AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 126:214–228 (2005)

New Method for Surname Studies of Ancient Patrilineal Population Structures, and Possible Application to Improvement of Y-Chromosome Sampling

Franz Manni,* Bruno Toupance, Audrey Sabbagh, and Evelyne Heyer

Equipe de Génétique des Populations, Département Hommes, Natures, Sociétés (UMR5145 CNRS), Unité d'Eco-Anthropologie, Musée de l'Homme MNHN, Paris 75016, France

KEY WORDS Kohonen maps; demography; population genetics; sampling design; census; migrations

ABSTRACT Several studies showed that surnames are good markers to infer patrilineal genetic structures of populations, both on regional and microregional scales. As a case study, the spatial patterns of the 9,929 most common surnames of the Netherlands were analyzed by a clustering method called self-organizing maps (SOMs). The resulting clusters grouped surnames with a similar geographic distribution and origin. The analysis was shown to be in agreement with already known features of Dutch surnames, such as 1) the geographic distribution of some well-known locative suffixes, 2) historical census data, 3) the distribution of foreign surnames, and 4) polyphyletic surnames. Thus, these results validate the SOM clustering of surnames, and allow for the generalization of the technique. This method can be applied as a new strategy for a better Y-chromosome sampling design in retrospective population genetics studies, since the idenfication of surnames with a defined geographic origin enables the selection of the living descendants of those families settled, centuries ago, in a given area. In other words, it becomes possible to virtually sample the population as it was when surnames started to be in use. We show that, in a given location, the descendants of those individuals who inhabited the area at the time of origin of surnames can be as low as ~20%. This finding suggests 1) the major role played by recent migrations that are likely to have distorted or even defaced ancient genetic patterns, and 2) that standard-designed samplings can hardly portray a reliable picture of the ancient Y-chromosome variability of European populations. Am J Phys Anthropol 126:214–228, 2005.

In societies that use a patrilineal transmission of surnames, family names simulate neutral alleles of a gene transmitted only through the Y-chromosome (Yasuda and Morton, 1967; Yasuda and Furusho, 1971; Yasuda et al., 1974; Zei et al., 1983, 1984), and therefore satisfy the expectations of the neutral theory of molecular evolution (Cavalli-Sforza and Bodmer, 1971; Crow, 1980), which is entirely described by random genetic drift, mutation, and migration (Kimura, 1983). This property of surnames, together with their prompt availability, made them useful for the study of population structure since 1965, when Crow and Mange (1965) published the quantitative relation existing between isonymy and inbreeding (e.g., see Lasker, 1968).[1]

Nowadays, in many countries, millions of surnames of telephone users, often available on CD-Roms or online, can be efficiently analyzed in a short time. As an example, the surname structure of Switzerland (Barrai et al., 1996), Germany (Barrai et al., 1997), Italy (Barrai et al., 1999), Austria (Barrai et al., 2002), France (Mourrieras et al., 1995), and the Netherlands (Manni, 2001a) were studied, taking into account, in total, more than 20 million surnames. It is the largest sample size ever used in human population studies. Investigated at different geographic scales, surname-inferred genetic structures were sometimes regarded with a certain suspicion because they are simulated markers for a single locus. A good example of the doubts about surname studies was expressed by Rogers (1991, p. 663): " The method . . . requires an assumption that

[1]Tables 1, 2, and 3 and Figures 2 and 4 refer to the Self-Organizing-Maps analysis of Figure 3. In the text we refer to the small maps of Figure 3 with X and Y coordinates according to the following notation: "1, 2 is equal to (X = 1; Y = 2). GSSGP means groups of surnames with a similar geographic distribution. GSSGP is a synonym of neuron/cell of the SOM clustering grid of Figure 3.

Grant sponsor: EGIDE; Grant sponsor: CNRS.

*Correspondence to: Dr. Franz Manni, Equipe de Génétique des Populations, Département Hommes, Natures, Sociétes, Musée de l'Homme MNHN, 17 Place du Trocadéro, 75016 Paris, France. E-mail: manni@mnhn.fr

Received 26 January 2003; accepted 19 September 2003.

DOI 10.1002/ajpa.10429
Published online 26 July 2004 in Wiley InterScience (www.interscience.wiley.com).

has not been appreciated: it is necessary to assume that all males in some ancestral generation, the founding stock, had unique surnames. Because this assumption is seldom justified in real populations, the applicability of the isonymy method is extremely limited. Even worse, the estimates it provides refer to an unspecified founding stock, and this implies that these estimates are devoid of information".

Recently, the isonymy method was applied to a genealogical database (Gagnon, 2001; Gagnon and Toupance, 2002), and consanguinity was estimated both from surnames and from true genealogies. Results indicate that random isonymy, estimated from family names, is not devoid of information; on the contrary, it fits well with consanguinity estimates obtained from genealogical records. These findings point to the validity and usefulness of quantifying migrations on the basis of surname data, and show that migration flow and consanguinity are inversely and closely linked (Darlu and Ruffié, 1992). It could be argued that the isonymy method does not take into account the cumulative effects of inbreeding, because founding groups are often from a small region, or even composed of relatives (Jobling, 2001). On the other hand, this issue would result in an underestimation of actual levels of consanguinity, implying that the isonymy method is conservative and parsimonious.

It is known that surnames provide no information for periods previous to the late Middle Ages (in early cases), when they originated and spread in most European countries. Nevertheless, their limited temporal depth represents a considerable advantage, because demographic phenomena of populations in the last six centuries (as migrations, drift, or isolation) can be identified and temporally distinguished from previous ones. The comparison of surname variability with genetic and linguistic data revealed, on small geographic scales, some dramatic changes in population structures after the origin of patronymic markers (Manni and Barrai, 2000, 2001; Manni, 2001b). In this sense, surnames are "myopic" markers, and their specific time depth can be of invaluable help in the identification of areas where recent migrations, consequent to the rural exodus of the last century, are likely to have significantly modified genetic structures of human populations. Their use can help minimize sampling errors by telling where and when a preexisting ancient genetic pattern is likely to have been distorted or even defaced (Manni, 2001b).

Analysis of the geographic variability of single surnames, even when applied to small groups of "interesting" family names (an excellent example can be found in Sokal et al., 1992), was never undertaken for thousands of them, mainly because of the overwhelming computing time required by classical multivariate techniques such as multidimensional scaling (MDS) (Seber, 1984; Torgerson, 1958) or principal component analysis (PCA) (Gabriel, 1968). Typically, in order to enable the visualization of overall surname variability, matrices of pairwise surname-derived distances between all the different localities were computed (Chen and Cavalli-Sforza, 1983; Lasker et al., 1977). These distances are based on the whole corpus of patronymic data and not on single surnames. If such an approach allows for classical multivariate analyses, it makes also very difficult the study of the geographic pattern of single family names: in other words, migrations of single or small groups of families can no longer be inferred, since the variability of all surnames in a certain area is turned into a single distance measure.

We present a new method to analyze, by means of neural networks, the variability of a large number of surnames and to identify their exact geographic origin.

A regular decreasing frequency gradient is usually observed around the area of origin of a given surname. The identification of such gradients is the criterion we followed in this study to attribute a geographic origin to family names (Fig. 3). The establishment of gradients implies that the individuals sharing a surname had time to migrate around the area of its origin, generation after generation, according to an isolation-by-distance model that has been confirmed for all European countries studied so far (Barrai et al., 1996, 1997, 1999, 2002; Mourrieras et al., 1995; Manni, 2001a).

We tested this new method, as a case study, on Dutch surnames, since they originated quite recently, if compared with other European countries (in 1811 in northern provinces, and in 1796 in southern provinces), and since the chances to observe a regular decreasing frequency gradient around the geographic origin of a surname are lower if compared to countries where surnames originated in earlier times. In the Netherlands, gradients had less time to be established; therefore, this country represents a good location to test the robustness of the above approach.

This paper also focuses on a new strategy for a better Y-chromosome sampling design in retrospective population genetics studies. Surnames with a defined geographic origin can be helpful in selecting only the descendants of families settled centuries ago in the area. In other words, it becomes possible to virtually sample the population as it was when surnames originated. In this way, the confounding effects of migrations of the last several centuries are minimized, enabling us to infer a more reliable picture of ancient and remote peopling phases in Europe and of a retrospective census of the population at the time of surname introduction.

MATERIALS AND METHODS

Surname selection

Analyzed surnames were chosen among the most frequent ones, since we were interested in the identification of their geographic origin, one by one, and we needed a frequency gradient to identify it. For

Fig. 1. Map of Netherlands, where administrative division in 12 provinces is shown. Province of Flevoland (asterisk) was excluded from analysis because of its recent creation (1963), out of Overijssel and Gelderland, on new lands. Dots show position of 226 localities where surnames were sampled. A detailed description of sample locations can be found in Barrai et al. (2002) according to Manni (2001a).

this reason, we chose a cutoff value corresponding to an absolute frequency of at least 40 individuals (telephone subscribers) sharing the same surname. Following this criterion, 9,929 surnames (corresponding to 1,642,354 telephone subscribers) were selected on a database derived from the official telephone book of the Netherlands (Topware©® 1996). Since the official listings on CD-Roms ignore surname prefixes (such as De, Van, or Van De), which very often precede Dutch surnames, we were forced to ignore them as well. While the absolute frequency of each surname was computed on a new database, the sampling grid of 226 localities (Fig. 1) is the same that was published by Barrai et al. (2002), according to Manni (2001a).

Self-organizing maps (SOMs)

Recent developments of cluster analysis through neural networks made available a specific application known as self-organizing maps (SOMs). A detailed description of this method is provided here since, to our knowledge, it has never been applied to population genetics. This category of tools is intended for a nonanalytic exploration of large corpora of vectors (inputs) that are mapped on a cell grid (map) according to their similarity. As an example, vectors can be the frequencies of different alleles or, as in this case, surnames. In this process, 1) identical vectors will be mapped at the same position of the map, 2) slightly different vectors will be close to each others, and 3) very different vectors will be mapped far from each other. The visual aspect of data representation obtained by SOMs is similar to a classical multidimensional scaling (MDS) (Torgerson, 1958; Seber, 1984) or principal component analysis (PCA) (Gabriel, 1968) plot, but in contrast with these techniques, SOMs can handle up to several thousand inputs on a standard computer. The difference between MDS (or PCA) and SOMs is that the former method provides a better description of the distances between items, while the latter gives a more accurate representation of their topology, i.e., their relative positions (Kaski, 1997). For this reason, SOMs are preferable to MDS or PCA when all the different inputs (vectors) slightly differ from one another, as is the case with family names. SOMs are based on "competitive learning," an adaptive process in which the cells (also called neurons) in a neural network gradually become sensitive to different input categories (Kohonen, 1982, 1984). A sort of a labor division emerges in the network, when different cells specialize to represent different types of inputs.

If there exists an ordering between the cells, i.e., if the cells are located on a discrete map, the competitive learning can be generalized; if not, then not only the winning neuron but also its *neighbors* on the map are allowed to learn. Neighboring cells will gradually specialize to represent similar inputs, and the representation of input data will become ordered. The degree of specialization of the map is enhanced by the competition among cells: when an input arrives, the neuron that is best able to represent it wins the competition and is allowed to learn even better. This is the essence of the SOM algorithm (Kaski, 1997).

Each cell of the map, indexed with i, represents a reference vector m_i whose components correspond to synaptic weights. In the exploration of data (inputs), the cell (indexed with c) whose reference vector is the nearest to the input vector x becomes the winner of the competition between all the different reference vectors:

$$c = c(x) = \arg\min_i \{\|x - m_i\|^2\}. \quad (1)$$

Usually the Euclidean metric is used as a measure of $\|x - m_i\|^2$.

The winning unit c and its neighbors adapt to represent the input even better by modifying their reference vectors towards the current input. The amount the units learn is governed by a neighborhood function, h, which decreases with the distance from the learning unit on the map and changes trough time. If the locations of cells i and j on the map-grid are denoted by the two-dimensional vectors r_i and r_j, respectively, then

$$h_{ij}(t) = h(\|r_i - r_j\|; t), \quad (2)$$

TABLE 1. Distribution of 9,929 surnames according to 15 × 15 Kohonen map

15	59	28	51	43	48	65	48	47	57	74	60	46	53	47	65
14	30	28	45	23	32	25	29	34	33	51	46	53	23	30	44
13	43	27	58	35	33	34	36	45	37	47	63	38	60	35	89
12	54	31	72	56	35	33	32	32	46	66	44	33	56	35	100
11	50	35	54	42	37	32	29	40	38	44	52	45	44	24	57
10	50	26	44	24	21	37	42	32	55	36	28	44	43	38	39
9	49	37	38	29	28	23	42	43	27	25	36	35	50	38	91
8	76	27	37	45	28	45	30	32	39	33	44	49	42	42	78
7	29	31	45	42	39	41	36	34	26	32	28	34	46	37	44
6	64	30	26	48	42	61	38	53	53	63	23	62	30	38	73
5	64	48	31	42	41	41	26	28	54	44	22	38	46	34	85
4	64	33	40	42	34	39	34	40	55	33	44	47	28	37	54
3	82	34	45	55	51	56	38	42	56	40	40	41	40	34	61
2	22	22	28	45	43	55	41	44	34	39	49	55	44	37	32
1	128	77	34	41	92	60	62	66	70	45	60	66	67	36	62
↑ Y/X→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

where t denotes time. During the learning process, at time t, the reference vectors are changed iteratively according to the following adaptation rule:

$$m_i(t + 1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)]. \quad (3)$$

where x(t) is the input at time t, and c = c (x(t)) is the index of the winning unit. In practice, the neighborhood function 1) is chosen to be wide at the beginning of the learning process to guarantee global ordering of the map, and 2) decreases (both its width and height) during learning.

The learning process, consisting of the winner selection by Equation (1) and the adaptation of synaptic weights by Equation (3), can be modeled with a neural network structure where the cells are coupled by inhibitory connections (Kaski and Kohonen, 1994; Kohonen, 1993). By virtue of its learning algorithm, the SOM forms a nonlinear regression of the ordered set of reference vectors into the input surface. The reference vectors form a two-dimensional "elastic network" that follows the distribution of data. It must be specified that, during the learning phase, the order of data-entry does not effect the final configuration of the map. Data vectors are randomly selected to initialize the map thousands of times.

SOMs can be rectangular or squared, and can consist of a number of cells (neurons) defined by the user (5 × 5, 6 × 6, 7 × 7, etc.). Data items will then be mapped on one of the neurons of the map according to their similarity to the reference vectors associated with each cell. In the case of very large datasets (as in the present study), several items can be mapped to the same neuron, giving rise to clusters. It may happen that some neurons may remain empty, meaning that there are no data vectors that correspond to them. The size of the map determines the number of different clusters that can be obtained; therefore, larger maps will classify items more accurately than smaller ones. If the user wants to obtain a limited number of clusters, each consisting of many data vectors, a small map-size should be specified. In this paper, we wanted to group surnames with a similar geographic distribution and origin, and so we chose a map size (15 × 15) that seems a good compromise between accuracy and the number of items mapped to each neuron (Table 1). A larger map, say 20 × 20 (not shown), would have provided a more detailed clustering of surnames, with patronymics having a more similar geographic origin in each group, but would also have been driven to several neurons associated with a single or very few surnames (when some surnames have a very peculiar spatial pattern), thus contradicting the purpose of identifying clusters.

Vectors, clustering, and software used

Each of the 9,929 surnames analyzed by a 15 × 15 SOM was entered in the analysis as a binary vector of 226 dimensions (for example, Heeringa: $0_{(city\ 1)}$; $1_{(city\ 2)}$; $0_{(city\ 3)}$; . . .; $0_{(city\ 226)}$), corresponding to its presence (1) or absence (0) in the 226 sampled localities (see Fig. 1 for distribution of localities). Input vectors were mapped to 225 (= 15 × 15) reference vectors (cells) (Tables 1 and 2) or to a group of surnames with a similar geographical pattern (GSSGP). It is important to note that the information on single surnames is still available, since the output file provides the list of family names clustered in each unit (cell) of the map.

As mentioned above, the computed map consists of a 15 × 15 lattice where each cell (or GSSGP) can be referred to by its row and column indexes n_x,n_y. For example, the cell in the left top corner of the map (see Tables 1 and 2, and Figs. 2, 3) will be reported as 1,15.

We used a recompiled version of the program koh.c to create Kohonen maps from a set of vectors. This software was written by Peter Kleiweg (State University of Groningen, Netherlands) and is freely available at: http://odur.let.rug.nl/~kleiweg/indexs.html.

Identification of geographic origin of surnames

Once the SOM is obtained, the subsequent step consists of the computation of the relative frequency of the 225 GSSGPs (in a map of 15 × 15 cells) in each of the 226 sampled localities. Their frequency is

TABLE 2. Distribution of 1,642,354 telephone subscribers according to clustering of 9,929 surnames on 15 × 15 Kohonen map (see Table 1)

15	266,761	50,074	42,280	34,045	24,476	23,426	12,028	8,505	7,561	6,019	3,616	3,416	5,481	6,241	23,783
14	48,907	33,138	25,006	14,016	13,729	7,956	6,183	5,288	3,243	3,642	2,439	3,187	1,549	2,846	5,557
13	38,971	23,193	31,583	17,237	12,320	10,280	8,266	7,644	4,400	3,103	3,434	2,032	3,401	2,507	8,292
12	32,123	15,477	19,018	26,455	11,464	8,300	5,646	4,972	4,564	4,300	2,465	1,694	3,001	4,428	6,014
11	16,905	11,906	11,442	14,526	10,924	7,489	5,586	5,289	3,418	2,976	3,009	2,541	2,979	1,245	2,911
10	13,567	6,272	7,463	7,957	5,656	8,686	7,801	4,198	5,470	3,109	3,136	2,994	2,405	1,927	1,895
9	10,894	8,305	9,672	6,668	6,132	4,421	6,047	5,574	2,732	3,121	3,709	2,993	3,038	1,971	4,629
8	15,117	5,403	6,663	9,280	5,769	6,273	3,120	3,627	5,048	4,034	5,485	4,854	2,880	2,487	4,542
7	4,170	4,071	6,264	6,650	5,584	3,794	3,079	3,343	3,006	3,323	2,387	4,846	3,001	2,206	2,248
6	7,310	3,724	5,890	5,829	4,476	5,199	2,973	3,460	3,285	4,204	1,583	3,561	9,486	2,265	3,707
5	6,142	4,317	4,614	3,642	2,937	2,946	1,464	2,686	2,796	2,477	1,653	3,091	3,803	2,180	5,096
4	4,382	1,954	4,556	2,616	2,079	2,258	1,825	2,078	3,073	2,067	2,834	3,159	2,466	3,568	4,716
3	4,309	1,788	2,967	2,840	2,800	3,157	2,241	2,389	3,282	2,384	2,212	2,309	3,527	3,908	9,061
2	1,395	1,177	2,219	2,313	2,366	3,370	2,381	2,751	2,126	2,223	2,773	2,903	3,728	4,605	5,404
1	26,228	4,819	2,519	2,158	6123	4117	4,299	4,483	3,951	2,627	3,140	4,154	5,854	5,157	14,044
↑ Y/X→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

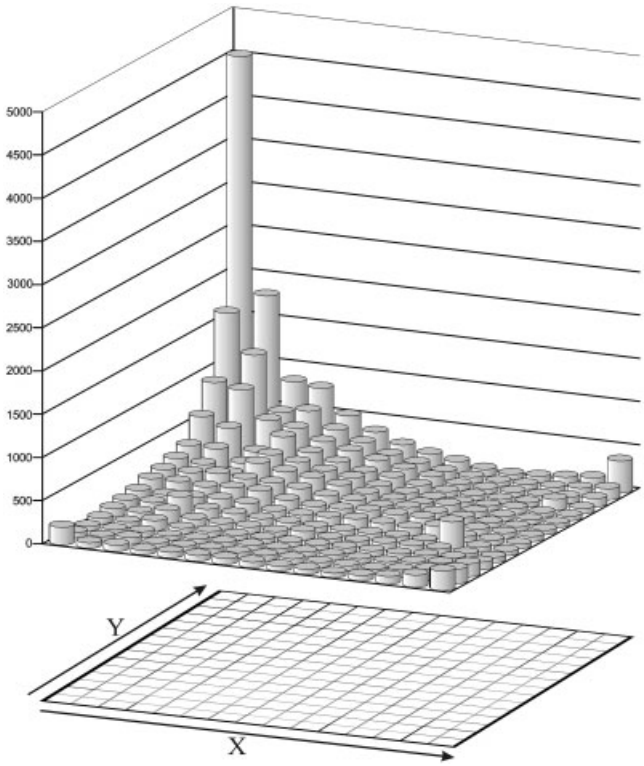

Fig. 2. Average of individuals sharing surnames clustered in each cell of 15 × 15 Kohonen map (below): 9,929 surnames represent 1,642,354 individuals in total.

computed, after the clustering, from the sum of absolute frequencies of all surnames clustered in the same cell. Subsequently, frequency values of all 225 GSSGPs were corrected by the size of the corresponding towns and cities. This correction is justified by the fact that, for example, the presence of 5 individuals sharing a given surname (in a village of 200 inhabitants) is more informative about the geographic origin of such a surname than the presence of 5 persons sharing this surname in a city of 100,000 inhabitants. Thereafter, the geographic distribution of each of the 225 GSSGPs was plotted on a geographic map, shown in Figure 3, thus making visible the areas of origin for the whole corpus of 9,929 surnames. By visual inspection of these maps, we assigned a geographic origin to the GSSGPs (Table 3), and we computed, from Table 1, the sum of the surnames that originated in each of the 11 Dutch provinces (Table 4).[2]

RESULTS

Clustering statistics

The number of surnames clustered in each cell of the Kohonen map (Table 1) ranges from 21 (5,10) to 128 (1,1), with a mean of 44.13 and a standard deviation (SD) of 15.52. Table 2 shows the total number of telephone subscribers corresponding to each cluster. This sum ranges from 266,761 (1,15) to 1,245 (14,11), with a mean of 7,299 and SD of 19,042. The average number of individuals sharing a given surname can be obtained by dividing the number of individuals corresponding to each cluster (Table 2) by the number of corresponding surnames (Table 1), as shown in Figure 2.

Polyphyletism

The average number of individuals sharing a given surname (Fig. 2) can be very different among the 225 clusters, and this difference is hardly explained by the number of corresponding surnames (Table 1), which may indicate that some clusters are associated with polyphyletic surnames. Further, all the most frequent family names of the Netherlands (Table 6) are clustered together in the cell 1,15 (Figs. 3, 4). Their geographic and linguistic analyses indicate, respectively, 1) the absence of any spatial pattern of distribution (Fig. 3) (see also Validation of the Method, below), and 2) a common-sense meaning that corresponds to body characteristics, professions, geographical features, etc. (Table 6). These properties also apply to the neighborhood of 1,15, i.e., the neurons 1,14; 2,14; and 2,15. A very conser-

[2]Officially, in the Netherlands, there are 12 provinces. Flevoland was excluded because of its recent creation (1963) out of Overijssel and Gelderland on new reclaimed lands.

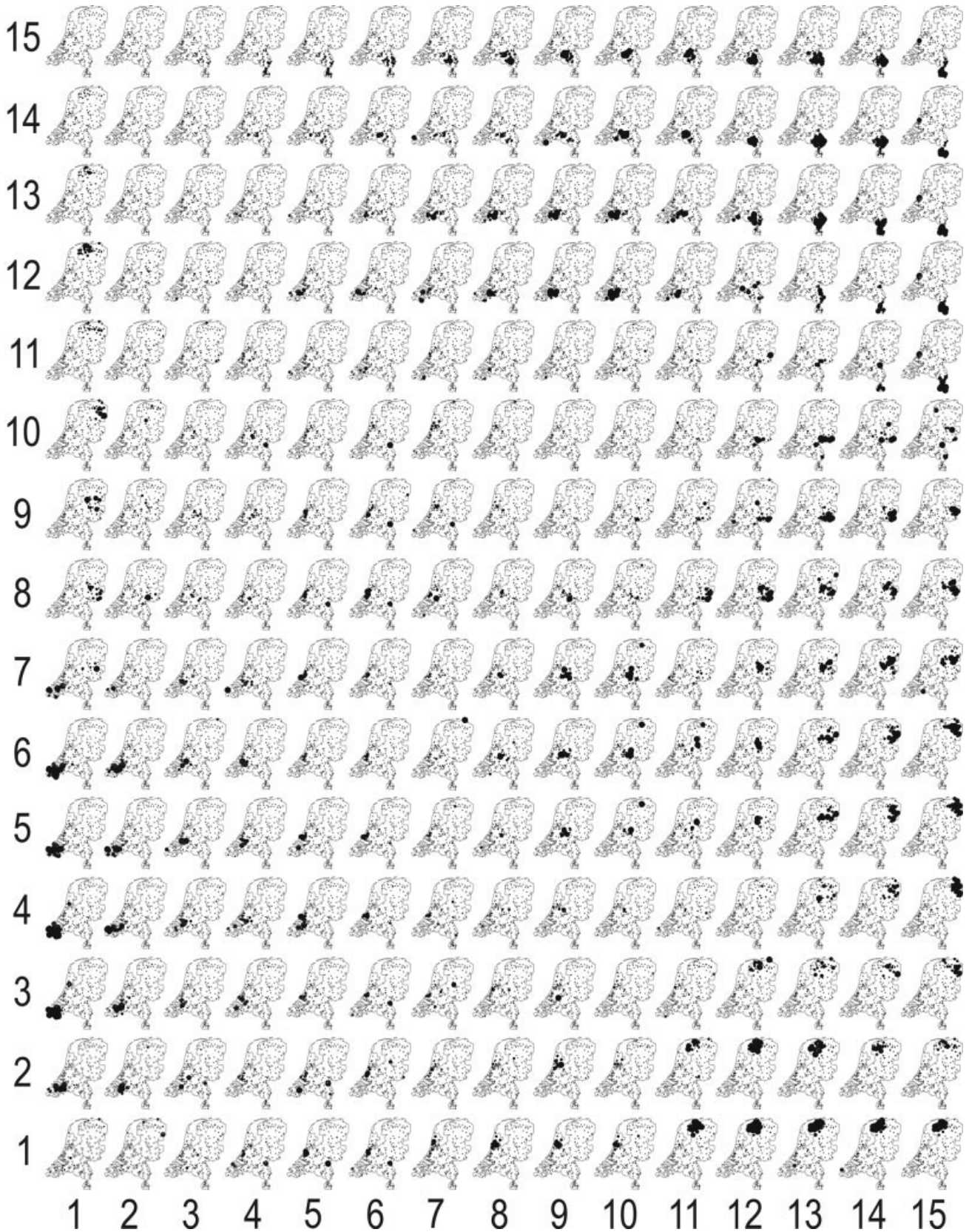

Fig. 3. Spatial patterns of surnames based on their frequency distribution, cluster by cluster, as obtained with a 15 × 15 SOM analysis. Frequencies are plotted on map of Netherlands to show their geographic origin. Each cell (cluster) corresponds to a group of surnames with a similar geographic pattern (GSSGP). Frequencies are reported after a correction by the population size of 226 localities investigated (9,929 surnames in total). Maps are ordered along X and Y axes visible in Figure 2.

TABLE 3. Geographic origin of surnames in each of 225 clusters of Kohonen map[1]

15	?	?	?	LB/NB	LB/NB	LB/NB	LB/NB	NB	NB	NB	NB	NB	LB/NB	LB	LB*
	0;1;0;0;0	0;0;1;0;0	0;0;0;2;0	0;0;0;0;0	0;0;0;1;0	0;0;0;3;0	0;0;0;2;0	0;0;0;3;0	0;0;0;2;1	0;0;0;3;1	0;0;0;3;0	0;0;0;3;0	0;0;0;3;0	0;0;0;3;0	0;0;0;2;0
14	?	?	?	NB	NB	NB	NB/ZL	NB	NB/ZL	NB	NB	LB/NB	LB/NB	LB	LB*
	0;0;0;0;0	0;0;00;0	0;0;0;0;0	0;0;0;0;0	0;0;0;2;0	0;0;0;1;0	0;0;0;2;0	0;0;0;0;0	0;0;0;3;0	0;0;0;4;0	0;0;0;1;0	0;0;0;4;0	0;0;0;2;0	0;0;0;1;1	0;0;0;5;0
13	FR	?	?	?	?	NB	NB	NB	NB	NB	NB/ZL	LB/NB	LB	LB	LB*
	0;0;0;1;0	0;0;0;0;0	0;0;0;0;0	0;0;0;0;1	0;0;0;0;1	0;0;0;0;0	0;0;0;4;0	0;0;0;2;0	0;0;0;4;0	0;0;0;40;1	0;0;0;5;0	0;0;0;2;0	0;0;0;3;0	0;0;0;3;0	0;0;0;2;3
12	FR	?	?	?	ZL	ZL	ZL	NB/ZL	NB	NB	NB	LB/GL/NB	LB	LB	LB*
	3;0;0;0;0	0;0;0;0;0	0;0;2;0;1	0;0;0;0;0	0;0;0;0;0	0;0;0;0;0	0;0;0;1;0	0;0;0;4;0	0;0;0;6;1	0;0;0;3;0	0;0;0;0;0	0;0;0;1;0	0;0;0;1;0	0;0;1;2;1	0;0;0;3;8
11	FR/GR	?	?	?	?	ZL	?	?	?	?	?	GL/OV	GL	LB/OV	LB*
	0;1;1;0;0	0;00;0;0	0;0;1;0;0	0;0;0;0;0	0;0;1;0;1	0;0;0;0;1	0;0;0;0;0	0;0;0;1;0	0;0;0;0;2	0;0;1;1;1	0;0;0;1;0	0;0;1;2;1	0;0;1;2;0	0;0;1;0;0	1;0;0;2;0
10	DR/GR	?	?	GL/UT	NH/ZH	LB/ZH	NH	?	?	?	?	GL	GL*	OV/GL	?
	2;0;0;0;0	0;0;1;0;0	0;0;1;0;0	0;0;0;0;0	0;0;0;0;0	0;0;0;0;2	0;0;0;0;2	0;0;0;0;1	0;0;0;0;0	0;0;0;0;0	0;0;0;0;1	0;0;1;1;0	0;0;9;0;0	0;0;0;0;1	0;0;2;2;0
9	GL/DR	?	GL/UT	GL/UT	ZH	LB/NH/ZH	NH/LB	NH	?	GL	GL/OV	GL/ZH	GL	OV/GL	OV
	0;0;4;0;0	0;0;1;0;1	0;0;0;0;0	0;0;0;0;0	0;0;0;0;0	0;0;0;0;0	1;0;0;0;2	0;1;0;0;1	0;0;0;1;1	0;0;3;0;0	1;0;1;0;0	0;0;1;0;1	1;0;14;0;0	0;0;7;0;0	0;0;15;1;0
8	GL/OV	GL	GL/UT	?	LB/ZH	LB/ZH	NH/ZH	UT	GL/UT	GL	GL/OV	GL/OV	DR/GL/OV	OV	OV
	0;0;13;0;0	0;0;3;0;0	0;0;0;0;0	0;0;0;0;0	0;0;0;0;1	0;0;0;0;0	0;0;0;1;1	0;0;0;0;0	0;0;0;0;0	0;0;3;0;0	0;0;17;0;0	0;0;16;0;0	0;0;12;0;0	0;0;13;0;1	0;0;10;0;0
7	OV/ZL	ZL	ZH	ZH/ZL	ZH	ZH	UT/ZH	UT	GL/UT	GL/UT	?	GL/OV	OV	GL/OV	OV/DR*
	0;0;1;0;0	0;0;0;0;0	0;0;0;1;1	0;0;0;1;1	0;0;0;0;1	0;0;1;1;0	0;0;0;1;0	0;0;0;1;0	0;0;0;0;0	0;0;2;0;0	0;0;4;0;0	0;0;11;0;0	0;0;3;0;0	1;0;5;0;0	0;0;6;1;0
6	ZL	ZL	NB/ZH	NB/Z H	ZH	ZH	?	GL/UT	UT	GL*	GL/OV*	OV	OV	DR/GR	DR/GR
	0;0;1;0;2	0;0;0;0;1	0;0;0;0;1	0;0;0;0;0	0;0;0;0;1	0;0;0;2;4	0;0;0;2;0	0;0;1;1;2	0;0;0;0;0	0;0;0;0;0	0;0;0;0;0	0;0;4;0;0	0;0;1;0;0	0;0;2;0;0	1;1;0;0;1
5	ZL	ZL	ZH/ZL	ZH	ZH/ZL	ZH	?	?	UT	GL/GR/UT	GL/OV/UT	OV	DR/OV	DR/GR	GR
	0;0;0;0;4	0;0;0;0;0	0;0;0;0;0	0;0;0;1;0	0;0;1;0;0	0;0;0;0;0	0;0;20;1	0;0;1;2;1	0;0;0;0;2	0;0;0;0;0	0;0;0;1;0	1;0;4;0;0	0;0;4;0;0	1;0;5;0;0	4;1;0;0;0
4	ZL	ZL	ZH/ZL	ZH/ZL	ZH/ZL	ZH	?	NH/UT	NH/UT	?	?	?	DR/GR/OV	DR/GR	DR/GR
	0;0;0;0;2	0;0;0;1;0	0;0;0;1;0	0;0;0;1;0	0;0;1;0;0	0;0;0;0;1	0;0;0;1;0	0;0;0;0;1	0;0;0;0;2	0;0;0;0;0	0;0;2;0;0	1;0;0;0;0	0;0;1;0;0	0;0;1;0;0	3;1;0;0;0
3	ZL*	ZL/NB	ZH	ZH/ZL	ZH	ZH/LB	?	?	NH/ZH	?	?	FR/GR	FR/GR	DR/GR	DR/FR/GR
	0;0;0;0;2	0;0;0;1;0	0;0;0;1;0	0;0;0;1;1	0;0;0;0;0	0;0;0;1;1	0;0;2;0;1	0;0;0;1;0	0;0;0;1;1	0;0;0;0;1	1;0;0;0;0	3;2;0;0;0	1;6;0;0;0	1;0;1;10	2;0;0;0;0
2	ZL	ZL/NB	LB/NB/ZH	ZH	LB/NB/ZH	NH/ZH*	NH/ZH	NH	NH	NH	FR	FR	FR	FR	FR/GR
	0;0;0;1;0	0;0;1;1;0	0;0;0;1;1	0;0;0;0;0	0;0;0;1;1	0;0;1;2;4	0;0;0;0;1	0;0;0;0;1	0;0;0;1;0	1;1;0;0;1	3;0;0;1;0	2;0;0;0;1	1;1;0;0;0	2;0;0;0;0	2;0;0;0;0
1	NB/ZL*	?	?	ZH/ZL*	LB/ZH	LB/NH/ZH	NH	NH	NH	NH	FR	FR	FR	FR	FR
	0;0;0;0;12	0;0;0;1;4	0;0;0;0;1	0;0;1;0;1	0;0;0;3;1	0;0;0;0;2	0;0;1;1;0	0;0;1;0;1	0;1;2;0;4	0;0;0;0;1	3;0;2;0;0	5;0;0;0;1	4;0;0;0;0	0;0;0;0;0	7;0;0;0;1
↑ Y/X→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

[1] Results are based on visual inspection of spatial patterns shown in Figure 3. Province are abbreviated as follows: DR, Drenthe; FR, Friesland; GL, Gelderland; GR, Groningen; LB, Limburg; NB = Noord Brabant; NH = Noord Holland; OV = Overijssel; UT = Uthrecht; ZH = Zuid Holland; ZL, Zeeland. Ambiguous geographic origin is reported as "?". Presence of asterisk indicates that an alternative origin may be found in pattern of spatial variation. Province of Flevoland was excluded from analysis owing to its recent creation. Geographic localization of provinces is provided in Figure 1. To illustrate some features of auto-organization process of map, we give, on second line of each cell, absolute frequencies of surnames ending respectively in -inga (locative of Friesland and Groningen); -sema (locative of Friesland and Groningen); -ink (locative of Drenthe, Gelderland, and Overijssel); and -mans (locative of Limburg). Last number corresponds to absolute number of surnames of French origin. For example, in cluster 9, 1, notation is "NH/0;1;2;0;4," where first line means that corresponding surnames originated in Noord Holland province; second line accounts for 1) presence of 0 surnames ending in -inga; 2) 1 surname ending in -sema; 3) 2 surnames ending in -ink; 4) 0 surnames ending in -mans; 5) 4 surnames of French origin.

TABLE 4. 1830 population census of Netherlands[1] and estimates of geographical origin of 7,436 Dutch surnames inferred by self-organizing map (Surnames, SOMs)

	1830 census	Surnames, SOM
Drenthe	21,328	273
Friesland	64,841	675
Gelderland	101,594	716
Groningen	50,161	345
Limburg	60,527	1,025
Noord Brabant	116,244	1,193
Noord Holland	117,971	610
Overijssel	58,344	699
Utrecht	41,814	399
Zeeland	42,824	864
Zuid Holland	141,410	1,035
Total	817,058	7,834

[1] Source: Centraal Bureau voor de Statistiek van Nederlaand (1904).

vative estimate of the number of Dutch polyphyletic surnames could be at least 145 (1.5% of the 9,929-surname database), which corresponds to 398,880 telephone users (24% of the whole sample). The clustering of all highly frequent surnames in the same area of the map reflects the division of labor in the SOM, where different cells specialize to represent different types of inputs, i.e., the auto-organization process of Kohonen maps.

Spatial patterns

Spatial patterns were obtained for the 9,929 most common surnames in the Netherlands (Fig. 3). Visual inspection of the frequency distribution of each of the 225 clusters (GSSGPs) indicates that 7,834 of them show a clear geographic origin. The remaining 2,095 surnames have either an ambiguous origin (for example, 7,3; 8,3; 15,10) or no origin at all (for example, 2,11; 3,11; 11,4) as visible in Figure 3. The superposition of a political map of the Netherlands (Fig. 1) over the maps of Figure 3 enables the identification of the province of origin for each GSSGP (Tables 3 and 4). Indeed, it is possible to distinguish between different areas of origin even within the same province. As an example, we can identify surnames originating in the southern (15,14) or eastern part (13,12) of Limburg (Fig. 4D,E).

Validation of the method

The Netherlands represent a case study of a "blind" automated approach to identify the geographic origin of surnames. In order to validate the method, we checked the reliability of the auto-organization process of the map in the description of some expected or well-known features of Dutch surnames.

Surname-variants. Surname variants, when originating by misspellings in official records, are expected to have a spatial pattern similar to surnames they originated from. We identified 107 spelling variants of surnames (1% of the total), and we show that the variants of a given surname cluster

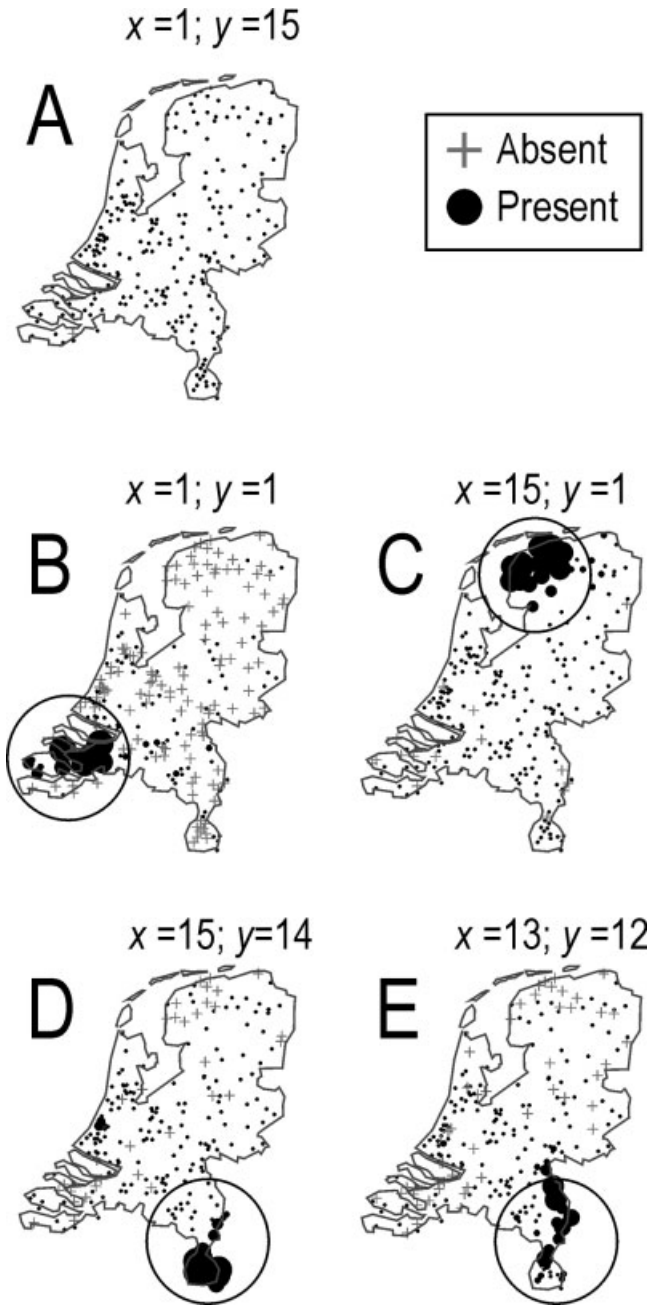

Fig. 4. Extension of some maps in Figure 3. **A:** Map of cell 1,15, corresponding to polyphyletic surnames listed in Table 6. **B, C:** Surnames respectively specific of provinces of Zeeland (cell 1,1) and Friesland (cell 15,1). **D, E:** Example of high-resolution geographic attribution of surnames, in southern (cell 15,14) and eastern (cell 13,12) parts, respectively, of Limburg province.

together in the same cell of the Kohonen map (data not presented). Note that, as mentioned, we considered only family names with a frequency of >40 individuals, thus restricting the analysis only to ancient transcription errors (since we assume that present frequency gradients of surnames needed several generations to establish themselves).

Surname suffixes. Some surname suffixes are well-known locatives of different provinces of the

TABLE 5. Correlations between population size inferred by SOMs at time of origin of surnames (Table 4) and different official population census[1]

	Census							
	1830	1840	1849	1859	1869	1879	1889	1899
Correlation	0.57 $P < 0.05$	0.56 $P < 0.05$	0.55 $P < 0.05$	0.51 $P < 0.05$	0.49 $P < 0.05$	0.46 $P < 0.05$	0.42 $P < 0.05$	0.40 $P < 0.05$

[1] Source: Centraal Bureau voor de Statistiek van Nederlaand (1904).

Netherlands. The idea was to check if the geographic origin of surnames, as given in Table 3 and Figure 3, was consistent with well-known locative suffixes. We chose surnames ending with -inga and -sekma (typical of the provinces of Friesland and Groningen); -ink (typical of the provinces of Drenthe, Gelderland, and Overijssel); and -mans (Limburg and Noord Brabant). Results show that 91%, 79%, 89%, and 69% of surnames ending in -inga, -sema, -ink, and -mans, respectively, fall in the cells of the Kohonen map corresponding to their traditionally known geographic origin (Table 3).

Surnames of French origin. Trade stimulated contacts and migrations between France and the Netherlands through Belgium (located at their southern border), where Vallons, a French-speaking community, constitutes one half of the population. For this reason, we expected the presence of a considerable number of surnames of French origin in the south of the Netherlands. One of us, a native French speaker (A.S.), identified (among the 9,929 surnames in the database) 120 surnames which show a clear French origin. The distribution of such surnames, according to their clustering in the map, indicates fairly well that French immigrants came from the south (Table 3), mainly from Zeeland (n = 26), Limburg (n = 18), and Holland (n = 32).

Historical census records. The identification of the geographic origin of surnames enables us to make inferences on the population size, province by province, at the time of surnames' origin (around the end of the 18th century in the Netherlands). According to the spatial patterns obtained (Fig. 3) and to the number of surnames for each cluster (Table 1), we computed the sum of surnames that originated in the 11 Dutch provinces (Table 4). By assuming that each surname corresponds to a family, we regard this estimate as proportional to a contemporary historical census of the population. As expected, the correlation (r) of our estimates with different historical census records of the Netherlands is shown to be significant (Tables 4 and 5). A progressive decrease of the correlation can be observed for more recent census records (Table 5), and the highest correlation is obtained with the 1830 census (0.57, $P < 0.05$).

DISCUSSION

Polyphyletic surnames

It is well-known that the most frequent surnames had polyphyletic origins. It can be easily shown that

TABLE 6. Surnames clustered in cell 1,15, corresponding to 59 most frequent in Netherlands

Surname	Translation
Bakker	Baker
Beek	Brook
Berg	Mountain
Bijl	Axe
Blom	Flower
Boer	Peasant
Bos	Forest
Bosch	Forest
Bosman	Man of the forest
Brink	Yard
Broek	Trousers
Brouwer	Brewer
Dam	Dam
Dekker	Thatcher
Dijk	Dike
Dijkstra	Of the dike (Frisian)
Graaf	Count, earl
Groot	Great, big
Haan	Cockerel, rooster
Hendriks	Son of Hendrik
Hock	Corner
Horst	Hurst
Huisman	Man of the house
Jager	Hunter
Jansen	Son of Jan (John)
Janssen	Son of Jan. (John)
Jong	Young
Jonge	The young one
Kamp	Camp, battle
Kok	Cook
Koning	King
Koster	Sexton, sacristan
Kramer	Hawker, pedlar
Kroon	Crown
Kuipers	Cooper
Laan	Tree-lined lane
Lange	The tall one
Leeuw	Lion
Leeuwen	Lions
Linden	Linden tree
Meer	Take
Meijer	Bailiff, landlord's steward
Mulder	Miller
Peters	(Family) of Peter
Post	Post, stake
Roos	Rose
Ruiter	Horseman
Smits	Smith; of the smith
Valk	Falcon
Veen	Peat
Velde	Field
Vermeulen	Variety of "to mill, to grind"
Visser	Fisherman
Vliet	Shoal, creek
Vonk	Spark
Vos	Fox
Vries	Freeze
Wal	Embankment, rampart, quay
Wijk	Settlement, quarter, neighborhood

the meaning of the first 100 most frequent surnames is very similar in European countries (Barrai, 2000, and personal communication). This datum reflects the similar natural and social environment of Europe, and implies the polyphyletic origins of corresponding patronymic markers.

In classical surnames analyses, i.e., studies based on surname distances (Chen and Cavalli-Sforza, 1983; Lasker et al., 1977), polyphyletic surnames decrease the value of pairwise distance measures between locations. This artifact arises because it has always been rather difficult to identify polyphyletic surnames. To avoid arbitrary exclusions of some family names, published studies were performed on the whole corpus of data by taking polyphyletic surnames as monophyletic. Such biased data can be considered a parsimonious estimate of the real degree of differentiation of studied populations, since surnames with polyphyletic origins give rise to an artificial kinship between different samples. If, until present times, surname analyses had to deal with this source of error, the problem can now be fixed by the use of SOMs (Kohonen, 1982, 1989, 1995) that make possible the identification of some clearly polyphyletic surnames whose signatures are 1) the absence of any geographic origin, 2) a high average number of people sharing them, and 3) a peculiar clustering in specific cells of the map (e.g., 1,15; 1,14; 2,15; 2,14; 3,14; 3,15, etc.) (Figs. 2, 3; Table 3).

We may mention here that their side position on the map (outliers) corresponds to an SOM property that reflects the auto-organization process of Kohonen maps, consisting of the specialization of different cells to represent different types of inputs. Figure 2 shows that the average number of individuals sharing a given surname gradually decreases in cells near the left top corner of the SOM. The explanation is that surnames, with a decreasing degree of polyphyletism, are clustered in such cells. Actually, even family names with polyphyletic origins can show a regional specificity that is justified by their linguistic context of origin (clusters 1,13; 1,12; 1,11; and 1,10), a datum that reflects the important dialect differentiation within the Netherlands (Nerbonne et al., 1996). This conclusion is reinforced by the finding of polyphyletic patronymics which are typical of the provinces of Friesland and Groningen, where the Frisian language is still in use or was recently spoken, respectively.

In conclusion, our results indicate that 1) the identification of polyphyletic surnames, or at least some of them, is possible thanks to the use of SOMs, and that 2) further classic patronymic analyses can be implemented by their exclusion. It will be of interest to compare isonymy levels, already computed in different European countries, with new estimates obtained after the withdrawal of SOM-identified polyphyletic surnames.

Historical census records

Different clues (suffixes, foreign surnames, and census data) suggest that SOMs are a convenient way to cluster surnames according to their geographic origin. In this way, it is possible to identify the geographic origin of a large part of the 9,929 studied surnames. Since there is no reason to expect a differential rate of surname extinction in the different provinces, our estimate must be proportional to the number of families established in the different provinces, as confirmed by its correlation with historical census records (Tables 4 and 5). The correlation between our SOM-based estimate for the number of surnames originating in each Dutch province and the historical population size is highest for the 1830 data, and thereafter decays. The highest correlation is the earliest because it was only 20–30 years earlier, in 1796 and 1811, that surnames originated. Indeed, the 1830 correlation (Table 4) increases from 0.57 to 0.66 ($P < 0.05$) by withdrawing Limburg province. This exclusion is justified because of its 1) very composite population and 2) particular geographic shape that favored migrations from and to Belgium and Germany. Moreover, linguistically, the dialect of Limburg is significantly different from the dialects of neighboring Dutch provinces (Nerbonne et al., 1996; Manni, 2001a).

We interpret the linear decay of the correlation (Table 5) between our estimates and more recent census records as the result of both regional migrations and a differential population growth in the different provinces. The observed linearity suggests that the change in population structuring occurs with a similar pace over time, resulting in the progressive defacement of ancient genetic structures.

Founding stock of the population two centuries ago

The geographic origin of the 225 GSSGPs can be traced by looking at their frequency distribution. The 15×15 adopted map provides a satisfactory clustering of surnames and portrays their geographic origins in detail. If the method is to be applied to a different database, a smaller or wider lattice may be more appropriate for the optimal identification of all the different GSSGPs (common geographic origin and spread).

The specialization of different areas of the map in the description of surnames with a specific geographic distribution reflects the good auto-organization of the SOM, since 7,834 of the 9,929 surnames (79%) show a spatial structure and a geographic origin (Tables 3 and 4). This confirms a similar result obtained by Sokal et al. (1992) on Welsh surnames, where three-quarters of them where found to be heterogeneously distributed over the investigated area. Therefore, the application of the SOM method enables a "blind" automatic analysis of surnames that provides evidence for a two-century-old founding stock of the population: it indicates which fami-

lies lived in different areas when patronymics were adopted. In this sense, this method answers one of the major criticisms concerning the use of surnames in population genetics (Rogers, 1991; see introduction), and can provide evidence for the Middle Ages founding stock of the population of those European countries where surnames are more ancient.

Application to improve the quality of Y-chromosome sampling

The sampling. The SOM analysis of surnames' frequency vectors of a country (or region) drives to a clustering that reflects the probable geographical origin of surnames. By the visual inspection of the resulting maps (Fig. 3), it is possible to identify which surnames originated in a given area and to list them. We will now provide a description of the practical way a high-quality sampling can be designed, once SOM analysis of surnames has been performed for that region or country.

Let's imagine that we are interested in the Y-chromosome variability of Friesland and Limburg, in order to study the past peopling phases of these two Dutch provinces. As a first step, we can list those surnames that, according to SOM analysis, unambiguously originated either in Friesland or in Limburg (Table 3 and Fig. 3). As a second step, the two lists of surnames can be sent to medical centers in Limburg and in Friesland, asking the personnel to sample only those individuals whose patronymics appear in the list specific for that area. As a result, only those individuals whose ancestors were settled in the same province where they presently live will be sampled. To correct any possible error in the SOM analysis, this "surname criterion" should be associated with the generally used "grandparents criterion," sampling only those donors whose grandparents lived in the area. As a conclusion, such two-step selection of Y-chromosome donors will be more stringent than currently adopted samplings, thus granting samples that are more representative of the ancient Y-chromosome genetic pool of these two areas (Friesland and Limburg in the example above).

We repeat that, since our SOM analysis was performed on the most frequent surnames of the country, the chances of finding Y-chromosome donors having one of the surnames of interest are maximized. As an example, according to our SOM analysis, we identified 476 surnames that unambiguously originated in Friesland (neurons 11,1; 12,1; 13,1; 14,1; 15,1; 11,2; 12,2; 13,2; and 14,2 in Table 3 and Fig. 3); since there are 46,358 individuals bearing such surnames (Table 2) individuals bearing them that still live in Friesland (20% of the Frisian population in our database), there is one individual who corresponds to our criterion out of every 5 donors presently living in Friesland, randomly selected from the data base. A similar ratio is expected on the ground (data not shown). In the same way, we found a percentage of 17% for the population of Limburg (whose specific surnames appear in cells 12,12; 12,13; 12,14; 12,15, 13,12; 13,13; 13,14; 13,15; 14,14; 14,15; 15,14; and 15,15 in Table 3 and Fig. 3).

To avoid any ambiguity, we stress that there is no particular reason to sample individuals having a given surname (of the list) instead of another one, since their Y-chromosome lineages are equally likely to be representative of the population (Friesland or Limburg) at the time of surname origin. At this time, each family name identified a single family; therefore, the higher or lower present frequency of surnames (besides the polyphyletic ones discussed above) only depends on the demographic history of corresponding families. Consequently, the multiple sampling of individuals sharing the same surname, according to its actual frequency in the population, will result in the overestimation of the past distribution of corresponding Y-chromosome lineages, since at the time of surname introduction, the frequency of all surnames was equal to one. Concerning the sample size, two possible strategies can be adopted: 1) either sampling a same sample size for each population (say, 30 individuals both in Friesland and in Limburg), or 2) sampling a number of individuals proportional to the population size at the time of surname origin. In the latter case, we can sample 30 individuals (30 different surnames) in Friesland and 60 (60 different surnames) in Limburg, according to SOM-inferred population sizes (Table 4) or historical census data, if available.

Because some confounding factors can diminish the power of the improved sampling, we discuss their possible role below.

Nonpaternity (adoption and surname-change). If this risk cannot be decreased with a SOM-designed sampling, it does not represent a real source of error, since it is highly probable that the vast majority of the Y-chromosome-transmitting fathers were from the same area of the surname-transmitting fathers and vice versa. We suggested in SOM-based surname analyses the way to sample those Y-chromosome lineages that are geographically more representative of ancient populations. Since we are not interested in forensic purposes, illegitimacy does not significantly counteract the advantage of using surnames in the first place. Obviously, nonpaternity can become a more serious limitation with surnames that are of very ancient origin, as with Han Chinese ones (patrilineally transmitted for ~2,500 years), since episodes of nonpaternity linearly increase with time. Nevertheless, it was shown that identical surnames are in close association with the same Y-chromosome haplotypes (Sykes and Irven, 2000).

Surnames that cannot be attributed to a place of origin. These can be either 1) polyphyletic surnames having a broad distribution, or 2) rare surnames that do not show a frequency distribution permitting the inference of their geographic origin

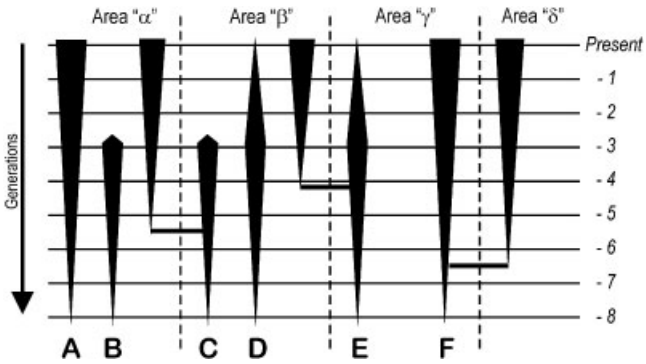

Fig. 5. Schematic representation of different kinds of surname distributions that can be expected. A, surname geographically specific to area α, whose geographic origin can be successfully identified by SOM algorithm. B, extinct surname corresponding to a Y-chromosome lineage that cannot be analyzed. C, surname whose present center of gravity differs from area of origin that can no longer be inferred, since founding family is extinct. D, rare surname whose geographic origin cannot be inferred, since too few individuals share it. E, surname corresponding to typology of C. It must be noted that correction for population size of village or town where this family name still exists may enable an inference of real geographic origin (γ), if enough descendants still live there. F, surname that underwent a geographical split in an earlier phase (close to time of origin of surnames) and that exhibits a present frequency distribution that leads to ambiguous results.

(example D in Fig. 5). On the one hand, polyphyletic surnames (we have shown how to identify them) should never be sampled, since they are neither geographically nor Y-chromosome specific. On the other hand, rare surnames should also not be sampled, because there is no way (with the exception of a genealogical or historical approach) to ascertain whether they are really representative of the area under investigation or not. According to the sampling guidelines defined above, only the "grandparents criterion" can partially warrant that their Y-chromosome lineages are specific to the area; therefore, their sampling should be undertaken only when time constraints make our improved strategy impossible.

The SOM-based approach works to systematically avoid the sampling of rare surnames, but the resulting bias is only apparent. There are no reasons that those Y-chromosomes linked to rare surnames should be more informative than those linked to frequent surnames, because the frequency of surnames depends on the stochasticity of demographic processes (when surnames had their origin in the same frame time). This stochasticity grants that the sampling of individuals having frequent surnames is as random as possible; otherwise it would mean that males without descendants are genetically different, on average, from those having some.

Drift and founder effect phenomena. It is possible that the geographic origin of some surnames (i.e., at the time of surname introduction) may be obscured by factors related to the migration of a branch of a family having a given surname. In the case of drift, one branch may have left a number of living descendants that, by chance, is higher than the original family branch, thus leading to ambiguities when assessing the real geographic origin of that surname. The founder effect is similar and relates to the establishment of some individuals in an unpeopled area (like one of the numerous islands of the Netherlands) and of the subsequent formation of a consanguineous community whose members still bear, in their majority, the surnames of the founders. In both cases, given the present frequency distribution of surnames (and if these migrations took place sufficiently far in the past), it may be impossible to trace the real geographic origin of a given surname (examples E and F in Fig. 5). Moreover, the family branch that did not "move" could be extinct (example C in Fig. 5). Given the specific spatial distribution of surnames that underwent these phenomena, SOM analysis will cluster together surnames that became specific to a given region at quite different times (Fig. 5). As a result, when inferring the past diversity of Y-chromosomes, different slices of time may be reflected in sampled Y-chromosome lineages with surnames that originally had their origin in the area and others that came from elsewhere. As assessed by Darlu et al. (2001), when they compared surname distributions of the 19th and 20th centuries in the Western Pyrénées region of France, there were ~4.5% of surnames that changed their geographical center of gravity in one generation. This rate, when applied to the eight-generation time-depth of Dutch surnames, would result in a ~30% error in the prediction of the true geographic origin of surnames (assuming 1 generation = 25 years). Actually, the real error on the two-century time-frame of the Netherlands has to be much lower, given that the estimate of Darlu et al. (2001) relates to the rural exodus period (1891–1965) when the mobility of people became much higher than in previous times. Moreover, our SOM analysis was performed on surname frequencies corrected by the size of the village/town, according to the assumption that recent migrations tended to be in the direction of towns/cities that had a bigger size than the departure area. This correction, which was not applied by Darlu et al. (2001) to their data, decreases the probability of inferring an untrue geographic origin of surnames and reinforces our approach, as confirmed by the high correlation we found between historical census data and a SOM-inferred retrospective census.

Surname variants. Several surname variants were shown to cluster in the same GSSGP. For example, in cell 1,1, we find variants such as Aart/Aarts, Bart/Baart, and Siemons/Simonis/Simons that could be grouped in a single surname, as suggested by Pollitzer et al. (1988) in a study concerning the bias of surname variants in the estimation of isonymy (see also, Legay and Vernay, 2000). Donors having surname variants, when they are grouped in

cells related to the same geographic area, can be considered descendants of the same family. As a consequence, only one variant should be selected (i.e., Aart or Aarts, Bart or Baarts, and Siemons or Simonis or Simons) to avoid sampling of the same Y-chromosome lineage.

Final remarks. The major point is that Y-chromosome donors, selected according to our method, have ancestors who lived in a given area much longer than the brief three-generation time warranted by the "grandparents criterion." Drift and founder-effect phenomena, since they reflect old migrations undetectable by our method, mean that some surnames, selected to be representative of a region, are "less ancient" and cannot be geographically traced as deep into the past as the time of their introduction (eight generations ago in the Netherlands). Even so, we are convinced that our method warrants more representative samples, since we can get closer, on average, to the genetic structures of old or ancient populations. It is as if we were defining the official borders of some European countries without any clue other than the 2100 AD spatial distribution of the Euro coins that were adopted in 2002 AD. It is obvious that the availability of an earlier geographic distribution of coins, say of 2070 AD, would enable a more detailed inference of the official borders of European countries. In this study, Y-chromosomes are the coins, and country borders are the ancient populations.

The major role played by regional migrations indicates that standard-design samples are representative of wider areas around them. For this reason, a well-spaced and wide sampling grid should be designed to study the genetic variability in macroregional and continental studies. On the contrary, on small scales, a regular and tight sampling grid seems insufficient to portray the genetic structures of populations, owing to the confusing effect of recent gene flow. Furthermore, when a SOM-designed sampling is adopted, there should be no ethical concerns, since the knowledge that a given DNA sequence belongs to someone whose surname appears in a list of 476 different Friesian family names (shared by ~20% of the present population) still guarantees a full anonymity that matches the UNESCO standards and guidelines on the consent of DNA donors. In this respect, we point out that an a posteriori sampling strategy consisting of 1) randomly sampling males, 2) recording the surname of each, and 3) attributing a geographic origin to surnames using the SOM approach would be less cost/time-effective, since almost half of the surnames of a country are very rare (meaning that their geographic origin cannot be inferred by SOMs), and more importantly, would be ethically problematic, since the link between the DNA sample and the surname of the donor has to be maintained.

The examples above (see The Sampling) should be sufficient to illustrate how recent population movements can affect the accuracy of regional anthropogenetic studies. We suggest that regional studies of the genetic variability of Y-chromosome markers are weakened by a considerable bias, since the present population of a given area is not representative of the surname-founding population, and as a consequence, is even less representative of the population of previous times. This discovery confirms, on a wider scale, a previous case study on a small, geographically isolated population (Valserine, France), where only 17.8% of the genes of the present population can be traced to the 16th century population (Bideau et al., 1992; Heyer, 1993). Similar values were obtained by Biraben (2002, personal communication), since only a low percentage of the present French population was demonstrated to descend from the population living in the same department in the 18th century.

CONCLUSIONS

The application of self-organizing maps to surnames, coded as frequency vectors, was tested on Dutch family names. The different results point to the validity of such an approach to identify surnames that have a similar geographic distribution and origin. Such groups of surnames (GSSGPs) correspond to families that had a similar geographic history of migration. The possibility of inferring the area were these families were settled, at the time of surname introduction, makes possible a retrospective census of the population for that historical time (two centuries ago in the Netherlands, and the Middle Ages in other European countries), since we show that the results are highly correlated with an almost contemporary historical census. A first conclusion is that, by an SOM approach of the geographic pattern of single surnames, it is possible to know in which proportion the population was settled in the different regions and areas of a country at that time, even when historical census data are unavailable.

Furthermore, the kind of geographic clustering of surnames provided by the SOM method permits the identification of polyphyletic family names that, in the Netherlands, correspond to ~24% of the whole population. Such identification makes possible their withdrawal from genetic analyses of populations, in order to realize a better inference of underlying genetic structures and differentiation processes in all surname-based microevolutionary studies.

With the use of the SOM approach, the geographic origin of almost three-quarters of surnames can be identified. In other words, it is possible to identify which individuals still live in the area where their ancestors were settled when surnames started to be in use (beginning of the 19th century in the Netherlands). To avoid the effects of recent migrations in Y-chromosome studies focusing on the past differentiation of human populations, such individuals should be preferred as DNA donors, thus virtually sampling the population as it was at the time of

surname introduction. This application of surname studies is very likely to provide a more accurate portrait of the Y-chromosome variability of ancient populations. The technique can be especially helpful in regional studies, since it was shown that patrilineal markers exhibit a larger geographic specificity than matrilineal or autosomic ones, because of the plausible reduced mobility of males compared to females (Seielstad et al., 1998).

ACKNOWLEDGMENTS

We thank Prof. L.L. Cavalli-Sforza and P. Darlu for providing useful suggestions. We are indebted to G. Van Driem for his help in surname translation, as well as to A. Gagnon for reading the manuscript. We warmly thank two anonymous reviewers for their comments and suggestions. F.M. was supported by an EGIDE grant in the framework of bilateral Italian/French scientific cooperation and by a CNRS contract. The quality of the English-language text was improved by L. Manni.

LITERATURE CITED

Barrai I, Scapoli C, Beretta M, Nesti C, Mamolini E, Rodriguez-Larralde A. 1996. Isonymy and the genetic structure of Switzerland I: the distributions of surnames. Ann Hum Biol 23:431–455.

Barrai I, Scapoli C, Beretta M, Nesti C, Mamolini E, Rodriguez-Larralde A. 1997. Isolation by distance in Germany. Hum Genet 100:684.

Barrai I, Scapoli C, Mamolini E, Rodriguez-Larralde A. 1999. Isolation by distance in Italy. Hum Biol 71:947–962.

Barrai I, Rodriguez-Larralde A, Manni F, Scapoli C. 2002. Isonymy and isolation by distance in the Netherlands. Hum Biol 74:263–283.

Bideau A, Brunet G, Heyer E, Plauchu H, Robert JM. 1992. An abnormal concentration of cases of Rendu-Osler disease in the Valserine valley of the French Jura: a genealogical and demographic study. Ann Hum Biol 19:233–247.

Cavalli-Sforza LL, Bodmer W. 1971. Human population genetics. San Francisco: Freeman.

Chen KH, Cavalli-Sforza L-L. 1983. Surnames in Taiwan: interpretations based on geography and history. Hum Biol 55:367–374.

Centraal Bureau voor de Statistiek van Nederlaand. 1904. Inleiding tot de uitkomsten der achtste algemeene tienjaarlijksche volkstelling van een en dertig december 1899 en daaraan verbonden beroepstelling en woningstatistiek in vergelijking zooveel mogelijk met de uitkomsten van vroegere tellingen. 'S-Gravenhage: Gebrs. Belinfante.

Crow JF. 1980. The estimation of inbreeding from isonymy. Hum Biol 52:1–4.

Crow JF, Mange AP. 1965. Measurements of inbreeding from the frequency of marriages between persons of the same surnames. Eugen Q 12:199–203.

Darlu P, Ruffié J. 1992. Relationships between consanguinity and migration rate from surname distributions and isonymy in France. Ann Hum Biol 19:133–137.

Darlu P, Degioanni A, Jakobi L. 2001. Les cloisonnement dans les Pyrénées occidentales. Évolution, du XIXe siècle à nos jours. In: Brunet G, Darlu P, Zei G, editors. Le patronyme: histoire, anthropologie, société. Paris: CNRS Editions. p 173–187.

Gabriel KR. 1968. The biplot graphical display of matrices with application to principal component analysis. Biometrika 58: 453–467.

Gagnon A. 2001. Patronymes, lignées et généalogies au Québec ancien. In: Brunet G, Darlu P, Zei G, editors. Le patronyme: histoire, anthropologie, société. Paris: CNRS Editions. p 333–349.

Gagnon A, Toupance B. 2002. Testing isonymy with paternal and maternal lineages in the early Quebec population: the impact of polyphyletism and demographic differentials. Am J Phys Anthropol 117:334–341.

Heyer E. 1993. Population structure and immigration; a study of the Valserine valley (French Jura) from the 17th century until the present. Ann Hum Biol 20:565–573.

Jobling M. 2001. In the name of the father: surnames and genetics. Trends Genet 17:353–357.

Kaski S. 1997. Data exploration using self-organizing-maps. Acta Polytechn Scand 82:1–57.

Kaski S, Kohonen T. 1994. Winner-take-all networks for physiological models in competitive learning. Neural Networks 7:973–984.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69.

Kohonen T. 1984. Self-organization and associative memory. Berlin: Springer.

Kohonen T. 1989. Self-organization and associative memory. 3rd ed. Berlin: Springer.

Kohonen T. 1993. Physiological interpretation of the self-organizing map algorithm. Neural Networks 6:895–905.

Kohonen T. 1995. Self-organizing maps. Berlin: Springer.

Lasker GW. 1968. The occurrence of identical, isonymous surnames in various relationships in pedigrees: a preliminary analysis of the relation of surname combinations to inbreeding. Am J Hum Genet 20:250–257.

Lasker GW. 1977. A coefficient of relationship by isonymy: a method for estimating the genetic relationship between populations. Hum Biol 49:489–493.

Legay JM, Vernay M. 2000. The distribution and geographical origin of some French surnames. Ann Hum Biol 27:587–605.

Manni F. 2001a. Strutture genetiche e differenze linguistiche: un approccio comparato a livello micro e macro regionale. Doctoral thesis. Ferrara: University of Ferrara.

Manni F. 2001b. Searching for ancient genetic patterns in Europe. Dialect, surname and genetic data, when combined, can tell us where. In: Testing the farming dispersals hypothesis (papers of the workshop: 24–27 August 2001). Cambridge: McDonald Institute for Archaeological Research.

Manni F, Barrai I. 2000. Patterns of genetic and linguistic variation in Italy: a case study. In: Refrew C, Boyle K, editors. Archaeogenetics: DNA and the population prehistory of Europe (McDonald Institute monographs). Cambridge: McDonald Institute for Archeological Research. p 333–338.

Manni F, Barrai I. 2001. Genetic structures and linguistic boundaries in Italy: a microregional approach. Hum Biol 73:335–347.

Mourrieras B, Darlu P, Hochez J, Hazout S. 1995. Surname distribution in France: a distance analysis by a distorted geographical map. Ann Hum Biol 22:183–198.

Nerbonne J, Heeringa W, van den Hout E, van der Kooi P, Otten S, van de Vis W. 1996. Phonetic distance between Dutch dialects. In: Durieux G, Daelemans W, Gillis S, editors. CLIN VI, papers from the sixth CLIN meeting. Antwerp: Center for Dutch Language and Speech, University of Antwerp. p 185–202.

Pollitzer WS, Smith MT, Williams WR. 1988. A study of isonymic relationships in Fylingdales parish from marriage records from 1654 through 1916. Hum Biol 60:363–382.

Rogers AR. 1991. Doubts about isonymy. Hum Biol 63:663–668.

Seber GAF. 1984. Multivariate analysis. New York: Wiley.

Seielstad MT, Minch E, Cavalli Sforza LL. 1998. Genetic evidence for a higher female migration rate in humans. Nat Genet 20: 278–280.

Sokal RR, Harding RM, Lasker GW, Mascie-Taylor CG. 1992. A spatial analysis of 100 surnames in England and Wales. Ann Hum Biol 19:445–476.

Sykes B, Irven C. 2000. Surnames and Y chromosome. Am J Hum Genet 66:1417–1419.

Torgerson W-S. 1958. Theory and methods of scaling. New York: Wiley.

Yasuda N, Furusho T. 1971. Random and non random inbreeding revealed from isonymy study. I. Small cities in Japan. Am J Hum Genet 23:303–316.

Yasuda N, Morton NE. 1967. Studies on human population structure. In: Crow JF, Neel JV, editors. Third International Congress of Human Genetics. Baltimore: Johns Hopkins Press. p 249–265.

Yasuda N, Cavalli-Sforza LL, Skolnick M, Moroni A. 1974. The evolution of surnames: an analysis of their distribution and extinction. Theor Popul Biol 5:123–142.

Zei G, Matessi RG, Siri E, Moroni A, Cavalli-Sforza L-L. 1983. Surnames in Sardinia. I. Fit of frequency distributions for neutral alleles and genetic population structure. Ann Hum Genet 47:329–352.

Zei G, Piazza A, Cavalli-Sforza LL. 1984. Geographic analysis of surname distributions in Sardinia: a test for neutrality. Atti Assoc Genet Ital 30:247.