Annals of
human genetics

doi: 10.1111/ahg.12001

So Close, So Far Away: Analysis of Surnames in a Town of Twins (Cândido Godói, Brazil)

Marcelo Zagonel De Oliveira[1,2], Lavínia Schüler-Faccini[1,2]*, Dario A. Demarchi[3], Emma L. Alfaro[4], José E. Dipierri[4], Mauricio R. Veronez[5], Marlise Colling Cassel[1,2], Alice Tagliani-Ribeiro[1,2], Ursula Silveira Matte[1,6] and Virginia Ramallo[1,2]

[1] *INAGEMP–Instituto Nacional de Genética Médica Populacional, Porto Alegre, Brazil*
[2] *Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil*
[3] *Museo de Antropología, FFyH, Universidad Nacional de Córdoba, Córdoba, Argentina*
[4] *Instituto de Biología de la Altura, Universidad Nacional de Jujuy, San Salvador de Jujuy, Argentina*
[5] *LASERCA–Laboratorio de Sensoriamento Remoto e Cartografia Digital – Programa de Pós-Graduação em Geologia, Universidade do Vale do Rio dos Sinos (UNISINOS), Brazil*
[6] *Unidade de Analise Molecular e de Proteínas, Centro de Pesquisas, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil*

Summary

Cândido Godói is a small Brazilian town known for high rates of twin birth. In 2011, a genetic study showed that this localized high rate of twin births could be explained by a genetic founder effect. Here we used isonymic analysis and surname distribution to identify population subgroups within 5316 inhabitants and 665 different surnames. Four clusters were constructed based on different twin rates ($P < 0.001$; MRPP test). Fisher´s α and consanguinity index showed low and high values, respectively, corresponding with observed values in isolated communities with high levels of genetic drift. Values of A and B estimators confirmed population isolation. Three boundaries were identified with Monmonier´s maximum difference algorithm ($P = 0.007$). Inside the isolated sections, surnames of different geographic origins, language, and religion were represented. With an adequate statistical methodology, surname analyses provided a close approximation of historic and socioeconomic background at the moment of colony settlement. In this context, the maintenance of social and cultural practices had strong implications for the population´s structure leading to drift processes in this small town, supporting the previous genetic study.

Keywords: Cândido Godói, twins, population structure, isonymy, founder effect

Introduction

Surnames represent cultural features that are transmitted from ancestors to their descendants through a vertical mechanism similar to that of genetic inheritance (Guglielmino et al., 1991). They are historic features of identity (Manrubia & Zanette, 2002) characteristic of a family, a population, or a group of related populations, and have been recognized as an inheritance system for our species

The analysis of surname distribution can substitute quantitative information on the genetic structure of a given population. All human communities are grouped into a certain structure, either because of their limited number of ancestors, or their acceptance or refusal of a specific kind of union. The selection of potential partners among a given group follows a pattern of preference based on a particular set of values that may vary according to opportunities, will, and the extension of the circle from which the selection is made.

Crow and Mange (1965) developed the isonymic method to calculate the consanguinity coefficient, including random and nonrandom components for large populations (for a review see Lasker, 1985). So far, studies on surname distribution

*Corresponding author: LAVÍNIA SCHÜLER-FACCINI, Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500 – Prédio 43323M, Caixa Postal 15053, CEP 91501–970, Porto Alegre, RS, Brasil. Tel: 55 51 3308-6722; Fax: 55 51 3308-7311; E-mail: lavinia.faccini@ufrgs.br

have become widely and successfully applied to estimate the structure and level of migration, admixture, or isolation (Jobling, 2001; King & Jobling, 2009; Graf et al., 2010) in a given community, region, country, or continent (Dipierri et al., 2005, 2007, 2011; Scapoli et al., 2007; Tarskaia et al., 2008; Bronberg et al., 2009; Cheshire et al., 2010).

This work shows results obtained for the population of Cândido Godói, Brazil (27°57' 07"S; 54°45' 07"W) from a Population Medical Genetics approach. The main objective of this multidisciplinary approach is the study and potential medical care of a whole population. This discipline is globally underdeveloped and quite recent in Latin America. The present findings provide the basis for future education projects on health prevention to inform the population about its historical and current reproductive structure.

Cândido Godói is a small town (246.28 km^2) in the northwest Rio Grande do Sul state, almost on the border between Argentina and Brazil, with a rural-based economy. This is a small municipality, divided into 24 sections (called "linhas"), with 6535 inhabitants in total and a demographic density of 26.54 persons/km^2 (Census 2010 IBGE- Brazilian Institute of Geography and Statistics). Almost 39% of the population is established in Linha Centro, the institutional seat and commercial center of Cândido Godói, 11.37 km^2 in size, and the rest of the population live in the peripheral areas. Most internal roads are in poor condition, making communication between people from different rural "linhas" difficult. The present population is largely of Polish or German ancestry and this characteristic is proudly emphasized in the oral memory of each family. Their ancestors were part of the great migration to the colonies of Rio Grande do Sul that took place during the 19th and beginning of the 20th century, and who then settled at Cândido Godói. At the time, the Brazilian government promoted immigration by allowing private undertakings. As a legacy of the recent past, many people still speak German, Polish, or dialects of both.

In this town, the most notable characteristic is the high rate of twin births (Tagliani-Ribeiro et al., 2011), close to 2%, considerably higher than that observed for the whole of Brazil [1% mean—Data obtained from DATASUS, national health records of the "Ministério da Saúde" for 1994–2006, see reference (Datasus 2010)]. It is known as the "TwinsTown" and this characteristic is a cause of celebration for the whole community. Every two years there is a Twins' Festival, a tourist event of economic relevance. This demographic phenomenon of a high twinning rate is not homogeneously distributed in Cândido Godói because, for instance, the prevalence rate in Linha São Pedro is between 7% and 10% (Matte et al., 1996; Tagliani-Ribeiro et al., 2011). Though most etiological aspects of these twin gestations have not yet been elucidated, several studies have been developed (Hall, 1996; Lummaa et al., 1998; Beemsterboer et al., 2006; Steinman, 2006), reporting geographic and time variation in twin birth rates (Hoekstra et al., 2008) involving both genetic and environmental factors (Al-Hendy et al., 2000; Hasbargen et al., 2000; Montgomery et al., 2004; Steinman 2006; Hoekstra et al., 2008, among many others).

Following the initiative of the community at Cândido Godói to ask for help from the University to understand the origins of the high incidence of twins, a research project involving baptism books and family stories on twin births was developed (Tagliani-Ribeiro et al., 2011). Authors supported the hypothesis of a genetic founder effect to explain the high prevalence rate of twin births.

Following this approach, this work analyzes the isonymic structure of the whole community at Cândido Godói and the surname distribution within the town. The purpose of this is to identify population subgroups and eventual cases of isolation that will in the end enable a scenario in which a founder effect could be said to have occurred.

Materials and Methods

Ethical Aspects

This work is part of the research project "Estudos genéticos e ambientais acerca da etiologia da gemelaridade em Cândido Godói, RS," already revised and approved by the Ethics Committee at Porto Alegre Clinical Hospital ("Hospital de Clínicas de Porto Alegre") under Protocol No.09–359. All data were analyzed preserving the identities of volunteers.

Data Sources Used

In Brazil, voting is mandatory for all citizens from 18 years of age. Each municipality in the country is divided into sectors and establishes a general voters' registry (Registro de Eleitores), comprising all residents within the sector over 18 years old. These registries constitute an excellent source of information in population studies because a large number of surnames are included at different levels of aggregation (electoral sector, municipality, state, and finally, an entire nation). In this work, the 2010 electoral registry for Cândido Godói and 23 sectors was analyzed.

Surnames can be indicative of migration, because they can be indicative of geographic origin. For this purpose, databases of the "Instituto de Imigração e Colonização da Polícia Marítima," 'Arquivo Histórico Nacional," and "Instituto Histórico e Geográfico Brasileiro," were consulted. These files are records of migration since the beginning of the 19th century, when the immigration process in the northern region of Rio Grande do Sul began. At the time of entering the country, the immigration office recorded the name of each family group, its members, and their ages, the ship on which they had traveled and which country they came from. These databases were the source for pinpointing a geographical origin for each surname analyzed in this study.

In addition, the National Census 2010 from IBGE (Brazilian Institute of Geography and Statistics) and health register databases from DATASUS (National Health Agency, Brazil) were also consulted. All geographic and demographic information was provided by municipal authorities of Cândido Godói and IBGE.

General Characteristics of the Sample

The total sample comprised 5316 adult inhabitants of Cândido Godói whose data were obtained from the 2010 Electoral Register, representing 81% of the total population.

Every section or linha represents an electoral district, except for Linha Cascata which was included in the neighboring district by the Regional Electoral Court because of its small number of residents (<30). Thus, data in this study were calculated over the total 23 linhas in the Register ("Registro de Eleitores").

Similar to many other countries in Latin America, every person in Brazil has two surnames: maternal surname first, followed by the paternal surname. Though Cândido Godói shows a significant number of double surnames, most of its inhabitants only use their paternal surname. This practice is frequently found in colonies of European descendants. To make our sample homogeneous for global calculations, we considered only the paternal surname for each individual. Double surnames were separately analyzed, as detailed in the following section.

Statistics and Isonymy

To analyze the variability of surnames in the community, their possible distributional patterns in space and a potential population structure, we calculated a set of statistical indices based on data distribution for each linha. Unbiased isonymy was calculated according to the following formula (Rodríguez-Larralde et al., 1993):

$$I_{NS} = \sum_k (N_{ki}/N_i)^2 - (1/N_i)m,$$

where N_{ki} represents the absolute frequency of surname k for linha i, and N_i is the size of the population in the linha. The sum includes all surnames.

Fisher´s α index was calculated according to Barrai et al. (1996):

$$\alpha = 1/I_{NS}.$$

This parameter reflects the effective number of surnames and was applied to evaluate diversity. High values were observed for communities with high immigration rates, whereas isolated communities with high genetic drift showed low values.

Consanguinity was assessed for the overall population and for each linha in particular. This inbreeding coefficient was computed as

$$F = Fn + Fr\,(1 - Fn).$$

In this formula, *Fr* is the random component, a measure of isonymy that occurs from random unions within the population and is therefore a function of surname frequency only:

$$Fr = \sum (p_k q_k)/4,$$

where p_k is the frequency of the *k*th surname in males, and q_k is the frequency of *k*th surname in females; summation is over all surnames.

Fn, the nonrandom component, resulting from the preference or intentional rejection of certain marriages.

$$Fn = (I - \sum p_k q_k)/4(1 - \sum p_k q_k)$$

Fn can be either positive or negative, and immediately becomes zero with random mating (Crow & Mange, 1965)

Isonymy was also calculated among linhas to identify the eventual common origin of individuals (Scapoli et al., 2007):

$$I_{ij} = \sum_k P_{ki}\, P_{kj},$$

where P_{ki} and P_{kj} are the relative frequencies of surname k for linhas i and j, respectively. The sum includes all surnames. When two groups show no surnames in common, isonymy is equal to 0 (zero).

A and *B* estimators of population sedentarism and isolation were calculated (Rodriguez-Larralde, 1990). The first represents the percentage of the population with only one representative per surname and the second is the percentage of the population with the seven most frequent surnames. High values of *B* estimator were observed for isolated or small communities with high emigration rates, in which only few surnames were found in the majority of the population.

Linhas were divided into four groups according to the different twin rates observed (Tagliani-Ribeiro et al., 2011) and surname distances were then calculated (distance formulas are presented in the following paragraph). The multiresponse permutation procedure (MRPP) was used to test significance of Linha grouping according to twin birth rates, using PC-ORD 4.0 software. The MRPP is a nonparametric procedure for testing the hypothesis of no differences between two or more groups of entities (in this case, populations); it is equivalent to discriminant analysis or one-way analysis of variance (McCune, 1991). As the probability value of an MRPP statistic is derived through a permutation argument, there are no distributional requirements on the data, such as multivariate normality and homogeneity of variances. A permutation is a specific arrangement or assignment of all N objects (in this case population samples) to the specified groups. The null hypothesis for the MRPP states that all the possible permutations are equally likely. The test statistic indicates the extent of differentiation between groups. The observed $\boldsymbol{\delta}$ (the average of the within-group distances) is compared to an expected $\boldsymbol{\delta}$, which is calculated to represent the mean $\boldsymbol{\delta}$ for all possible partitions of the data. Small values of $\boldsymbol{\delta}$ indicate a tendency for clustering, whereas larger values of $\boldsymbol{\delta}$ indicate a lack of clustering. The variance and skewness of $\boldsymbol{\delta}$ are descriptors of the distribution of all possible values of $\boldsymbol{\delta}$ corresponding to the possible partitions of the items. The probability value expresses the likelihood of obtaining a $\boldsymbol{\delta}$ as extreme or more extreme than the observed, given the distribution of all possible deltas. For details, see Zimmerman et al. (1985).

Euclidean distance formula (Cavalli-Sforza & Edwards, 1967) between groups I and J is defined as:

$$E = \sqrt{(1 - \sum_k \sqrt{P_{ki}\,P_{kj}})},$$

and Sørensen´s distance formula (Sørensen, 1948):

$$QS = 2C/A + B$$

where A and B are the total number of surnames for A and B linhas respectively, and C is the number of surnames shared by both linhas. The Sørensen index used as a distance measure was calculated as $1 - QS$. Though other distance formulas are available (Nei´s, Laker´s), those used in this work obtain results between 0 (absolute similarity) and 1 (no similarity). Matrixes of paired isonymic distances were calculated for all linhas, following both distance formulas.

The relationship of surname distribution among linhas was graphically expressed by means of a dendogram based on the UPGMA (unweighted pair group method with arithmetic mean) algorithm using NTSYSpc 2.11S® software (Rohlf, 2000).

A matrix of geographic distances was constructed from relative central points inside each linha. Significance of the correlation between this matrix and that of surname distances was evaluated by the Mantel test (Mantel, 1967) and PASSaGE 2® software (Rosenberg & Anderson, 2011).

After describing the distribution of population variability in space, discontinuous zones indicate eventual potentially geographic, biological, or cultural barriers. A geometric method based on Monmonier´s maximum differentiation algorithm (Monmonier, 1973) was applied to a net connecting all linhas—relative central spots previously defined—by means of Delaunay´s triangulation (Brassel & Reif, 1979) to identify potential barriers. According to the matrix of surname distances, barriers were calculated using BARRIERS 2.2® software (Manni et al., 2004), and MRPP was also used to evaluate the statistical significance of these clusters of linhas within boundaries.

According to the historical databases mentioned before, each surname from electoral lists was assigned to a certain geographic origin and the frequencies obtained were compared between linhas. These results were also compared with the distribution of barriers obtained by the Monmonier algorithm.

The Electoral Register of Cândido Godói includes 546 individuals with double surnames. As mentioned before, the first surname is inherited from the mother and the second one from the father. After marriage, women can take the family names of their husbands, adding it as a second surname, discarding her mother's surname and retaining the paternal surname. This inheritance system is useful for identifying marital preferences or residence patterns of the population under study. Every double surname refers to a particular type of union, either contemporary to the individual (her own marriage, for the case of women) or to the previous generation (parents´ marriage). This subset of the sample (546 individuals with double surnames, 491 different family names in total) was analyzed separately, establishing the geographical origin of each surname, to identify possible trends toward a preferential kind of union.

The pattern of change of residence, even within the same community, may be different for men and women. If the migration rate varies according to sex, it is possible that the continuity of a certain pattern changes random isonymic values in the following generations, making maternal or paternal surnames more diverse (Pinto-Cisternas et al., 1990; Herrera Paz et al., 2010a; Herrera Paz et al., 2010b). The percentage of two variability models for the lists of doubles surnames was separately determined to estimate the trend in residence patterns.

Maternal percentage was calculated as follows:

$$(Ia1SN/(Ia1SN + Ia2SN))100.$$

Paternal percentage was calculated as follows:

$$(Ia2SN/(Ia1SN + Ia2SN))100.$$

where Ia 1SN and Ia 2SN represent random isonymy for the first and second surname, respectively.

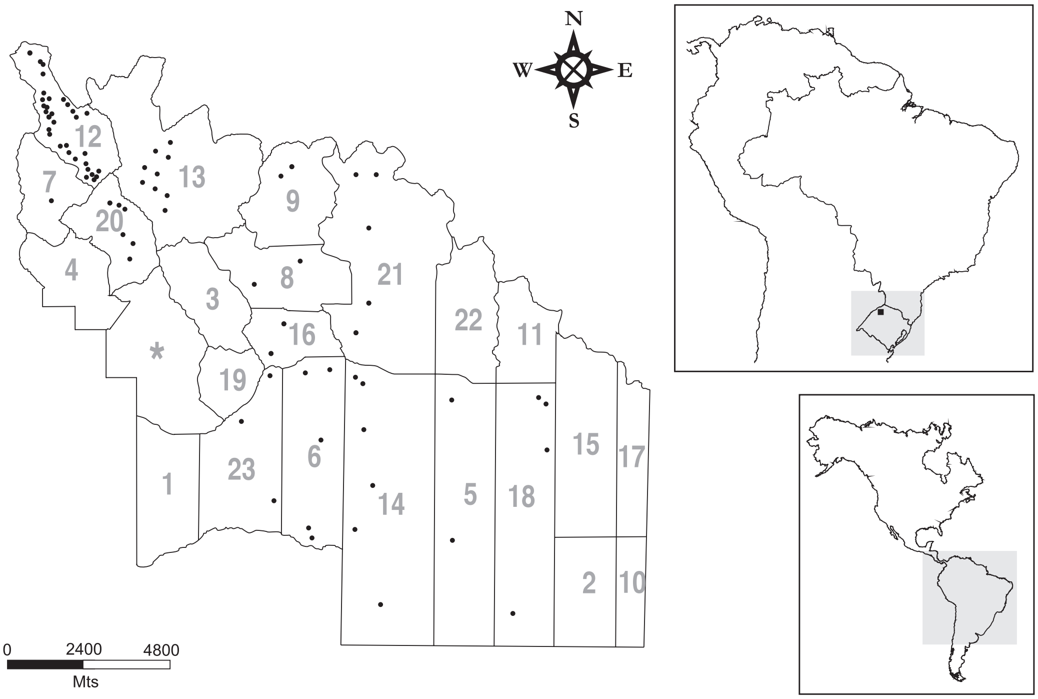

Figure 1 Districts of Cândido Godói and their respective twin birth registration. The distribution is based on twin's baptism records from 1959 to 2006 (Tagliani-Ribeiro et al., 2011). Each twin record is marked with a dot. * Symbol denotes Linha Cascata, not included in the general analysis. See details in "General characteristics of the sample."

Maps and Graphics

Geographic maps and data were obtained with the ArcGis 9.3® software and the cartographic base was geo-referenced to the SIRGAS geodesic system (Geocentric Reference System for the Americas) in UTM projection (Transverse Universal Mercator). Corel-DRAW X3® software (Corel Corporation, Ottawa, Canada) was used to edit all images obtained.

Results and Discussion

Possibly because of erroneous transcription, the electoral register shows alternative spellings for many surnames, with a mean of five variants for some of them. There were a total of 246 surnames with orthographic variants, organized into 49 groups according to graphic or phonetic similarity. This procedure requires careful and homogeneous criteria because the change in the final number of surnames to analyze may lead diversity to be under- or overestimated. Further interviews with local inhabitants enabled confirmation of group assignment. The interest shown by the Cândido Godói community was essential for this research study. Thus, differences were determined as historic mistakes at the time of registering migrant families into Brazil. From a initial total of 862 surnames, an effective final number of 665 different surnames were recorded among 5316 individuals. Figure 1 shows a map of the town and its location in Rio Grande do Sul, Brazil, the different linhas in it and their respective twin birth registration (Tagliani-Ribeiro et al., 2011). Each twin record is marked with a dot.

The distribution by linha of the surname used in the analysis, with the isonymic parameters calculated, is given in Table 1. Data by linha are shown according to increasing Fisher´s α values. In general, alpha shows low values, which suggest high consanguinity within sections, except for Centro and Timbauva, the two sections densely populated. Linha Centro, institutional seat and commercial center of Cândido Godói is the one which shows the highest variety of different surnames and number of inhabitants. This linha also shows the lowest isonymy values and therefore the lowest level of consanguinity. Linha Doze Norte is on the opposite side, with 159 inhabitants. Though population concentration in this linha is not the lowest, it shows the lowest α value. This parameter is calculated in direct relationship with unbiased isonymy rate and is a useful tool for the assessment of surname variety. Barrai et al. (2000) have reported the α value to be an effective indicator of the population structure and

Table 1 ID for linha, names, number of residents N, number of different surnames S, A estimator, and B estimator in Cândido Godói.

ID	Name	N	S	I	α	F	A	B
1	Doze Norte	159	28	0.0729	13	0.0198	2.6	62.3
2	Abrantes	128	24	0.0685	14	0.0191	5.5	63.3
3	Castelo Branco	61	19	0.0618	16	0.0196	9.8	67.2
4	Secção C	138	30	0.0541	18	0.0153	7.2	56.5
5	Paranagua	179	36	0.0516	19	0.0143	5.6	52.5
6	Natal	147	36	0.0485	20	0.0138	7.5	49.7
7	São Bonifácio	113	38	0.0428	23	0.0129	17.7	52.2
8	São João	107	33	0.0416	24	0.0127	10.3	52.3
9	La Salle	101	31	0.0394	25	0.0123	10.9	54.5
10	Dr. Pederneira	143	35	0.0380	26	0.0113	4.9	44.1
11	Dos Louros	156	43	0.0357	27	0.0105	9.6	51.3
12	São Pedro	214	45	0.0338	29	0.0084	3.7	41.6
13	Boa Vista	155	49	0.0332	30	0.0099	9	43.9
14	Acre	213	50	0.0307	32	0.0088	4.2	40.4
15	Dr.P. Toledo	159	41	0.0272	36	0.0084	5.7	39
16	Esquina União	167	46	0.0267	37	0.0082	9	41.3
17	Treze de Maio	114	33	0.02647	37	0.0088	9.6	48.2
18	Silva Jardim	274	56	0.02627	38	0.0075	2.9	35
19	Secção B	134	41	0.02562	39	0.0083	7.5	37.3
20	São Miguel	223	49	0.02477	40	0.0073	4	37.7
21	Secção A	217	57	0.02323	43	0.0070	8.8	38.7
22	Timbauva	342	88	0.01395	71	0.0042	6.4	26.3
23	Centro	1672	368	0.00545	183	0.0015	7.3	12.7
	Total	5316	665					

Comparison of isonymy parameters (Data by linha according to increasing Fisher´s α values)

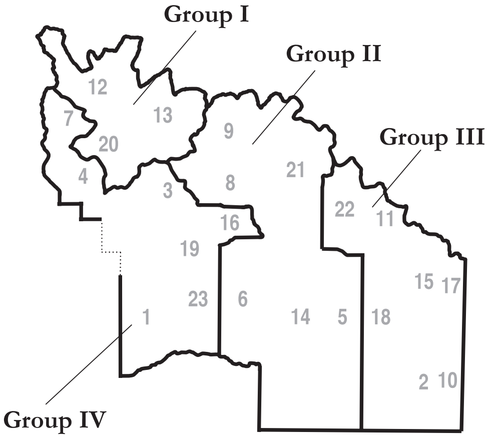

Figure 2 Groups of linhas classified according to different historical rates of twin births. ID numbers for linhas are given in Table 1.

predictor of inheritance dynamics. In the same way, Linha Castelo Branco has only 61 inhabitants, a low α value and the higher value of B estimator. As detailed in the "Materials and Methods" section, this estimator is the percentage of the population with the seven most frequent surnames. In small communities, this is directly and strongly influenced by emigration rates. Low average values of both indices were found in seven linhas with high records of twin births (Acre, Boa Vista, La Salle, Natal, São Joao, São Pedro, São Miguel; ID numbers 14, 13, 9, 6, 8, 12, 20). Inside these linhas, between 38% and 55% of the inhabitants are related as the seven most frequent surnames are found here, indicating an important degree of isolation. A low α value in the present findings would be the consequence of a significant genetic drift process. In Linha São Bonifacio, the A estimator was 17.7%, i.e., people included in this percentage are the only representatives of their surname. In situations of expansion and rapid population growth (e.g., by factors of population attraction, such as an economic or industry improvement), a high frequency of unique representatives of surnames can be registered, because the new immigrants would introduce variation into the population. These are people who do not yet have descendants and tare as yet the only representatives of their surnames. However, in this rural community, with a significant trend toward migration to urban areas (Censo 2010, IBGE), the value of the A estimator denotes the percentage of last representatives of family names still living in the linha.

Linhas were classified according to different historical rates of twin births, information on which was obtained obtained from baptism records available in Cândido Godói churches (Tagliani-Ribeiro et al, 2011). In Figure 2 the four groups (I, II, III, and IV) obtained are shown. To test whether this grouping, based on a differential characteristic of the population (twinning), was accompanied by a random or non-random distribution of surnames, the isonymic distances within each group were measured, using the Euclidean and Sørensen formula.

Table 2 summarized the MRPP analysis for these four groups. The results reveal that the average distances observed between populations, within groups, are significantly lower than the distances expected for groups of populations randomly generated, both for Euclidean and Sørensen distances. A more detailed analysis shows a lack of clustering for group IV, because within-group distances were higher than those expected from a random distribution. This group is the most heterogeneous because it includes Linha Centro.

Sørensen´s coefficient was observed to sensibly correspond with these groups of heterogeneous data and was slightly affected by extreme values. Group I showed the lowest isonymic average distances and included two adjacent linhas with high twin birth rates (São Miguel and São Pedro). Inhabitants in these sections were mostly interrelated; possibly the same lineage groups related by a common history. In fact, historical

Table 2 Results of MRPP test for linha grouping according to twin birth rates, based on Euclidean and Sørensen distance surnames.

Euclidean Distance

Group	Average distance within group	Delta value	p
I	0.0568		
II	0.0651		
III	0.0627		
IV	0.0674	0.0630	0.000014

Sørensen Distance

Group	Average distance within group	Delta value	p
I	0.8625		
II	0.9263		
III	0.8815	0.9065	0.000009
IV	0.9334		

Small values of delta indicate a tendency for clustering, and small *P* values suggest that the null hypothesis (all the possible permutations are equally probable) is unlikely to be true. As compared to Euclidean distance, Sørensen distance measures retain sensitivity in more heterogeneous data sets.

Group I: Boa Vista, São Miguel, São Pedro.

Group II: Acre, La Salle, Natal, Paranaguá, São João, Secção A.

Group III: Treze de Maio, Abrantes, Dos Louros, Pederneiras, Dr. Pedro Toledo, Silva Jardim, Timbauva.

Group IV: Castelo Branco, Centro, Doze Norte, Esquina União, São Bonifacio, Secção B, Secção C.

registers confirm that at the beginning of the 20th century Linha São Pedro was occupied by eight families of German descent who had first settled at Vale dos Sinos, located in the south of the state of Rio Grande do Sul and then moved to Cândido Godói. So far, this population has kept a stable size, a rural-based economy, and only a few other families have joined after its initial colonization. As mentioned before, the tendency to a low population density may be related to the lack of economic incentive in these peripheral regions. Over the years, this situation could be altered by differential migration rates, with individuals in the population moving outside the boundaries of the linha. This phenomenon is not restricted just to Linha São Pedro. Demographic data from the national census show a general trend toward population decrease and genetic drift caused by emigration from rural to urban areas during the last 30 years. In the 1980s, the population of Cândido Godói reached a peak of 8008 inhabitants and then fell steadily to the current population of 6535 (source IBGE).

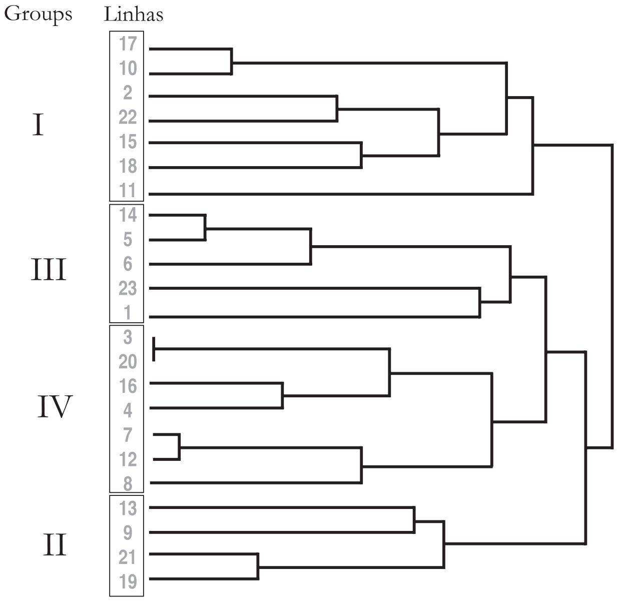

Figure 3 Dendrogram obtained from matrix of Sorensen's pairwise surname distances for 23 linhas. Calculated by UPGMA algorithm (matrix based on Sørensen´s distance formula). ID numbers for linhas are given in Table 1.

Sørensen´s distance formula was applied to compare pairwise linhas, only based on the surname distribution. From the distance matrix obtained, a dendrogram was constructed using the unweighted pair group method with arithmetic mean (UPGMA) method (Fig. 3), in which the relationship between surnames from different sections was graphically represented. The same Sørensen´s distance matrix was then compared with a geographic distance matrix (measured in kilometers between linhas) by the Mantel test. The correlation coefficient obtained was 0.421 ($P = 0.001$; after 1000 permutations), showing that both distances have a significant degree of relationship. However, geographical distance would not be the most important factor to explain the variability of surnames found from one linha to another. Some linhas belonging to the same cluster are not close in distance, e.g., Secção B (ID 19), which belongs to Group II, is mainly concentrated to the north of Cândido Godói. The information is more accessible if the typology of the dendogram is transferred to the map of town, as is showed in Figure 4. Four linha groups can be observed again, though their components (linhas) are somewhat different from those found when they are grouped by twin birth rates (Fig. 2). Groups were numbered according to a decreasing order of inclusiveness. Significantly, São Miguel and São Pedro (ID numbers 20 and 12), two adjacent linhas with high records of twin birth, remained here in the same cluster, now named Group IV, and Group I contains the same linhas as Group III in Figure 2. This cluster remains unchanged in both analyses.

The discontinuities in the isonymic structure of Cândido Godói might be the consequence of directional migration, relocalization of a major group (Sokal, 1992), or diverse barriers (linguistic, economic, religious, etc) interrupting genetic

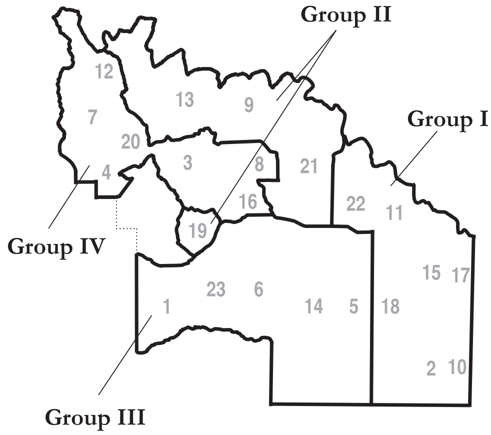

Figure 4 Groups of linhas classified according to dendogram UPGMA. ID numbers for linhas are given in Table 1. Group numbers correspond to Figure 3. Note cluster II does not contain conterminous linhas.

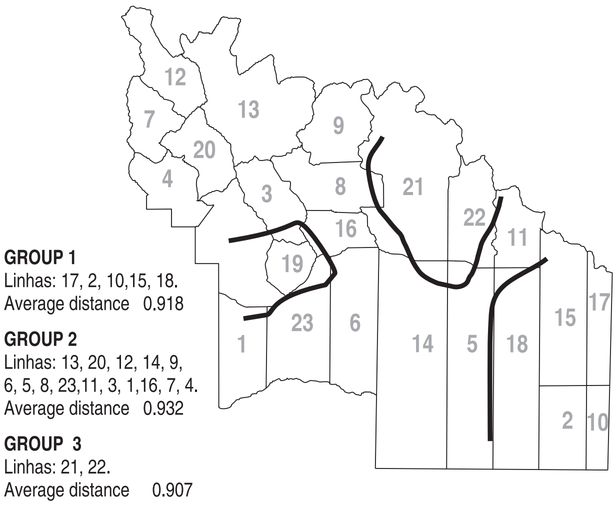

Figure 5 Barriers computed by Monmonier´s maximum difference algorithm, based on matrix of Sorensen's pairwise surname distances for 23 linhas.
Delta = 0.927, $P = 0.007$. Values of delta indicate a tendency for clustering, and small P values suggest that the null hypothesis (all the possible permutations are equally probable) is unlikely to be true. ID numbers for linhas are given in Table 1.

flow. Based on the same Sørensen´s pairwise distance matrix, Monmonier´s maximum difference algorithm enabled the identification of three barriers, defining three different groups of linhas (Fig. 5). The statistical accuracy of this differentiation was also checked by the MRPP test. The null hypothesis states that all the possible permutations (specific arrangement of linhas) are equally likely. The delta value obtained (0.927) was highly significant ($P = 0.007$) and the null hypothesis rejected. Variability between inhabitants from different linhas was not only associated with inherited surnames but also with cultural characteristics related to those markers. Though quite difficult to adjust, these qualitative data dramatically affect any demographic phenomena involving population displacement or admixture (Barbujani & Sokal, 1990). One of these qualitative data types is religion. Based on the governmental records of Cândido Godói, eight linhas: Secção B, Secção A, Timbauva, Silva Jardim, Pedro Toledo, Abrantes, Treze de Maio, and Pederneiras (ID numbers 19, 21, 22, 18, 15, 2, 17, and 10, respectively, see Fig. 5), contain the highest percentage of Protestants. The three identified barriers separate this fraction from the rest of town, which is principally Catholic. Remarkably, five of these linhas clustered in the dendogram using the UPGMA algorithm, belonging to Group I, and the same result was found in both analyses (see Figs 4 and 2).

Among the other qualitative data is the geographic origin of the families. German and Polish descendents comprise the two most numerous groups, unevenly distributed across the town. To elucidate this aspect, all the surnames in the sample were classified according to their geographic origin into six major groups: German, Polish, Portuguese, Italian, Minority (United Kingdom, France, Spain, Belgium, Croatia, Ukraine, Hungary), or Undetermined. The source for the designation of a geographical origin is described in "Materials and Methods." The results of frequencies by linhas are shown in Table 3. German surnames registered high values (between 98.2% and 74.2%) in 21 linhas. This pattern changes only in Linhas Secção A and Secção B, where the majority of surnames are of Polish origin (63.6 and 47.8%, respectively). The remaining four categories (Portuguese, Italian, Minority, and Undetermined) presented considerably smaller frequencies, all less than 12%. Interestingly, barriers calculated before also isolate those linhas showing the highest percentage of Polish descendants, particularly Secção A, Secção B and Timbauva (ID numbers 21, 19, and 22, respectively). The results of these analyses are summarized in Figure 6, in which the frequency of surnames by geographical origin and the most common religious affiliation for each linha are shown. These differences at the level of religion and/or geographical origin could be potentially related to patterns of marriage.

On this topic, 546 individuals with double surnames were analyzed and a total of 491 different surnames were detected among the subset. Table 4 shows all maternal and paternal surnames, classified by geographical origin. The total includes everyone with two surnames. Each combination of surnames marks a particular type of marital union. The values presented in Table 4 indicate a tendency to establish marriage between individuals with surnames of the same geographical origin. It is remarkable to note that people with both German and Polish surnames are less likely to be related to people with Portuguese surnames, especially in unions involving women

Table 3 Frequencies of surnames by geographic origin in 23 linhas of Cândido Godói

ID	Name	Germany	Poland	Portugal	Italy	Minority	Indeterminate
1	Doze Norte	96.2	2.5	1.3	–	–	–
2	Abrantes	95.3	4.7	–	–	–	–
3	Castelo Branco	95.1	3.3	–	1.6	–	–
4	Secção C	81.9	2.9	3.6	2.9	8.7	–
5	Paranagua	73.7	21.2	1.1	1.1	2.8	–
6	Natal	87.1	10.2	1.4	0.7	–	0.7
7	São Bonifácio	98.2	0.9	0.9	–	–	–
8	São João	86	12.1	0.9	–	0.9	–
9	La Salle	75.2	5.9	8.9	–	9.9	–
10	Dr. Pederneira	91.6	5.6	–	–	–	2.8
11	Dos Louros	74.4	9	3.2	–	9.6	3.8
12	São Pedro	93	2.3	2.8	–	1.9	–
13	Boa Vista	80.6	13.5	1.3	–	4.5	–
14	Acre	75.1	18.8	0.5	4.7	0.9	–
15	Dr.P. Toledo	74.2	22	3.8	–	–	–
16	Esquina União	91.6	3.6	2.4	1.2	–	1.2
17	Treze de Maio	79.8	7.9	5.3	7	–	–
18	Silva Jardim	74.8	21.2	–	–	2.2	1.8
19	Secção B	32.1	47.8	11.9	0.7	7.5	–
20	São Miguel	92.8	–	1.8	–	5.4	–
21	Secção A	24	63.6	4.1	0.9	7.4	–
22	Timbauva	77.5	16.4	4.1	1.5	0.6	–
23	Centro	77.6	6.2	6.6	2	4.8	2.8

Minority (United Kingdom, France, Spain, Belgium, Croatia, Ukraine, and Hungary).

with German/Polish surnames and men with a Portuguese surname. In this community, the surnames of Portuguese origin represent the fraction of the population that could be considered local, i.e., citizens of Brazilian descent of former migrants from Portugal, not related to migration events during the 19th and beginning of the 20th century, as opposed to the surnames of other origins that we might consider "foreign," which have arrived more recently into the country. However, these data in Table 4 should be carefully evaluated. As presented in Table 3, most of the inhabitants of Cândido Godói have German names. Because this subset of double surnames is a fragment of the total sample, the frequencies may be biased.

The estimate of possible patterns of residence differently associated with each sex, show that the predominant residence pattern was masculine (56.9%). Maternal surnames are more diverse. This may be because of two causes: (i) women change their place of residence after marriage, which increases surname variability in the linha of her husband; or (ii) the general economy encourages male migration, which affects paternal surname variability. Both scenarios are compatible with a trend of marriage between individuals from the same group of origin.

This has probably been the tendency followed since the very foundation of Cândido Godói. During the 19th and 20th centuries, the settlement pattern of colonies in Rio Grande do Sul was very particular because not only was the economic development of the area expected but also the tightening of Brazilian limits with Paraguay and Argentina was a factor. German immigration started in 1824. Foreigners were received as rural settlers: they were granted a piece of land and tools, and established colonial settlements quite different from Brazilian land holdings (Neumann, 2008).

As mentioned in the Introduction, the authorities promoted immigration by allowing private undertakings. This policy generated two types of communities: "Public Colonies" gathering populations of different origins and religions under government administration, and those in which private companies were mostly hired by religious institutions to organize "Homogeneous Communities" with faith and language as cohesive identity principles. Thus, after several religious cooperatives were organized by German immigrants (Catholic or Protestant-Lutheran) with the purpose of strengthening faith and providing social care and education (Rambo, 1988), a major education and religious structure was created. These associations made their own teaching material for rural schools and classes were given in German (Kreutz, 2008).

During the 20th century colonists moved to different areas in the state in search of better economic prospects, as

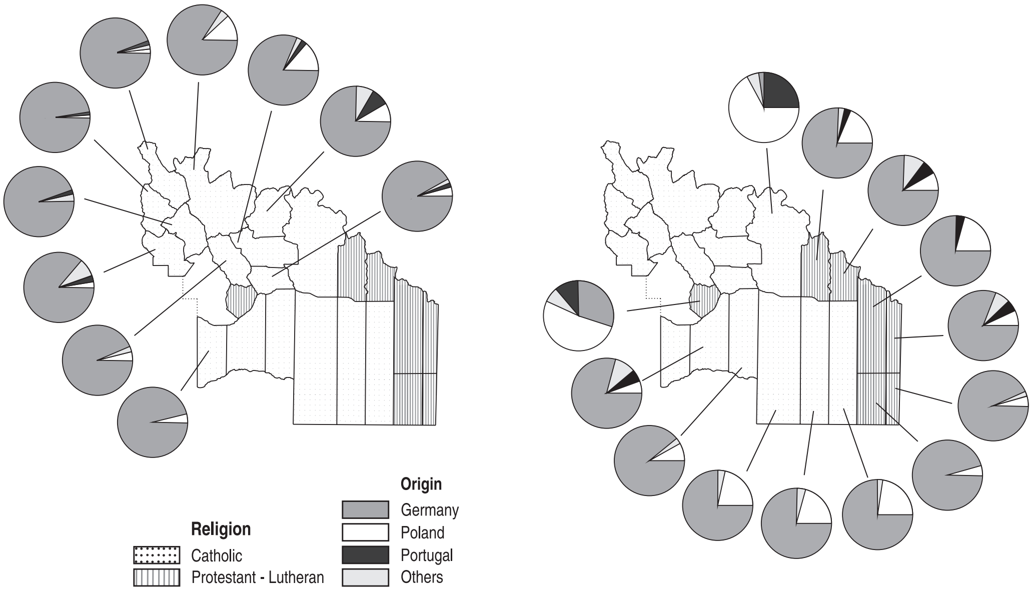

Figure 6 Surname origins and religion by linha.
Most common religious affiliation for each linha is also shown.

Table 4 Marriage pattern by surname origin

	Paternal surname					
Maternal surname	G	P	Po	It	M	Un
G	282	26	23	4	14	2
P	19	19	4	1	–	–
Po	19	2	71	1	4	–
It	6	2	3	3	2	–
M	14	3	8	6	2	–
Un	3	2	1	–	–	–

Country codes: G, Germany; P, Poland; Po, Portugal; It, Italy; M, Minority (United Kingdom, France, Spain, Belgium, Croatia, Ukraine, and Hungary); Un, Undetermined-

mentioned for São Pedro, whose population derived from a displacement of families from Vale dos Sinos. This internal migration, mainly organized by religious associations, expanded colonial limits. Particularly since 1914, these new population clusters became more homogeneous and isolated to maintain their ancestral culture and religion. The period between both World Wars, in which Brazil and Germany were on opposing sides, seriously affected these colonies. School classes were taught in Portuguese, and education in the German language and using German texts was prohibited.

Effects of these global changes were also evident in the municipal town of Santa Rosa and consequently Cândido Godói, which was historically included in it. The new administrative status of Cândido Godói was defined only in 1972. However, its current population has its roots far back in history and is not limited to only one territory or country.

Conclusions

Different research studies on migration, population genetics, and admixture dynamics have also used the isonymic method. There are examples of surname analyses in previously published research findings on the general population of Brazil (Azevêdo et al., 1969, 1980, 1982; Stueber-Odebrecht et al., 1985; Cabello & Krieger, 1991; Leal Barbosa et al., 2006) and on Rio Grande do Sul in particular (Dornelles et al., 1999), where the clustering of mesoregions and a similarity in the frequencies of German surnames was also observed. The low diversity of surnames indicated that the state's population was becoming more homogeneous and still reflected the effect of more recent European migrations.

The analysis of family names is also quite useful for Medical and Population Genetics studies, providing significant available information from state registers and documents, and allowing a large sample size that may even include the whole population for study. Supported by adequate statistical methodology, surname analyses are useful to identify the genetic-demographic structure of any population and enable accurate testing of different hypotheses on micro-evolutive processes.

When applied to Cândido Godói these analyses provided a close approximation of the historic and socio-economic background at the moment settlers established this rural colony and the rules they followed to keep their identity. Results obtained show that the maintenance of these rules over time has strong implications for the structure of the population, generating marked phenomena of isolation. In this scenario, processes of genetic drift can lead to significant changes in the variability of the population and the findings support the conclusions of previous investigations: the hypothesis of a genetic founder effect, which occurred during the colonization of Cândido Godói, is a valid alternative explanation of the high prevalence of twin births.

Acknowledgements

The authors wish to thank Dr. Eduardo Castilla for interesting discussions and ideas on surname analysis.

We also thank Mrs. Selia Heck from Cândido Godói municipality for her help with the voters' registry.

The authors acknowledge INAGEMP—National Institute of Population Medical Genetics (grant CNPq 573993/2008–4) for the support provided to this project

Conflict of Interest

The authors declare no conflict of interest.

References

Al-Hendy, A., Moshynska, O., Saxena, A. & Feyles, V. (2000) Association between mutations of the follicle-stimulating-hormone receptor and repeated twinning. *Lancet* **356**, 914, doi:10.1016/S0140-6736(00)02687-8.

Azevêdo, E. E., Morton, N. E., Miki, C. & Yee Am, S. (1969) Distance and kinship in northeastern Brazil. *Am J Hum Genet* **21**, 1–22.

Azevêdo, E. S. (1980) The anthropological and cultural meaning of family names in Bahia, Brazil. *Curr Anthrop* **21**, 360–363.

Azevêdo, E. S., Fortuna, C. M., Silva, K. M., Sousa, M. G., Machado, M. A., Lima, A. M., Aguiar, M. E., Abé, K., Eulálio, M. C., Conceição, M. M., Silva, M. C. & Santos, M. G. (1982) Spread and diversity of human populations in Bahia, Brazil. *Hum Biol* **54**, 329–341.

Barbujani, G. & Sokal, R. R. (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc Natl Acad Sci* **87**, 1816–1819.

Barrai, I., Scapoli, C., Beretta, M., Nesti, C., Mamolini, E. & Rodriguez-Larralde, L. (1996) Isonymy and the genetic structure of Switzerland. The distribution of surnames. *Ann Hum Biol* **23**, 431–455.

Barrai, I., Rodríguez-Larralde, A., Mamolini, E., Manni, F. & Scapoli, C. (2000) Elements of the surname structure of Austria. *Ann Hum Biol* **27**, 607–622.

Beemsterboer, S. N., Homburg, R., Gorter, N. A., Schats, R., Hompes, P. G. & Lambalk, C. B. (2006) The paradox of declining fertility but increasing twinning rates with advancing maternal age. *Hum Reprod* **21**, 1531–1532.

Brassel, K. E. & Reif, D. (1979) A procedure to generate Thiessen polygons. *Geogr Anal* **325**, 31–36.

Bronberg, R., Dipierri, J. E., Alfaro, E. L., Barrai, I., Rodríguez Larralde, A., Castilla, E., Colonna, V., Rodríguez Arroyo, G. & Bailliet, G. (2009) Isonymy structure of Buenos Aires city. *Hum Biol* **8**, 447–461.

Cheshire, J. A., Longley, P. A. & Singleton, A. D. (2010) The surname regions of Great Britain. *J Maps* **2010**, 401–409.

Cabello, P. H. & Krieger, H. (1991) Note on estimates of the inbreeding coefficient through study of pedigrees and isonymous marriages. *Hum Biol* **63**, 719–723.

Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967) Phylogenetic analysis models and estimation procedures. *Am J Hum Genet* **19**, 233–257

Crow, J. E. & Mange, A. P. (1965) Measurements of inbreeding from the frequency of marriages between persons of the same surnames. *Eugen Q* **12**, 190–203.

Datasus. (2010) http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinasc/cnv/nvrs.def. Accessed November 28, 2010.

Dipierri, J. E., Alfaro, E. L., Scapoli, C, Mamolini, E., Rodriguez-Larralde, A. & Barrai, I. (2005) Surnames in Argentina: A population study through isonymy. *Am J Phys Anthropol* **28**, 199–299.

Dipierri, J. E., Rodríguez-Larralde, A., Alfaro, E. L. & Barrai, I. (2007) Isonymy structure of the Argentine northwest. *Ann Hum Biol* **34**, 498–503.

Dipierri, J. E., Rodríguez-Larralde, A., Alfaro, E. L., Scapoli, Ch., Mamolini, E., Salvatorelli, G., Caramori, G., De Lorenzi, S., Sandri, M., Carrieri, A. & Barrai, I. (2011) A study of the population of Paraguay through isonymy. *Ann Hum Genet* **75**, 678–687

Dornelles, C. L., Callegari-Jacques, S. M., Robinson, W. M., Weimer, T. A., Franco, M. H., Hickmann, A. C., Geiger, C. J. & Salzano, F. M. (1999) Genetics, surnames, grandparents' nationalities, and ethnic admixture in Southern Brazil–do the patterns of variation coincide? *Genet Mol Biol* **22**, 151–161.

Graf, O. M., Zlojutro, M., Rubicz, R. & Crawford, M. H. (2010) Surname distributions and their association with Y-Chromosome markers in the Aleutian Islands. *Hum Biol* **82**, 745–757.

Guglielmino, C. R., Zei, G. & Cavalli-Sforza, L. L. (1991) Genetic and cultural transmission in Sicily as revealed by names and surnames. *Hum Biol* **63**, 607–627.

Hall, J. G. (1996) Twinning: Mechanisms and genetic implications. *Curr Opin Genet Dev* **6**, 343–347.

Hasbargen, U., Lohse, P. & Thaler, C. J. (2000) The number of dichorionic twin pregnancies is reduced by the common MTHFR 677C–>T mutation. *Hum Reprod* **15**, 2659–2662.

Herrera Paz, E. F. & Mejía de Herrera, D. (2010a). Apellidos en Gracias a Dios: Estructura poblacional y patrones de residencia en la Moskitia Hondureña inferidos por el método de isonimia. http://lahondurasvaliente.blogspot.com/2010/09/investigacion-apellidos-moskitia.html. Accessed September 26, 2012.

Herrera-Paz, E. F., Matamoros, M. & Carracedo, A. (2010b). The Garífuna (Black Carib) people of the Atlantic coasts of Honduras: Population dynamics, structure, and phylogenetic relations inferred from genetic data, migration matrices, and isonymy. *Am J Hum Biol* **22**, 36–44.

Hoekstra, C., Zhao, Z. Z., Lambalk, C. B., Willemsen, G., Martin, N. G., Boomsma, D. I. & Montgomery, G. W. (2008) Dizygotic twinning. *Hum Reprod Update* **14**, 37–47.

Jobling, M. A. (2001) In the name of the father: Surnames and genetics. *Trends Genet* **17**, 353–357.

King, T. E. & Jobling, M. A. (2009) Founders, drift, and infidelity: The relationship between Y chromosome diversity and patrilineal surnames. *Mol Biol Evol* **26**, 1093–1102.

Kreutz, L. (2008) Escolas de imigração alemã na Argentina, no Brasil e no Chile. In: *Campos múltiplos: Identidade, Cultura e História* (eds A. Sidekum, I. Grützmann & I. C. Arendt), pp. 153–168. São Leopoldo, Brazil: Oikos press.

Lasker, G. W. (1985) *Surnames and Genetic Structure*. Cambridge University Press, Cambridge, United Kingdom.

Leal Barbosa, A. A., Bispo Sousa, S. M., Abé-Sandes, K., Alonso, C. A., Schneider, V., Costa, D. C., Cavalli, I. J. & Azevêdo, E. E. (2006) Microsatellite studies on an isolated population of African descent in the Brazilian state of Bahia. *Genet Mol Biol* **29**, 23–30.

Lummaa, V., Haukioja, E., Lemmetyinen, R. & Pikkola, M. (1998). Natural selection on human twinning. *Nature* **394**, 533–534

McCune, B. (1991) *Multivariate Analysis on the PC-ORD System*. Oregon, United States of America: Corvallis press.

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

Manrubia, S. C. & Zanette, D. H. (2002) At the boundary between biological and cultural evolution: The origin of surnames distributions. *J Theor Biol* **216**, 461–477.

Mantel, N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res* **27**:209–220.

Matte, U., Le Roux, M. G., Bénichou, B., Moisan, J. P. & Giugliani, R. (1996) Study on possible increase in twinning rate at a small village in south Brazil. *Acta Genet Med Gemellol* **45**, 431–437.

Monmonier, M. (1973) Maximum-difference barriers: An alternative numerical regionalization method. *Geogr Anal* **3**, 245–261.

Montgomery, G. W., Zhao, Z. Z., Marsh, A. J., Mayne, R., Treloar, S. A., James, M., Martin, N. G., Boomsma, D. I. & Duffy, D. L. (2004) A deletion mutation in GDF9 in sisters with spontaneous DZ twins. *Twin Res* **7**, 548–555.

Neumann, R. M. (2008) A iniciativa privada na colonização do donoreste do Rio Grande do Sul: A colonizadora Meyer. In: *Campos múltiplos: Identidade, Cultura e História*. (eds A. Sidekum, I. Grützmann & I. C. Arendt), pp. 123–140. São Leopoldo, Brazil: Oikos press.

Pinto-Cisternas, J., Rodriguez-Larralde, A. & Castro de Guerra, D. (1990) Comparison of two Venezuelan populations using the coefficient of relationship by isonymy. *Hum Biol* **62**, 413–419.

Rambo, A. B. (1988) Associativismo teuto-brasileiro e os primórdios do cooperativismo no Brasil. *Persp Econ* **23**, 263–276.

Rodriguez-Larralde, A. (1990) Distribución de los apellidos y su uso en la estimación de aislamiento y sedentarismo en los municipios del Estado Lara, Venezuela. *Acta Cientifica Venezolana* **41**, 163–170.

Rodríguez-Larralde, A., Formica, G., Scapoli, C., Baretta, M., Mamolini, E. & Barrai, I. (1993) Microevolution in Perugia: Isonymy, 1890–1990. *Ann Hum Biol* **20**, 261–274.

Rohlf, F. J. (2000) NTSYS-pc: Numerical Taxonomy and Multivariate Analysis System, version 2.11s. *Exeter Software, Setauket*, New York.

Rosenberg, M. S. & Anderson, C. D. (2011) PASSaGE: Pattern analysis, spatial statistics and geographic exegesis. Version 2. *Meth Ecol Evol* **2**, 229–232.

Scapoli, C., Mammolini, E., Carrieri, A., Rodriguez Larralde, A. & Barrai, I. (2007) Surnames in Western Europe: A comparison of the subcontinental populations through isonymy. *Theor Popul Biol* **71**, 37–48.

Sokal, R. R., Harding, R. M., Lasker, G. W. & Mascie-Taylor, C. G. N. (1992) A spatial analysis of 100 surnames in England and Wales. *Ann Hum Biol* **19**, 445–476.

Sørensen, T. (1948) A method for establishing groups of equal amplitude in plant sociology based on similarity of species content. *K Dan Vidensk Selsk Bio Skr* **5**, 1–34.

Steinman, G. (2006) Can the chance of having twins be modified by diet? *Lancet* **367**, 1513–1519.

Stueber-Odebrecht, N., Chautard-Freire-Maia, E. A., Primo-Parmo, S. L. & Carrenho, J. M. X. (1985) Studies on the CHE1 locus of serum cholinesterase and surnames in a sample from Santa Catarina (Southern Brazil). *Rev Brasil Genet* **8**, 535–543.

Tagliani-Ribeiro, A., Oliveira, M., Sassi, A. K., Rodrigues, M. R., Zagonel-Oliveira, M., Steinman, G., Matte, U., Fagundes, N. J. R. & Schuler-Faccini, L. (2011) Twin town in south Brazil: A nazi's experiment or a genetic founder effect? *PLoS One* **6**, e20328.

Tarskaia, L., El'chinova, G. I., Scapoli, C., Mamolini, E., Carrieri, A., Rodriguez-Larralde, A. & Barrai, I. (2008) Surnames in Siberia: A study of the population of Yakutia through isonymy. *Am J Phys Anthropol* **138**, 190–198.

Zimmerman, G., Goetz, H. & Mielke, P. (1985) Use of an improved statistical method for group comparisons to study effects of prairie fire. *Ecology* **66**, 606–611.

Received: 5 April 2012
Accepted: 16 September 2012