

ARTICLE

Y-chromosome diversity in Catalan surname samples: insights into surname origin and frequency

Neus Solé-Morata¹, Jaume Bertranpetit¹, David Comas¹ and Francesc Calafell^{*1}

The biological behavior of the Y chromosome, which is paternally inherited, implies that males sharing the same surname may also share a similar Y chromosome. However, socio-cultural factors, such as polyphyly, non-paternity, adoption, or matrilineal surname transmission, may prevent the joint transmission of the surname and the Y chromosome. By genotyping 17 Y-STRs and 68 SNPs in ~2500 male samples that each carried one of the 50 selected Catalan surnames, we could determine sets of descendants of a common ancestor, the population of origin of the common ancestor, and the date when such a common ancestor lived. Haplotype diversity was positively correlated with surname frequency, that is, rarer surnames showed the strongest signals of coancestry. Introgression rates of Y chromosomes into a surname by non-paternity, adoption, and transmission of the maternal surname were estimated at 1.5–2.6% per generation, with some local variation. Average ages for the founders of the surnames were estimated at ~500 years, suggesting a delay between the origin of surnames (twelfth and thirteenth centuries) and the systematization of their paternal transmission. We have found that, in general, a foreign etymology for a surname does not often result in a non-indigenous origin of surname founders; however, bearers of some surnames with an Arabic etymology show an excess of North African haplotypes. Finally, we estimate that surname prediction from a Y-chromosome haplotype, which may have interesting forensic applications, has a ~60% sensitivity but a 17% false discovery rate.

European Journal of Human Genetics advance online publication, 18 February 2015; doi:10.1038/ejhg.2015.14

INTRODUCTION

In most societies, surnames are inherited in the same way as the male-specific region of the Y chromosome, which creates a correlation between the two. However, several factors will decrease this correlation: surnames with more than one founder, such as patronymic surnames (surnames originating from former first names) or surnames of occupation or status, non-paternity events, child adoption, and matrilineal surname transmission. In recent years, Y-chromosome variants are being used to investigate surname history.¹ The correspondence is simple: men sharing surnames might be expected to share similar Y chromosomes. Obviously, this link can also be affected by nonpatrilial transmissions of the surname.

Two types of polymorphic markers are commonly used to distinguish Y chromosomes from one another.¹ Binary markers such as SNPs (single-nucleotide polymorphisms) have low mutation rates, ~10⁻⁸ per base per generation,² and represent often unique events in human evolution. STRs (short tandem repeats) are multiallelic markers and most have a mutation rate of 10⁻³–10⁻⁴ per generation.³ Mutation diversifies Y-chromosome haplotypes associated with a specific surname. However, SNP mutation rates are low enough not to be represented in the timescale of surname inheritance systematization (surnames appeared 500–1000 years ago). As a result, mutation among tested SNPs is highly improbable among the descendants of a surname founder. Hence, we expect biological descendants of the same founder of a surname to have the same SNP haplotype (called a haplogroup) but that might not be the case for Y-STR haplotypes due to the high mutation rate of STRs.

The pioneering study of the surname Sykes⁴ revealed low Y-haplotype diversity among unrelated carriers of the name, suggesting that this was compatible with a single founder. Other, more comprehensive studies, have considered sets of surnames in relation to the Y-chromosome diversity of their bearers. King and Jobling⁵ analyzed British surnames and estimated haplotype diversity and the number of founders. The results showed that, in Britain, the more frequent surnames tended to have more diverse Y-chromosome haplotypes, as if surname frequency was driven by polyphyly. On the contrary, an analog study carried out in Ireland⁶ showed no significant correlation between the frequency of a surname and the diversity of the Y chromosomes within it. Moreover, the finding that frequent Irish surnames such as O'Sullivan and Ryan present a single founder may be accounted for by male social and reproductive success in the past. Consequently, they considered that the Y chromosome could have suffered the effect of natural selection due not to its genetic content but to a cultural marker: the surname.

Similar studies have pointed out that the link between surname and Y-chromosomal haplotype suggests that it would be feasible to predict a surname in forensic investigations.^{5,7} Having a list of surnames with associated Y-STR haplotypes could be helpful to match a Y profile with one or more surnames, because this would provide a means to prioritize a suspect list. However, surname prediction would be used only as an investigative tool because, to exclude or match individuals with a high degree of likelihood, an autosomal profile is needed.

¹Departament de Ciències Experimentals i de la Salut, Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Universitat Pompeu Fabra, CEXS-UPF-PRBB, Barcelona, Catalonia, Spain

*Correspondence: Professor F Calafell, Departament de Ciències Experimentals i de la Salut, Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Universitat Pompeu Fabra, CEXS-UPF-PRBB, Doctor Aiguader 88, Barcelona, Catalonia 08005, Spain. Tel: +34 93 316 08 42; Fax: +34 93 316 09 01; E-mail: francesc.calafell@upf.edu
Received 16 September 2014; revised 26 November 2014; accepted 13 January 2015



Figure 1 Map of NE Spain.

To the best of our knowledge, surname systems have only been analyzed in Britain and Ireland, with, as we have stated above, different results. We have investigated the driving forces behind the origin, systematization, and diffusion of surnames by using a set of 50 Catalan surnames. Catalonia is a distinct region in the NE corner of the Iberian Peninsula (Figure 1); its origins can be traced to the aftermath of the Islamic invasion of Iberia in AD 711. A narrow strip of land in the North of the Peninsula remained under Christian control and fragmented into multiple kingdoms; the earldom of Barcelona, in the northern half of current Catalonia, was initially a fief of the Carolingian empire that broke up in AD 897. In 1162, it joined the Kingdom of Aragon, which, in the thirteenth to fifteenth centuries, expanded southward to València and eastward to the Balearic Islands, S Italy and as far east as Athens. The Catalan language originated from the fragmentation of vulgar Latin, and its earliest known texts date from the twelfth century. It has remained a distinct language, spoken by 4.5 million people as their first language and by an additional 5 million as a second language, in an area including Roussillon

(in S France), Andorra, a strip in E Aragon, Catalonia, the Balearic Islands, València, and the Sardinian town of Alghero.⁸

Catalan surnames are based on the Catalan language, and thus, they are easily recognizable⁹ from Spanish- or Basque-based surnames. They are diffused throughout the Catalan-speaking lands, and carried by migration to the Americas. Catalan surnames present a higher diversity compared with Spanish surnames: for instance, the top 10 most frequent Spanish surnames have a total frequency of 18.1% (National Statistics Institute of Spain, www.ine.es), while the 10 most frequent Catalan surnames are carried by just 8% of the total population of Catalan-surname bearers (see also Rodríguez-Larralde *et al*¹⁰). Most people seem to have had a surname in Catalonia (as in most of Western Europe) by the thirteenth century, although the Council of Trent (1545–1563), which mandated that each parish keep birth, marriage, and death records, seems to mark the universal, systematic transmission of the paternal surname.^{11,12}

For the present project, we have genotyped 17 Y-chromosome STRs and 68 SNPs in ~2500 samples of men carrying each one of the 50

Catalan surnames that we previously selected (see Methods section) with the following aims: (i) To discover and quantify the processes that drive surname frequency. Surname polyphyly (which can be measured from the number of founders detected from the Y-chromosome diversity) can drive surname frequency, as was found in a sample of British surnames, but drift and cultural selection (associated with high-status surnames) may also have a role, as shown in the analysis of Irish surnames.⁶ (ii) To quantify the rate of irregular transmission of surnames, a compound measure of adoption, false paternity, and maternal surname transmission. (iii) To explore the forensic application of surname studies, by quantifying the sensitivity to predict a surname from a forensic sample. And (iv), to verify whether the origin of the founders of some surnames matched the origin suggested by the etymology of those surnames.

MATERIALS AND METHODS

Surnames in the project

We selected 50 surnames (Table 1) to address the objectives of this project. Most were chosen from different surname categories and because of their frequency; ie, we selected the 2–3 most frequent surnames in patronymic, place name, trade, landscape feature, and byname surnames, as well as 2–3 rarer surnames in each category (~500 carriers of both sexes, according to population registry data compiled by the National Statistics Institute of Spain, and available at www.ine.es). We aggregated spelling variants when obtaining surname frequencies. Other surnames were selected for their Arabic or Hebrew linguistic roots or because they were apparent demonyms (surnames indicating geographic origin, such as the English surname French, or the French surname Langlais) (Table 1). Overall, carriers of these 50 surnames represent ~3% of the Catalan population or ~9% of the population in Catalonia that carries linguistically Catalan surnames.

Sample collection

After publicizing the project in the local media, volunteers were asked to fill and submit a data collection form to verify whether they fulfilled inclusion criteria for the project: carrying one of the 50 surnames as first (paternal) surname, and not having a relative (first-degree cousin or closer) already volunteering. If these criteria were met, the volunteers were sent a Falcon tube (filled with 2 ml of a stabilizing buffer: 0.67 M LiCl, 33 mM Tris base, 0.67 M Urea, 0.6% SDS (W/V), 3.3 mM EDTA) to collect a 2-ml sample of saliva, along with instructions and an informed consent form to be signed and mailed back with the sample. A total of 1375 samples were collected with this procedure. To complete sampling for all but the most common surnames, the collection kit was mailed to 6857 candidate addresses obtained from the telephone directory; 1185 samples were returned (17.3%). In this case, pairs of samples corresponding to close relatives were detected and filtered *a posteriori*, but 135 individuals who were, by the genealogies they provided, second cousins or more distant relations with another volunteer were retained. The volunteers were also asked to provide birthplaces for as many of their paternal ancestors as they knew; the birthplace of the oldest paternal ancestor was recorded as the origin of that particular sample (Supplementary Table S1). This project was reviewed and approved by the Institutional Review Board of the Comitè Ètic d'Investigació Clínica-Institut Municipal d'Assistència Sanitària (CEIC-IMAS) in Barcelona.

DNA extraction and genotyping

DNA was extracted using a standard organic method with proteinase K digestion, followed by phenol–chloroform extraction. Samples were typed for the 17 Y-chromosome STRs contained in the AmpFISTR YFiler PCR Amplification Kit (Life Technologies, Carlsbad, CA, USA)¹³ and for 68 Y-chromosome SNPs (Supplementary Table S2). Sixty-four SNPs were typed with the OpenArray technology, adapting the SNPs more likely to be informative in a Western European population from the 128-SNP Open Array devised by Martínez-Cruz *et al*¹⁴ and including five SNPs (Z195, Z209, Z220, Z268, and Z278) mapping within haplogroup R1b-P312 that have recently been found to be polymorphic in populations of Spanish descent.¹⁵ Additionally,

four SNPs (M123, M78, M26, and M423) were typed using the single Taqman (Life Technologies) assays (Supplementary Table S3); M123 and M78 were typed in E-M35*(xV12, V13, V22, V65) individuals, and M26 and M423 were typed in I-P37 cases. Finally, SNP U152 gave inconsistent results in the Open Array and was typed also in a Taqman assay in R1b-P312*(xZ195, L21) samples.

The data set obtained has been deposited in YHRD (the Y-Chromosome Haplotype Reference Database; www.yhrd.org) with accession numbers YA004020–YA004025. Note that YHRD does not allow the submission of haplotypes with missing values, which affected 16.7% of our final data set. The complete data set can be found in Supplementary Table S4. Haplogroup nomenclature follows the basic conventions of Van Oven *et al*¹⁶; note, however, that SNPs Z209, Z268, Z220, and Z278¹⁵ are not present in Van Oven *et al*.¹⁶ Further information can be found in <http://isogg.org/tree/index.html>.

Data analysis

Basic descriptive statistics of the genetic diversity in the different samples such as haplotype diversity, average number of pairwise differences, and population pairwise R_{ST} for Y-STR haplotypes were calculated using Arlequin 3.1.¹⁷ Median-joining networks were drawn with Network 4.5.1.6 (www.fluxus-engineering.com); alleles at DYS389I were subtracted from those at DYS389II, and DYS385a,b was used in the within-haplogroup networks, as, at short genealogical distances, repeat differences at the duplicated DYS385 STR can be reasonably approximated to mutation events. Multidimensional scaling (MDS) based on R_{ST} matrices was carried out using the R software.¹⁸

We expect the biological descendants of the medieval founder of a surname to carry the same Y-chromosome haplogroup and the same Y-chromosome haplotype or, at least, to accumulate just a few STR mutations compared with the founder. We determined descent clusters, ie, groups of men likely to have a common ancestor that was more recent than the age when surnames were established following with a modified version of the heuristic in Martínez-González *et al*¹⁹ (for further details on determining descent clusters and their ages, see Supplementary Text S1). Several external references were used for comparison as necessary: Catalans and Italians (only for STRs),¹⁹ Gascons,²⁰ Bavarians,²¹ Moroccans,²² Swedes (www.yhrd.org), and Sephardic Jews.²³ Local haplogroup frequencies were estimated from our data by using descent clusters as sampling units, thus avoiding related individuals. For particular surnames, the local vs external contributions were estimated as admixture proportions with the least-square mY estimate, computed with the ADMIX 2.0 software (<http://cmpg.unibe.ch/software/admix>).²⁴

RESULTS AND DISCUSSION

We have genotyped 17 Y-chromosome STRs and 68 SNPs in 2560 men each bearing one of the 50 Catalan surnames. After removing one of each pair of first cousins and closer relatives, we obtained results for 2309 samples. A description of the results for each individual in the data set can be found in Supplementary Table S4 and a summary by surname is included in Table 1.

Surname frequency and polyphyly

Haplotype diversity was highly correlated with surname frequency (Spearman's $\rho = 0.756$, $P \sim 10^{-10}$; Figure 2a), which implies that frequent surnames are not likely to have expanded suddenly, as some Irish surnames did.⁶ We next tried to determine descent clusters for each surname, that is, groups of one or more men that shared a surname, and whose haplotypes indicated that they had a common ancestor in the thirteenth century (the date Catalan surnames spread; see Methods section) or later. Across all 50 surnames, we found 1151 descent clusters, of which 751 (65.2%) consisted of just one man each. In 731 instances (across all surnames), different descent clusters were found in the same haplogroup. In order to verify the consistency of our descent cluster definition, we computed the median distance in absolute number of STR allele differences between descent clusters of the same haplogroup, which was six, and only in 11.1% of the

Table 1 Surnames in the study

Surname	Type	Meaning	<i>f</i>	<i>N</i>	<i>D</i> (SD)	<i>nl</i>	FMDC
Adell (ADL)	L Patr.	Adellius	2480	38	0.7824 (0.0733)	12	0.7105
Albert (ALB)	G Patr.	Albert	7095	58	0.9734 (0.0084)	22	0.5517
Alemaný (ALE)	Demonym	German	5554	68	0.921 (0.0236)	21	0.6324
Armengol (ARM)	G Patr.	Ermengaud	4105	71	0.9394 (0.0185)	23	0.5352
Balash (BAL)	Byname?	Black crow?	559	25	0.7733 (0.0876)	6	0.8000
Bonastre (BON)	Byname	Fortunate	871	21	0.9158 (0.0462)	12	0.5238
Bosch (BOS)	Landscape	Forest	14 403	55	0.9951 (0.0048)	41	0.1455
Cabot (CAB)	Byname	Bighead	1303	29	0.918 (0.035)	9	0.5862
Calafell (CAL)	Place name		975	24	0.9094 (0.037)	6	0.7500
Cardona (CAR)	Place name		9639	69	0.9837 (0.0065)	32	0.4638
Casajuana (CAS)	Habitat	John's house	616	21	0.8246 (0.0648)	5	0.8095
Codina (COD)	Landscape	Barren rock	5851	56	0.9935 (0.0048)	34	0.0714
Comas (COM)	Landscape	Dales	8984	68	0.9968 (0.0032)	48	0.0588
Danés (DAN)	Demonym	Danish	526	13	0.5641 (0.1117)	3	0.9231
Estruch (EST)	Byname	Fortunate	2418	38	0.9289 (0.0243)	9	0.8684
Ferrer (FER)	Occupation	Smith	55 356	90	0.999 (0.0019)	68	0.0000
Fortuny (FOR)	L Patr.	Fortunius	2389	52	0.9736 (0.0141)	23	0.4615
Gasull (GAS)	Arabic	Wazul tribe?	761	27	0.9345 (0.0326)	8	0.7407
Grau (GRA)	G Patr.	Gairoald	15 723	68	0.9912 (0.0044)	38	0.1912
Gual (GUL)	G Patr.	Wadald	3006	61	0.961 (0.0171)	23	0.6066
Guasch (GUS)	Demonym	Gascon	3914	68	0.9833 (0.0065)	23	0.6029
Llach (LLA)	Landscape	Lake	616	24	0.942 (0.0309)	8	0.5833
Mas (MAS)	Habitat	Manor	20 602	50	0.9949 (0.0056)	36	0.1000
Massot (MST)	Arabic	Mas'ud	893	27	0.9886 (0.0131)	10	0.4444
Maymó (MAY)	Hebrew	Maimun	892	28	0.906 (0.0398)	12	0.5357
Melis (MEL)	Arabic	Mahli	616	21	0.8474 (0.0786)	6	0.7143
Miquel (MIQ)	L Patr.	Michael	4544	64	0.9804 (0.0086)	26	0.5781
Moragues (MOR)	Arabic	Muħraqa	2735	42	0.9872 (0.0096)	16	0.5714
Nàcher (NAC)	Arabic	Nağğar	1735	23	0.9802 (0.0198)	9	0.6522
Nadal (NAD)	Byname	Christmas	7088	66	0.9935 (0.0034)	41	0.2424
Oriol (ORI)	L Patr.	Aureolus	1302	23	0.9565 (0.022)	10	0.4348
Pitarch (PIT)	Place name		3097	46	0.8636 (0.0403)	3	0.9565
Pons (PON)	L Patr.	Pontius	24 793	67	0.993 (0.0046)	41	0.2388
Raga (RAG)	Basque	Unknown	1687	18	0.8758 (0.0628)	6	0.7222
Reixach (REI)	L Patr.	Rixa	1347	25	0.9967 (0.0125)	16	0.2400
Ricart (RIC)	G Patr.	Richard	2736	61	0.9921 (0.0052)	28	0.4918
Robert (ROB)	G Patr.	Robert	2339	49	0.9694 (0.034)	23	0.3061
Roig (ROI)	Byname	Red	17 489	64	0.994 (0.041)	44	0.0625
Roma (ROM)	Place name		1781	28	0.9392 (0.0307)	10	0.7500
Ros (ROS)	Byname	Blond	11 989	61	0.982 (0.0107)	42	0.1803
Sabater (SAB)	Occupation	Shoemaker	10 223	67	0.9905 (0.0051)	37	0.4030
Sala (SLA)	Habitat	Hall	13 398	64	0.9936 (0.0046)	48	0.0938
Salom (SLM)	Hebrew	Shalom	2596	40	0.941 (0.0204)	13	0.8000
Santacana (STC)	Saint name	St Candida	640	27	0.9117 (0.032)	6	0.8148
Serra (SER)	Landscape	Range	26 855	54	0.9958 (0.0044)	37	0.0741
Soler (SOL)	Habitat	Land plot	44 447	88	0.9979 (0.0022)	63	0.1477
Taberner (TAB)	Occupation	Publican	1495	26	0.9323 (0.033)	8	0.6538
Tió (TIO)	Landscape	Log	627	18	0.9542 (0.0301)	8	0.5000
Vidal (VID)	Byname	Lively	35 144	67	0.9986 (0.003)	57	0.0000
Vives (VIV)	Byname	May you live	9534	51	0.9758 (0.011)	22	0.4902

Abbreviations: *D*, haplotype diversity; *f*, population frequency; FMDC, proportion of men in the sample belonging to a MDC (ie, with $n \geq 4$); G. patr., German patronym; L. patr., Latin patronym; *N*, sample size; *nl*, number of descent clusters (sets of men descended from a common ancestor).

cases the nearest haplotype was three presumed mutation steps away. Thus our definition of descent clusters seems to delineate clear sets of haplotypes of close coancestry.

The relative number of descent clusters (number of descent clusters over sample size) is also strongly correlated with surname frequency

($\rho = 0.527$, $P = 1.0 \times 10^{-4}$). The number of likely medieval founders plus subsequent Y-chromosome introgressions in a surname clearly grows with surname frequency, and this result is not explained by differences in sample size (partial correlation controlling for sample size $r = 0.734$, $P \sim 10^{-9}$). To try to separate initial founders from later

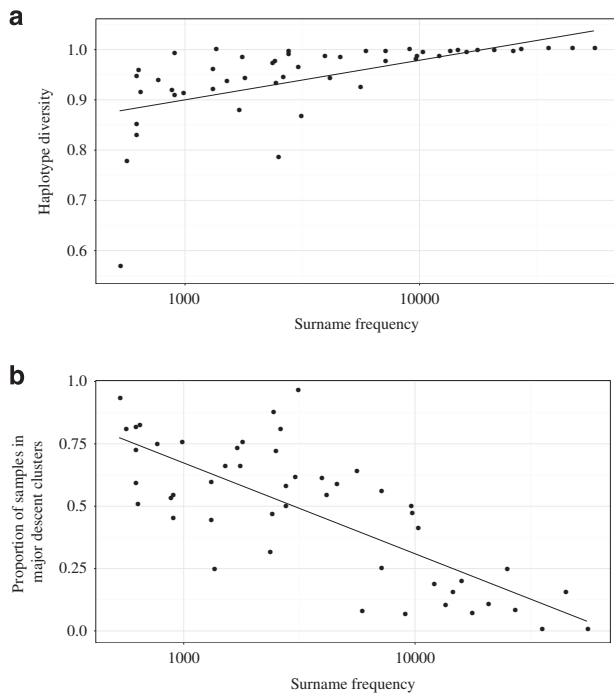


Figure 2 Scatter plots showing (a) haplotype diversity and (b) the proportion of the sample in a MDC as a function of the surname frequency.

introgressions, we considered that more frequent descent clusters would be more likely to represent the former than the latter. We arbitrarily defined major descent clusters (MDCs) as those with $n \geq 4$, where n is the number of individuals in the cluster. MDCs (and thus, putative independent medieval founders) by surname ranged from 0 to 6, with an average 2.64; the proportion of men in a surname that belonged to a MDC ranged from 0 to 95.7%, with an average of 40.6%. Both the relative number of MDCs and the proportion of men in a surname that belonged to a MDC were negatively correlated with surname frequency ($\rho = -0.679$, $P \sim 10^{-8}$ and $\rho = -0.711$, $P \sim 10^{-9}$, respectively) (Figure 2b). The latter correlation remained significant when controlling for sample size (partial correlations $r = -0.277$, $P = 0.059$; and $r = -0.461$, $P = 8.7 \times 10^{-4}$, respectively). That is, less common surnames comprise fewer descent clusters, which are more frequent and presumably older, while the extremely high number of descent clusters in frequent surnames means that each descent cluster is less frequent, and fewer reach the $n \geq 4$ threshold. These results seem robust to the arbitrary definition of MDC: we repeated the correlations with $n \geq 3$ and ≥ 5 thresholds, with very similar results: the correlation between the relative number of MDCs and surname frequency was $\rho = -0.524$ ($P = 1.8 \times 10^{-4}$; partial correlation controlling for sample size, $r = -0.335$, $P = 0.024$) for descent clusters with $n \geq 3$ and $\rho = -0.720$ ($P \sim 10^{-8}$; $r = -0.209$, $P = 0.163$ controlling for sample size) for descent clusters with $n \geq 5$. The correlation between MDC coverage and surname frequency was $\rho = -0.662$ ($P = 1.8 \times 10^{-6}$; partial correlation controlling for sample size, $r = -0.529$, $P = 1.12 \times 10^{-4}$) for descent clusters with $n \geq 3$ and $\rho = -0.660$ ($P \sim 10^{-7}$; $r = -0.410$, $P = 0.004$ controlling for sample size) for descent clusters with $n \geq 5$.

Next, we measured R_{ST} genetic distances among surnames and with a sample of the general population.¹⁹ Genetic distance from the general population was negatively correlated with surname frequency ($\rho = -0.664$, $P = 1.5 \times 10^{-7}$) (Supplementary Figure S1), and, when

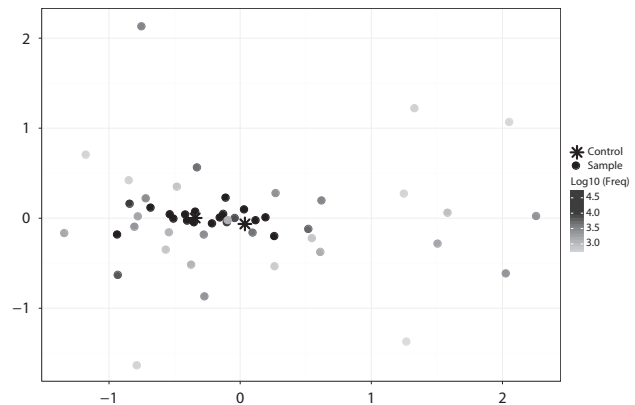


Figure 3 Multidimensional scaling (MDS) analysis based on pairwise R_{ST} calculated from Y-STR haplotype frequencies in 50 surnames (abbreviations as in Table 1) and in the Catalan and Italian (black stars) general populations.

the distances were represented with MDS (Figure 3), common surnames tended to cluster in the center of the plot, while peripheral surnames tended to be less frequent. An AMOVA based on R_{ST} showed that 12.49% of the overall genetic variation was explained by differences among surnames; this is a very large value, on the order of magnitude of autosomal genetic differences among worldwide populations.²⁵ For reference, R_{ST} between general Catalan and North Italian populations was 0.75%. Thus Y-chromosome genetic diversity in Catalan surname samples shows clear signs of founder effects that are more dramatic in rarer surnames. A sample consisting of the median haplotype for each descent cluster was not significantly different from the general Catalan population ($R_{ST} = -0.00123$, $P = 0.486$).

Median-joining networks based on Y-STR haplotypes and SNP data were used to study the haplotype diversity associated with a surname in more detail and to identify groups of men descending from a common founder. Figure 4 illustrates the differences between six surnames that are present at different frequencies within the population, which, again, show a trend for increased haplotype and haplogroup diversity and fewer MDCs in more frequent surnames (See also Supplementary Figures S2 and S3).

Y-chromosome haplotypes and haplogroups in samples of men bearing one of the 50 Catalan surnames show increased diversity with surname frequency and an increased number of founders but a decreased frequency of well-represented (and presumably older) descent clusters. Moreover, frequent surnames were closer to a general population sample, while rarer surnames tended to be more peripheral in an MDS plot of genetic distances, which implies that founder effects were much more pronounced for rarer surnames. These results are extremely similar to those found in Britain,⁵ and, as in that case, point to polyphyly as the major driver for surname frequency; that is, frequent surnames became abundant because they were founded multiple times, while rarer surnames tended to have fewer origins. This also implies that the Irish case,⁶ in which some surnames increased their frequencies but maintained low genetic diversities, probably because of the prestige and wealth associated with those particular surnames, may be an exception in Western Europe. Further studies in other European populations are needed to confirm this point.

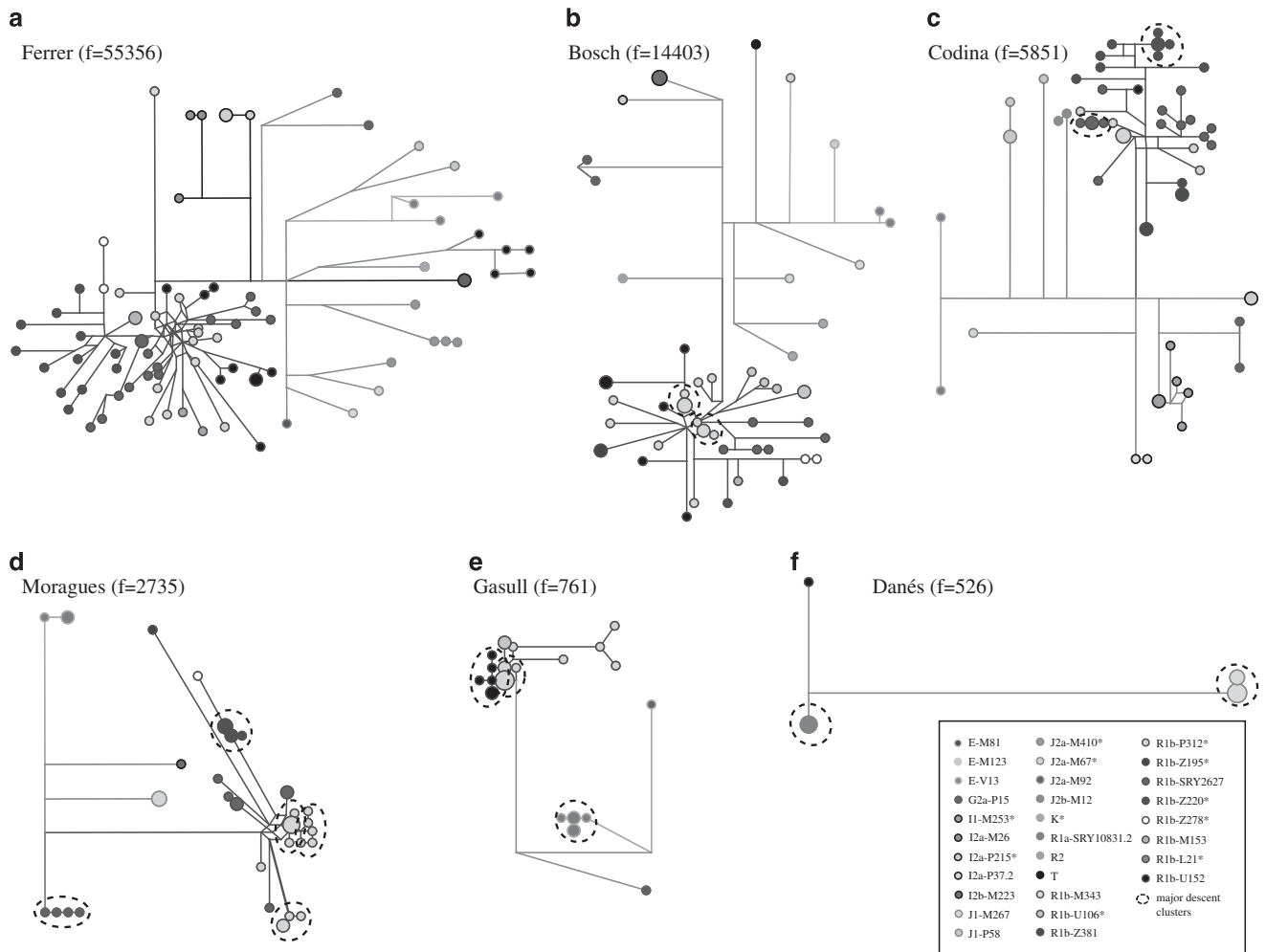


Figure 4 Haplogroup and Y-STR haplotype diversity within six selected surname samples represented by median-joining networks presented in decreasing surname frequency. Each circle represents a haplotype with its size proportional to frequency. Each circle has been colored according to haplogroups (see key), and the lines between them indicate the Y-STR mutational steps. MDCs are indicated with dotted ellipses. A full colour version of this figure is available at the *European Journal of Human Genetics* online.

Descent cluster ages

Ages were calculated only for the 131 MDCs (see Methods section) and are shown in Supplementary Figure S4. The average date of origin for MDCs is 1479 (SD 232) if computed with ρ and 1513 (SD 235) if computed with ASD_0 . Descent cluster age did not correlate with surname frequency ($\rho = -0.051$, $P = 0.565$ for the ρ method and $r = -0.044$, $P = 0.617$ if computed with ASD_0). Thirteen of the 131 MDCs comprised a single haplotype each. As the MDCs contained 44 patrilineal relative pairs (more distant than first cousins, and as ascertained from the known genealogies that each volunteer provided), we tested their possible contribution by counting the minimum average relatedness within each descent cluster. Individuals who were not known to be related to each other were assumed to share an ancestor in the generation prior to their known genealogy. Men in monomorphic MDCs (and thus an estimated age of 0) were thus related by a minimum of 6.94 ± 0.25 meioses, against 6.85 ± 0.09 for the MDCs with some haplotype diversity ($F = 0.102$, $P = 0.75$). Thus known relatedness does not seem to significantly lower estimated descent cluster age.

Catalan descent cluster ages are slightly more recent though than British and Irish surnames, which were about 500, 650, and 1100 years

old, respectively, although the Catalan and British surname age differ by < 1 SD. Surnames appeared in Catalonia and Britain in the twelfth and thirteenth centuries but were older in Ireland, where they are recorded in early tenth century. It should be noted that these ages' estimates depend on the definition of the clusters that are presumed to be descendants from the same medieval founder and on the mutation rate employed. The criteria for inclusion in a descent cluster we used (see Methods section) are less restrictive than those adopted by King and Jobling²; this would actually result in older age estimates in our case, as we included in descent clusters some individuals that would be excluded by King and Jobling and that are more peripherally related to the rest of the cluster. On the contrary, we adopted a faster mutation rate (1 mutation per 777 years, based on the tens of thousands of meioses compiled by www.yhrd.org) than King and Jobling did, who, for a partly overlapping set of STRs, used an estimate of 1 mutation per 1373 years, based on their own genotyping of pairs of related individuals totaling 274 transmissions of the Y chromosome. A possible factor that may have biased downward the age estimate for Catalan surnames is that surname inheritance may have been relatively lax in the Late Middle Ages and that the mandate of the Council of Trent, in

the sixteenth century, that parishes keep adequate birth, marriage and death records may have crystallized the universal, systematic transmission of the paternal surname. A comparison with other Catholic countries may be useful to test this hypothesis.

Introgession rates

Non-paternity, adoption, multiple founding events for names and matrilineal transmission of the surname can affect the link between Y chromosome and surnames, triggering the appearance of new chromosomes linked to a certain surname. If a surname is assumed to be monophyletic, the presence (and quantity) of descent clusters other than the presumed founder descent clusters can be used to estimate the rate of Y-chromosome introgession per generation.⁵ We have applied the equations by King and Jobling⁵ not to whole monophyletic surnames as in the original formulation but to locally monophyletic surnames, that is, to surnames in which a given descent cluster predominated in a particular, well-defined geographical area and provided that the frequency of that descent cluster was $n \geq 8$. Thus we could increase the sample size for the estimation of the introgession rate to 42 local surname origins. The introgession rate range estimate was 1.5–2.6% per generation (King and Jobling⁵ provided lower and upper bounds for the actual value) (Supplementary Table S5), which are similar values to those found in Britain. Additionally, other studies that included genealogical information in Flanders²⁶ and N Italy²⁷ estimated introgession rates, respectively, at 0.9 and 1.2% per generation; however, given the fact that these studies contained the genealogy of the participants, they estimated practically the non-paternity rate.

We could also estimate local introgession rates. Although differences were not statistically significant, a clear pattern emerged: introgession estimates were lower in the south of the studied area (Lower Ebro River Valley, 0.8–1.7%; València, 0.9–2.2%; Castelló, 1.2–2.7%) and in the island of Majorca (0.4–1.2%) than in the north (Central Catalonia, 2.8–4.2%, Penedès 2.4–3.9%, Barcelona region 2.2–3.3%). At least two factors could explain this pattern: the areas with higher introgession rate coincide both with regions where in rural areas often the surname was linked more to the manor than to the family (and if a woman inherited a large estate, she transmitted her surname²⁸), and to areas that received a larger fraction of Occitan migrants from southern France in the sixteenth and seventeenth centuries, who carried surnames that often were similar to (and assimilated with) Catalan surnames.^{29,30}

Surname origins vs founder origins

We next focused on groups of surnames in which the etymology of the surname suggested a particular origin for the founder of the surname and verified whether that was the case. For instance, we selected five Germanic patronymic surnames, that is, surnames that were originally Germanic first names, such as Armengol, Ricart, Gual, Albert, and Robert, and compared both with a frequency-matched set of Latin patronymic surnames (Adell, Fortuny, Miquel, Oriol, Reixach) and with the rest of our surname sample. Although Germanic patronymic surnames have a different haplogroup composition when compared with Latin patronymic surnames ($F_{CT} = 0.0141$, $P = 0.042$, mostly driven by the increased frequency of R1b-P312* in Germanic surnames; however, this haplogroup is much more frequent in Catalonia than in Germany), they are not different when compared with all other surnames ($F_{CT} = 0.0019$, $P = 0.225$). Next, we performed an admixture analysis using haplogroup frequencies and with Catalans and Bavarians as parental populations (Supplementary Figure S5A). Using ADMIX 2.0,²⁴ the admixture proportions for the Germanic

patronymic surname bearers were $m = 1$ for the Catalan contribution and $m = 0$ for the Bavarian contribution. That is, Germanic patronymic surname bearers did not carry Central European haplogroups in a proportion different than what was found in the general Catalan population, and it can be inferred that the founders of these surnames were not more of Germanic origin than the founder of other surnames.

We also investigated surnames with a linguistic Arabic or Hebrew background to ascertain whether their founders were North Africans or Jews themselves. An Arabic etymology has been suggested for the Gasull, Moragues, Nàcher, Massot, and Melis surnames (Supplementary Figure S5B), although alternative etymologies have been proposed for the latter two.^{11,12,31} Although as a category, these surnames did not show North African admixture ($m = 0$ for the North African contribution by using Moroccans as a parental population), statistically significant excesses of haplogroups common in North Africa (such as E-M81 or J1-M267) were found in Nàcher and Massot. Similarly, surnames Maymó and Salom may have a Hebrew etymology, Estruch was a common first name among medieval Jews in Catalonia, and Vidal is often mentioned (even though without much evidence) as a surname taken by Jews who converted to Christianity (Supplementary Figure S5C). By using Sephardic Jews as a parental population, the admixture proportion was $m = 0.202 \pm 0.0068$, likely due to the fact that about two-thirds of Estruch men carried J2 haplogroups, which are found in 22.7% of Jews³² but only in 8% of Catalans. No such signature of a possible Jewish origin was found in the other Jewish-related surnames, although Salom showed an excess of North African haplogroups.

Finally, we turned to ethnonymic or demonymic surnames. Some Catalan surnames (Alemany, Danés, Guasch) denote geographic origin (they mean German, Dane, Gascon, respectively). By comparing haplogroup frequencies between the carriers of those surnames and their putative populations of origin, we have seen that Alemany and Danés surnames do not reflect the origin of their founders (Supplementary Figures S6A and B). Note though that alternative etymologies have been suggested for both surnames. Haplogroup frequency similarity between the Gascon and Catalan populations (Supplementary Figure S6C) prevented any meaningful inference about the origin of the founders of the Guasch surname.

Surname prediction from a Y-chromosome genotype

As proposed by King *et al*,³³ surname prediction from a sample of unknown origin could be a useful forensic application of surname studies. A simple method to predict a surname can be based on MDCs as defined in this study. Thus if a Y-chromosome haplotype from a forensic sample belongs to a given MDC, we attribute that surname to the sample. We can estimate the sensitivity of this prediction as the number of samples in our study that were placed in a MDC, namely 40.8%. However, as discussed above, MDC coverage correlates negatively with surname frequency. We have used this correlation (and the linear regression line we calculated between MDC coverage and surname frequency) to estimate sensitivity for each Catalan surname. Weighing each predicted sensitivity by the frequency of the surname, we estimate that sensitivity would be 62.3% for the entire set of ~ 3300 Catalan surnames. However, we can also estimate a false discovery rate (FDR) of 17% by counting how often a sample that did not belong to a MDC in their own surname was two or fewer STR mutation steps away from a MDC in a different surname. FDR by surname does not correlate with surname frequency ($\rho = 0.044$, $P = 0.762$; Supplementary Figure S7).

A haplotype may be captured by a MDC of another surname for, at least, two reasons: the combination of 17 STRs and 68 SNPs we have genotyped may lack sufficient genetic resolution to differentiate unrelated Y chromosomes that may have similar haplotypes (especially taking into account the low haplotype diversity within the abundant R1b-M269 haplogroup^{34,35}), or the predicted surname was indeed carried by an ancestor of the donor, and, at some point in the genealogy, by false paternity, adoption, or transmission of the maternal surname, Y chromosome, and surname decoupled. In the latter scenario, it is plausible that the ancestors of the man carrying the queried haplotype and those of the MDC of the surname he should have carried could be found in closer geographic proximity; this would not be case, given the genetic homogeneity of the studied area (R_{ST} among geographical subregions, computed by taking one haplotype per descent cluster, was 0.0014, $P=0.22$), if the surname misprediction is due to homoplasy. We measured geographic proximity in travel hours by road and ferry as provided by the Google Distance Matrix API, in order to take into account the presence of mountain ranges and the fact that the Balearics are islands, which are factors that would not be captured by as-the-crow-flies distances. Although travel times have certainly shortened in the past generations, we can assume that the decrease has been roughly proportional throughout the studied area. For each mispredicted haplotype, we computed the average travel time to the ancestral birthplaces of the individuals in the MDC that captured it and compared this distribution to the average travel times to a different MDC that this haplotype did not belong to, selected for having the same frequency or the most similar to the descent cluster that captured the haplotype (Supplementary Figure S8). Average travel times are shorter to the MDC of the mispredicted surname (3.95 vs 4.61 h, $P=0.013$, paired Student's t -test), which suggests that a few of the ancestors of these individuals may have indeed carried the predicted surname and that it changed later by false paternity, adoption, or transmission of the maternal surname.

For surname prediction to have any practical application beyond the 50 surnames in this study, a general database should be established. We have estimated the size of such a database. With a multinomial distribution, we estimated first the sample size necessary to identify all the MDCs in each surname in our sample with an 80% probability, found that this value is strongly correlated with the population size of each surname, and used the regression line between sample size necessary and population frequency of each surname to extrapolate values for all Catalan surnames. We found that a total sample size of 37 368 males sampled from the 1 848 486 men carrying the 3173 most frequent Catalan surnames would be sufficient to establish a database in which all MDCs (with the sensitivity of our original samples) would be identified with an 80% probability. This is not an impossible sample size to reach, but these figures were derived under the assumption that the sampling was structured by surname *a priori*.

We have tried to dissect the intricate relationships between surnames and Y-chromosome diversity in Catalonia, which are the result not only of their shared inheritance but also of history, linguistics, and culture. Although similar studies have now been performed for three countries, it remains to be seen which details of this relationship are shared in other European societies and which of those societies harbor new levels of complexity.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We especially thank the >2500 participants in the study for giving their samples and genealogies, and, in many cases, for providing relevant details about the history of their surnames. Cristina de Vasconcelos superbly organized the logistics of sample collection. Cristina Junyent (www.cienciaensocietat.org) put her media contact list at our disposal, which was crucial for raising awareness for the project. Osmel Companioni (Institut Català d'Oncologia) advised us about the DNA sampling and extraction protocols. Irina Badell and Isshak Mrabet collaborated in DNA extraction. Advice on study design was provided by Joan Peitavi (Université de Perpignan), August Bover (Universitat de Barcelona), Enric Guinot (Universitat de València), Jordi Lleonart (Institut de Ciències del Mar), Anna Cabré and Joana Maria Pujades (Centre d'Estudis Demogràfics, project 'Five centuries of marriages', ERC-2010-AdG-26976), and Albert Turull (Universitat de Lleida). Funding was provided by the Institut d'Estudis Catalans and by the Spanish Ministry of Economy and Innovation (grant CGL2013-44351-P).

- King TE, Jobling MA: What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends Genet* 2009; **25**: 351–360.
- Poznik GD, Henn BM, Yee MC *et al*: Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* 2013; **341**: 562–565.
- Gusmao L, Sanchez-Diz P, Calafell F *et al*: Mutation rates at Y chromosome specific microsatellites. *Hum Mutat* 2005; **26**: 520–528.
- Sykes B, Irven C: Surnames and the Y chromosome. *Am J Hum Genet* 2000; **66**: 1417–1419.
- King TE, Jobling MA: Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. *Mol Biol Evol* 2009; **26**: 1093–1102.
- McEvoy B, Bradley DG: Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. *Hum Genet* 2006; **119**: 212–219.
- Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: Identifying personal genomes by surname inference. *Science* 2013; **339**: 321–324.
- Grimes BF: *Ethnologue. Languages of the world*. Summer Institute of Linguistics: Dallas, TX, USA, 1988.
- Mateos P, Tucker D: Forenames and surnames in Spain in 2004. *Names* 2008; **54**: 156–184.
- Rodriguez-Larralde A, Gonzalez-Martin A, Scapoli C, Barrai I: The names of Spain: a study of the isonymy structure of Spain. *Am J Phys Anthropol* 2003; **121**: 280–292.
- Moll F de B: *Els Llinatges Catalans*. Ed. Moll: Mallorca, Spain, 1982.
- Moreu-Rey E: *Antroponímia: Història dels Nostres Prenoms, Cognoms i Renoms*. Edicions Universitat de Barcelona: Barcelona, Spain, 1993.
- Mulero JJ, Chang CW, Calandro LM *et al*: Development and validation of the AmpFISTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. *J Forensic Sci* 2006; **51**: 64–75.
- Martínez-Cruz B, Ziegler J, Sanz P *et al*: Multiplex single-nucleotide polymorphism typing of the human Y chromosome using TaqMan probes. *Investig Genet* 2011; **2**: 13.
- Rocca RA, Magoon G, Reynolds DF *et al*: Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach. *PLoS One* 2012; **7**: e41634.
- Van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MHD: Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. *Hum Mutat* 2014; **35**: 187–191.
- Excoffier L, Lischer HEL: Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 2010; **10**: 564–567.
- R core team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2012.
- Martínez-González LJ, Martínez-Espín E, Álvarez JC *et al*: Surname and Y chromosome in Southern Europe: a case study with Colom/Colombo. *Eur J Hum Genet* 2012; **20**: 211–216.
- Martínez-Cruz B, Ioana M, Calafell F *et al*: Y-chromosome analysis in individuals bearing the Basarab name of the first dynasty of Wallachian kings. *PLoS One* 2012; **7**: e41803.
- Rebala K, Martínez-Cruz B, Tönjes A *et al*: Contemporary paternal genetic landscape of Polish and German populations: from early medieval Slavic expansion to post-World War II resettlements. *Eur J Hum Genet* 2013; **21**: 415–422.
- Fadhlaoui-Zid K, Haber M, Martínez-Cruz B, Zalloua P, Benammar Elgaaiet A, Comas D: Genome-wide and paternal diversity reveal a recent origin of human populations in North Africa. *PLoS One* 2013; **8**: e80293.
- Zhong H, Shi H, Qi X-B *et al*: Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. *Mol Biol Evol* 2011; **28**: 717–727.
- Dupanloup I, Bertorelle G: Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol Biol Evol* 2001; **18**: 672–675.

- 25 Romualdi C, Balding D, Nasidze IS *et al*: Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* 2002; **12**: 602–612.
- 26 Larmuseau MHD, Vanoverbeke J, Van Geystelen A *et al*: Low historical rates of cuckoldry in a Western European human population traced by Y-chromosome and genealogical data. *Proc Biol Sci* 2013; **280**: 20132400.
- 27 Boattini A, Sarno S, Pedrini P *et al*: Traces of medieval migrations in a socially stratified population from Northern Italy. Evidence from uniparental markers and deep-rooted pedigrees. *Heredity (Edinb)* 2015; **114**: 155–162.
- 28 Pérez-Molina I: *Honour and Disgrace: Women and the Law in Early Modern Catalonia*. Edicions Universitat de Barcelona: Barcelona, Spain, 2001.
- 29 Sahlins P: *Boundaries: The Making of Spain and France in the Pyrenees*. University of California Press: Berkeley and Los Angeles, CA, USA, 1989.
- 30 Cabré A: *El Sistema Català de Reproducció*. Editorial Proa: Barcelona, Spain, 1999.
- 31 Corominas J: *Onomasticon Cataloniae*. Curial, Caixa d'Estalvis i Pensions de Barcelona, La Caixa: Barcelona, Spain, 1989.
- 32 Hammer MF, Behar DM, Karafet TM *et al*: Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. *Hum Genet* 2009; **126**: 707–717.
- 33 King TE, Ballereau SJ, Schurer KE, Jobling MA: Genetic signatures of coancestry within surnames. *Curr Biol* 2006; **16**: 384–388.
- 34 Larmuseau MHD, Vanderheyden N, Van Geystelen A, van Oven M, de Kniff P, Decorte R: Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. *Ann Hum Genet* 2014; **78**: 92–103.
- 35 Solé-Morata N, Bertranpetit J, Comas D, Calafell F: Recent radiation of R-M269 and high Y-STR haplotype resemblance confirmed. *Ann Hum Genet* 2014; **78**: 253–254.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)