Science and Society

In the name of the father: surnames and genetics

Mark A. Jobling

Hereditary surnames contain information about relatedness within populations. They have been used as crude indicators of population structure and migration events, and to subdivide samples for epidemiological purposes. In societies that use patrilineal surnames, a surname should correlate with a type of Y chromosome, provided certain assumptions are met. Recent studies involving Y-chromosomal haplotyping and surname analysis are promising and indicate that genealogists of the future could be turning to records written in DNA, as well as in paper archives, to solve their problems.

In most human populations, surnames, like DNA, pass down from generation to generation. People who share surnames might therefore be expected to share more of their DNA than people who do not. However, people sharing a most recent common ancestor a mere ten generations ago are only expected to share around a millionth of their DNA by direct descent, and there's no telling which millionth that might be. In societies where surnames are passed from fathers to children, though, matters are simpler. One part of our genome, the Y chromosome, is passed down in the same way as a surname; the distinction is that although all of the children get the name, only half – the sons – get the chromosome. Recent molecular studies have begun to explore the relationships between Y chromosomes and surnames, and interest is growing among amateur genealogists in the use of DNA to trace their ancestors. This article reviews the history of genetic studies of surnames and examines what the molecular analyses of the future are likely to tell us.

Origins of surnames

Most populations now use hereditary surnames, although the date of their establishment varies greatly around the world, from almost 5000 years ago in China, to only 68 years ago in Turkey. There is also variation among regions within countries and among social classes. In Japan, for example, the governing classes took hereditary surnames from the 13th century AD, but prohibited their use by other people until 1868 (Ref. 1). Some societies still do without them and use, for example, names based on father's forename (e.g. in Iceland), which therefore change each generation.

In the British Isles, at least, the origins of particular surnames usually fall into one of a few classes[2], albeit with some ambiguity: toponymic (from a specific place name; e.g. York, Lancashire), topographical (from a natural or man-made feature of the landscape, e.g. Hill, Townsend), from a personal name (e.g. Jones, Richardson), from a nick-name or characteristic (e.g. Grey, Wellbeloved) or from an occupation (e.g. Fletcher, Sawyer). Clearly, some surnames had multiple origins: Smith is the commonest surname in England and Wales, at about 1.3% (Refs 3,4), originating many times from the occupation of blacksmith. Most surnames are rare, however – 61% of the 32 457 different surnames within a sample of 165 510 individuals in England and Wales were unique within the sample[3] – and although some of these are fixed spelling variants of other names, and therefore relatively recent, many are likely to have had a unique origin at the time of surname establishment 500–700 years ago.

Surnames and genetics

The use of surnames in genetic studies dates back to 1875, when George Darwin, son of the more famous Charles, used them to estimate the frequency of first-cousin marriages[5]. He calculated the expected proportion of marriages between people of the same surname, based simply on surname frequency, and then ascribed the observed excess above this figure to marriages between cousins sharing surnames. He then adjusted this to account for cousins marrying but not sharing surnames and came up with the figures 4.5% for first-cousin marriages among the upper classes and ~2.25% for the general rural population. George had a personal interest in such liaisons, as his own parents were themselves first cousins. He was presumably reassured by the second part of his study, which suggested that the proportion of inmates of lunatic asylums who were the children of first cousins was not greatly different from the average proportion in the general population.

This work was extended much later and led to the development of a method for estimating inbreeding coefficients (the average probability that an individual inherits two copies of a gene that are identical by descent) from 'isonymy' – the frequency of marriages between individuals having the same surname[6]. This method became popular and widely used, because of the ease and cheapness of collecting large datasets relating to past and present populations from Births, Marriages and Deaths registers – 'the poor man's population genetics'[7].

Although the method sometimes agreed with estimates from other data, sometimes it did not, and this inconsistency is owing to the assumptions that underlie it. The isonymy method works when a founding population is a small group derived from a large outbred population and each founder has a different name. This is not likely to be so for most real cases[8], where founding groups are often from a small region, or even composed of relatives[9]. Furthermore, the isonymy method does not take into account the cumulative effects of inbreeding – sibs have the same surname whether or not their parents are themselves inbred.

Less controversial is the use of general information about place of origin contained within surnames: in studies of populations where a component represents an immigrant subpopulation (such as the Irish in Scotland[10], or Hispanics[11] or Chinese[12] in the USA) surnames provide an easy and effective means of subdivision for epidemiological purposes. Surnames have also been used to estimate population migration rates[13] and, in combination with genetic, linguistic and geographical data, to define barriers to gene flow[14].

http://tig.trends.com 0168–9525/01/$ – see front matter PII: S0168-9525(01)02284-3

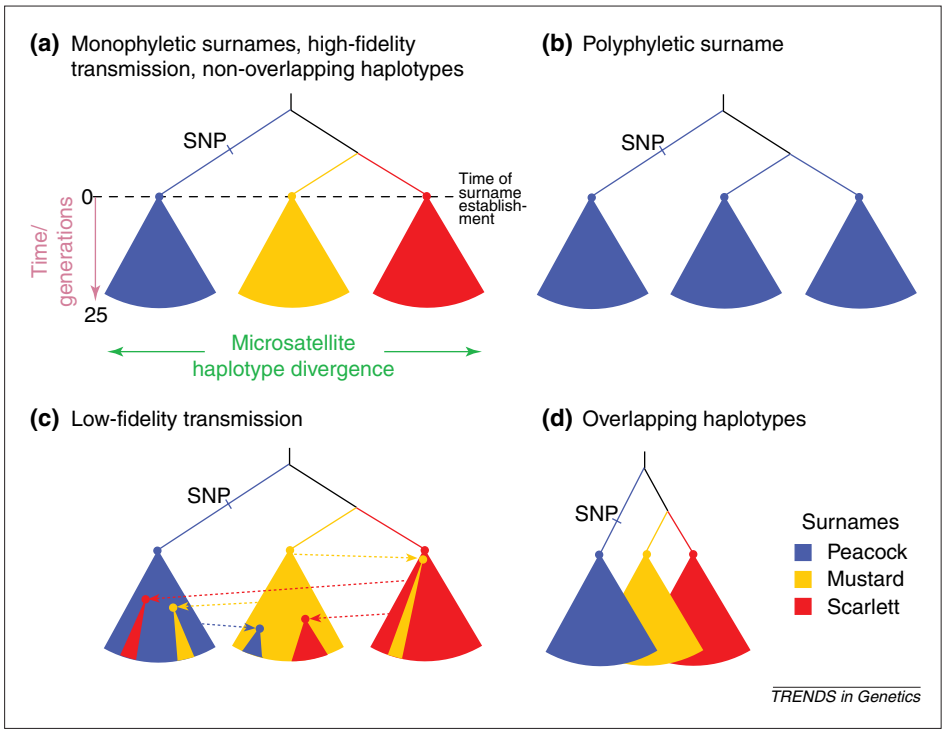

Fig. 1. Relationships between Y-chromosomal haplotypes and surnames. Each panel represents a simple hypothetical relationship between three surnames (Peacock, Mustard and Scarlett) and Y haplotypes under different founding and subsequent conditions. Small coloured circles represent founders, and coloured cones represent microsatellite haplotype divergence during descent from these founders. Surname foundation is taken to be 25 generations ago (time 0). (a) An ideal situation. Each surname has a unique founder whose haplotypes are highly diverged from the others. One (Peacock) is distinguished by a single-nucleotide polymorphism (SNP) as well as by microsatellite haplotype. There is no illegitimacy or other disturbance in transmission of surnames. (b) Polyphyletic Peacocks. (c) The effects of illegitimacy, surname adoption or maternal inheritance events (dotted lines) on the correlation between surnames and haplotypes. (d) The effects of close haplotype relationship of the three founders. Microsatellite haplotypes in the descendants overlap between all three surnames, but in the case of the Peacocks the SNP allows them to be distinguished easily.

Surnames and the Y chromosome

Apart from this use of surnames to infer general aspects of the genetic structures of populations, a more direct connection can be made between surnames and genetics in societies that have patrilineal surnames: in principle, a patrilineal surname should correlate with a type of Y chromosome (Fig. 1a).

Associating Y chromosomes with surnames is not a new idea and began before any DNA polymorphisms were available. Some very rare males carry a 'satellited' Y chromosome[15] (Yqs) bearing, on the tip of its long arm, a translocation from the short arm of an acrocentric autosome such as chromosome 15 or 22. This structure contains an active nucleolar organizer region, from which ribosomal DNA genes are transcribed. Such translocations are without any apparent deleterious effect and are detectable cytogenetically, so they provide easily scored neutral markers for particular rare Y chromosome lineages.

The most impressive case of a family bearing a Yqs is a 12-generation French-Canadian pedigree, in which the translocation arose at least 300 years ago and which provides the first example of a link between a surname and a Y-chromosomal lineage[16]. The chromosome was discovered during a karyotypic analysis of the father of a daughter with trisomy 21. A search among 50 men sharing the same surname ('R.') identified 17 who shared the same Yqs; genealogical research showed that all of these men were descended from Antoine R., a French barrel-maker who emigrated to Canada in 1665. Interestingly, a link could be made between the R. family and another French-Canadian family carrying an apparently identical Yqs but a different surname, through illegitimacy around 1830 (Ref. 16). In a separate study[17], a Yqs was found in four Colombian families, three of which shared a surname, although they did not know themselves to be related.

Over the past few years, a large number of more convenient polymorphic markers have been identified on the non-recombining portion of the Y chromosome[18,19]. These include slowly mutating binary polymorphisms [such as single nucleotide polymorphisms (SNPs)] that define monophyletic lineages [here called haplogroups (hg)], and more rapidly mutating multi-allelic markers (such as microsatellites) that define very large numbers of haplotypes within haplogroups and which can be used to estimate the ages of the most recent common ancestors of groups of chromosomes (e.g. Refs 20,21). The resolution of these systems is sufficient to distinguish most unrelated males within European populations from each other[22], so in principle they offer a means to identify a lineage that can be associated with a surname.

The general validity of the principle of associating Y chromosomes with surnames is shown by a study of Y-chromosomal haplogroup diversity in Ireland[23]. First, the Y chromosomes of 221 Irishmen were assigned to haplogroups. Then, to remove the effects of incursions from outside Ireland, chromosomes associated with surnames having English (e.g. Harrison, Kent), Scottish (e.g. Boyd, Knox), Norman (e.g. Bourke, Fitzgerald) or Norse (Doyle) origins were removed. The Y-chromosomal haplogroup composition of the remaining men with Gaelic surnames was different from the undivided set, with a higher frequency of one haplogroup, hg 1, in particular. When the Gaelic surnames were further subdivided according to the four counties of Ireland in which they originated about 1000 years ago, further structure was revealed: the four groups were significantly different and the westernmost group (Connaught) showed near fixation (98.5%) of hg 1 chromosomes.

In a different study[24], men belonging to the Jewish Cohanim priesthood, supposed to share patrilineal descent from Aaron who lived about 3000 years ago, were studied using Y-specific microsatellites and binary polymorphisms. A common modal haplotype was found in the Cohanim – the frequency of this haplotype and its one-step microsatellite mutation neighbours was >60%, compared with <15% in control groups. Not all of the Cohanim males shared surnames, although many were called Cohen, or related names such as Kahn and Kane, but this study does show that Y-chromosomal analysis can be used to demonstrate patrilineal descent within a group. A study from Korea[25] might be

considered a counter-example to this, provided we take stories about population origins literally: according to legend, the entire Korean population descended from a single man, 'Tangoon', who lived 5000 years ago, but analysis using a moderately variable Y-chromosomal polymorphism indicates that Korean Y chromosomes actually belong to multiple lineages.

Although the Irish study shows that a group of surnames carries a strong genetic signal of its place of origin, the correlation between an individual surname and a Y-chromosomal lineage is more problematic. For such a correlation to hold, a number of conditions must apply:

(1) The surname must have a unique origin. As is clear from the Smiths, this is likely to be untrue for many surnames (Fig. 1b).

(2) There must have been no illegitimacy, which would introduce chromosomes from other surname groups (Fig. 1c), as was demonstrated in the case of the French-Canadian Yqs (surname adoption or instances of matrilineal surname inheritance would have the same effect).

(3) Chromosomes associated with different surnames must have been unrelated at the time of surname establishment (Fig. 1d).

A study of a famous surname, that of Thomas Jefferson, third president of the United States, shows that males sharing the Jefferson surname share a high-resolution Y-chromosome haplotype, as defined by six binary markers, 11 microsatellites and a minisatellite; it also shows that paternity testing can be done with a time delay of several generations[26], suggesting that Jefferson fathered at least one child by his slave, Sally Hemings.

One recent study[27] has addressed some of the issues raised above more generally by examining the diversity of haplotypes defined by four Y-specific microsatellites in 48 men bearing the surname 'Sykes'. An unrooted network of these haplotypes, together with those found in 160 control non-Sykes men, is shown in Fig. 2. Twenty-one of the Sykes Y chromosomes belonged to a single 'core' haplotype, which was not found in any of the controls. The remaining Sykes chromosomes belonged to a wide range of different haplotypes, including two that were one mutational step away from the core haplotype. This pattern was interpreted as reflecting a single origin for the name Sykes (about 700 years ago), which seems surprising given its topographical derivation from a Yorkshire name for a stream or boundary ditch. Those Sykes chromosomes not belonging to the core haplotype originate either from illegitimacy or from mutation of the Y-specific microsatellites. Neglecting the latter, the average rate of illegitimacy was calculated at 1.3% per generation[27].

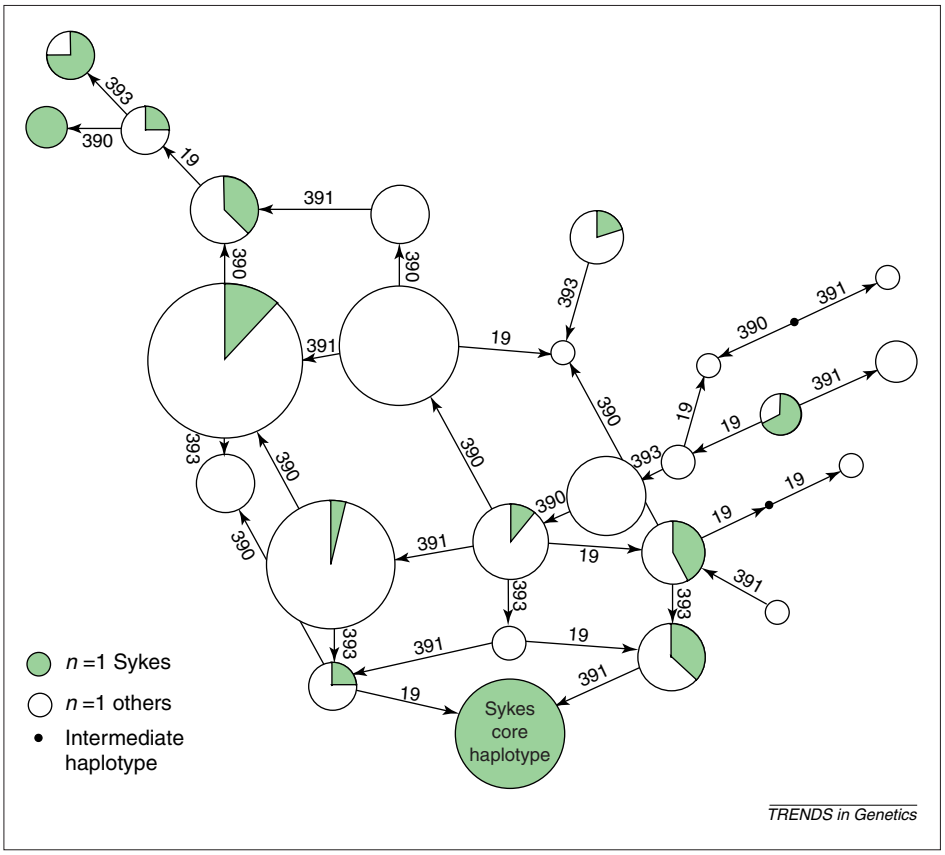

Fig. 2. Y-chromosomal haplotype diversity within the surname 'Sykes'. Median-joining network[43] of four-locus (*DYS19, DYS390, DYS391, DYS393*) microsatellite haplotypes. Haplotypes are indicated by circles, whose areas are proportional to haplotype frequency. The areas of green sectors within circles indicate the relative frequency of the surname Sykes. An arrow between circles indicates a single-step increase at a particular microsatellite, indicated next to the arrow (e.g. 19 indicates *DYS19*). Drawn from data in Ref. 27 using Network 2.0c (Ref. 43).

Haplotypes defined by only four microsatellites are of low resolution, and it is therefore something of a surprise that the core haplotype in the Sykes case is Sykes-specific. A clue comes from the distribution of this haplotype in a database of over 5000 European Y-specific microsatellite haplotypes[28] – it is common in the Baltic States, but rare or absent elsewhere. Also, its only occurrence in a set of 586 haplotypes from the British Isles, Scandinavia and Iceland[29] is in the single Irish chromosome belonging to hg 16, very rare in Western Europe, but common east of the Baltic[30], so this lineage might be unusual enough in Britain to be resolved by only a few markers.

The future

Further studies of individual surnames will show whether the Sykes are unusual and also how many markers we will need to discriminate between lineages sufficiently finely. At least 20 useful microsatellite markers are currently available[22,31], and the availability of the complete sequence of the 30 Mb Y-chromosomal euchromatin[32,33] provides a resource for the easy *in silico* identification of many new ones[31,34]. About 150 useful tri- to hexanucleotide repeat loci are estimated to be on the Y chromosome[31]. Binary markers can be used in combination with microsatellites: they can sometimes resolve cases where microsatellite haplotypes are overlapping (Fig. 1d), and their population-specificity in Europe[30] and elsewhere[35] will be useful in analysing surnames that are thought to reflect origins outside a particular region[2], such as the British names Fleming, Flanders and Brabant (from Belgium), or Gascoigne, Burgoyne and Champness (from France).

Interest in using DNA analysis as an adjunct to traditional historical research in

genealogy will continue to grow – there are already a number of commercial companies offering such services. Their customers will often hope that DNA evidence can provide a link between putative branches of a family that cannot be connected by other means. However, the finding of haplotypes matched between individuals needs careful interpretation and, in particular, a consideration of the frequency of the haplotype in a relevant population sample (often unavailable information), which can be high, even when as many as 16 microsatelllites are analysed[36]. Forensic geneticists have already learned these lessons[37].

The finding of mismatched haplotypes needs to be interpreted in terms of what is known about the mutation rates of the markers under consideration. Binary markers have low mutation rates ($\sim 2 \times 10^{-8}$ per generation[38]) and so a mismatch will almost always exclude a common ancestor in the past 25 generations. Microsatellites, however, have average mutation rates some five orders of magnitude higher[39,40], and we therefore expect to see mutations on this timescale.

A study[39] aiming to estimate Y-specific microsatellite mutation rates used nine microsatellites to analyse contemporary members of several deep-rooting pedigrees, each typically representing 6–8 generations of male-to-male transmission from a common ancestor and totalling 257 independent transmissions of the Y chromosome. Four instances of single-step differences at single microsatellite loci were observed within this sample, and these were interpreted as mutations. However, three cases had differences at more than one microsatellite locus. These three were interpreted as the results of non-paternity, rather than of multiple mutation, and this conclusion was supported when an independent polymorphic system, a minisatellite, was used to analyse the same individuals[41].

Although the use of, say, 50 microsatellites would greatly increase haplotype resolution and allow surname-specific haplotypes to be discerned, mutations would frequently be observed within families, and should be taken into account: a pair of males sharing a common great-grandfather have an approximately 50% chance of a mutation in at least one microsatellite in a 50-locus haplotype, assuming a mean mutation rate of 2×10^{-3} per locus[39]. Ideally, the question of whether a mismatch is owing to mutation, or represents an exclusion of a recent common ancestor, should be considered in terms of locus-specific mutation rates, and the statistics of such issues are not trivial. When a 'paper-trail' seems irrefutable, the DNA evidence might sometimes conflict with it, providing unwelcome news of an illegitimacy.

When all that is available from a crime-scene is what forensic scientists fastidiously refer to as a 'stain' (a sample of body fluid), the police are keen to know as much as possible about the person who left it. Although advances in genetics will allow the identification of genes directly involved in certain phenotypic traits such as pigmentation, a better understanding of Y-chromosome–surname relationships might one day also allow the surname of the depositor of the stain to be deduced[42], or at least a pool of suspects to be defined. Although Mrs Peacock and Miss Scarlett, lacking Y chromosomes, can relax, Colonel Mustard should await a knock at the door with trepidation and would be well advised to hide his lead piping.

Acknowledgements

I am a Wellcome Senior Research Fellow in Basic Biomedical Science (grant no. 057559). I thank E.W. Hill, T.E. King and Z.H. Rosser for assistance, and D.G. Bradley, M.E. Hurles and C. Tyler-Smith for helpful comments on the manuscript.

References

1 Yasuda, N. (1983) Studies of isonymy and inbreeding in Japan. *Hum. Biol.* 55, 263–276
2 McKinley, R.A. (1990) *A History of British Surnames*, Longman
3 Lasker, G.W. (1983) The frequencies of surnames in England and Wales. *Hum. Biol.* 55, 331–340
4 Lasker, G.W. (1985) *Surnames and Genetic Structure*, Cambridge University Press
5 Darwin, G.H. (1875) Marriages between first cousins in England and their effects. *J. Statist. Soc.* 38, 153–184
6 Crow, J.F. and Mange, J.F. (1965) Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Eugenics Q.* 12, 199–203
7 Crow, J.F. (1983) Surnames as biological markers – discussion. *Hum. Biol.* 55, 383–397
8 Rogers, A.R. (1991) Doubts about isonymy. *Hum. Biol.* 63, 663–668
9 Fix, A.G. (1999) *Migration and Colonization in Human Microevolution*, Cambridge University Press
10 Abbotts, J. *et al.* (1999) Association of medical, physiological, behavioural and socio-economic factors with elevated mortality in men of Irish heritage in West Scotland. *J. Publ. Health Med.* 21, 46–54
11 Stewart, S.L. *et al.* (1999) Comparison of methods for classifying Hispanic ethnicity in a population-based cancer registry. *Am. J. Epidemiol.* 149, 1063–1071
12 Choi, B.C.K. *et al.* (1993) Use of surnames to identify individuals of Chinese ancestry. *Am. J. Epidemiol.* 138, 723–734
13 Piazza, A. *et al.* (1987) Migration rates of human populations from surname distributions. *Nature* 329, 714–716
14 Zei, G. *et al.* (1993) Barriers to gene flow estimated by surname distribution in Italy. *Ann. Hum. Genet.* 57, 123–140
15 Schmid, M. *et al.* (1984) Satellited Y chromosomes: structure, origin and clinical significance. *Hum. Genet.* 67, 72–85
16 Genest, P. (1973) Transmission héréditaire, depuis 300 ans, d'un chromosome Y à satellites dans une lignée familiale. *Ann. Génét.* 16, 35–38
17 Giraldo, A. *et al.* (1981) A family with a satellited Yq chromosome. *Hum. Genet.* 57, 99–100
18 Jobling, M.A. and Tyler-Smith, C. (1995) Fathers and sons: the Y chromosome and human evolution. *Trends Genet.* 11, 449–456
19 Jobling, M.A. and Tyler-Smith, C. (2000) New uses for new haplotypes: the human Y chromosome, disease, and selection. *Trends Genet.* 16, 356–362
20 Zerjal, T. *et al.* (1997) Genetic relationships of Asians and northern Europeans, revealed by Y chromosomal DNA analysis. *Am. J. Hum. Genet.* 60, 1174–1183
21 Hurles, M.E. *et al.* (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am. J. Hum. Genet.* 65, 1437–1448
22 Kayser, M. *et al.* (1997) Evaluation of Y chromosomal STRs: a multicenter study. *Int. J. Legal Med.* 110, 125–133
23 Hill, E.W. *et al.* (2000) Y chromosomes and Irish origins. *Nature* 404, 351–352
24 Thomas, M.G. *et al.* (1998) Origins of Old Testament priests. *Nature* 384, 138–140
25 Kim, Y.J. *et al.* (1999) 49a/*Taq*I haplotypes according to the surname groups in Korean population. *Korean J. Genet.* 21, 181–192
26 Foster, E.A. *et al.* (1998) Jefferson fathered slave's last child. *Nature* 396, 27–28
27 Sykes, B. and Irven, C. (2000) Surnames and the Y chromosome. *Am. J. Hum. Genet.* 66, 1417–1419
28 Roewer, L. *et al.* (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forens. Sci. Int.* 118, 103–111
29 Helgason, A. *et al.* (2000) Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. *Am. J. Hum. Genet.* 67, 697–717
30 Rosser, Z.H. *et al.* (2000) Y-chromosomal diversity within Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.* 67, 1526–1543
31 Ayub, Q. *et al.* (2000) Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic Acids Res.* 28, 8e
32 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
33 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
34 Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580

35 Underhill, P.A. *et al.* (2000) Y chromosome sequence variation and the history of human populations. *Nat. Genet.* 26, 358–361
36 Mohyuddin, A. *et al.* (2001) Y-chromosomal STR haplotypes in Pakistani populations. *Forens. Sci. Int.* 118, 145–151
37 Balding, D.J. and Donnelly, P. (1995) Inferring identity from DNA profile evidence. *Proc. Natl. Acad. Sci. U. S. A.* 92, 11741–11745
38 Thomson, R. *et al.* (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7360–7365
39 Heyer, E. *et al.* (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* 6, 799–803
40 Kayser, M. *et al.* (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* 66, 1580–1588
41 Jobling, M.A. *et al.* (1999) Y-chromosome-specific microsatellite mutation rates re-examined using a minisatellite, MSY1. *Hum. Mol. Genet.* 8, 2117–2120
42 Jobling, M.A. *et al.* (1997) The Y chromosome in forensic analysis and paternity testing. *Int. J. Legal Med.* 110, 118–124
43 Bandelt, H-J. *et al.* (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48

M.A. Jobling
Dept of Genetics, University of Leicester, University Road, Leicester, UK LE1 7RH.
e-mail: maj4@leicester.ac.uk

Reprinted from *Trends in Genetics*, Volume 17,
MA Jobling, 'In the name of the father: surnames and genetics',
Pages 353-357