which most sampled individuals will be heterozygous even under panmictic expectations. In summary, the estimates using biallelic markers such as autosomal SNPs seem to be more reliable than the ones using microsatellites or the set of all markers. Therefore, our analysis using adequate molecular markers (SNPs) indicates average figures of the mean inbreeding coefficient ranging from about 0.036 (using data from all sampled individuals) to 0.055 (using the more homogeneous data from individuals that were genotyped for at least 27 different markers).

Population Substructure Analysis

Genealogical relations among individuals from different quilombo communities of the Ribeira Valley exist to a certain degree, since the founders of some of these population aggregates are likely to be the same, as indicated by the sharing of some common surnames. This fact and the physical proximity of the different communities (as Figure 1 shows, most are contiguous, within walking distance, with the farthest < 20 km away) suggest a priori a modest level of substructure among these communities.

Table 7 presents the values of the fixation indexes (F_{IT}, F_{ST}, and F_{IS}) obtained from all 30 loci for the set of all quilombo communities. Simulations by means of bootstrap techniques, using all data (but also excluding outliers), generated reliable estimates of the 95% confidence interval for each one of these fixation indexes. When the lower and upper limits of a 95% confidence interval of F_{IT} or F_{IS} thus constructed have different signs, it is assumed that the corresponding fixation indexes are not significantly different from zero at the rejection level of 5%. Since F_{ST} indexes are always obtained from the relation $\mathrm{var}(p)/(pq)$, and all three quantities in the formula belong to the domain of positive numbers, the numerical value of the parameter, as well as all the values contained in its corresponding confidence interval, will be positive. Inferences regarding the significance of F_{ST} (is F_{ST} significantly different from zero?) are then obtained indirectly from the behavior of the corresponding confidence intervals of both F_{IT} and F_{IS}: in all instances in which F_{IS} is not different from zero, F_{IT} is not different from F_{ST}; therefore, in all cases in which both F_{IT} and F_{IS} are not different from zero, F_{ST} is also not statistically different from zero. The very few instances in which this did not take place are indicated by F_{ST} values in boldface in Table 7 and should be interpreted as cases in which we can assume unambiguously that the index is different from zero.

The F_{ST} values were in general very small, a finding already detected for these same populations in a study by Kimura et al. (2013) using indel molecular markers. This suggests the existence of a significant amount of gene flow or recent shared ancestry, with little time for differentiation between the subpopulations.

What is important and immediately assumed from the mere inspection of Table 7 is that, with exception of locus *ACE* (rs1799752), in the few instances in which the F_{ST} was significantly different from zero, the proportional contribution of F_{ST} to the F_{IT} index was always much smaller than that of F_{IS}. The dubious results obtained in relation to locus *PLIN1* were caused by extremely high *F*-values in three of the seven communities that resisted the process of outlier cleaning, a behavior for which we have no logical explanation.

In spite of the difficulties brought about by the sets of genealogical as well as molecular data, our results indicate that the levels of substructure among the quilombo communities are negligible or at least very small, probably a consequence of gene flow and shared history among communities. This finding legitimizes the genealogical and molecular estimations of the fixation index we performed by considering the set of communities as a whole.

Acknowledgments

This work was partially supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo and Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil. The comments, suggestions, and corrections from three anonymous referees from the editorial board of *Human Biology* are gratefully acknowledged. Many of them were included in the final version of this article. We also thank Maria Teresa B. M. Auricchio for technical support and Lilian Dluhosch for reading critically the manuscript. The helpful assistance of Lilian Kimura in the collection of genealogical information is also gratefully acknowledged.

Received 10 July 2014; revision accepted for publication 19 January 2015.

Literature Cited

Abney, M., M. S. McPeek, and C. Ober. 2000. Estimation of variance components of quantitative traits in inbred populations. *Am. J. Hum. Genet.* 66:629–650.

Angeli, C. B., L. P. Capelli, M. T. B M. Auricchio et al. 2005. AGG interspersion patterns in the CGG repeat of the FMR1 gene and linked DXS548/FRAXAC1 haplotypes in Brazilian populations. *Am. J. Med. Genet.* 132A:210–214.

Angeli, C. B., L. Kimura, M. T. B. M. Auricchio et al. 2011. Multilocus analyses of seven candidate genes suggest interacting pathways for obesity-related traits in Brazilian populations. *Obesity* 19:1,244–1,251.

Auricchio, M. T. B. M., J. P. Vicente, D. Meyer et al. 2007. Frequency and origins of hemoglobin S mutation in African-derived Brazilian populations. *Hum. Biol.* 79:667–678.

Bittles, A. H. 2002. Endogamy, consanguinity and community genetics. *J. Genet.* 81:91–98.

Cannings, C., and A. W. Edwards. 1969. Expected genotypic frequencies in a small sample: Deviation from Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* 21:245–247.

Cavalli-Sforza, L. L., and W. F. Bodmer. 1971. *The Genetics of Human Populations.* San Francisco: W. H. Freeman.

Cotrim, N. H., M. T. Auricchio, J. P. Vicente et al. 2004. Polymorphic ALU insertion in six Brazilian African-derived populations. *Am. J. Hum. Biol.* 16:264–277.

Dorsten, L. E., L. Hotchkiss, and T. M. King. 1999. The effect of inbreeding on early childhood mortality: Twelve generations of an Amish settlement. *Demography* 36:263–271.

Ellis, W. S., and W. T. Starmer. 1978. Inbreeding as measured by isonymy, pedigrees, and population size in Törbel, Switzerland. *Am. J. Hum. Genet.* 30:366–376.

Freire-Maia, N. 1957. Inbreeding in Brazil. *Am. J. Hum. Genet.* 9:284–298.

Freire-Maia, N. 1990. Genetic effects in Brazilian populations due to consanguineous marriages. *Am. J. Med. Genet.* 35:115–117.

Freire-Maia, N., and H. Krieger. 1963. A Jewish isolate in southern Brazil. *Ann. Hum. Genet. Lond.* 27:31–39.

Fyfe, J. L., and N. T. Bailey. 1951. Plant breeding studies in leguminous forage crops I. Natural cross-breeding in winter beans. *J. Agric. Sci.* 41:371–378.

Hamamy, H., L. Jamhawi, J. Al-Darawsheh et al. 2005. Consanguineous marriages in Jordan: Why is the rate changing with time? *Clin. Genet.* 67:511–516.

Jackson, C. E., W. E. Symon, E. L. Pruden et al. 1968. Consanguinity and blood group distribution in an Amish isolate. *Am. J. Hum. Genet.* 20:522–527.

Kimura, L., C. B. Angeli, M. T. B. M. Auricchio et al. 2012. Multilocus family-based association analysis of seven candidate polymorphisms with essential hypertension in an African-derived semi-isolated Brazilian population. *Int. J. Hypertens.* 2012:859219.

Kimura, L., E. M. Ribeiro-Rodrigues, M. T. B. M. Auricchio et al. 2013. Genomic ancestry of rural African-derived populations from southeastern Brazil. *Am. J. Hum. Genet.* 25:35–41.

Mingroni-Netto, R. C., C. B. Angeli, L. Kimura et al. 2009a. Doenças modernas nos antigos quilombos: A obesidade e a hipertensão no Vale do Ribeira. In *Saúde nos quilombos*, A. Volochko and L. E. Batista, orgs. São Paulo, Brazil: Instituto da Saúde, 179–191.

Mingroni-Netto, R. C., M. T. B. M. Auricchio, and J. P. Vicente. 2009b. Importância da pesquisa do traço e da anemia falciforme nos remanescentes de quilombos do Vale do Ribeira-SP. In *Saúde nos quilombos*, A. Volochko and L. E. Batista, orgs. São Paulo, Brazil: Instituto da Saúde, 169–177.

Pasinato, R., and K. I. Rettl. 2009. Desenvolvimento local sustentável: A contribuição das comunidades quilombolas do Vale do Ribeira. In *Saúde nos quilombos*, A. Volochko and L. E. Batista, orgs. São Paulo, Brazil: Instituto da Saúde, 43–56.

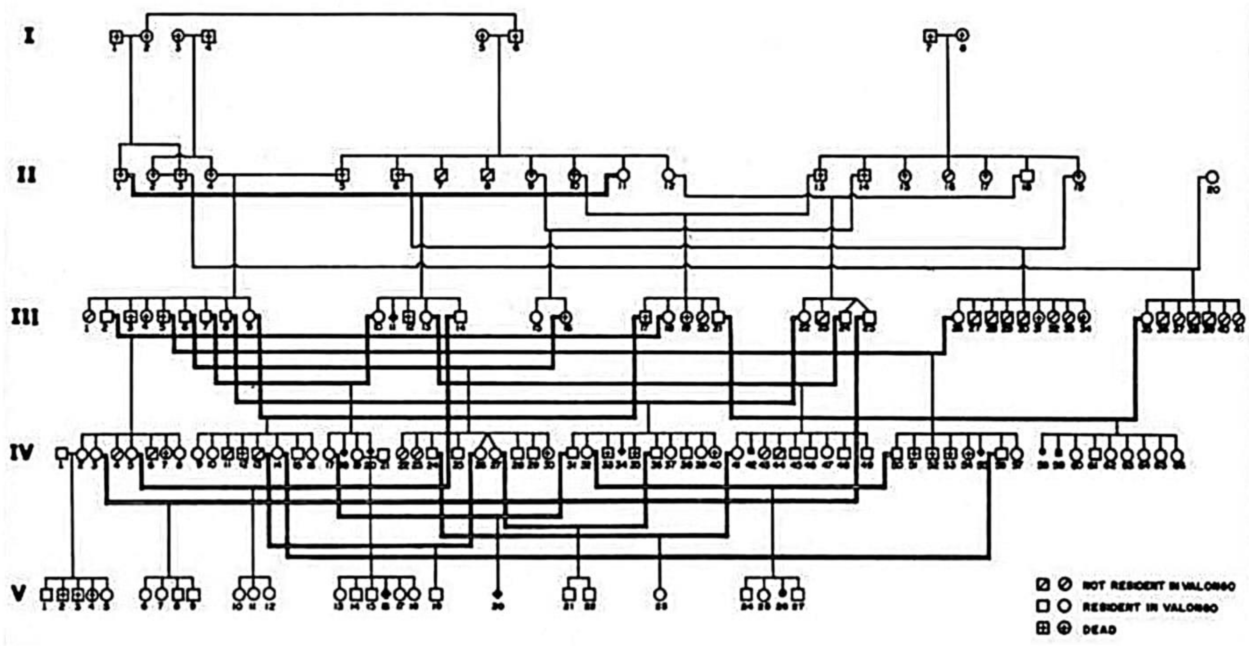
Rozen, S., and H. J. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, S. Krawetz and S. Misener, eds. Totowa, NJ: Humana Press, 365–386.

Santos, K. M. P., and N. Tatto, eds. 2008. *Agenda socioambiental de comunidades quilombolas do Vale do Ribeira.* São Paulo, Brazil: Ipsis Gráfica e Editora.

Souza, I. R., and L. Culpi. 1992. Valongo, an isolated Brazilian black community. I. Structure of the population. *Braz. J. Genet.* 15:439–447.

Wright, S. 1922. Coefficients of inbreeding and relationship. *Am. Nat.* 56:330–338.

Yeh, E., L. Kimura, F. I. Errera et al. 2008. Association of polymorphisms at the ADIPOR1 regulatory region with type-2 diabetes and body mass index in a Brazilian population with European or African ancestry. *Braz. J. Med. Biol. Res.* 41:468–472.



SUPPLEMENTARY FIGURE S1. Genealogy of quilombo from Valongo located in the state of Santa Catarina, Brazil (from Souza and Culpi 1992).

Supplementary Table S1. Primer Sequences and Fluorescence Types of All Microsatellite Loci

Locus	Chr	Forward Primer 5′–3′	Reverse Primer 5′–3′	Fluorescence[a]	Multiplex[b]
D1S551	1	TGGTGATCTGCCCCTATTCTA	TGGGAGTGTGCTCATTTTTAAC	FAM	II
D4S3248	4	CACACAGACAGAAAGCGTTACA	AATGCAGTGGGCCTATGTATCTA	FAM	II
D5S816	5	GAGCTATTGCCACTGAAAATCA	CTACTTGGCATCCCTGATGG	FAM	II
D6S1040	6	ATTGGATGAGGCTGGTGAGA	GGAAATGGCCAGAAAATCAG	FAM	IV
D7S821	7	TTTAAGATGGTGTGTGAAGCAGTAG	GGGGCAATAGGTAGGGAACTATAA	HEX	I
D7S3061	7	CCTGGCCTACTATAGGATTTTATCA	GGAAGAGTGGGTGAGGAAAGTA	FAM	II
D8S2324	8	GCAGGTGTTCCTGTCCATAATC	TGACGGAATGAGACTCCATCTAA	FAM	IV
D9S922	9	GAATTCACTCACGGAGCATACA	TCACAGCCACACAAGGACATA	HEX	I
D9S301	9	TTCAAGACAGACAGGCAGACA	GGAAGGTGTGCAAGGATGTT	HEX	III
D10S1426	10	TTTGCTTGGCACCAACTATTC	GTTGAAAACAGGGGCCTACAC	HEX	I
D13S317	13	GAAGTCTGGGATGTGGAGGA	TCCTTCAACTTGGGTTGAGC	FAM	IV
D16S539	16	CAAGCTCTTCCTCTTCCCTAGAT	GTGTGTGCATCTGTAAGCATGTAT	HEX	I
D18S535	18	GACAAAAGCCACACCCATAACT	GCAGTTCCTTTCCTGGGATAAT	HEX	III
D19S559	19	ACCAGCCTGACCAACATAGTG	GGAGGTCGATTTGGGACATA	FAM	IV
D20S482	20	ATCAGAGGACAGCCTCCATATC	CAGAGACACCGAACCAATAAGA	HEX	III
D21S1437	21	GGTTGATTCCATGTCTTTGCT	TGAGGTGCTCCCAAACTCTT	HEX	III

[a]HEX, hexachloro-6-carboxyfluorescein; FAM, 6-carboxyfluorescein.
[b]Roman numerals represent how microsatellites were grouped in the PCR.

Supplementary Table S2. Number of Individuals Genotyped for All 30 Loci (N_{G-30})

Community	N_{G-30}	N	Proportion
AB	17	573	0.0297
AN	8	320	0.0250
GA/SP	16	266	0.0602
IV	9	270	0.0333
MR/PS	8	184	0.0435
NH	7	447	0.0157
PC	16	286	0.0599
TU	6	295	0.0203
Total	*87*	*2,641*	*0.0329*

Communities: AB, Abobral; MR, Maria Rosa; PS, Pilões; GA, Galvão; SP, São Pedro; PC, Pedro Cubas; IV, Ivaporanduva; TU, Sapatu; AN, André Lopes; NH, Nhunguara. *N*, total number of inhabitants of each community. The proportion of genotyped individuals per community is also given.

Supplementary Table S3. Number of Individuals Genotyped for at Least 27 of 30 Loci (N_{G-27})

Community	N_{G-27}	N	Proportion
AB	26	573	0.0454
AN	20	320	0.6250
GA/SP	31	266	0.1165
IV	35	270	0.1296
MR/PS	25	184	0.1359
NH	24	447	0.0537
PC	29	286	0.1014
TU	17	295	0.0576
Total	*207*	*2,641*	*0.0784*

Communities: AB, Abobral; MR, Maria Rosa; PS, Pilões; GA, Galvão; SP, São Pedro; PC, Pedro Cubas; IV, Ivaporanduva; TU, Sapatu; AN, André Lopes; NH, Nhunguara. *N*, total number of inhabitants of each community. The proportion of genotyped individuals per community is also given.

2. CHAPTER 2

Chapter 2 is a paper dealing with the calculation of the variance of the estimated inbreeding coefficient in the generalized case of multiple alleles segregating at an autosomal locus. This theoretical study, performed in collaboration with Professor Paulo A. Otto, showed that reliable simple approximations, obtained by applying basic statistical methods, can be used to estimate the variance of **f**. The estimates obtained with our approximation methods were fully validated by computer simulation methods we developed. The article was published in the periodical Journal of Genetics (under the reference Otto PA, Lemes RB. A note on the variance of the estimate of the fixation index F. **J. Genet. 94, 759–763.** 2015).

© Indian Academy of Sciences

RESEARCH NOTE

A note on the variance of the estimate of the fixation index F

PAULO A. OTTO* and RENAN B. LEMES*

Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, Caixa Postal (P.O. Box) 11.461, 05422-970 São Paulo, SP, Brazil

[Otto P. A. and Lemes R. B. 2015 A note on the variance of the estimate of the fixation index F. *J. Genet.* **94**, 759–763]

Introduction

In the two-allele case, the formulas for the estimated variances of allelic frequency $p = 1 - q$ and fixation index (average inbreeding coefficient) F are known in the specialized literature of statistical genetics. Besides presenting here an alternative manner to estimate the variance of both parameters, we also derive a very simple approximation for the estimate of the variance of F. The approximation, with adequate validity, can be applied not only to the two-allele case but also to the generalized case of any number of alleles segregating at an autosomal locus.

The variance of F has many practical applications in population genetics. For example, if geneticists are interested in a precise determination of its value, commonly the parameter is estimated from sets of data obtained from the genotypic analysis of several independent autosomal loci of the same population. If the estimates of F for loci $1, 2, \ldots, k$ are $\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_k$, the method of averaging these estimates is obtained usually by weighing them by the reciprocal of their corresponding variances:

$$\bar{F} = \frac{\sum \frac{\hat{F}_i}{var\left(\hat{F}_i\right)}}{\sum \frac{1}{var\left(\hat{F}_i\right)}} \quad .$$

Our paper deals with the population as specified by formulas (2.22) on page 65 of Weir's monograph (Weir 1996). The virtue of the resulting approximation for the estimate of *var* (F) we provide is a simple formula with adequate validity for multiple alleles, whereas Weir does leave his reader with details to be supplied.

Our results are presented below in three different sections: the first one deals with the case of two alleles, leading naturally to a second section on multiple alleles; a third section deals with simulation studies we performed to validate the approximations derived here.

The special case of two autosomal alleles

The generic population genotype frequencies in relation to an autosomal biallelic locus can be represented by equations

$$P(AA) = p^2 + pqF,$$

$$P(Aa) = 2pq(1 - F),$$

and

$$P(aa) = q^2 + pqF,$$

that represent a special case of Weir's population formulas referred to in the previous section, and where $p = P(A)$ is the frequency of allele A, $q = 1 - p = P(a)$ the frequency of its alternative allele a, and F the fixation index normally obtained from the formula,

$$F = 1 - \frac{h}{2pq},$$

where h is the heterozygous frequency $h = \frac{NAa}{N}$, NAa the observed number of heterozygous individuals, and N the total number of sampled subjects.

Since the expected values corresponding to observed numbers NAA, NAa and Naa of individuals AA, Aa and aa, respectively in a sample with size N and to a fixation index (average inbreeding coefficient) $F \neq 0$ are

$$N\left(p^2 + pqF\right),$$

$$2Npq(1 - F),$$

and

$$N\left(q^2 + pqF\right)$$

*For correspondence. E-mail: Paulo A. Otto, otto@usp.br; Renan B. Lemes, lemes.rb@usp.br.

Keywords. inbreeding; inbreeding coefficient; fixation index; variance estimation; variance of the inbreeding coefficient.

respectively, the likelihood function in logarithmic form is given by expression:

$$L = NAA \log\left[p^2 + p(1-p)F\right] + NAa \log\left[2p(1-p)(1-F)\right] + Naa \log\left[(1-p)^2 + p(1-p)F\right].$$

Maximum likelihood estimates of both p and F are obtained from the system $\{\frac{\partial L}{\partial p} = 0,\ \frac{\partial L}{\partial F} = 0\}$ and it is not difficult to determine that these solutions are identical to the estimates of p and F obtained through the application of intuitive direct counting methods: $\hat{p} = \hat{d} + \frac{\hat{h}}{2}$ and $\hat{F} = 1 - \frac{\hat{h}}{2\hat{p}(1-\hat{p})}$.

In the formulas above (and in many equations that follow) symbols like $\hat{p} = 1 - \hat{q}$ and $\hat{F}$ have carets because they are not unknown population (true) values but estimates of the corresponding parameters from the population, obtained from simple random sampling of a large population with genotype proportions occasionally different from Hardy–Weinberg ratios.

The determination of the values for the variances of p and F using iterative numerical procedures such as the usual generalized Newton–Raphson method is a complicated issue since it is practically impossible to get convergence to the estimation points p and F (Weir 1996), but values of $var(\hat{p})$ and $var(\hat{F})$, the variances of the estimated values of p and F can be taken directly from the variance–covariance matrix obtained by inverting the information matrix of second derivatives evaluated at estimation points $\{\hat{p}, \hat{F}\}$:

$$var(\hat{\mathrm{p}}) = \frac{a_{22}}{a_{11}.a_{22} - a_{12}.a_{21}}$$

and

$$var(\hat{\mathrm{F}}) = \frac{a_{11}}{a_{11}.a_{22} - a_{12}.a_{21}},$$

where $a_{11} = -\frac{\partial^2 L}{\partial p^2}$, $a_{12} = -\frac{\partial^2 L}{\partial p \partial F}$, $a_{21} = -\frac{\partial^2 L}{\partial F \partial p}$, and $a_{22} = -\frac{\partial^2 L}{\partial F^2}$, with all four second derivatives evaluated at estimation points

$$\hat{p} = \hat{d} + \frac{\hat{h}}{2}$$

and

$$\hat{F} = 1 - \frac{\hat{h}}{2\hat{p}(1-\hat{p})}.$$

In the case of the variance of the estimated value of p, we obtain $var(\hat{p}) = \frac{\hat{p}\hat{q}(1+\hat{F})}{2N}$, as expected. This formula coincides with the expression obtained by Curie-Cohen (1982) and other authors (references of the many papers on the variances of p and F by Cockerham, Weir, and Cockerham and Weir, in Weir 1996) using different alternative methods.

Since

$$\frac{var(\hat{F})}{var(\hat{p})} = \frac{a_{11}}{a_{22}},$$

we get straightforwardly

$$var(\hat{F}) = \frac{(1-\hat{F})\left[2\hat{p}\hat{q} + 2\hat{F}(1-3\hat{p}\hat{q}) - \hat{F}^2(\hat{p}-\hat{q})^2\right]}{2N\hat{p}\hat{q}}, \tag{1}$$

a result that is algebraically equivalent to the formulas derived by Fyfe and Bailey (1951) and Curie-Cohen (1982) using alternative methods.

In the two-allele case, an approximate value of the variance of the estimate F can be obtained in a simple and straightforward way if we treat p, that can be directly calculated from the sample through $\hat{p} = \hat{d} + \frac{\hat{h}}{2}$, as an independently estimated parameter. Then the variance of $\hat{F}$ is obtained directly from $(a_{22})^{-1}$, taking form

$$var(\hat{F}) = \left\{\frac{NAA(1-\hat{p})^2}{\left[\hat{p} + (1-\hat{p})\hat{F}\right]^2} + \frac{NAa}{(1-\hat{F})^2} + \frac{Naa(1-\hat{q})^2}{\left[\hat{q} + (1-\hat{q})\hat{F}\right]^2}\right\}^{-1}. \tag{2}$$

This formula works as well as the one derived in this paper or other expressions from the literature.

The generalized case of any number of autosomal alleles

When the number of alleles (k) segregating at an autosomal locus is larger than two, estimates obtained through intuitive counting methods (and that correspond to maximum likelihood estimates under stringent conditions) are given by

$$\hat{p}_i = \frac{2N(a_i a_i) + \sum N(a_i a_j)}{2N},$$

$\hat{p}_j = \ldots, \ldots, \hat{p}_{k-1} = \ldots$, with i fixed and $j \neq i$ varying from 1 to k, that is $\sum N(a_i a_j)$ in the formula above represents the total number of heterozygous individuals as to the i allele, and

$$F = 1 - \frac{\sum\sum N(a_i a_j)}{2N\left(\sum\sum p_i p_j\right)},$$

with i varying from 1 to k and $j > i$, that is $\sum\sum N(a_i a_j)$ in the formula above represents the total number of heterozygous individuals as to alleles i and j.

In spite of being generally impossible to obtain convergence to the values shown above using numerical iterative procedures and to get the value of the variance of $\hat{F}$ by means of variations of Fisher's variance method (a rigorous argumentation on the subject is presented by Weir on pages 49–51 of his 1996 book), numerical values of $var(\hat{F})$ can be obtained either from large series of computer simulations or from the inspection of the main diagonal of the variance–covariance matrix evaluated at estimation points $\hat{p}_1, \ldots, \hat{p}_{k-1}, \hat{F}$. The variance of $\hat{p}_i$ in the multiallelic case can be determined independently through the formula (Curie-Cohen 1982; Weir 1996)

$$var\left(\hat{p}_i\right) = \frac{\hat{p}_i\left(1-\hat{p}_i\right)\left(1+\hat{F}\right)}{2N}.$$

Literal expressions for the variance of the estimated value of F when the number of alleles is larger than two can be obtained from the matrix method we used in the previous section (two-allele case), but they are however much more complicated; reliable, easily handled approximations should be preferred instead on practical grounds. Curie-Cohen (1982) and Robertson and Hill (1984) derived some of them under stringent statistical assumptions.

The real importance of the approximate formula derived for the two-allele case, however, stems from the fact that it is very easy to generalize it for the generic case of any number of alleles segregating at an autosomal locus. In fact, for the three-allele case, by treating the estimates $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3 = 1-\left(\hat{p}_1+\hat{p}_2\right)$ as independently estimated parameters, each obtained by means of the intuitive formula

$$\hat{p}_i = \frac{2N\left(a_ia_i\right)+\sum N\left(a_ia_j\right)}{2N},$$

with i fixed and $j \neq i$ varying from 1 to k-1, that is $\sum N\left(a_ia_j\right)$ in the formula above represents the total number of heterozygous individuals as to the i allele, the corresponding formula for the variance of $\hat{F}$ is taken from

$$\left(-\frac{\partial^2 L}{\partial \hat{F}^2}\right) = \frac{1}{var\left(\hat{F}\right)} = \frac{N\left(a_1a_1\right)\left(1-\hat{p}_1\right)^2}{\left[\hat{p}_1+\left(1-\hat{p}_1\right)\hat{F}\right]^2} + \frac{N\left(a_2a_2\right)\left(1-\hat{p}_2\right)^2}{\left[\hat{p}_2+\left(1-\hat{p}_2\right)\hat{F}\right]^2} + \frac{N\left(a_3a_3\right)\left(1-\hat{p}_3\right)^2}{\left[\hat{p}_3+\left(1-\hat{p}_3\right)\hat{F}\right]^2} + \frac{N\left(a_1a_2\right)}{\left(1-\hat{F}\right)^2} + \frac{N\left(a_1a_3\right)}{\left(1-\hat{F}\right)^2} + \frac{N\left(a_2a_3\right)}{\left(1-\hat{F}\right)^2}$$

so that in the k-allele case we have

$$\left(-\frac{\partial^2 L}{\partial \hat{F}^2}\right)^{-1} = var\left(\hat{F}\right) = \left\{\sum \frac{N\left(a_ia_i\right)\left(1-\hat{p}_i\right)^2}{\left[\hat{p}_i+\left(1-\hat{p}_i\right)\hat{F}\right]^2} + \frac{\sum\sum N\left(a_ia_j\right)}{\left(1-\hat{F}\right)^2}\right\}^{-1}, \quad (3)$$

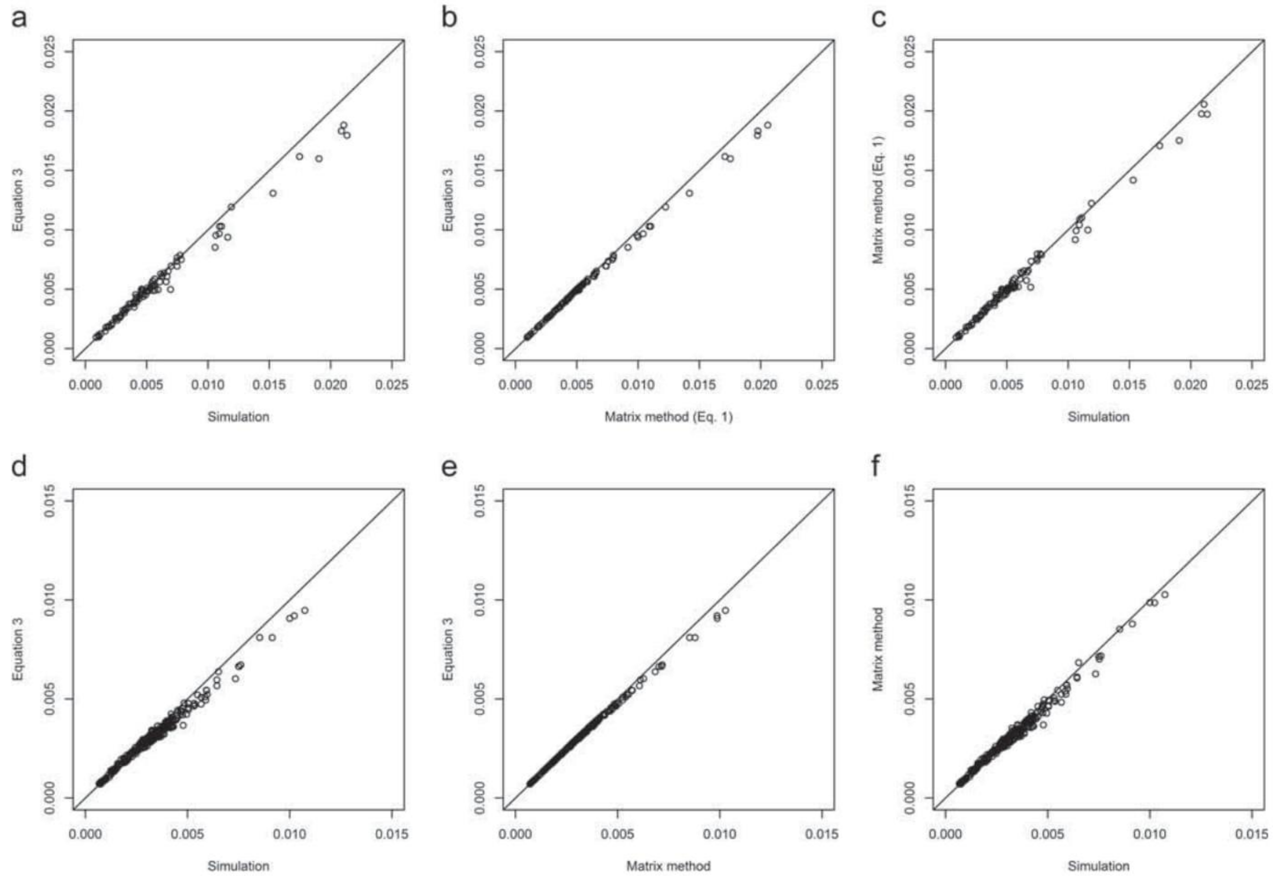


Figure 1. Comparison of values of $var(F)$ corresponding to different combinations of values of p and F. In all cases p varied from 0.05 to 0.95 in intervals of 0.05, F varied from 0.1 to 0.9 in intervals of 0.1, and $N = 200$ in the cases of two alleles (graphs a,b,c) and three alleles (graphs d,e,f).

where $N(a_i a_i)$ indicates the observed number of homozygous individuals as to allele a_i and $N(a_i a_j)$ (with $j > i$) the observed number of heterozygous individuals as to both alleles a_i and a_j. This formula is valid for any value of $k \geq 2$, i.e. the case $k = 2$ (equation 2) is just a special case of equation 3.

Computer simulations

We also obtained values of *var* (*F*) using computer simulation methods, in which we proceeded as follows: from a relatively large number of sets of known values of F and allele frequencies $\{p_1, p_2, \ldots\}$, we determined the quantities $\{p_{11} = p_1 F + p_1^2 (1 - F),\ p_{12} = 2p_1 p_2 (1 - F),\ \ldots\}$, that were used to generate, through computer bootstrap simulations with replacement, for each combination of $\{p_1, p_2, \ldots, F\}$, 200 genotypes $\{a_1 a_1\ , a_1 a_2, \ldots\}$; from the genotype and allele frequencies estimated from each set of 200 genotypes so generated, we calculated the value of the fixation index *F*. The process was repeated 1000 times for each combination $\{p_1, p_2, \ldots, F\}$, and from the set of 1000 values of F so obtained we determined the value of *var* (*F*) after the usual formula $var\ (F) = \frac{\sum F_i^2}{1000} - \left(\frac{\sum F_i}{1000}\right)^2$. The values of *var* (*F*) obtained with different combinations of $\{p_1, p_2, \ldots, F\}$ could then be compared with the values calculated using the matrix method (detailed for the 2-allele case) or their corresponding approximations given by generalized equation 3.

The results we got when the values obtained (in the cases of two to six alleles) with either the simulation or the matrix method were compared to the values obtained with the approximation given by equation 3 were virtually the same beyond any reasonable doubt, as the graphs of figure 1 show for the cases of two or three alleles.

Taking into account the facts presented above, we studied, in the 2-allele case, the behaviour of the relative error, defined as $\frac{|v_1 - v_2|}{v_1}$, where v_1 and v_2 are respectively corresponding values of *var* (*F*) with same *p* and *F* obtained using equations 1 and 2. Extensive numerical analysis of the relative error showed that it is on average a bit large (its maximum value is around 11%) only when *F* has intermediate values (near 0.5) and the frequencies of the two alleles are very uneven. For other combinations of *p* and *F* the relative error is small, generally much less than 10%. For extreme *F* values (near 0 or 1) the relative error is very small (less than 2%) for any combination of allele frequencies and practically negligible when the allelic frequencies are approximately equal. The surface graph of figure 2, corresponding to the situation above discussed of two alleles and to a population size of $N = 200$, shows this in a straight forward manner. When the number of alleles was larger than two, the corresponding analyses were performed directly using the results shown by graphs as in figure 1 and the larger deviations from the diagonal line occurred exactly in the situations described for the case of two alleles, i.e. when *F* had intermediate values and allele frequencies were very uneven.

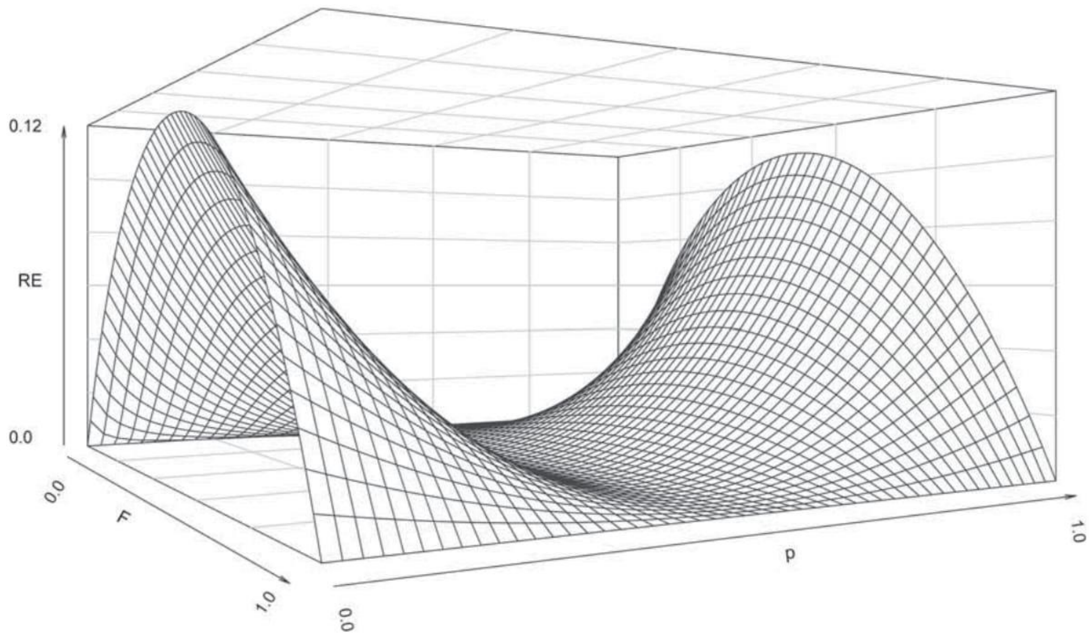


Figure 2. Relative error (RE) of *var* (*F*) values obtained using equations 1 and 2 in relation to all possible combinations of *p* and *F* for the case of two alleles. $RE = \frac{|v_1 - v_2|}{v_1}$, where v_1 and v_2 are corresponding values of *var* (*F*) obtained using equations 1 and 2, respectively.

Acknowledgements

We thank Drs Bruce S. Weir, Alan E. Stark and Richard M. Single for their valuable comments and suggestions on preliminary drafts of this manuscript. The very useful comments, suggestions, and important corrections made by two anonymous referees are also fully appreciated and acknowledged; most of them were incorporated (some of them literally) into the revised (final) version of the paper. This work was partly funded by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), all from Brazil.

References

Curie-Cohen M. 1982 Estimates of inbreeding in a natural population: a comparison of sampling properties. *Genetics* **100**, 339–358.

Fyfe J. L. and Bailey N. T. J. 1951 Plant breeding studies in leguminous forage crops I. Natural cross-breeding in winter beans. *J. Agric. Sci.* **41**, 371–378.

Robertson A. and Hill W. 1984 Deviations from Hardy-Weinberg proportions: sampling variances and use in the estimation of inbreeding coefficients. *Genetics* **107**, 703–718.

Weir B. S. 1996 *Genetic data analysis II.* Sinauer Associates Inc, Sunderland, MA, USA.

Received 11 February 2015, in final revised form 11 May 2015; accepted 12 May 2015
Unedited version published online: 28 May 2014
Final version published online: 20 October 2015

3. CHAPTER 3

This chapter deals with inbreeding levels and demographic events in the quilombo isolate, inferred from the analysis of high density SNP datasets. All genotyping laboratory procedures of the quilombo dataset we used were performed by members of the laboratories of Professors Regina C. Mingroni Netto and Diogo Meyer.

The analyses performed focused mainly on (1) estimating reliable inbreeding levels by means of traditional methods applied to high density data, using a novel (as far as we know) approach that combines the information of two datasets (a complete one and another with no linkage disequilibrium); and (2) making inferences from demographic events based on the distribution of runs of homozygosity of different sizes in quilombo individuals.

A detailed description of the data cleaning process is preseted in Annex I.

The manuscript is in its final stage of review in order to be submitted to a specialized genetics periodical.

INBREEDING ESTIMATES USING DIFFERENT ALTERNATIVE METHODS: AN EXAMPLE APPLIED TO THE STUDY OF AN ADMIXED BRAZILIAN ISOLATE

Renan B. Lemes, Kelly Nunes, Juliana E. P. Carnavalli, Lilian Kimura, Regina C. Mingroni Netto, Diogo Meyer, Paulo A. Otto *

Department of Genetics and Evolutionary Biology, Instituto de Biociências, Universidade de São Paulo, Caixa Postal (P. O. Box) 11461, 05422-970 São Paulo SP Brazil.

* Corresponding author (otto@usp.br, phone 55-11-3091-7591, fax 55-11-3091-7553)

Keywords: Inbreeding; Inbreeding coefficient; Wright's fixation index; Runs of homozygosity; Population isolate.

ABSTRACT

The analysis of high density genomic data (~400.000 autosomal SNPs) enabled the reliable estimation of inbreeding levels in 541 individuals sampled from a highly admixed Brazilian population isolate (an African-derived quilombo in the State of São Paulo). To achieve this, different alternative methods were applied to the joint information of two sets of markers (one complete and another excluding loci in patent linkage disequilibrium). This strategy allowed the detection and further exclusion of markers that biased the estimation of the average population inbreeding coefficient (Wright's fixation index **f**), which value was eventually estimated in ~0.01 using any of the methods we applied. Quilombo demographic inferences were made by means of runs of homozygosity (ROHs) analyses, which were adapted to cope with a highly admixed population with a complex foundation history.

INTRODUCTION

Measures of population inbreeding levels have been traditionally obtained from (1) the direct genotyping of population samples through estimates of heterozygous frequency deviations from the proportions expected under random-mating assumptions (Hardy–Weinberg or HW expectations) or (2) from the analysis of sets of individual or grouped genealogies, that in rare instances may include precise relationship information on more than three or four generations.

The situation has changed dramatically with the recent use of large datasets of genomic autosomal single nucleotide polymorphisms (SNPs), allowing the identification of long tracts of consecutive homozygosity (runs of homozygosity or ROHs) in human population samples. Studies using these approaches have revealed high levels of autozigosity even in cosmopolitan non-inbred populations, showing that there exists, as expected by the out-of-Africa model of human origins, an increment of inbreeding levels and a significant reduction of genetic diversity which proportional to the distance from Africa (Kirin *et al.*, 2010, Leutenegger *et al.*, 2011, Pemberton *et al.*, 2012). An important mechanism responsible for a large portion of genomic homozygosity levels, composed mainly by short and intermediate ROHs, is the background relatedness, which results from the combined effects of demographic and evolutionary events, such as remote inbreeding, geographic isolation, small population size with bottleneck and founder effects, and long-lasting and stable systems of endogamous marriages due to the persistence of cultural traditions (McQuillan *et al.*, 2008; Kirin *et al.*, 2010; Pemberton *et al.*, 2012; Teo *et al.*, 2012; Pemberton and Rosenberg, 2014).

Population isolates are powerful tools for medical and evolutionary studies, since many of them have well documented pedigrees, high prevalence of individuals affected by rare genetic conditions, high degree of inbreeding due to cultural practices or limited population size, and a demographic history of foundation consisting of bottlenecks followed by founder effects (Arcos-Burgos and Muenke, 2002). Even in the case of population isolates with absence of well documented pedigrees and a paucity of historical records, reliable genetic information can be obtained from the analysis of large SNP datasets. Several studies on inbreeding and demographic history have been successfully performed around the world in isolated populations with variable amounts of genealogical documentation and historical records of population-based evolutionary phenomena (Carothers *et al.*, 2006; McQuillan *et al.*, 2008; Abdellaoui *et al.*, 2015; Ben Halim *et al.*, 2015; Jalkh *et al.*, 2015; Karafet *et al.*, 2015).

The admixture of populations with different genetic backgrounds can create high levels of linkage disequilibrium (LD), which besides taking many generations to disappear, will interfere with the distribution of ROHs lengths, thus enabling the recovery of genetic information on important historic events, including the dynamics of the admixture process (Templeton, 2006, Kirin, 2010).

By means of the analysis of a high-density dataset of genomic autosomal single nucleotide polymorphisms (SNP), we make inferences on inbreeding levels and demographic history of a Brazilian isolate with about 40% African, 39% European and 21% Amerindian contribution (Kimura *et al.*, 2013). This study presents: (1) an alternative way to estimate the population inbreeding coefficient (Wright’s fixation index **f**), based solely on the analysis of a high-density SNP array; (2) the application of a likelihood-based approach to identify genomic ROHs in a population that underwent a complex demographic history with tri-hybrid ancestral contribution; (3) a comparison between individual estimates of the inbreeding coefficient obtained from SNP genotypes through different methods.

Based on our results of the distribution of ROHs lengths, we discuss its relation to the process of population admixture and in the combination of background relatedness and recent inbreeding events.

SUBJECTS AND METHODS

The Brazilian *Quilombo* (QUI) Admixed Population

The present study was performed in an admixed Brazilian isolate located in the Ribeira River Valley, in the southern part of the State of São Paulo, Brazil (Figure 1). This isolate, known in Brazil as *quilombo*, was founded by runaway, abandoned and freed slaves, who created small rural settlements in isolated areas inside the Atlantic rainforest for several generations; other details of interest are described in Kimura *et al.* (2013) and Lemes *et al*. (2014). The isolate aggregates 12 communities that were treated as a single one, since the degree of differentiation among its communities is very low, with $\mathbf{f}_{\mathbf{ST}}$ indexes generally smaller than 0.05 (Lemes *et al.*, 2014).

This quilombo population was founded around 1890 mainly by runaway, freed or abandoned African-descendant slaves (some of them being the mixed offspring of white farmer owners and African female slaves) and a few pure or mixed native Americans (for other details on the quilombo population structure and demography, see Kimura *et al.*, 2013 and Lemes *et al.*, 2014). Some fifty years ago a road was built near the communities and a significant migration flow of some neighbor populations began to take place. Because of this recent history of admixture, some people argue that the quilombo reported here does not represent a true isolate anymore. In order to warrant or preserve the isolate condition with which we classify this population aggregate, however, all individuals selected for this study, aged between 17-65 years, have at least two generations of quilombo ancestors.

DNA samples were extracted from peripheral blood and genotyped with the high density SNP array Axiom Genome-Wide Human Origins (~600,000 SNPs) according to the manufacturer's standards (Affymetrix/Thermo-Fisher Scientific). We analyzed DNA samples from 541 individuals (Table S1) from the Ribeira River Valley, 365 of them having already been genotyped in a previous study (Nunes *et al.*, 2016) and the remaining 176 samples of this study. The research was approved by the Ethics Committee, Instituto de Ciências Biomédicas, Universidade de São Paulo (111/CEP, Feb. 14th 2001), and an informed consent was obtained from all its participants or their legal guardians.

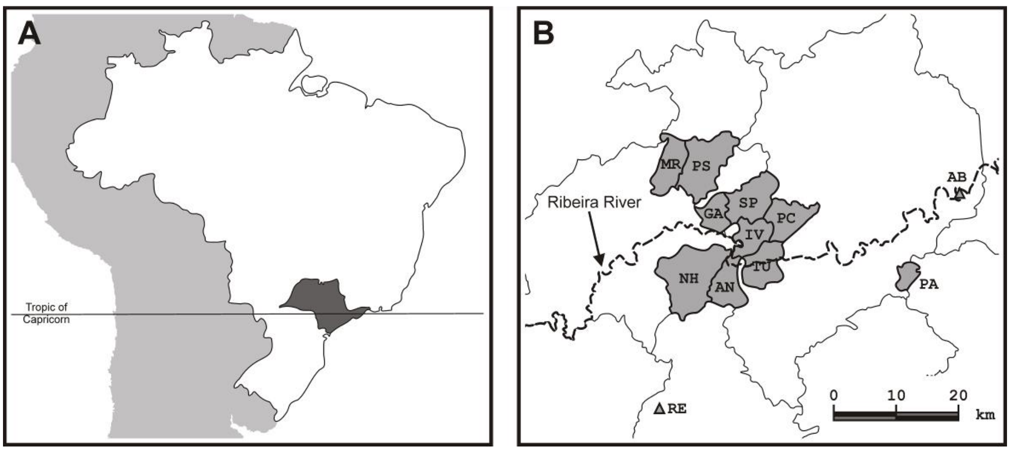


Figure 1: (A) State of São Paulo (grey) within Brazilian territory in South America. (B) Location of quilombo communities. AB, Abobral; AN, André Lopes; GA, Galvão; IV, Ivaporanduva; MR, Maria Rosa; NH, Nhunguara; PA, Poça; PC, Pedro Cubas; PS, Pilões; RE, Reginaldo; SP, São Pedro; TU, Sapatu.

Data Preparation (Data Cleaning and Filter)

The data cleaning excluded systematically: (1) all markers with low quality scores, using the software Genotype Console Software v.4.2 according to the manufacturer's standards parameters (Genotype Console Workflow – Affymetrix/Thermo Fisher Scientific); (2) all markers that presented significant differences in missing data proportions between groups (sexes, batches, and subpopulations) using the R package GWASTools v. 3.5 (Gogarten *et al.*, 2012); (3) all data from mitochondria and X and Y chromosomes; (4) all genotyped loci with more than 1% of missing values; (5) all markers with minor allele frequency MAF = 0, that is, all alleles that were fixed; (6) all data from loci that extremely deviated from Hardy-Weinberg proportions ($P \leq 10^{-4}$), using the asymptotic exact test (Wigginton *et al.*, 2005) by means of the software PLINK v1.07 (Purcell *et al.*, 2007); (7) all data corresponding to 300 markers located in the extremities of all chromosome arms. The final set consisted of data from 485,957 autosomal SNPs.

HGDP Samples

Populations from different geographic regions have distinct and well-established distribution of ROHs sizes (Kirin *et al.*, 2010; Pemberton *et al.*, 2012). In order to identify without significant biases the ROHs in our quilombo samples (QUI), we selected three populations belonging to different geographic sources, available from the public Human Genome Diversity Panel databank (HGDP): African Yoruba (YRI), European French (FRE), and Asian Han Chinese (CHB), containing respectively data from 22, 28 and 34 individuals. As in the case of the QUI sample, markers with extreme deviations from Hardy-Weinberg proportions ($P \leq 10^{-4}$) and with 1% or more missing values were excluded.

The HGDP information was merged with the QUI dataset, resulting in a set of 402,142 commonly shared markers. Data for 29,897 SNPs corresponding to genotypes coded for opposite strands (i.e. forward and reverse) were converted in order to match QUI dataset.

Estimation of the Inbreeding Coefficient

We estimated both the average population inbreeding coefficient **f** (Wright’s fixation index) and the average individual inbreeding coefficient **f’** (Purcell *et al.*, 2007).

To obtain the average estimates (across all loci of all individuals) of both **f** and **f′** we used the information from (1) all 485,957 SNPs (complete dataset) and (2) 11,642 SNPs with no LD (no-LD dataset), obtained from the first one by means of the software PLINK v1.07, considering a threshold of $r^2 = 0.0071$, which corresponds to a critical 5% chi-square value of $\chi^2 = 3.841$ pairwise estimated in sliding windows of 50 SNPs incremented in steps of 5.

*ESTIMATION OF WRIGHT'S **f** COEFFICIENT*

The inbreeding coefficient $\mathbf{f_k}$ was obtained for each biallelic locus by means of the formula

$$\mathbf{f_k = 1 - \frac{P_k(Aa)}{2p_kq_k}} , \qquad (1)$$

where $\mathbf{P_k(Aa)}$ and $\mathbf{2p_kq_k}$ are respectively the observed and HW expected frequencies of heterozygous genotypes at the k-th locus. The mean population inbreeding coefficient ($\bar{\mathbf{f}}$) was obtained weighing the per locus $\mathbf{f_k}$ estimates by the reciprocals of their corresponding variances:

$$\bar{\mathbf{f}} = \sum \mathbf{x_k . f_k} \,, \tag{2}$$

with

$$\mathbf{x_k} = \mathbf{var}^{-1}\left(\mathbf{f_k}\right) / \sum_{j=1}^{n} \mathbf{var}^{-1}\left(\mathbf{f_j}\right) \,, \tag{3}$$

where $\mathbf{n}$ is the number of loci and $\mathbf{var(f_k)}$ is the estimate of the variance of $\mathbf{f_k}$, obtained for each biallelic locus by the formula (Fyfe and Bailey, 1951; Curie-Cohen, 1982; Otto and Lemes, 2015):

$$\mathbf{var(f_k)} = \frac{\left(1 - f_k\right)\left[2p_kq_k + 2f_k\left(1 - 3p_kq_k\right) - f_k^{\,2}\left(1 - 4p_kq_k\right)\right]}{2Np_kq_k} \,, \tag{4}$$

where $\mathbf{N}$ is the sample size, and $\mathbf{p_k}$ and $\mathbf{q_k}$ are the frequencies of the alleles segregating at the k-th bialellic SNP locus.

On the long run, one expects that the estimates of $\mathbf{f_k}$ thus obtained should be normally distributed around the average value of $\bar{\mathbf{f}}$, with the limits of the usual 95% confidence interval being given approximately by $\bar{\mathbf{f}} \pm 1.96\sqrt{\mathbf{var}\left(\mathbf{f}\right)}$, where $\mathbf{var}\left(\mathbf{f}\right)$ is now given by

$$\mathbf{var}\left(\mathbf{f}\right) = \sum \mathbf{x_k f_k^2} - \bar{\mathbf{f}}^2 \,, \tag{5}$$

with $\mathbf{x_k}$ as defined in formula (3) (Lemes *et al.*, 2014).

We also ranked the values of $\mathbf{f_k}$ in order to obtain the median and its 95% confidence interval corresponding to the set of all values between the limits of the 2.5th and 97.5th percentiles.

ESTIMATION OF THE AVERAGE INDIVIDUAL INBREEDING COEFFICIENT The estimate of the inbreeding coefficient for each individual of the sample, referred here as $\mathbf{f'_i}$, was obtained by means of the function **– – het** of the software PLINK v1.07 using the expression:

$$\mathbf{f'_i} = \frac{(\mathbf{O_i} - \mathbf{E_i})}{(\mathbf{L_i} - \mathbf{E_i})} \quad , \tag{6}$$

where $\mathbf{O_i}$ and $\mathbf{E_i}$ are the observed and expected numbers of homozygous genotypes considering all $\mathbf{L_i}$ genotyped autosomal SNPs of individual $\mathbf{i}$ (Purcell *et al.*, 2007). Average and median estimates of $\mathbf{f'}$ and the corresponding 95% confidence interval of the whole observed distribution of $\mathbf{f'_i}$ values were obtained as before, by ranking the individual values or using the normal approximation indicated above.

Identification of Runs of Homozygosity (ROHs)

The identification of ROHs was performed in the four samples (QUI, YRI, FRE, and CHB), using a sliding window of $\mathbf{n}$ markers SNP-wise-incremented along the whole genome across all individuals (Pemberton *et al.*, 2012). The windows' autozygosity LOD-scores were estimated adding the LOD-score values obtained from each marker, which is, in turn, calculated according to the expression

$$\mathbf{LOD} = \mathbf{log_{10}} \left[\frac{\mathbf{P}(\mathbf{g_i} \mid \textbf{autozygous at i})}{\mathbf{P}(\mathbf{g_i} \mid \textbf{alozygous at i})} \right] \quad , \tag{7}$$

where $\mathbf{g_i}$ is the observed genotype at a given locus of the individual $\mathbf{i}$. Both conditional probabilities take into account the allele frequencies estimated from the population, considering the occurrence of mutations and genotyping errors at a combined rate of $\boldsymbol{\varepsilon} = 0.001$ (Broman and Weber, 1999; Wang *et al.*, 2009).

The distribution of Gaussian Kernel density estimates (GKDE) of LOD-score values, calculated across all windows including all individuals for the four populations, was then obtained, first considering window sizes of $\mathbf{n} = \{10, 15, \ldots, 100\}$ SNPs and then using unity steps inside the interval $15 < \mathbf{n} < 20$. The optimal $\mathbf{n}$ (a window of size 18 markers) based on the figure that produced a clear bimodal distribution of GKDE in all four populations (**Figure 2**). This is important because the common anti-mode represents the optimal statistical boundary between alozygous (at left) and autozygous (at right) windows for the four populations. The periodicity pattern presented in the distributions (mainly in the African-derived ones) are due to the resampling procedure described below.

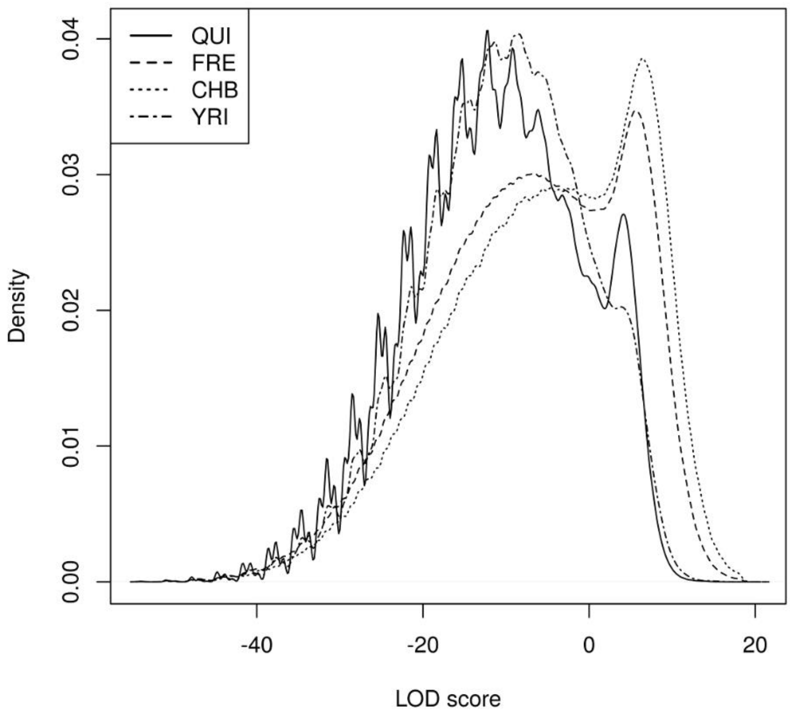


Figure 2: LOD score distribution for QUI, FRE, CHB and YRI datasets considering a window of size n = 18.

The windows corresponding to LOD-values above the anti-mode threshold (minima between the modes) were defined as autozygous; overlapping autozygous windows were grouped together to form ROHs (Pemberton *et al.*, 2012).

In order to enable the comparison of ROHs among the populations, their sample sizes were adjusted as follows: for all loci, 40 independent alleles (20 genotypes) were sampled with replacement according to their estimated frequencies; new allele frequencies were then calculated from the computer-generated sets. For all SNP markers with MAF = 0 as a consequence of the resampling procedure, the corresponding LOD estimate was set to 1 (Pemberton *et al.*, 2012).

Estimation of inbreeding Coefficient from ROHs

Individual and population inbreeding coefficients were also estimated using ROHs data. The $\mathbf{F_{ROH}}$, defined as the genomic autosomal proportion of ROHs of an individual, was estimated by the expression (McQuillan *et al.*, 2008):

$$\mathbf{F_{ROH}} = \frac{\sum \mathbf{L_{ROH}}}{\mathbf{L_{auto}}} , \qquad (8)$$

where $\sum \mathbf{L_{ROH}}$ corresponds to the length of ROHs and $\mathbf{L_{auto}}$ corresponds to the total genomic region covered by the SNP array.

The individual $\mathbf{F_{ROH}}$ figures, their population average values as well as their corresponding 95% confidence intervals were estimated considering either the total set of ROHs or only the set of class C ROHs.

Demographic Inferences from ROHs

As proposed by Pemberton *et al.* (2012), the ROHs were classified according to their lengths, considering their distribution as a mixture of three Gaussian distributions, using the function Mclust of the package mclust v.5.2 (Fraley *et al.*, 2012) of R v.3.3.0 (R Core Team, 2016), and treating the number of components, means and variances as free parameters (**Figure 3**). The distributions were then categorized in three classes: **A**, short ROHs resulting from ancient homozygous state contributing to population LD patterns; **B**, intermediate ROHs resulting from background relatedness; **C**, long ROHs reflecting recent inbreeding. The boundaries between classes A and B and classes B and C were obtained averaging the largest and the smallest values of the shorter and longer ROHs classes respectively, that is through $\frac{(A_{max} + B_{min})}{2}$ and $\frac{(B_{max} + C_{min})}{2}$, that correspond approximately to the values of pairs of A and B and of B and C with the same ordinate value (probability density function).

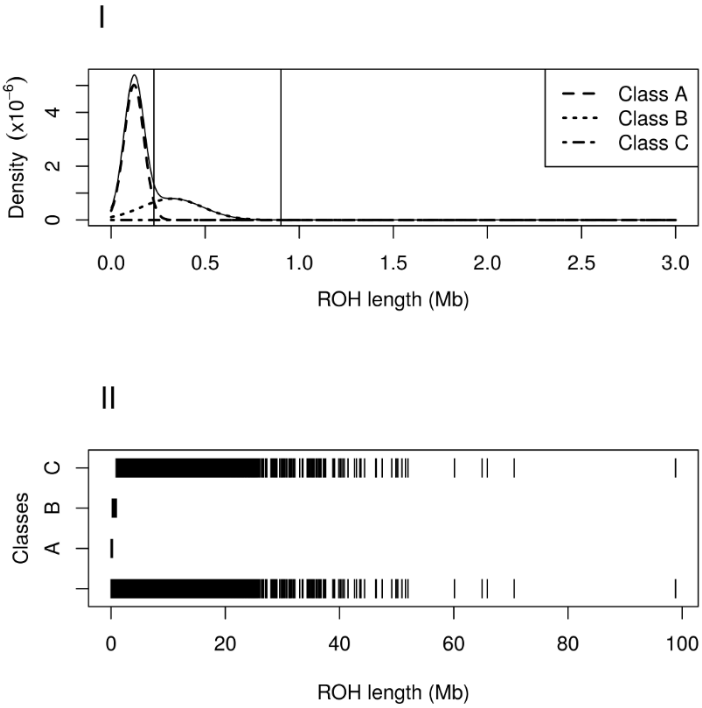


Figure 3: Classification of ROHs classes in QUI. I, Distribution of ROHs lengths according to the classes A, B, and C; II, ROHs classes distributed according to their lengths.

RESULTS

Population Inbreeding Coefficient f

The average estimates of **f** using the information of both complete and no-LD SNP datasets were -0.00397 and -0.00108, respectively. These negative values were not expected in a population with a structure like the quilombo, since they imply an overall excess of heterozygosity. Because previous results (Bhatia, *et al.*, 2013) show that MAF constrains the values and influences the variances of inbreeding and $\mathbf{f_{ST}}$ metrics, we therefore re-examined the behavior of $\bar{\mathbf{f}}$ according to the minor allele frequency (MAF).

We performed the analysis of complete and no-LD datasets using two approaches: (1) obtaining the $\bar{\mathbf{f}}$ estimates for subsets of markers above different MAF thresholds; and (2) observing the behavior of per locus estimates of $\mathbf{f_k}$.

Average $\bar{\mathbf{f}}$-values were estimated for subsets of markers according to thresholds of MAF ≥ {0, 0.01, ..., 0.49}, shown in **Figure 4**. The mere inspection of the graph enables the identification of a predictable pattern on the behavior of $\bar{\mathbf{f}}$-values, with a large distortion (shift to negative values) for markers with MAF ≤ 0.1 and a tendency to reach a constant plateau for higher MAF values.

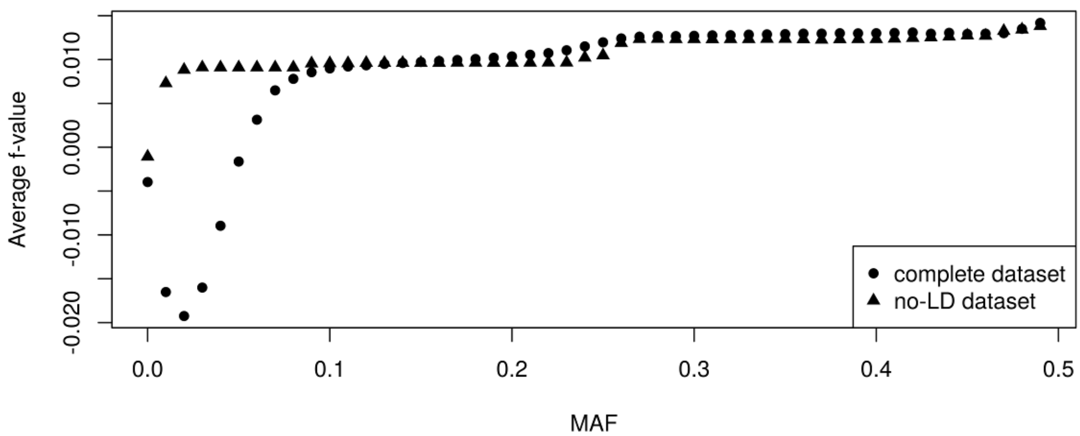


Figure 4: Average **f**-values corresponding to subsets of markers with MAF value equal or above the value shown in the abscissa axis.

Considering now the behavior of $\mathbf{f_k}$ estimates across all loci (Figure 5), we notice that two regions (MAF < 0.1 and 0.1 ≤ MAF ≤ 0.3) of both graphs should be highlighted.

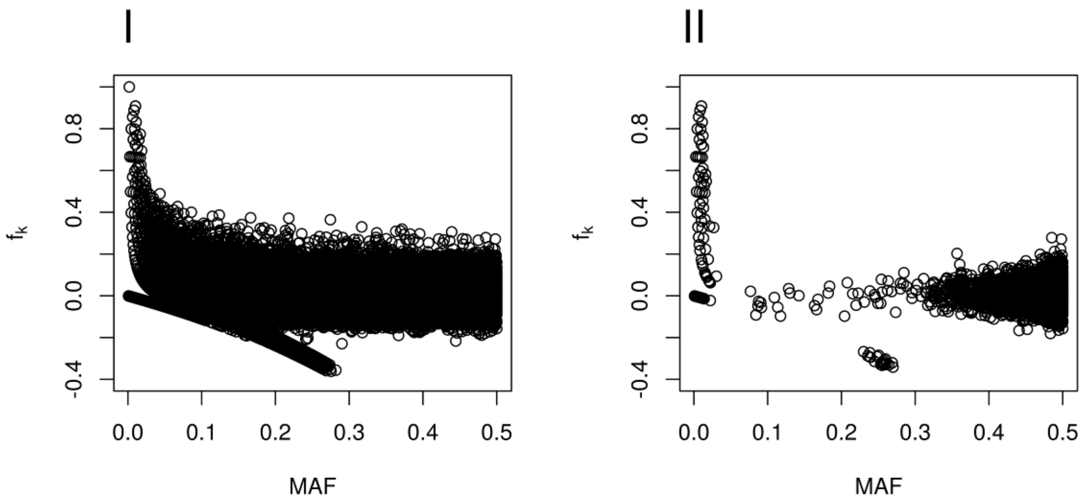


Figure 5: Estimates of per locus inbreeding coefficient values. I, complete dataset; II, no-LD dataset.

In spite of a huge amount of individual $\mathbf{f_k}$ estimates obtained from markers with MAF < 0.1 holding positive values, they are associated with larger $\mathbf{var(f_k)}$ in both complete and no-LD datasets. On the other hand, almost half of $\mathbf{f_k}$ estimates have near zero and negative values associated to much smaller values of $\mathbf{var(f_k)}$.

Since the average value of $\mathbf{f_k}$ is calculated after $\bar{\mathbf{f}} = \sum \mathbf{x_k . f_k}$, where $\mathbf{x_k} = \dfrac{\mathbf{var^{-1}(f_k)}}{\sum_{j=1}^{n} \mathbf{var^{-1}(f_j)}}$ (formulae 2 and 3), negative values of $\mathbf{f_k}$ with very small variance values strongly influence the $\bar{\mathbf{f}}$ estimate, when loci corresponding to MAF < 0.1 are considered.

Figures 6 and 7 show, respectively, the values of $\mathbf{var(f_k)}$ as a function of $\mathbf{f_k}$ and the distributions of $\mathbf{var(f_k)}$ estimated for the MAF intervals 0-0.1, ..., 0.4-0.5, making it clear that the smallest values of MAF are associated with highly heterogeneous $\mathbf{var(f_k)}$ values, many of them being very small and responsible for creating biased average $\bar{\mathbf{f}}$-values.

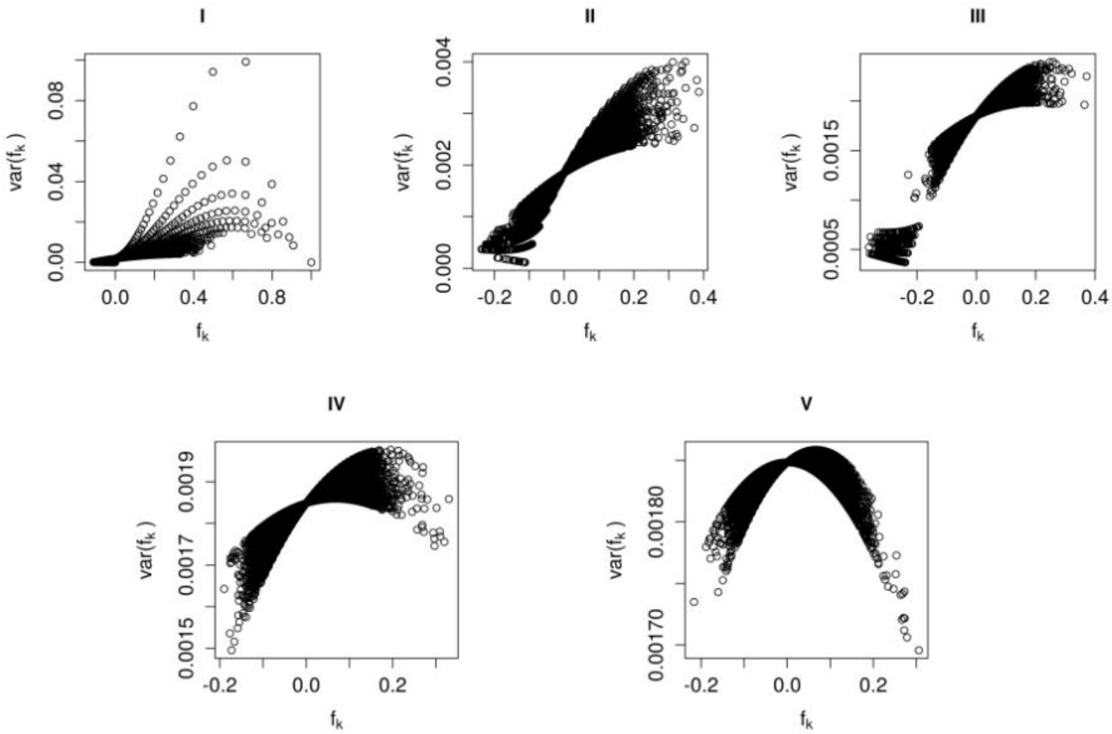


Figure 6: Scatter plot of per locus **var(f_k)** estimates and their corresponding f_k values according to MAF intervals for the complete dataset. I, 0-0.1; II, 0.1-0.2; III, 0.2-0.3; IV, 0.3-0.4; V, 0.4-0.5.

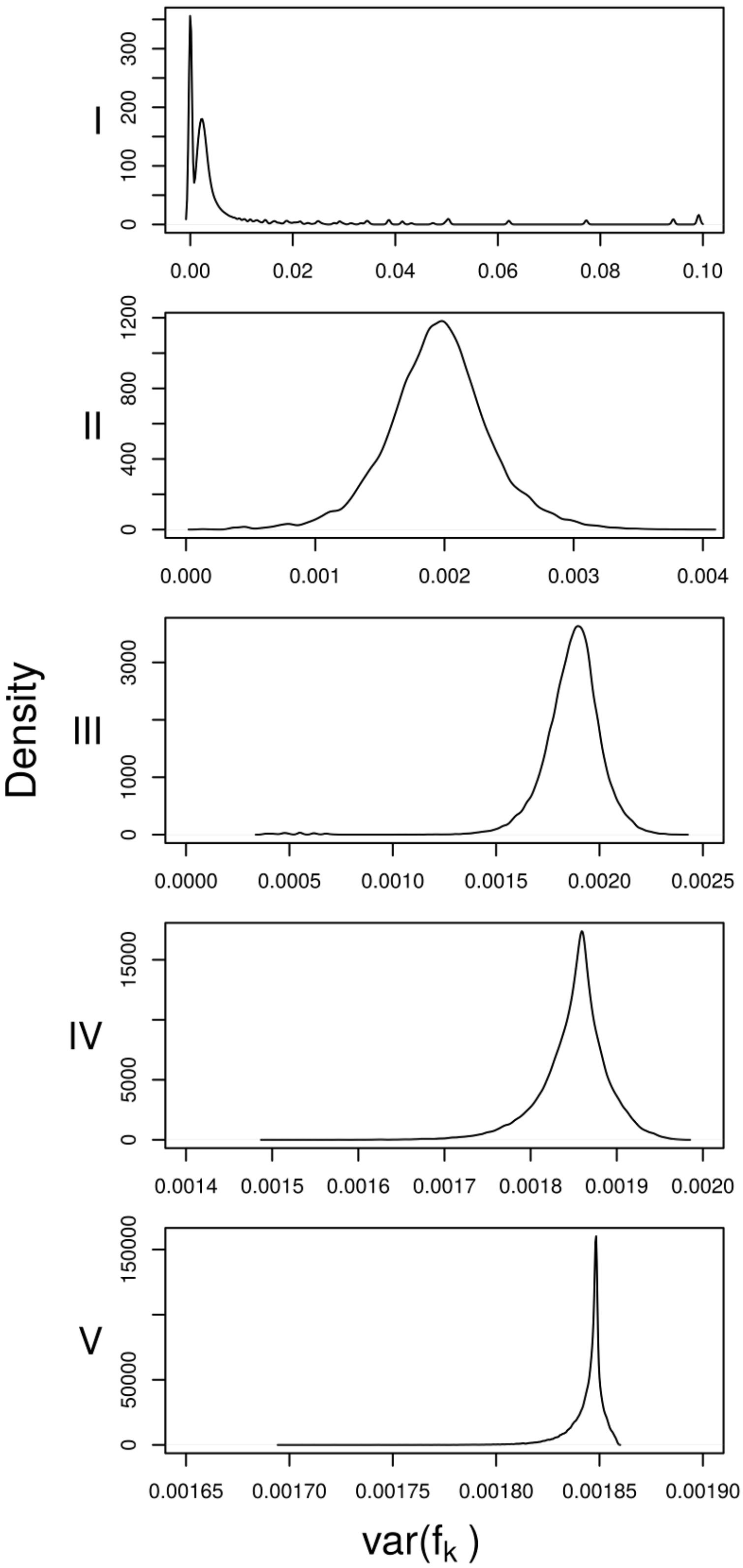


Figure 7: Distribution of per locus **var(f_k)** estimates according to MAF intervals for the complete dataset. I, 0-0.1; II, 0.1-0.2; III, 0.2-0.3; IV, 0.3-0.4; V, 0.4-0.5.

From the regions of the graphs (Figure 5) with $0.1 \leq MAF \leq 0.3$, it is easy to notice, mainly in the complete dataset, the existence of a subset of markers with very low $\mathbf{f_k}$-values ($\mathbf{f_k} < -0.17$), that clearly deviates from the distributions of most estimates. This behavior explains, for example, some extremely anomalous sets (559 and 22 loci for complete and no-LD datasets, respectively) of observed {AA, Aa, aa} genotype absolute frequencies of the order of {~250, ~250, ≤5} respectively, that are very unlikely to occur.

The presence of these anomalous genotype frequencies might be explained simply by the occurrence of systematic errors in the process of machine genotyping that resisted the data cleaning procedure. If we consider, for example, the genotyping error rate $\delta = P(AA \rightarrow Aa)$ and $p = P(A)$, $q = 1-p = P(a)$, $d = P(AA)$, $h = P(Aa)$, and $r = P(aa)$, we obtain $d' = d(1-\delta)$, $h' = d\delta + h$, and $r' = r$, so that estimated allele frequencies and corresponding inbreeding coefficient become $p' = d' + h'/2$, $q' = h'/2 + r'$, and $f' = 1 - h'/(p'q')$. It is then clear that the genotyping error is directly correlated with an increase in the estimation of heterozygous frequency. The numerical analysis of the simple expressions above shows also that the higher the value of p the lower the value of $\mathbf{f'}$. For example, for a $\mathbf{f}$-value of 0.01, if δ is set to 0.05, the estimates of f' will have negative values for all loci with $p > 0.2$; and all typed loci with $p \geq 0.5$ will produce estimates of $\mathbf{f'} \leq -0.15$.

Considering an alternative model, in which the typing error rate is associated with the identification of an allele instead of a genotype, f' is always smaller than $\mathbf{f}$, but unlike to the first model, large deviations from Hardy-Weinberg proportions take place only when the typing error is very large.

The machine average genotyping error is declared as of the order of 1/1000 by their manufacturers, and at this level only loci at the edge of fixation would led to significant negative f-values. However, the occurrence of genotyping errors is an important factor to be taken into account, because negative $\mathbf{f_k}$-values so generated are always associated with very small $\mathbf{var(f_k)}$ values that create a significant bias in the estimation of the population average value of the inbreeding coefficient.

Taking into account the facts above and the results shown in Figures 4 and 5, in order to avoid the use of markers associated with obvious biases in the estimation of the average inbreeding coefficient $\bar{\mathbf{f}}$, we considered in our final analysis, presented in the paragraph below, only loci with MAF ≥ 0.3.

In spite of having their original datasets dramatically reduced in size (the complete one from 485,957 to 147,200 SNPs and the no-LD one from 11,642 to 9,208 SNPs), the $\mathbf{f_k}$-values virtually retained their original properties of being symmetrically and normally distributed around their mean and median estimates. Taking into account that both sets were cleaned from most of their biases and errors, the parameters extracted from them (shown in Table 1 below) surely constitute now much more reliable estimates.

Table 1: Average f-values, medians, corresponding variances and 95% confidence intervals obtained for the two cleaned datasets. The (approximate) theoretical 95% confidence intervals were constructed under Gaussian assumptions and the (empirical) observed ones, as well as their medians, were obtained by ranking all individual f_k-values.

Dataset	$\bar{\mathbf{f}}$	**var(f)**	**theoretical 95% c.i.**	$\mathbf{f_{median}}$	**observed 95% c.i.**
Complete	0.0127	0.00248	(-0.0848, 0.1102)	0.0126	(-0.0816, 0.1121)
no-LD	0.0123	0.00249	(-0.0855, 0.1101)	0.0123	(-0.0832, 0.1114)

Individual Inbreeding Coefficient f'

The average population $\bar{f}'$ value (obtained averaging the estimates of f'_i obtained from QUI sample by means of the software PLINK v1.07) was 0.0075; the median, obtained from the whole f'_i distribution, was 0.0028, with corresponding 95% confidence interval limits of -0.2219 and 0.2098 (Figure 8). Interestingly, these estimates are not very different from those obtained using the traditional methods mentioned above.

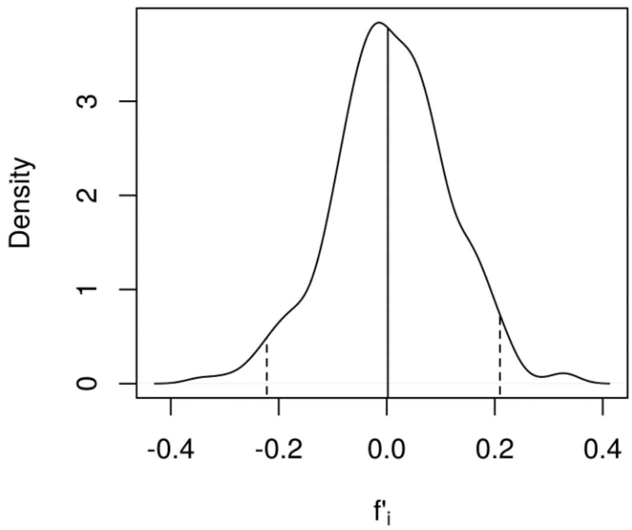


Figure 8: Distribution of f'_i values. The median and the limits of its 95% confidence interval correspond respectively to the intersections of the vertical black and dotted lines with the abscissa axis of the graph.

Identification of ROHs

The LOD of autozigosity was estimated from a window with size n = 18 SNPs, sliding SNP-wise across the genome of all individuals. As already pointed out, the anti-mode of each GKDE distribution (**Figure 2**) is considered as the population specific LOD-score threshold above which a window is assumed to be autozygous; all overlapping windows with LOD figures above the threshold were grouped to form the ROHs.

It can be observed from **Figure 2** that the areas of the right-hand portion of LOD distributions are proportionally larger as the distances from Africa increase, suggesting that autozygous regions are more frequent in these populations, as already noticed by Pemberton *et al.* (2012). For the admixed quilombo population we observed a pattern of distribution similar to that from YRI, in spite of the tri-hybrid composition of QUI.

The ROHs obtained from autozygous stretches were classified using a Gaussian mixture model of three components according to their lengths (A, short; B, intermediate; and C, long). The boundaries between classes A-B and B-C are population specific (**Table 2**) and related to both LD patterns and the amounts of inbreeding.

Table 2: Population specific boundaries in base pairs between ROHs classes A and B and classes B and C.

Sample	Boundary A-B	Boundary B-C
QUI	227,789.5	902,739.5
FRE	262,947.5	893,750.5
CHB	223,641.5	661,562.0
YRI	184,325.0	581,563.0

Inbreeding and Demographic Inferences from ROHs

The inbreeding coefficients $\mathbf{F_{ROH}}$ of all individuals of the four populations were assessed considering ROHs of all classes together as well as those belonging to classes A, B and C separately (Table 3). The mean $\mathbf{F_{ROH}}$ estimates from the QUI population were smaller than those from the FRE and CHB samples and higher than that from YRI, taking into account all ROHs together or separated by class.

Table 3: Mean, median and corresponding observed 95% confidence intervals of individual inbreeding coefficients F_{ROH} per population, considering all ROHs together and separately.

Class		QUI	FRE	CHB	YRI
A+B+C	mean	0.2678	0.5639	0.6153	0.2193
	median	0.2624	0.5672	0.6175	0.2277
	95% c.i.	0.2270-0.3384	0.4749-0.5759	0.5534-0.6227	0.1521-0.2267
A	mean	0.1031	0.1573	0.1157	0.0735
	median	0.1028	0.1579	0.1156	0.0751
	95% c.i.	0.0928-0.1161	0.1433-0.1632	0.1107-0.1198	0.0444-0.0769
B	mean	0.1280	0.2980	0.2860	0.1118
	median	0.1237	0.3007	0.2874	0.1142
	95% c.i.	0.1052-0.1741	0.2559-0.3070	0.2685-0.2920	0.0656-0.1171
C	mean	0.0367	0.1086	0.2137	0.0339
	median	0.0285	0.1080	0.2140	0.0332
	95% c.i.	0.0120-0.1177	0.0757-0.1404	0.1699-0.2276	0.0272-0.0421

The GKDE distributions of F_{ROH} estimates for the four populations are shown in **Figure 9.** As expected for ROHs of classes A and B and for all ROHs together, QUI $\mathbf{F_{ROH}}$-values are intermediate when compared to the African and European ones, because the estimates for the admixed quilombo population reflect aspects of the demographic histories of the populations which contributed to it. On the other hand, the quilombo class C $\mathbf{F_{ROH}}$-values were very low, a surprising finding given that the population remained isolated for several generations and was founded by a small group of individuals.

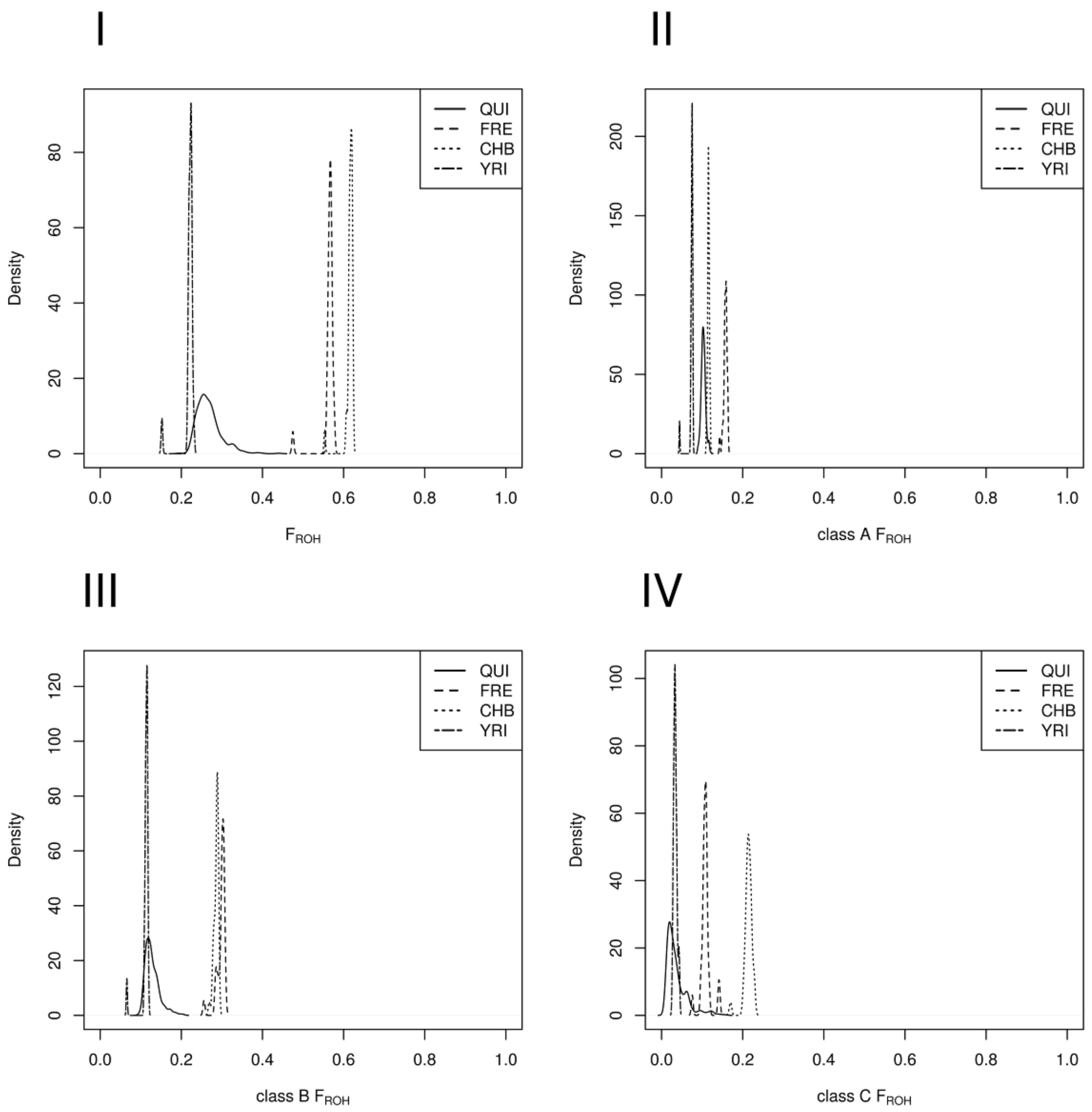


Figure 9: Distribution of F_{ROH} in the four populations. I, F_{ROH} of classes A, B and C; II, class A F_{ROH}; III, class B F_{ROH}; IV, class C F_{ROH}.

To better understand the behavior of $\mathbf{F_{ROH}}$ values in the quilombo, we analyzed the distribution of mean values of total lengths of ROHs per individual by class (A, B, and C) and subclasses of class C ROHs according to arbitrary length intervals (Figures 10 and 11).

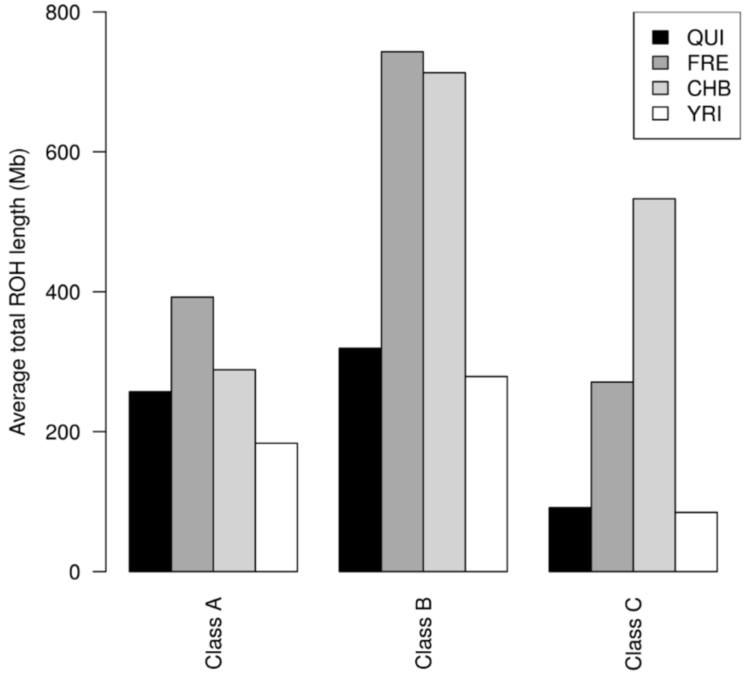


Figure 10: Distribution of individual average total ROHs lengths per class per population.

Considering ROHs of A and B classes, African YRI individuals showed the lowest genomic ROHs proportions, with a total of less than 500Mb of the genome composed by short or intermediate ROHs (Figure 10). Conversely, European FRE and Asian CHB showed an average total length of short and intermediate ROHs of approximately 1Gb. For the QUI population, we obtained an intermediate value of approximately 600Mb, which is expected since the isolate was founded by individuals of three different ancestries and the amount of genomic ROHs should be approximately proportional to the genomic contribution of the parental populations. This result suggests that LD patterns of admixed populations are strongly influenced by the LD patterns of the populations from which founder individuals originated.

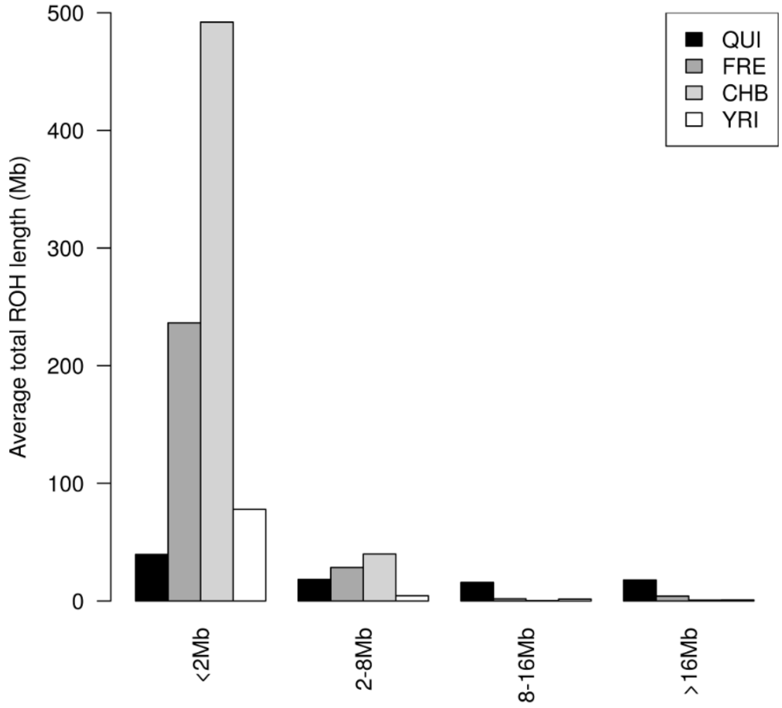


Figure 11: Distribution of individual average total ROHs lengths, considering subclasses of class C ROHs.

Considering now the subclasses of class C ROHs (Figure 11), we observed in the HGDP samples high amounts of ROHs <2MB followed by a drastic reduction in the larger subclasses, which suggests low levels of very recent inbreeding in the three cosmopolitan populations. In the QUI sample, on the other hand, subclasses with larger sizes were far more common, highlighting the occurrence of close inbreeding for at least part of the population and, less probably, the contribution of Native American ancestry components, that are likely to also harbor comparatively large portions of class C ROHs.

Single ROHs larger than 50Mb were found in eight (out of the total of 541) individuals, including a segment of almost 100Mb. Checking the genealogical data available in our laboratory, we found out that three of them are the offspring of first cousins while another one is the son of double first cousins. As for the other four individuals, the paucity of reliable historical records prevented us from establishing the degree of biological relationship between their parents, who however share same surnames and might be closely related.

Relationship Between f′ and F_{ROH} Estimates

The quilombo values of **f′** and $\mathbf{F_{ROH}}$ were estimated using two different techniques that should be correlated, since they are somehow associated to the inbreeding levels of the population. The scatter plots of Figure 12 show the dispersion of individual values of corresponding pairs of **f′** and $\mathbf{F_{ROH}}$ considering the set of all ROHs (Pearson's r = 0.496, Spearman's ρ = 0.460) and the subset of class C ROHs (Pearson's r = 0.542, Spearman's ρ = 0.550); all four correlation coefficients differ significantly from zero at a rejection level of $P < 2.2\times10^{-16}$.

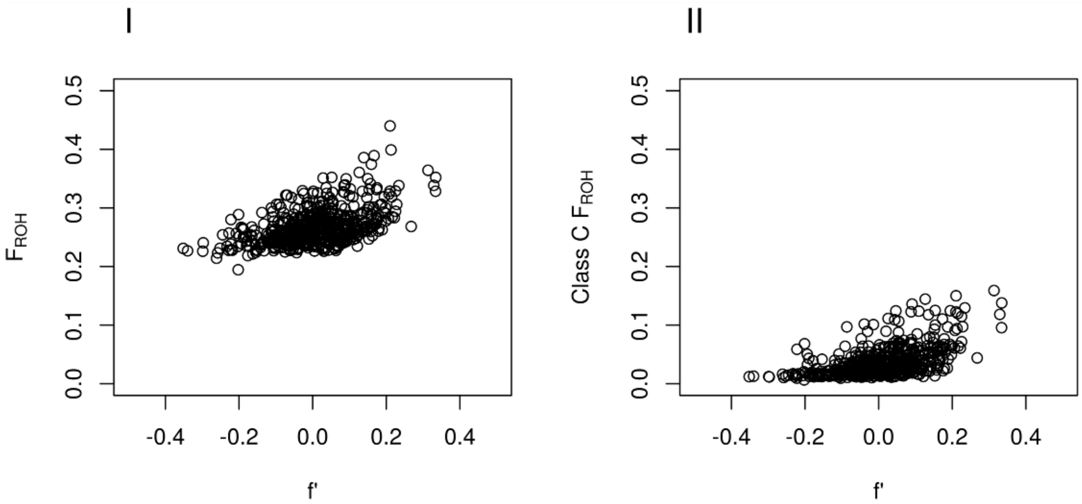


Figure 12: Scatter plots of individual estimates of inbreeding coefficient **f′** and $\mathbf{F_{ROH}}$ of all ROHs classes together (left) and of class C ROHs (right).

As expected, the correlation coefficients estimated for the set of class C ROHs are a bit higher than the ones obtained for the whole set of ROHs, since class C ROHs are more influenced by the occurrence of events of recent inbreeding than classes A and B.

DISCUSSION AND CONCLUSIONS

This study dealt with the issue of estimating parameters related to the system of marriages, endogamy levels, and population/demographic events of a complex tri-hybrid admixed population.

Using information from both complete and no-LD datasets, we presented novel (as far as we know) procedure to cope efficiently with biases associated to the estimation of average value of Wright's fixation index **f**. Our analyses showed (1) that systematic machine genotyping errors might be pivotal in originating spurious negative values of **f**; and (2) that the optimal range of MAF for using in the estimation process in the QUI sample is in the range of $0.3 \le MAF \le 0.5$. We suggest that this should also be investigated using available large SNP dataset for other populations. It is possible that this range might vary significantly among populations, since it is reasonable to admit that it can be dependent both on sample sizes and number of available dataset SNP markers.

The $\bar{\mathbf{f}}$ estimates obtained from the complete and no-LD SNP datasets agree with the values of estimates obtained with other procedures that we describe in the paper. The $\bar{\mathbf{f}}$ estimates obtained here are not significantly different from zero, a fact that can be explained by consanguineous marriages taking place mainly as a consequence of the relatively small population size of the quilombo isolate.

In relation to the ROHs study, we used a reliable method that allowed us (1) to identify autozygous segments of different lengths resulting from evolutionary forces acting in multiple time scales and (2) to separate them in three categories according to their sizes

(Pemberton *et al.*, 2012; Rosenberg *et al.*, 2013). The quilombo population has an intermediate average total length of ROHs of A and B classes, suggesting that the amount of shorter ROHs should be somehow proportional to the amount of corresponding ROHs inherited from its parental populations. Due to a complex admixture of individuals from different genomic sources, a factor that introduces genetic variability into the population, its average proportion of shorter ROHs in admixed populations should be lower than the fractions contributed directly from each parental stock. A similar behavior was observed in the quilombo genome proportion made up of ROHs ($\mathbf{F_{ROH}}$): its average $\mathbf{F_{ROH}}$-value is lower when compared to European and Asian populations and a bit higher than the African one.

The class C ROHs results suggest that the smallest sizes are influenced by both background relatedness and cryptic inbreeding, that is by multiple distant parental relationship, whereas longer ROHs reflects the presence of very recent inbreeding levels. As expected, the quilombo isolate showed the greatest average total length of very long ROHs, reflecting its condition of recent endogamy due mainly to its low effective population size.

Consistent with previous results from the literature, we detected significant positive correlation coefficients between the individual estimates of $\mathbf{F_{ROH}}$ and $\mathbf{f'}$.

ACKNOWLEDGMENTS

We thank Doctors Gabriel Marroig for the use of his computer server, Tabita Hünemeier for critically reading a previous version of this article, and Trevor Pemberton for his suggestions as to the analysis of ROHs. This paper was funded by grants from CAPES, CNPq and FAPESP, all from Brazil.

REFERENCES

Abdellaoui A; Hottenga JJ; Willemsen G; Bartels M; van Beijsterveldt T; Ehli EA; Davies GE; Brooks A; Sullivan PF; Penninx BWJH; Geus EJ; Boomsma DI. Educational attainment influences levels of homozygosity through migration and assortative mating. **PLoS One. 10: e0118935, 2015.**

Arcos-Burgos M; Muenke M. Genetics of population isolates. **Clinical Genetics. 61: 233-247, 2002.**

Auricchio MTBM; Vicente JP; Meyer D; Mingroni-Netto RC. Frequency and origins of hemoglobin S mutation in African-derived Brazilian populations. **Human Biology. 79: 667-677, 2007.**

Bhatia G; Patterson N; Sankararaman S; Price AL. Estimating and interpreting FST: the impact of rare variants. **Genome Research. 23: 1514-1521, 2013.**

Ben Halim N; Nagara M; Regnault B; Hsouna S; Lasram K; Kefi R; Azaiez H; Khemira L; Saidane R; Ammar SB; Besbes G; Weil D; Petit C; Abdelhak S; Romdhane L. Estimation of Recent and Ancient Inbreeding in a Small Endogamous Tunisian Community Through Genomic Runs of Homozygosity. **Annals of Human Genetics. 79: 402-417, 2015.**

Broman KW; Weber JL. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. **American Journal of Human Genetics. 65: 1493-1500, 1999.**

Carothers AD; Rudan I; Kolcic I; Polasek O; Hayward C; Wright AF; Campbell H; Teague P; Hastie ND; Weber JL. Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. **Annals of Human Genetics. 70: 666-676, 2006.**

Curie-Cohen M. Estimates of inbreeding in a natural population: a comparison of sampling properties. **Genetics. 100: 339-358, 1982.**

Fraley C; Raftery AE; Scrucca L. Normal mixture modeling for model-based clustering, classification, and density estimation. **Department of Statistics, University of Washington, 23, 2012.**

Fyfe JL; Bailey NTJ. Plant breeding studies in leguminous forage crops I. Natural cross-breeding in winter beans. **The Journal of Agricultural Science. 41: 371, 1951.**

Gogarten SM; Bhangale T; Conomos MP; Laurie CA; McHugh CP; Painter I; Zheng X; Crosslin DR; Levine D; Lumley T; Nelson SC; Rice K; Shen J; Swarnkar R; Weir BS; Laurie CC. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. **Bioinformatics, 28: 3329-3331, 2012.**

Jalkh N; Sahbatou M; Chouery E; Megarbane A; Leutenegger AL; Serre JL. Genome-wide inbreeding estimation within Lebanese communities using SNP arrays. **European Journal of Human Genetics. 23: 1364-1369, 2015.**

Karafet TM; Bulayeva KB; Bulayev OA; Gurgenova F; Omarova J; Yepiskoposyan L; Savina OV; Veeramah KR; Hammer MF. Extensive genome-wide autozygosity in the population isolates of Daghestan. **European Journal of Human Genetics. 23: 1405-1412, 2015.**

Kimura L; Ribeiro-Rodrigues EM; De Mello Auricchio MT; Vicente JP; Batista Santos SE; Mingroni-Netto RC. Genomic ancestry of rural African-derived populations from Southeastern Brazil. **American Journal of Human Biology. 25: 35-41, 2013.**

Kirin M; McQuillan R; Franklin CS; Campbell H; McKeigue PM; Wilson JF. Genomic runs of homozygosity record population history and consanguinity. **PLoS One. 5: e13996, 2010.**

Lemes RB; Nunes K; Meyer D; Mingroni-Netto RC; Otto PA. Estimation of inbreeding and substructure levels in african-derived brazilian quilombo populations. **Human Biology. 86: 276-288, 2014.**

Leutenegger AL; Sahbatou M; Gazal S; Cann H; Genin E. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? **European Journal of Human Genetics. 19: 583-587, 2011.**

McQuillan R; Leutenegger AL; Abdel-Rahman R; Franklin CS; Pericic M; Barac-Lauc L; Smolej-Narancic N; Janicijevic B; Polasek O; Tenesa A; MacLeod AK; Farrington SM; Rudan P; Hayward C; Vitart V; Rudan I; Wild SH; Dunlop MG; Wright AF; Campbell H; Wilson JF. Runs of homozygosity in European populations. **American Journal of Human Genetics. 83: 359-372, 2008.**

Nunes K; Zheng X; Torres M; Moraes ME; Piovezan BZ; Pontes GN; Kimura L; Carnavalli JEP; Mingroni-Netto RC; Meyer D. HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set. **Human immunology. 77: 307-312, 2016.**

Otto PA; Lemes RB. A note on the variance of the estimate of the fixation index F. **Journal of Genetics. 94: 759-763, 2015.**

Pemberton TJ; Absher D; Feldman MW; Myers RM; Rosenberg NA; Li JZ. Genomic patterns of homozygosity in worldwide human populations. **American Journal of Human Genetics. 91: 275-292, 2012.**

Pemberton TJ; Rosenberg NA. Population-genetic influences on genomic estimates of the inbreeding coefficient: a global perspective. **Human Heredity. 77: 37-48, 2014.**

Purcell S; Neale B; Todd-Brown K; Thomas L; Ferreira MA; Bender D; Maller J; Sklar P; de Bakker PI; Daly MJ; Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics. 81: 559-575, 2007.**

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. **URL https://www.R-project.org/. 2016.**

Rosenberg NA; Pemberton TJ; Li JZ; Belmont JW. Runs of homozygosity and parental relatedness. **Genetics in Medicine. 15: 753-754, 2013.**

Templeton AR. **Population genetics and microevolutionary theory.** John Wiley & Sons, Inc, Hoboken, New Jersey, 2006.

Teo SM; Ku CS; Salim A; Naidoo N; Chia KS; Pawitan Y. Regions of homozygosity in three Southeast Asian populations. **Journal of Human Genetics. 57: 101-108, 2012.**

Wang S; Haynes C; Barany F; Ott J. Genome-wide autozygosity mapping in human populations. **Genetic Epidemiology. 33: 172-180, 2009.**

Wigginton JE; Cutler DJ; Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. **American Journal of Human Genetics. 76: 887-893, 2005.**

SUPPLEMENTARY TABLE

Table S1: Numbers of Genotyped Individuals at a Given Community

	AB	*AN*	*GA*	*IV*	*MR*	*NH*	*PA*	*PC*	*PS*	*RE*	*SP*	*TU*	Total
N	573	320	134	270	56	447	220	286	128	250	132	295	3111
N_G	95	75	37	44	10	39	26	55	34	28	43	55	541
n_G	16.6	23.4	27.6	16.3	17.9	8.7	11.8	19.2	26.6	11.2	32.6	18.6	17.4

Communities are as defined in Figure 1; N, estimated number of adult individuals (Auricchio *et al.*, 2007); N_G, number of genotyped individuals; n_G, percentage of genotyped individuals.

4. GENERAL DISCUSSION AND CONCLUSIONS

This dissertation dealt with issues related to the estimation of average population inbreeding levels and includes two manuscripts already published in specialized international journals and another one yet to be submitted.

Chapter 1 shows how the inbreeding coefficient is estimated by using genealogical and marker (molecular) information. The genealogical (direct) estimation of inbreeding coefficient **F** is complicated due to the usual lack of complete pedigree information and to the arbitrary choice of the number of generations to take into account in its estimation. In spite of these limitations, **F**-values so estimated are used to make valid comparisons of autozygous levels among populations.

Quilombo **F**-values were obtained using all available pedigree information and averaging the individual inbreeding coefficients from all individuals. The values thus obtained were compared with others estimates from the literature (Table 4, Chapter 1). Quilombo **F**-values (and the frequencies of consanguineous marriages) showed to be significantly lower than the values obtained for most isolates from the literature, except in relation to a Brazilian Jewish isolate (Freire-Maia and Krieger, 1963). In any case, the value we estimated is about three times higher than the corresponding one from the Brazilian population (**F** = 0.00088; Freire-Maia, 1990).

As to the quilombo **f** (molecular) estimates of Chapter 1, we used a highly heterogeneous set of 30 molecular markers (14 biallelic SNPs and 16 multiallelic microsatellites). Seven SNPs markers were obtained from a sample of 700 individuals in an association study of hypertension (Kimura *et al.*, 2012) and another seven SNPs from 400 sampled individuals in an obesity association study (Angeli *et al.*, 2011); the remaining 16 microsatellites were genotyped from a sample of 300 individuals especially selected for the study described in Chapter 1.

We analyzed SNPs and microsatellites data separately and together (Tables 5 and 6, Chapter 1), obtaining average population **f**-values by weighing **f** estimates from each community by the reciprocal of the corresponding variances. Given that the sample sizes required to obtain **f**-values significantly different from zero are extremely high (Figure 2, Chapter 1) in tests that verify departures from HW proportions, no **f** estimate obtained from SNP markers was significantly different from zero. In two instances of microsatellite markers we found **f**-values significantly lower than zero, a result that might result from the combination of small sample sizes and multiallelic nature of these markers.

Historical records collected by members of Dr. Regina Mingroni's laboratory account for the presence of intense migration among all subpopulations analyzed, indicating an absence of genetic isolation. Using our molecular markers data, we estimated Wright's fixation indexes. The estimates of $\mathbf{f}_{\mathbf{ST}}$ values obtained were in general lower than 5%, which is according to results previously obtained from the analysis of INDEL markers data for the same subpopulations (Kimura *et al.*, 2013). These results indicate, as expected from the historical records mentioned, the absence of significant population substructure levels in the whole quilombo aggregate.

The second article presented in Chapter 2 dealt with the estimation of **var(f)**. The very simple approximation we provided could be applied to a locus with any number of alleles, producing estimates very similar to those obtained using simulations or approximations already known in the literature for two allele case (Fyfe and Bailey, 1951; Curie-Cohen, 1982). Given that the formal estimation of **var(f)** is (mathematically) a very complicated issue, our work resulted in a very simple and efficient method to obtain reliable f-variance estimates.

The third chapter is represented by an umpublished manuscript dealing with the estimation of the coefficient **f** (in the same quilombo population) using high density SNP array data and presenting a new manner to estimate the index, by using the joint information from two sets of markers (complete and no-LD datasets).

It is known from population genetics theory that the unbiased estimation of the average inbreeding coefficient $\bar{\mathbf{f}}$ should consider only completely independent loci, that is, loci with no linkage disequilibrium. The main problem in excluding linked data is the drastic reduction of dataset information.

With the aim of seeking for markers with more reliable information, we considered in our analysis both datasets (complete and no-LD), observing that: (1) markers with MAF < 0.3 introduced a bias underestimating the average **f**-values, since they might include data with errors in genotype determination that resisted to the filtering process; (2) no statistically significant difference between the **f** average estimates from both datasets was found, since their 95% confidence intervals overlapped.

We made also some inferences from the quilombo demographic history, as we were dealing with a highly admixed tri-hybrid population with a complex foundation history. Both the total ROHs lengths and the

F_{ROH} values were lower in the quilombo than in the European and Asian population datasets and a bit higher than in the African one selected for comparison. The results we obtained suggest that the patterns of ROH and **F_{ROH}** of an admixed population such as the quilombo reported here should be somehow proportional to the contribution of the parental (stock) populations, but lower, given that the admixture process inserted some degree of variability in the gene pool of the hybrid population.

5. ABSTRACT

Endogamy levels are usually estimated using genealogical or molecular markers data. By means of both type of data from a traditional Brazilian tri-hybrid quilombo population aggregate (located at the Ribeira River Valley in the State of São Paulo), the aim of this work, using different methods, was to obtain reliable estimates of its average inbreeding coefficient, as well as to establish pertinent demographic inferences.

The results we obtained are presented in three chapters.

The first one, represented by the offprint of a published paper, deals with the estimation of the inbreeding coefficient using both a complete genealogical and comprehensive molecular information. **F**-values were estimated for each community using all available *pedigree* information and averaging the inbreeding coefficients from all individuals represented in the genealogies. Molecular **f**-values were estimated from the analysis of 30 highly heterogenous sets of molecular markers (14 biallelic SNPs and 16 multiallelic microsatellites), genotyped in different groups of individuals from the population.

The second chapter (a research paper already published), presents a simplified method to estimate the variance of the inbreeding coefficient. The simple approximations we provided can be applied to a locus with any number of alleles, producing estimates fully validated by computer simulations.

The last chapter is a manuscript yet to be published that deals with inbreeding and demographic inferences, obtained from the information of hundreds of thousands of biallelic SNP markers. A new

manner to obtain estimates of Wright’s fixation index **f** is presented, consisting in the use of the joint information of two sets of markers (one complete and another excluding markers in patent linkage disequilibrium). Quilombo demographic inferences were obtained by means of ROHs analyses, which were adapted to cope with a highly admixed population with a complex foundation history.

6. RESUMO

Os níveis de endogamia de uma população são comumente estimados por meio do coeficiente de endocruzamento, que pode ser obtido de dados genealógicos (**F**) ou dados provenientes da análise de marcadores moleculares (**f**).

O objetivo do trabalho foi obter estimativas confiáveis do coeficiente de endocruzamento populacional, bem como realizar inferências demográficas, usando dados de um agregado populacional quilombola miscigenado com ancestralidade complexa tri-híbrida, localizado no Vale do Rio Ribeira, na região sul do estado de São Paulo.

No trabalho é apresentado em três capítulos. No primeiro (um trabalho já publicado), estimamos o coeficiente de endocruzamento usando dados genealógicos e moleculares. As estimativas genealógicas de **F** foram obtidas para cada comunidade por meio da média dos coeficientes individuais de todos os indivíduos representados nas genealogias da população. Os valores de **f** foram estimados por meio dos dados de 30 marcadores moleculares altamente heterogêneos (14 SNPs e 16 microssatélites), genotipados em diferentes grupos de indivíduos com diferentes finalidades.

O segundo capítulo, representado por um trabalho também já publicado, apresenta um método simples para estimar a variância do coeficiente de endocruzamento **f**. As aproximações obtidas, validadas devidamente por simulações em computador, podem ser aplicadas a lóci multialélicos, produzindo estimativas que não diferem significativamente de outras aproximações complicadas descritas na literatura.

O último capítulo (um manuscrito a ser submetido para publicação) apresenta inferências a respeito dos processos de endogamia e demografia no isolado quilombola, utilizando a informação de centenas de milhares de marcadores moleculares bialélicos. É apresentada uma nova maneira de se estimar o índice de fixação **f** de Wright, usando a informação combinada de dois conjuntos de marcadores (o conjunto completo de marcadores e um outro contendo apenas marcadores não ligados significativamente entre si). Também foram feitas inferências sobre a história demográfica do isolado por meio do estudo das regiões genômicas em homozigose (ROHs), uma contribuição inédita e importante do trabalho, adaptada à análise de um isolado populacional altamente miscigenado com contribuição tri-híbrida e uma história de fundação complexa.

7. REFERENCES

Abdellaoui A; Hottenga JJ; Willemsen G; Bartels M; van Beijsterveldt T; Ehli EA; Davies GE; Brooks A; Sullivan PF; Penninx BWJH; Geus EJ; Boomsma DI. Educational attainment influences levels of homozygosity through migration and assortative mating. **PLoS One. 10: e0118935, 2015.**

Abney M; McPeek MS; Ober C. Estimation of variance components of quantitative traits in inbred populations. **American Journal of Human Genetics. 66:629–650, 2000.**

Abou Jamra R; Wohlfart S; Zweier M; Uebe S; Priebe L; Ekici A; Giesebrecht S; Abboud A; Al Khateeb MA; Fakher M; Hamdan S; Ismael A; Muhammad S; Nöthen MM; Schumacher J; Reis A. Homozygosity mapping in 64 Syrian consanguineous families with non-specific intellectual disability reveals 11 novel loci and high heterogeneity. **European Journal Human Genetics. 19: 1161-1166, 2011.**

Ahmad B; Rehman AU; Malik S. Consanguinity and Inbreeding Coefficient in Tribal Pashtuns Inhabiting the Turbulent and War-Affected Territory of Bajaur Agency, North-West Pakistan. **Journal of Biosocial Science. 48: 113-128, 2016.**

Alkuraya FS. The application of next-generation sequencing in the autozygosity mapping of human recessive diseases. **Human Genetics. 132: 1197-1211, 2013.**

Angeli CB; Capelli LP; Auricchio MTBM; Vianna-Morgante AM; Mingroni-Netto RC; Leal-Mesquita ER; Ribeiro-dos-Santos AKC; Ferrari I; Oliveira SF; Klatau-Guimarães MN. AGG interspersion patterns in the CGG repeat of the FMR1 gene and linked DXS548/FRAXAC1 haplotypes in Brazilian populations. **American Journal of Medical Genetics. 132A: 210–214, 2005.**

Angeli CB; Kimura L; Auricchio MT; Vicente JP; Mattevi VS; Zembrzuski VM; Hutz MH; Pereira AC; Pereira TV; Mingroni-Netto RC. Multilocus analyses of seven candidate genes suggest interacting pathways for obesity-related traits in Brazilian populations. **Obesity. 19: 1244-1251, 2011.**

Arcos-Burgos M; Muenke M. Genetics of population isolates. **Clinical Genetics. 61: 233-247, 2002.**

Auricchio MTBM; Vicente JP; Meyer D; Mingroni-Netto RC. Frequency and origins of hemoglobin S mutation in African-derived Brazilian populations. **Human Biology. 79: 667-677, 2007.**

Bhatia G; Patterson N; Sankararaman S; Price AL. Estimating and interpreting FST: the impact of rare variants. **Genome Research. 23: 1514-1521, 2013.**

Ben Halim N; Nagara M; Regnault B; Hsouna S; Lasram K; Kefi R; Azaiez H; Khemira L; Saidane R; Ammar SB; Besbes G; Weil D; Petit C; Abdelhak S; Romdhane L. Estimation of Recent and Ancient Inbreeding in a Small Endogamous Tunisian Community Through Genomic Runs of Homozygosity. **Annals of Human Genetics. 79: 402-417, 2015.**

Bittles AH. Endogamy, consanguinity and community genetics. **Journal Genetics. 81: 91-98, 2002.**

Bittles AH; Black ML. Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex

diseases. **Proceedings of the National Academy of Sciences USA. 107: 1779-1786, 2010.**
Bittles AH; Black ML. Global patterns & tables of consanguinity. **URL http://consang.net, 2015.**
Broman KW; Weber JL. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. **American Journal of Human Genetics. 65: 1493-1500, 1999.**
Cannings C; Edwards AW. 1969. Expected genotypic frequencies in a small sample: Deviation from Hardy-Weinberg equilibrium. **American Journal of Human. Genetics. 21: 245–247, 1969.**
Carothers AD; Rudan I; Kolcic I; Polasek O; Hayward C; Wright AF; Campbell H; Teague P; Hastie ND; Weber JL. Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. **Annals of Human Genetics. 70: 666-676, 2006.**
Cavalli-Sforza LL;Bodmer WF. **The Genetics of Human Populations.** W. H. Freeman., San Francisco, 1971.
Chakraborty, R. Comments on 'A note on the variance of the estimate of the fixation index F'. **Journal of genetics, 95: 229-230, 2016.**
Christodoulou K; Tsingis M; Deymeer F; Serdaroglu P; Ozdemir C; Al-Shehab A; Bairactaris C; Mavromatis I; Mylonas I; Evoli A; Kyriallis K; Middleton LT. Mapping of the familial infantile myasthenia (congenital myasthenic syndrome type Ia) gene to chromosome 17p with evidence of genetic homogeneity. **Human Molecular Genetics. 6: 635-640, 1997.**
Cockerham CC. Variance of gene frequencies. **Evolution. 23: 72–84, 1969.**
Cockerham CC. Analyses of gene frequencies. **Genetics. 74: 679–700, 1973.**
Cotrim NH; Auricchio MT; Vicente JP; Otto PA; Mingroni-Netto RC. Polymorphic Alu insertion in six brazilian african-derived populations. **American Journal of Human Biology. 16: 264-277, 2004.**
Cotterman CW. **A Calculus for Statistico-genetics.** Unpublished thesis, Ohio State University, Columbus, Ohio. 190
Crow JF; Felsenstein J. The effect of assortative mating on the genetic composition of a population. **Eugenics Quarterly. 15: 85-97, 1968.**
Crow JF, Kimura M. **An introduction population genetics theory.** Alpha Editions, Madison, 1970.
Curie-Cohen M. Estimates of inbreeding in a natural population: a comparison of sampling properties. **Genetics. 100: 339-358, 1982.**
Dorsten LE; Hotchkiss L; King TM. The effect of inbreeding on early childhood mortality: twelve generations of an Amish settlement. **Demography. 36: 263-271, 1999.**
Ellis WS; Starmer WT. Inbreeding as measured by isonymy, pedigrees, and population size in Törbel, Switzerland. **American Journal of Human Genetics. 30: 366-376, 1978.**
Fraley C; Raftery AE; Scrucca L. Normal mixture modeling for model-based clustering, classification, and density estimation. **Department of Statistics, University of Washington, 23, 2012.**
Fisher RA. **The theory of inbreeding.** Oliver and Boya; London, 1949.
Freire-Maia N. Inbreeding in Brazil. **American Journal of Human Genetics. 9: 284–298, 1957.**

Freire-Maia N. Genetic efffects in Brazilian populations due to consanguineous marriages. **American Journal of Medical Genetics, 35:115–117, 1990.**

Freire-Maia N; Krieger H. A Jewish isolate in southern Brazil. **Annals of Human Genetics. 27:31–39, 1963.**

Fyfe JL; Bailey NTJ. Plant breeding studies in leguminous forage crops I. Natural cross-breeding in winter beans. **The Journal of Agricultural Science. 41: 371, 1951.**

Ghadami S; Mohammadi HM; Malbin J; Masoodifard M; Sarhaddi AB; Tavakkoly-Bazzaz J; Zeinali S. Frequencies of Six (Five Novel) STR Markers Linked to TUSC3 (MRT7) or NSUN2 (MRT5) Genes Used for Homozygosity Mapping of Recessive Intellectual Disability. **Clinical Laboratory. 61: 925-932, 2015.**

Gogarten SM; Bhangale T; Conomos MP; Laurie CA; McHugh CP; Painter I; Zheng X; Crosslin DR; Levine D; Lumley T; Nelson SC; Rice K; Shen J; Swarnkar R; Weir BS; Laurie CC. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. **Bioinformatics, 28: 3329-3331, 2012.**

Hamamy H; Jamhawi L; Al-Darawsheh J; Ajlouni K. Consanguineous marriages in Jordan: why is the rate changing with time? **Clinical Genetics. 67: 511–516, 2005.**

Haldane JBS; Moshinsky P. Inbreeding in mendelian populations with special reference to human cousin marriage. **Annals of Eugenics. 9: 321-340, 1939.**

Hartl DL; Clark AG. **Principles of Population Genetics.** Sinauer Associates, Inc, Sunderland, MA, 2007.

Hina S; Malik S. Pattern of Consanguinity and Inbreeding Coefficient in Sargodha District, Punjab, Pakistan. **Journal of Biosocial Science. 47: 803-811, 2015.**

Jackson CE, Symon WE, Pruden EL, Kaehr IM, Mann JD. Consanguinity and Blood Group Distribution in an Amish Isolate. **American Journal of Human Genetics. 20: 522-527, 1968.**

Jalkh N; Sahbatou M; Chouery E; Megarbane A; Leutenegger AL; Serre JL. Genome-wide inbreeding estimation within Lebanese communities using SNP arrays. **European Journal of Human Genetics. 23: 1364-1369, 2015.**

Karafet TM; Bulayeva KB; Bulayev OA; Gurgenova F; Omarova J; Yepiskoposyan L; Savina OV; Veeramah KR; Hammer MF. Extensive genome-wide autozygosity in the population isolates of Daghestan. **European Journal of Human Genetics. 23: 1405-1412, 2015.**

Kimura L; Angeli CB; Auricchio MT; Fernandes GR; Pereira AC; Vicente JP; Pereira TV; Mingroni-Netto RC. Multilocus family-based association analysis of seven candidate polymorphisms with essential hypertension in an african-derived semi-isolated brazilian population. **International Journal of Hypertension. 2012: 859219, 2012.**

Kimura L; Ribeiro-Rodrigues EM; De Mello Auricchio MT; Vicente JP; Batista Santos SE; Mingroni-Netto RC. Genomic ancestry of rural African-derived populations from Southeastern Brazil. **American Journal of Human Biology. 25: 35-41, 2013.**

Kirin M; McQuillan R; Franklin CS; Campbell H; McKeigue PM; Wilson JF. Genomic runs of homozygosity record population history and consanguinity. **PLoS One. 5: e13996, 2010.**

Lander E; Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. **Science. 236: 1567-1570, 1987.**

Lemes RB; Nunes K; Meyer D; Mingroni-Netto RC; Otto PA. Estimation of inbreeding and substructure levels in african-derived brazilian quilombo populations. **Human Biology. 86: 276-288, 2014.**

Lencz T; Lambert C; DeRosse P; Burdick KE; Morgan TV; Kane JM; Kucherlapati R; Malhotra AK. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. **Proceedings of the National Academy of Sciences USA. 104: 19942-19947, 2007.**

Leutenegger AL; Sahbatou M; Gazal S; Cann H; Genin E. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? **European Journal of Human Genetics. 19: 583-587, 2011.**

Lewontin R; Kirk D; Crow J. Selective mating, assortative mating, and inbreeding: definitions and implications. **Eugenenics Quarterly. 15: 141-143, 1968.**

Long JC. The allelic correlation structure of Gainj- and Kaam-speaking people. I. The estimation and interpretation of Wright's F-statistics. **Genetics. 112: 629–647, 1986.**

Malécot G. **Les mathématiques de l'hérédité.** Masson et Cie, Paris, 1948.

McQuillan R; Leutenegger AL; Abdel-Rahman R; Franklin CS; Pericic M; Barac-Lauc L; Smolej-Narancic N; Janicijevic B; Polasek O; Tenesa A; MacLeod AK; Farrington SM; Rudan P; Hayward C; Vitart V; Rudan I; Wild SH; Dunlop MG; Wright AF; Campbell H; Wilson JF. Runs of homozygosity in European populations. **American Journal of Human Genetics. 83: 359-372, 2008.**

Mingroni-Netto RC; Angeli CB; Kimura L; Auricchio MTBM; Vicente JP. Doenças modernas nos antigos quilombos: a obesidade e a hipertensão no Vale do Ribeira. In: Volochko A; Batista LE. **Saúde nos quilombos.** Instituto da Saúde, São Paulo, 2009a, pp. 179-191.

Mingroni-Netto RC; Auricchio MTBM; Vicente JP. Importância da pesquisa do traço e da anemia falciforme nos remanescentes de quilombos do Vale do Ribeira-SP. In: Volochko A; Batista LE. **Saúde nos quilombos.** Instituto da Saúde, São Paulo, 2009b, pp. 169-177.

Nalls MA; Guerreiro RJ; Simon-Sanchez J; Bras JT; Traynor BJ; Gibbs JR; Launer L; Hardy J; Singleton AB. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. **Neurogenetics. 10: 183-190, 2009.**

Nei M. Analysis of gene diversity in subdivided populations. **Proceedings of the National Academy of Sciences USA. 70: 3321–3323, 1973.**

Nei M. F-statistics and analysis of gene diversity in subdivided populations. **Annals of Human Genetics. 41: 225–233, 1977.**

Nei M; Chesser R. K. Estimation of fixation indices and gene diversities. **Annals of Human Genetics. 47: 253–259, 1983.**

Nunes K; Zheng X; Torres M; Moraes ME; Piovezan BZ; Pontes GN; Kimura L; Carnavalli JEP; Mingroni-Netto RC; Meyer D. HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set. **Human immunology. 77: 307-312, 2016.**

Otto PA; Frota-Pessoa O. Genetic risks of consanguineous marriages. In: Mayo O; Leach C (Orgs.). **Fifty years of human genetics: A Festschrift and liber amicorum to celebrate the life and work of**

George Robert Fraser. Adelaide: Wakefield Press, Australia, 2007, pp. 436-442.
Otto PA; Lemes RB. A note on the variance of the estimate of the fixation index F. **Journal of Genetics. 94: 759-763, 2015.**
Pasinato R; Rettl KI. Desenvolvimento local sustentável: a contribuição das comunidades quilombolas do Vale do Ribeira. In: Volochko A; Batista LE. (Orgs.). **Saúde nos quilombos.** Instituto da Saúde, São Paulo, 2009, pp. 43-56.
Parvari R; Hershkovitz E; Kanis A; Gorodischer R; Shalitin S; Sheffield VC; Carmi R. Homozygosity and linkage-disequilibrium mapping of the syndrome of congenital hypoparathyroidism, growth and mental retardation, and dysmorphism to a 1-cM interval on chromosome 1q42-43. **American Journal of Human Genetics. 63: 163-169, 1998.**
Pemberton TJ; Absher D; Feldman MW; Myers RM; Rosenberg NA; Li JZ. Genomic patterns of homozygosity in worldwide human populations. **American Journal of Human Genetics. 91: 275-292, 2012.**
Pemberton TJ; Rosenberg NA. Population-genetic influences on genomic estimates of the inbreeding coefficient: a global perspective. **Human Heredity. 77: 37-48, 2014.**
Purcell S; Neale B; Todd-Brown K; Thomas L; Ferreira MA; Bender D; Maller J; Sklar P; de Bakker PI; Daly MJ; Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics. 81: 559-575, 2007.**
R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. **URL https://www.R-project.org/. 2016.**
Riaz HF; Mannan S; Malik S. Consanguinity and its socio-biological parameters in Rahim Yar Khan District, Southern Punjab, Pakistan. **Journal of Health, Population and Nutrition. 35: 14, 2016.**
Robertson A; Hill W. Deviations from Hardy-Weinberg proportions: sampling variances and use in the estimation of inbreeding coefficients. **Genetics. 107, 703–718, 1984.**
Rosenberg NA; Pemberton TJ; Li JZ; Belmont JW. Runs of homozygosity and parental relatedness. **Genetics in Medicine. 15: 753-754, 2013.**
Rozen S; Skaletsky HJ. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S; Misener S (Eds.). **Bioinformatics Methods and Protocols: Methods in Molecular Biology.** Humana Press, Totowa, NJ, 2000, pp. 365-386.
Salzano FM, Freire-Maia N. **Populações brasileiras: Aspectos demográficos, genéticos e antropológicos.** Editora da Universidade de São Paulo, São Paulo, 1967.
Santos KMP; Tatto N (Eds.). **Agenda socioambiental de comunidades quilombolas do Vale do Ribeira.** Ipsis Gráfica e Editora, São Paulo, 2008.
Sheffield VC; Carml R; Kwltek-Black A; Rokhlina T; Nishlmura D; Duyk GM; Elbedour K; Sunden SL; Stone EM. Identification of a Bardet-Biedl syndrome locus on chromosome 3 and evaluation of an efficient approach to homozygosity mapping. **Human Molecular Genetics. 3: 1331-1335, 1994.**
Souza IR; Culpi L. Valongo, an isolated Brazilian Black community. I. Structure of the population. **Brazilian Journal of Genetics. 15: 439-447, 1992.**
Speed D; Balding DJ. Relatedness in the post-genomic era: is it still useful? **Nature Reviews Genetics. 16: 33-44, 2015.**

Templeton AR. **Population genetics and microevolutionary theory**. John Wiley & Sons, Inc, Hoboken, New Jersey, 2006.

Teo SM; Ku CS; Salim A; Naidoo N; Chia KS; Pawitan Y. Regions of homozygosity in three Southeast Asian populations. **Journal of Human Genetics. 57: 101-108, 2012.**

Wang S; Haynes C; Barany F; Ott J. Genome-wide autozygosity mapping in human populations. **Genetic Epidemiology. 33: 172-180, 2009.**

Weir BS. Genetic data analysis II. **Sinauer Associates Inc, Sunderland, MA, 1996.**

Weir BS. Interpreting Whole-Genome Marker Data. **Statistics in Biosciences. 5: 2013.**

Weir BS; Cockerham CC. Estimating F-statistics for the analysis of population structure. **Evolution. 38: 1358–1370, 1984.**

Wigginton JE; Cutler DJ; Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. **American Journal of Human Genetics. 76: 887-893, 2005.**

Winick JD; Blundell ML; Galke BL; Salam AA; Leal SM; Karayiorgou M. Homozygosity mapping of the Achromatopsia locus in the Pingelapese. **American Journal of Human Genetics. 64: 1679-1685, 1999.**

Wright S. Coefficients of inbreeding and relationship. **American Naturalist. 56: 330-338, 1922.**

Wright S. Isolation by distance. **Genetics. 28: 114–138, 1943.**

Wright S. The genetical structure of populations. **Annals of Eugenics. 15:323-354, 1951.**

Yang HC; Chang LC; Liang YJ; Lin CH; Wang PL. A genome-wide homozygosity association study identifies runs of homozygosity associated with rheumatoid arthritis in the human major histocompatibility complex. **PLoS One. 7: e34840, 2012.**

Yeh E; Kimura L; Errera FI; Angeli CB; Mingroni-Netto RC; Silva ME; Canani LH; Passos-Bueno MR. Association of polymorphisms at the ADIPOR1 regulatory region with type 2 diabetes and body mass index in a Brazilian population with European or African ancestry. **Brazilian Journal of Medical and Biological Research. 41: 468-472, 2008.**

A. ANNEX 1

This section details the process of data cleaning (briefly summarized on Chapter 3) performed on data obtained from all 541 sampled individuals genotyped with a commercial ~600,000 SNPs array high density platform.

The whole process consisted of five different steps described below, the first two of them having been performed by Dr. Kelly Nunes.

A.1. Step 1

The software *Genotype Console 4.2* was used to exclude all markers presenting low quality scores, according to the manufacturer's standard parameters.

A.2. Step 2

All markers presenting significant pairwise differences in missing data proportions between gender, batch and subpopulation groups were excluded by means of the R package *GWASTools v. 3.5*.

A.3. Step 3

Markers located within the pericentromeric and peritelomeric regions were excluded, because these segments normally have a small number of SNPs, responsible for gaps with lack of genetic information, and are enriched in repetitive DNA sequences, thus reducing the accuracy of the genotype determination process. Three different exclusion methods were tested, taking into account: (M1) all markers located within the first 2Mb starting from the outermost genotyped marker across all chromosomal arms; (M2) the 300 outermost genotyped markers across all chromosomal arms; (M3) the largest number of genotyped markers contained in the chromosomal segments by applying methods (M1) or (M2).

Figure A1 shows, in graphs A to D (A: raw data, B to D: datasets selected by methods M1 to M3) the pairwise distances between consecutive SNPs (ordinate axis) as function of their physical order (abscissa axis), separated by the 21 chromosomal boundaries.

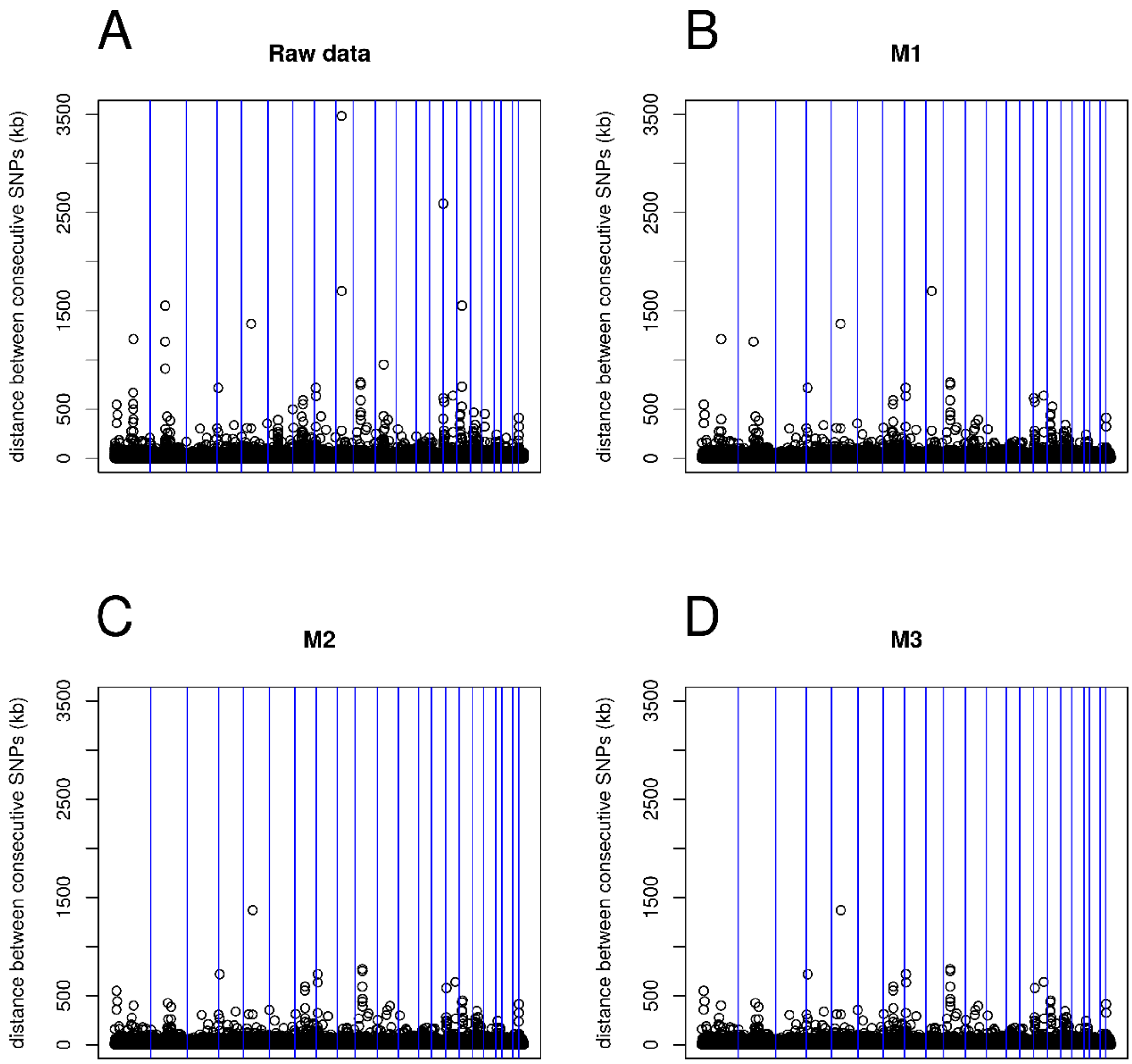


Figure A1: Distances between consecutive SNPs (y axis), according to its order in genomic physical position. Blue lines: boundaries between chromosomes, organized in ascending order.

The inspection of the four graphs of Figure A1 shows clearly that the vast majority of points are in the range below 500kb, indicating that the SNP coverage is relatively homogenous across the genome, and that both graphs C and D contain fewer points with higher values. Based on the quantitative results shown in Table A1, the M2 method was selected for further analyses, because its resulting dataset, besides having the lowest values of mean, median, and variance (of distances between consecutive SNPs), is more conservative as to the number of loci retained.

Table A1: Distances of consecutive SNPs (descriptive statistics).

Dataset	Raw data	M1	M2	M3
Number of loci	591228	565140	567789	560601
Genomic coverage (Gb)	2.670	2.509	2.494	2.469
Mean distances (kb)	4.517	4.440	4.392	4.404
Median distances (kb)	2.243	2.233	2.221	2.227
Maximum distance (Mb)	3.484	1.702	1.369	1.369
Variance ($\times 10^6$)	128.403	77.079	63.334	63.737
Standard deviation ($\times 10^3$)	11.331	8.779	7.958	7.984

A.4. Step 4

Loci exhibiting highly significant deviations ($P < 0.0001$) from HW proportions were excluded using the software PLINK v. 1.07, which estimates P-values by means of the exact test of Wigginton *et al.* (2005). Such deviations might result from low quality genotyping process, mainly when widespread across the genome, or from the effects of evolutionary selection processes, especially when limited to specific genomic regions (Weir, 2013). Of course, **f**-values significantly different from zero could also be the result of inbreeding, but P-values obtained for each locus would never attain the level above.

We also excluded data from loci with lack of information, testing empirically four different thresholds (5%, 1%, 0.5%, and 0%), as suggested by Weir (2013). The graphs of Figure A2 show the observed (ordinate axis) and expected (abscissa axis) P-values obtained when testing the null hypothesis of panmixia in all five resulting datasets. The inspection of graph A shows clearly that P-values from the filtered data are closer to the expected ones than the raw data.

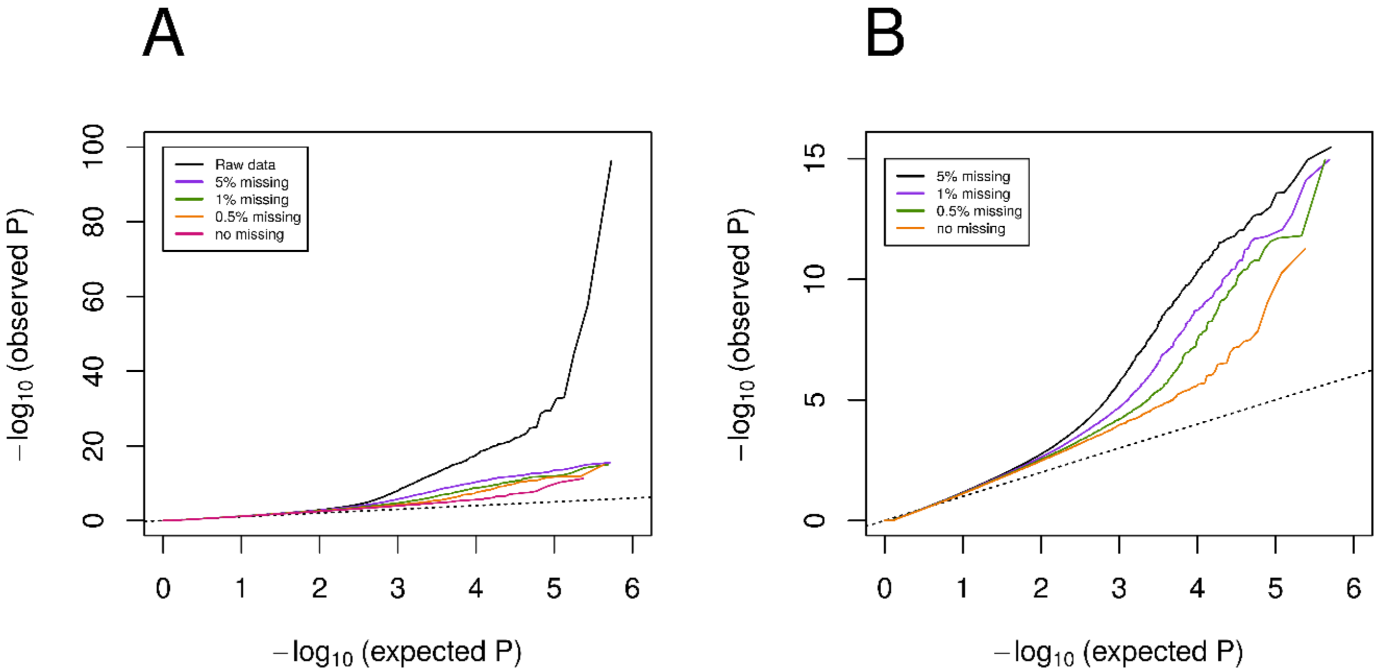


Figure A2: QQ-plots of quilombo exact tests, using raw and filtered datasets (graph A) or only filtered datasets (graph B).

The missing data threshold of 1% was considered for further analysis because the corresponding filtering process eliminated a huge number of markers with biased genotype frequencies while retaining approximately 95% of all markers from the cleaning step 3, as shown in Table A2.

Table A2: Number and proportion of loci left after data filtering.

Dataset	Number of loci	Proportion of remaining loci
Raw data (step 2)	591,228	-
Trimmed data (step 3)	567,789	1.0000
HW + 5% missing data (step 4)	566,000	0.9968
HW + 1% missing data (step 4)	538,981	0.9493
HW + 0.5% missing data (step 4)	481,284	0.8476
HW + 0% missing data (step 4)	273,143	0.4811

HW: HW testing with $P < 0.0001$.

A.5. Step 5

Some loci were excluded according to their minor allele frequencies (MAF), after testing empirically three different MAF thresholds:

(1) MAF = 0 (monomorphic markers, which are non-informative, thus introducing a significant bias on the identification of ROHs);

(2) MAF ≤ 1/(2N), where N is the sample size (loci that might include genotypes containing *de novo* mutations or genotyping errors);

(3) MAF ≤ 1% (idiomorphic markers).

Datasets resulting after the application of the three criteria preserves respectively 485,957, 478,327, and 454,988 loci. In order to keep the largest number of markers for the final analysis, the MAF = 0 threshold was selected.