Current Biology *16*, 384–388, February 21, 2006 ©2006 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2005.12.048

Report

Genetic Signatures of Coancestry within Surnames

Turi E. King,[1] Stéphane J. Ballereau,[1] Kevin E. Schürer,[2] and Mark A. Jobling[1,*]
[1] Department of Genetics
University of Leicester
University Road
Leicester LE1 7RH
United Kingdom
[2] Department of History
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom

Summary

Surnames are cultural markers of shared ancestry within human populations. The Y chromosome, like many surnames, is paternally inherited, so men sharing surnames might be expected to share similar Y chromosomes as a signature of coancestry. Such a relationship could be used to connect branches of family trees [1], to validate population genetic studies based on isonymy [2], and to predict surname from crime-scene samples in forensics [3]. However, the link may be weak or absent due to multiple independent founders for many names, adoptions, name changes and nonpaternities, and mutation of Y haplotypes. Here, rather than focusing on a single name [4], we take a general approach by seeking evidence for a link in a sample of 150 randomly ascertained pairs of males who each share a British surname. We show that sharing a surname significantly elevates the probability of sharing a Y-chromosomal haplotype and that this probability increases as surname frequency decreases. Within our sample, we estimate that up to 24% of pairs share recent ancestry and that a large surname-based forensic database might contribute to the intelligence-led investigation of up to ~70 rapes and murders per year in the UK. This approach would be applicable to any society that uses patrilineal surnames of reasonable time-depth.

Results and Discussion

150 men carrying different British surnames were recruited through local and national advertisement. In the 1996 UK electoral registers, a total of ~5.75 million people (~13% of the population) carry these 150 surnames, and names of English, Scottish, and Welsh origin are represented in the set in approximately the same proportion as in the national population of Great Britain. A second cohort of 150 men, matching the surnames of the first and chosen randomly from electoral rolls, was also recruited.

*Correspondence: maj4@leicester.ac.uk

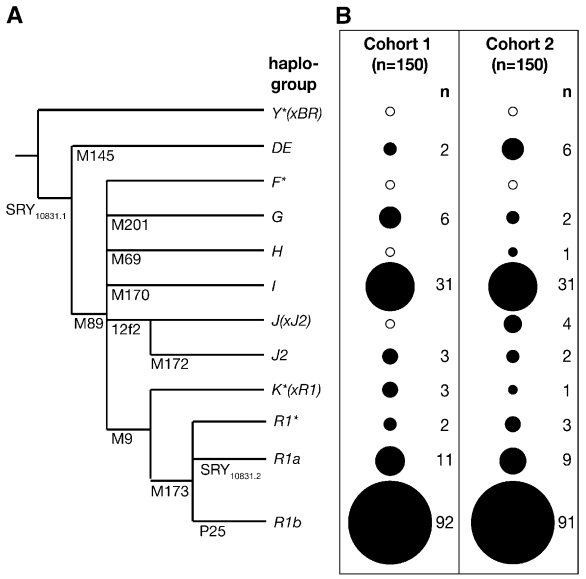


Figure 1. Y-Chromosomal Haplogroups in the Two Surname Cohorts

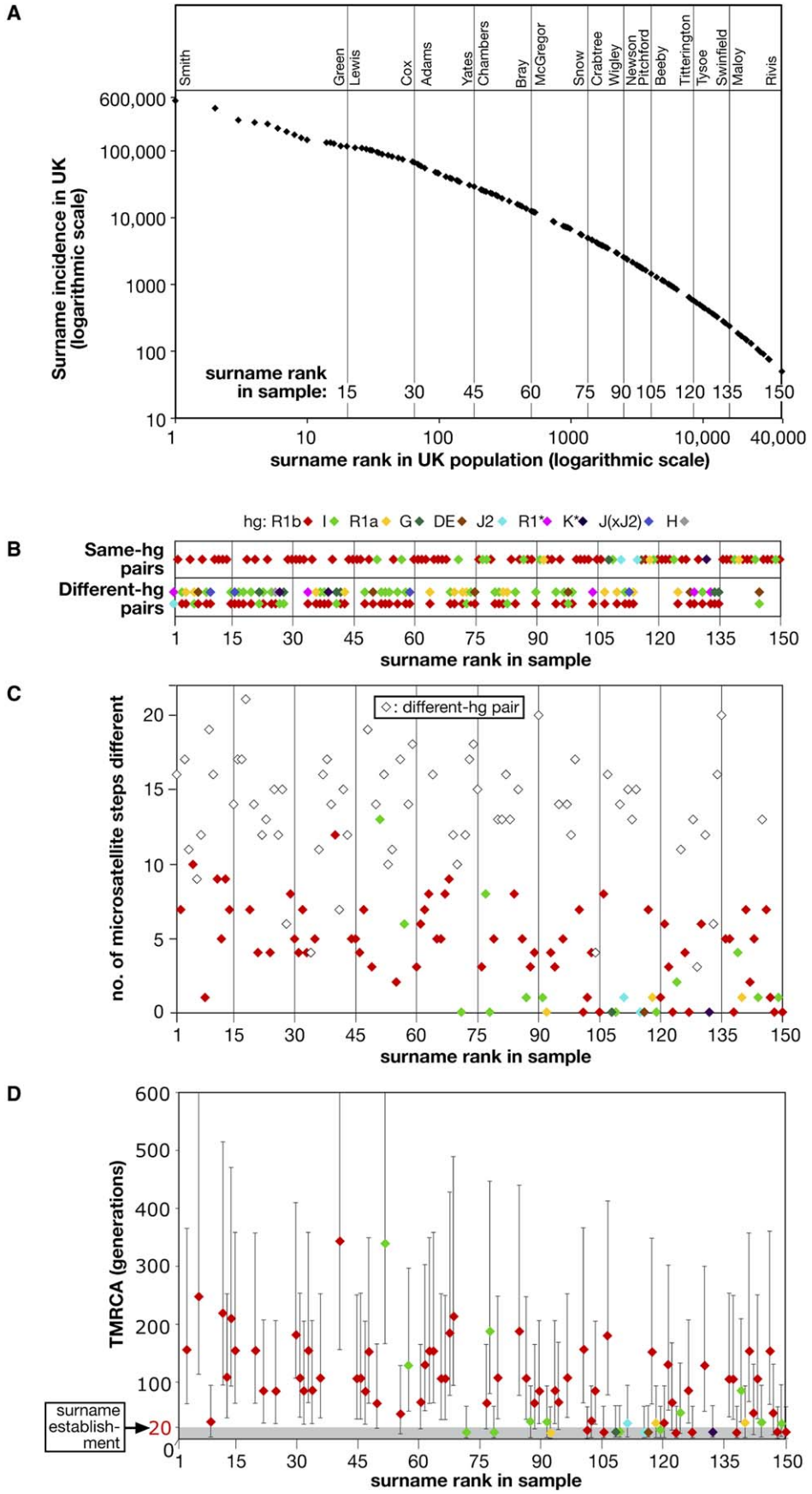
(A) Binary marker phylogeny of the Y chromosome, showing the typed mutations on the branches of the tree, and haplogroup names [7, 10] to the right.

(B) Haplogroup profiles of surname cohorts 1 and 2. Areas of filled circles are proportional to haplogroup frequency, and unfilled circles indicate unobserved haplogroups. Numbers of individuals are given to the right of circles.

We analyzed the nonrecombining region of the Y chromosomes of the two cohorts by using a set of 11 binary markers, defining a maximum number of 12 haplogroups (Figure 1A; see Table S1 in the Supplemental Data available with this article online), of which ten were observed (Figure 1B). Under a hypothesis of a perfect surname-Y chromosome correlation, the cohorts should have identical haplogroup compositions. Unsurprisingly, this is not so (Figure 1B), although, judged by a population differentiation test, they are not significantly different ($p = 0.325 \pm 0.022$).

The ages of mutations defining the haplogroups greatly predate the time of surname establishment (which is ~700 years [5]), as demonstrated by their widespread geographical distributions [6, 7] and by time-to-most-recent-common-ancestor (TMRCA) estimates [8]. Therefore, if the Y chromosomes of two men sharing a surname belong to different haplogroups, they cannot share recent common paternal ancestry. Of the 150 same-surname pairs, only 43% (65 pairs) fall within different haplogroups, while the average figure when pairs are permuted 1000 times is 57% (85 pairs). Thus, even for randomly ascertained pairs, sharing a surname significantly ($p < 0.001$) elevates the probability of sharing a haplogroup.

Since many rare surnames probably had single founders while common surnames had multiple founders


A
Smith
Green
Lewis
Cox
Adams
Yates
Chambers
Bray
McGregor
Snow
Crabtree
Wigley
Newson
Pitchford
Beeby
Titterington
Tysoe
Swinfield
Maloy
Rivis
600,000
100,000
10,000
1000
100
10
Surname incidence in UK (logarithmic scale)
surname rank in sample: 15 30 45 60 75 90 105 120 135 150
1
10
100
1000
10,000
40,000
surname rank in UK population (logarithmic scale)
B
hg: R1b I R1a G DE J2 R1* K* J(xJ2) H
Same-hg pairs
Different-hg pairs
1 15 30 45 60 75 90 105 120 135 150
surname rank in sample
C
◇: different-hg pair
no. of microsatellite steps different
20
15
10
5
0
1 15 30 45 60 75 90 105 120 135 150
surname rank in sample
D
TMRCA (generations)
600
500
400
300
200
100
0
surname establish-ment
20
1 15 30 45 60 75 90 105 120 135 150
surname rank in sample

[5], we next asked how surname frequency affects the probability of haplogroup sharing. The 150 surnames can be ranked by frequency from *Smith* (carried by 560,000 people) to *Rivis* (50 people) and cover a broad range of the frequencies found in the commonest 40,000 British surnames (Figure 2A).

Rare names are strikingly more likely to share haplogroups than are common names (Figure 2B). In the highest-frequency decile, only 7/15 surname pairs share a haplogroup, as opposed to 14/15 for the lowest-frequency decile ($p = 0.001$). In the high-frequency half of all surname pairs, 47% share a haplogroup, while in the low-frequency half the figure is 69% ($p < 0.01$). Furthermore, a greater proportion of the sharing observed within the high-frequency half probably occurs by chance, since it is overwhelmingly (91%) in hg R1b (Figure 2C), the most prevalent haplogroup in the population (Figure 1). By contrast, in the low-frequency half, only 65% of sharing is within hg R1b, and there are examples of sharing within the rare haplogroups (R1a, G, DE, J2, and K*), which strongly suggests that the sharing is due to common ancestry.

Haplotypes based on multiple Y-specific microsatellites represent more sensitive indicators of recent coancestry. They are highly variable and have much lower average population frequencies than do haplogroups, so chance sharing is less likely. We therefore determined 17-locus microsatellite haplotypes for the two cohorts (Table S1) and compared haplotypes within surname by considering the number of mutational steps between each member of a pair (Figure 2C). As expected, the mean number of mutational steps in the 65 different-haplogroup surname pairs is greater than that in the 85 same-haplogroup pairs (13.67 versus 4.05), although the ranges overlap.

There are 16 examples of same-surname identical haplotype pairs (zero mutational steps difference). All are also same-haplogroup pairs, and all but one (surname *Major*, carried by ~5600 people) lie in the lower-frequency half of the sample. These are all likely to indicate shared recent coancestry associated with surname: a permutation test in different-surname males (within-cohort) finds no examples of identical haplotype pairs ($p < 0.001$).

Haplotype pairs that differ by only a small number of microsatellite steps might also reflect surname-related coancestry, with divergence due to mutation. To estimate what proportion of the sample this represents, we calculated TMRCA [9] for each within-haplogroup pair (Figure 2D; Table S1). Assuming a 35-year generation time (see Experimental Procedures) and that surnames were established 700 years ago, the maximum expected time to a common surname ancestor for two men is 20 generations. From the proportion of the probability distribution for each TMRCA estimate lying below 20 generations, we estimate the most probable value for the proportion of surname pairs sharing coancestry as 11%. For the 16 surname pairs having identical haplotypes, the median value for TMRCA (11 generations) lies well within the time of surname establishment. For an additional 20 surname pairs, the median lies outside this time, but the lower bound of the TMRCA 95% credible region is less than 20 generations, and so up to 24% of the sample of surname pairs plausibly shares coancestry through shared surname.

We have taken the most conservative possible approach to examining surname-Y chromosome links by choosing the smallest sample size—a pair—and sampling completely randomly. The strong signal of coancestry observed in up to a quarter of the pairs under these conditions suggests that studies of larger samples within surnames (particularly the rarer names) are worthwhile. They should reveal clear associations with Y haplotypes, allowing more precise estimates of TMRCA, and inferences about founder numbers and historical nonpaternity rates. Our findings indicate that inbreeding coefficients estimated from isonymy [2] should be modified to reflect departure from a perfect relationship between surnames and genetics and that this modification could be made in a surname frequency-dependent manner. DNA-based genealogical research is a burgeoning area of privately commissioned genetic testing, and this study both validates the general approach and suggests a caveat: if two randomly ascertained men who share a surname often share a 17-locus microsatellite haplotype, then many more markers will need to be tested to support a more specific historical link.

Finally, this study allows a first judgement to be made about the feasibility of drawing forensically useful conclusions about surnames from Y haplotypes. From our preliminary data, we can ask what the chance is of correctly predicting a surname from a Y profile (we ignore haplogroup here, since routine forensic profiling utilizes only microsatellites). If we use Cohort 1 to represent a database of surnames and associated Y profiles and assign each Cohort 2 haplotype the surname(s) of the nearest Cohort 1 match(es), the correct surname is among the predicted names in 28 cases (~19%), with a mean number of only 1.3 predicted names (range 1–6), and a mean of 0.54 microsatellite step difference (range 0–3). The approach is most successful for less common names, with 27/28 correct predictions being made between *Major* (rank 71) and *Rivis* (rank 150), corresponding to a 34% chance of correct prediction in this subsample of 80 names. Individuals carrying the ~39,000 names within this frequency range represent ~42% of the population, and if we extrapolate from our estimated success rate to a potential ~25–65

Figure 2. Relationship between Y-Chromosomal Haplotype and Surname Frequency

(A) Distribution of frequencies of the 150 sampled surnames. Surnames bounding each frequency decile are shown at the top.

(B) Haplogroup sharing within surname pairs. The top panel shows same-haplogroup pairs ranked by frequency, with each symbol representing a pair, and color indicating haplogroup. The lower panel shows different-haplogroup pairs, with two differently colored symbols representing each pair. hg, haplogroup.

(C) Microsatellite mutational steps between haplotypes within surname pairs. Symbols for same-haplogroup pairs are colored as in (B); different-haplogroup pairs are indicated by unfilled symbols.

(D) TMRCA estimates for same-haplogroup surname pairs. Colored symbols (as in [B]) indicate the median, and bars indicate the limits of the 95% credible region for each estimate. The gray-shaded area represents approximate time since surname establishment (the past 20 generations).

no-suspect murders and ~300–400 no-suspect rapes that include unidentified DNA samples and remain undetected for a significant period each year in the UK, an idealized database containing the ~39,000 names with associated Y profiles could contribute to the intelligence-led investigation of up to 10 murders and 57 rapes per year. While we do not expect a perfect prediction system to emerge, surnames suggested by a Y-DNA profile could be combined with existing intelligence to allow a pool of suspects to be identified; Bayesian statistical adjustments could also be applied to predictions based on local demographic factors—for example, a requirement for local knowledge in a perpetrator could prioritize local surnames. The approach has additional benefits: it would have deterrent value because it targets individuals whose profiles are not on DNA databases, and it should therefore allow perpetrators to be apprehended early in their criminal careers. Though our sampling was in Great Britain, DNA-based surname prediction is in principle applicable to any society having diverse patrilineal surnames of reasonable time-depth.

Experimental Procedures

DNA Samples

In sampling, we avoided: (1) individuals for whom information collected about their paternal grandfather's surname, first language, and birthplace indicated recent name changes or origin outside the UK; (2) surnames that represent recognized spelling variants of other names in the set; and (3) surnames with incidence <50 (of which there are very many), because of the high probability of accidentally sampling closely related men; a specific questionnaire was also used to exclude close patrilineal relatives. Sampling was with informed consent and followed ethical review by the Leicestershire Research Ethics Committee (ref. 5796). DNA donors self-sampled buccal cells by using a cytology brush (Rocket Medical) and suspended them in 0.75 ml NDS (0.5 M EDTA, 10 mM Tris-HCl, 1%[w/v] sodium lauroyl sarkosine [pH 9.5]). At this stage, samples could be stored apparently indefinitely at ambient temperature without loss or degradation of DNA. DNA was extracted from 200 μl of suspension with the QiaAmp kit (QiaGen) according to manufacturer's instructions.

Y Haplotyping

Binary markers shown in Figure 1A [10] were typed in two multiplexes by the SNaPshot minisequencing procedure (Applied Biosystems) and an ABI3100 Genetic Analyzer (Applied Biosystems). Primer sequences were as described [11]. Note that the five chromosomes classified here as belonging to hg R1* have been previously shown [12] to be derived for the marker M269 and therefore to carry a reversion of the marker P25. 17 Y-specific microsatellites (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS434, DYS435, DYS436, DYS437, DYS438, DYS439, DYS460, DYS461, and DYS462) were typed in three multiplexes as described [13].

Analysis

Surnames were ranked by frequency according to information from the 1996 UK electoral registers, covering those aged 18 and over who register themselves to vote and including 43,776 million persons (a population coverage of ~96%). Geographical origins of surnames were estimated by K.S. from their historical distributions. Population differentiation tests were carried out with Arlequin [14]. Significance testing by permutations was done with programs written in PERL, and a p value of <0.001 was assigned when an observed value was not attained in 1000 permutations. TMRCA estimates (median value, and bounds of the 95% credible region) based on microsatellite haplotype differences for within-haplogroup pairs were calculated by the method of Walsh [9], implemented in Mathematica 4.2, with a mean per-locus, per-generation mutation rate of 0.002 [15, 16] and $\lambda = 1/5000$ under a single-step mutation model. Mean male generation time for the period after 1550 in England was estimated as 35 years, by adding the mean difference between ages of marriage partners to the mean age at maternity [17]. Generation time prior to 1550 is likely to be lower, but is difficult to estimate [18], so we use the conservative value of 35 years for the entire period since 1300. The approximate number of murders and rapes where unidentified DNA material might benefit from surname prediction analysis was estimated from current offending and laboratory submission rates and the success rate of the National DNA Database; in practice, the individual circumstances of each case affect whether samples are taken beyond initial database and/or suspect comparison.

Supplemental Data

One supplemental table can be found with this article online at http://www.current-biology.com/cgi/content/full/16/4/384/DC1/.

Acknowledgments

We thank all DNA donors, Elena Bosch and Susan Adams for assistance with SNP typing assays, Robert Feakes for help with programming, Malcolm Wells and Orlando Elmhirst for assistance with estimating offense numbers, Experian Ltd (Nottingham) for kindly supplying surname data from the 1996 Electoral Registers for the UK, and Chris Tyler-Smith and Jon Wetton for very helpful comments on the manuscript. T.E.K. was supported by a Wellcome Prize Studentship (grant no. 061129), S.J.B. by the Wellcome Trust, and M.A.J. by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (grant no. 057559).

Received: November 29, 2005
Accepted: December 21, 2005
Published: February 21, 2006

References

1. Jobling, M.A. (2001). In the name of the father: surnames and genetics. Trends Genet. *17*, 353–357.
2. Lasker, G.W. (1985). Surnames and Genetic Structure (Cambridge: Cambridge University Press).
3. Jobling, M.A., Pandya, A., and Tyler-Smith, C. (1997). The Y chromosome in forensic analysis and paternity testing. Int. J. Legal Med. *110*, 118–124.
4. Sykes, B., and Irven, C. (2000). Surnames and the Y chromosome. Am. J. Hum. Genet. *66*, 1417–1419.
5. McKinley, R.A. (1990). A History of British Surnames (London: Longman).
6. Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P., et al. (2000). Y chromosome sequence variation and the history of human populations. Nat. Genet. *26*, 358–361.
7. Jobling, M.A., and Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. Nat. Rev. Genet. *4*, 598–612.
8. Hammer, M.F., and Zegura, S.L. (2002). The human Y chromosome haplogroup tree: nomenclature and phylogeny of its major divisions. Annu. Rev. Anthropol. *31*, 303–321.
9. Walsh, B. (2001). Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics *158*, 897–912.
10. Y Chromosome Consortium (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. *12*, 339–348.
11. Bosch, E., Calafell, F., González-Neira, A., Flaiz, C., Mateu, E., Scheil, H.-G., Huckenbeck, W., Efremovska, L., Mikerezi, I., Xirotiris, N., et al. (2006). Male and female lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. Ann. Hum. Genet., in press. Published online December 22, 2005. 10.1111/j.1529–8817.2005.00251.x.
12. Adams, S.M., King, T.E., Bosch, E., and Jobling, M.A. (2006). The case of the unreliable SNP: recurrent back-mutation of Y-chromosomal marker P25 through gene conversion. Forensic Sci. Int., in press.

13. Bosch, E., Lee, A.C., Calafell, F., Arroyo, E., Henneman, P., de Knijff, P., and Jobling, M.A. (2002). High resolution Y chromosome typing: 19 STRs amplified in three multiplex reactions. Forensic Sci. Int. *125*, 42–51.
14. Schneider, S., Roessli, D., and Excoffier, L. (2000). Arlequin ver. 2.0: a software for population genetics data analysis, Genetics and Biometry Laboratory, University of Geneva, Geneva, Switzerland.
15. Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., and de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum. Mol. Genet. *6*, 799–803.
16. Gusmão, L., Sanchez-Diz, P., Calafell, F., Martin, P., Alonso, C.A., Alvarez-Fernandez, F., Alves, C., Borjas-Fajardo, L., Bozzo, W.R., Bravo, M.L., et al. (2005). Mutation rates at Y chromosome specific microsatellites. Hum. Mutat. *26*, 520–528.
17. Wrigley, E.A., Davies, R.S., Oeppen, J.E., and Schofield, R.S. (1997). English Population History from Family Reconstitution 1580–1837 (Cambridge: Cambridge University Press).
18. Goldberg, P.J.P. (2004). Medieval England. A Social History, 1250–1550 (London: Hodder Arnold).