What's in a name? Y chromosomes, surnames, and the genetic genealogy revolution

Turi E. King and Mark A. Jobling

Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Correspondence can be addressed to either author:

Dr Turi E. King, Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Tel.: +44 (0)116 252 3377. Fax: +44 (0)116 252 3378. Email: tek2@le.ac.uk

Prof Mark A. Jobling, Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Tel.: +44 (0)116 252 3427. Fax: +44 (0)116 252 3378. Email: maj4@le.ac.uk

Keywords: surnames; Y chromosome; haplotype; haplogroup; genealogy

Running head: Surnames and Y chromosomes

This is the authors' revised personal version

Published in *Trends in Genetics* **25**, 351-360; http://www.trends.com/tig/

Access to published version: http://dx.doi.org/10.1016/j.tig.2009.06.003

Abstract

Heritable surnames are highly diverse cultural markers of coancestry in human populations. A patrilineal surname is inherited in the same way as the non-recombining region of the Y chromosome, and there should therefore be a correlation between the two. Studies of Y haplotypes within surnames, mostly of the British Isles, reveal high levels of coancestry among surname cohorts, as well as the influence of confounding factors including multiple founders for names, non-paternities and genetic drift. Combining molecular genetics and surname analysis illuminates population structure and history, has potential applications in forensic studies, and in the form of 'genetic genealogy' is an area of rapidly growing interest for the public.

Cultural markers of ancestry

Before Darwin, humans accorded themselves a special place in the kingdom of life. Now, 150 years after the publication of the *Origin of Species*, we can appreciate that we are part of the continuum of the evolution of all species, but our unique qualities remain undeniable. *Homo sapiens* literally means 'knowing man', but Linnaeus might equally have called us *Homo nominans* - 'naming man' – because of our capacity for complex language and our innate need to apply names to things, and to ourselves. Some of these names are heritable, and are recorded and persist through the generations. So, uniquely among organisms, many of us carry a cultural marker of coancestry, a surname, to go with the biological marker of coancestry common to all organisms, DNA.

In this review we examine the relationship between these two kinds of information: surnames and DNA. Because most heritable surnames pass from father to son, we focus on the relationships between surnames and paternally inherited Y-chromosomal haplotypes. Together with the recent revolution in the power of DNA analysis, the internet has introduced a new dimension in the way that this power can be made easily available to the public, and the way that surname information can be shared, exploited and understood. Most studies have focused on surnames in the developed world, and the British Isles in particular [1-4], and even reliable data on surname diversity are difficult to come by for many countries. Although this leads to an inevitable geographical and cultural bias, we hope that our description of principles and case studies will help to stimulate studies of a greater diversity of populations in the future.

History, inheritance and diversity of surnames

In human societies, having a name, and thus being identifiable, is essential. The addition of a heritable element facilitates identification, and also marks lineages, providing a label of regional and familial membership. Although some societies (such as that of Iceland) continue to eschew heritable surnames, governments like them, and in some countries have imposed them quite recently. For example, in Turkey all citizens were obliged to adopt a heritable surname in 1934, and in Mongolia a compulsory surname law was introduced in 1997. The earliest heritable surnames are those of China, dating back ~5000 years; time-depths for other nations vary (*Table 1*).

The diversity of heritable surnames also varies considerably; in China it is inconveniently low, as anyone who has carried out a PubMed search for a particular *Li* (the world's commonest surname) can testify [5], but in most countries it is amazingly high, with the mean number of bearers of any one surname well below 100 (*Table 1*). Some populations have high surname diversity because of a long history of admixture - this is certainly true of the USA. The current population of Great Britain has ~1.6 million surnames, but this value is much greater than that in the past, owing to recent immigration – the number listed in the 1881 census of England and Wales was only some 420,000. Though the derivations of surnames are often debatable, many fall into a limited number of classes, including patronyms (*son of…*) and those related to occupation, status or place-names (*Box 1*).

Patrilineal surnames and the Y chromosome

Given that DNA passes down to us from our ancestors together with surnames, people sharing surnames should have a greater than average

chance of sharing segments of DNA by descent than the general population. Although most DNA is inherited from both parents there is one segment, the non-recombining region of the Y chromosome, which is only passed down from father to son [6]. We might therefore expect that a surname should correlate with a type of Y chromosome, inherited from a shared paternal ancestor – perhaps the surname's original founder. A plethora of polymorphic DNA markers for distinguishing between Y chromosomes allows this idea to be tested; the types of marker and their properties are described in *Box 2*.

The simple expectation of a correlation between Y chromosome type and surname is complicated by several confounding factors. Some surnames are likely to have been founded independently more than once (*Figure 1*); this will result in more than one Y type being associated with a given surname. Non-paternity events, the adoption of male children and deliberate surname change will have the same consequence (*Figure 1*).

Mutation also acts to diversify the Y chromosome types associated with a particular surname, but, unlike the factors described above, its impact is relatively predictable. The mutation rates of single nucleotide polymorphisms (SNP) are low, so within the time-depths of surnames in most populations (~500-1000 years in most European populations) the widely typed SNPs are not expected to undergo mutations. By contrast, short tandem repeats (STRs) mutate rapidly, so mutations are relatively likely to be observed – indeed, our knowledge of their rates comes from identifying mutations within pedigrees [7] and father–son pairs [8]. The probability of detecting mutations within lineages depends on the number of STRs analysed, and also their individual properties.

Genetic drift – the random changes in haplotype frequencies over the generations – is the final factor that acts against the influences described above by reducing the diversity of haplotypes within surnames. For example, the stochastic variation in the number of sons fathered by different men can, over many generations, lead to the extinction of some Y chromosome lineages and the increase in the frequency of others within surname cohorts. Indeed, genetic drift (known in genealogical circles as 'daughtering out') is responsible for the complete extinction of some British surnames (such as *Campinot*) that had persisted for many generations [9].

Y chromosome diversity within surnames of the British Isles

Most detailed studies have focused on surnames of the British Isles. The pioneering and eponymous study of the surname *Sykes* [4] indicated low Y haplotype diversity among unrelated carriers of the name, suggesting that this was compatible with a single founder. However, its haplotype resolution (4 Y-STRs) was low.

The availability of more STRs and haplogroup-defining SNPs (*Box 2*) has allowed higher-resolution studies to be performed. A general link between surnames and Y-haplotypes was revealed in a study of 150 pairs of randomly ascertained men, each sharing a different British surname [1]. Sixteen of the 150 pairs shared identical 17-STR haplotypes, and 20 more pairs shared sufficiently similar haplotypes to suggest coancestry within the past 700 years – the average time since British surnames were established. Overall, the link is stronger the rarer the surname, with all pairs that show a strong signal of coancestry being found among the less common surnames (<5 600

bearers); this suggests that the commoner surnames had relatively large numbers of founders.

Two studies, in Britain [2] and Ireland [3], have collected and analysed larger groups of men with fewer surnames, using the same set of 17 Y-STRs, plus a number of haplogroup-defining SNPs. Both studies used networks to display and analyse diversity, with different approaches to defining 'descent clusters' of related haplotypes (*Box 2*). Both also estimated the time to most recent common ancestor (TMRCA) for clusters, finding ages compatible with the known time-depths of surname establishment. British control males carrying different surnames show very little haplotype sharing (*Figure 2a*), and the same is true of men carrying the commonest surname, *Smith* (*Figure 2b*). However, less common names show decreasing haplogroup diversity, and increasing degrees of STR haplotype sharing (*Figure 2c,d*): rare names (such as *Attenborough*) can be dominated by a single descent cluster (*Figure 2d*), which might indicate a single founder. However, the shallow time depth of many clusters within names, the absence of an effect of surname type on diversity, and computer simulations, together suggest a strong influence of genetic drift, such that current diversity is a poor reflection of the initial founder number [2].

Irish Y chromosomes show much lower haplogroup diversity than those of Britain, ~90% belonging to a single haplogroup, so most information is provided by Y-STRs [3]. Based on the same set of 17 STR markers [2], Irish controls carrying different surnames (like British ones) show very few shared haplotypes. However, within surname cohorts descent clusters are again evident, with an average of 61% of haplotypes within a surname lying in descent clusters – a very similar value to the British proportion of 62% [2].

Most of the variation between names was attributed to differences in founder numbers.

Comparison of the two studies reveals a striking difference between these neighbouring islands: the surname frequency-dependence of coancestry proportions evident in British names is absent from Ireland [2]. Some common Irish names such as *Ryan* (*Figure 2e*), borne by as much as 1% of the population, are dominated by single descent clusters, and, unlike in Britain, there is no significant correlation between a surname's rarity and the diversity of the Y chromosomes within it. The difference could be due to an amplification of genetic drift in Ireland, as a result of the prevalence of medieval patrilineal dynasties that linked male social and reproductive success in the past (discussed further below), but could also reflect other demographic historical differences, such as greater urbanisation in Britain and different impacts of epidemic disease.

These studies also highlight several factors that should be considered when systematic surname studies are carried out in other populations: (i) Sampling strategy needs to be planned carefully to avoid sampling related individuals; (ii) Geographical structure could affect diversity within sampled surnames, and its extent needs to be assessed [3]; (iii) Use of a standard set of Y-STRs would facilitate comparisons between studies, and, because of their convenience and high resolution, the commercially available profiling kits such as Y-filer (ABI) seem appropriate; (iv) The criteria for membership of descent clusters need careful consideration, since the boundary of a cluster is often not obvious. Our recommendation is to type binary markers as well as STRs, which will allow the definition of clusters within haplogroups that are rare in the population, and which therefore have relatively clear boundaries

[2]. The observed pattern of STR divergence within such clusters can be used to define a set of rules for cluster definition that can be more generally applied to common haplogroups. In some populations (e.g. Ireland) haplogroup diversity is currently inconveniently low for this approach [3], but new marker discovery should soon alleviate this problem; (v) Deduction of relevant generation times [10], perhaps from genealogical research in the populations under study, would aid in the accuracy of dating; (vi) Standardisation of Y-STR mutation rates would help in the estimation of TMRCAs across studies. The mutation rate derived from direct observation in father-son pairs (the 'pedigree rate'; ~2 x 10^{-3} per STR per generation [8]) is about three-fold greater than that derived from consideration of accumulated diversity within populations (the 'evolutionary rate' [11]), and studies have differed in which of these they apply leading to challenges in comparing studies [2, 3].

Applications of surname studies

The first application of surnames in genetics was in 'isonymy' studies, a field originated by Charles Darwin's son George, where they were used to estimate the degree of inbreeding in populations, based on the frequency of same-surname marriages [12], or on surname frequencies alone [13]. The underlying assumption, that a shared surname implies shared ancestry, has not been tested in most of the surveyed populations, and, as our previous discussion indicates, is often likely to be incorrect [14]. Despite such objections, the field of isonymy studies remains active; for a review, see Ref. 15.

Here we focus on three areas in which surname information has been combined with molecular genetic analysis to yield new insights.

Past population structure and history

Surnames tend to be specific to particular indigenous populations, and to show geographical specificity within regions. This property means that they find wide application as convenient proxies for ethnic origin [16] in health-care [17], epidemiological studies [18] and directed marketing [19]. However, combining surnames with Y-chromosome analysis has also allowed them to be used in genetic studies of historical migrations and admixture.

Much of this work has been carried out in the Irish population. For example, removal of individuals with non-Gaelic surnames in an analysis of Irish Y chromosomes leads to a significant change in haplogroup frequencies [20], and likely access to a more 'indigenous' sample and its population structure. A further link with the distant past is suggested by a common haplotype [21], interpreted to reflect the demographic impact of a medieval patrilineal dynasty, the *Uí Néill*. This 17-STR haplotype accounts for ~17% of Y chromosomes in the northwest of Ireland and is proposed to be the Y-lineage of a 5th century warlord, Niall of the Nine Hostages. This interpretation is supported by the over-representation of a descent cluster centred on the haplotype in 25 Irish surnames thought to originate in the *Uí Néill* dynasty.

The high reproductive success of this lineage seems to provide support for the idea of an amplification of genetic drift through social selection in the history of Ireland, adduced above to explain differences in haplotype diversity between Irish and British surnames. However, studies of multiple

surname groups thought to descend from two other patrilineal clans (*Eóganacht* and *Dál Cais*) show much less evidence of coancestry within either clan [22] and this suggests either that not all clans were really established by eponymous founders, or that in these cases the link between modern surnames and early origins has been severed. A broken link might also be suggested by an analysis of males with names of Norse Viking derivation (e.g. *Thunder*, *Doyle* and *Hanrick*), which reveals no difference from a general Irish sample [23], although this could also simply indicate that the Norse contribution in the Viking period (800-1200 CE) was very low.

The geographical differentiation of Y haplotypes is particularly marked in intercontinental comparisons. An association of a clearly African Y lineage with a rare English surname [24] provides evidence of a past African presence in Britain, and genealogical research connecting men carrying the surname and the exotic chromosome together allow a lower limit to be placed on its time-depth, during the mid-eighteenth century. In a different geographical context, observation of the low diversity of Y haplotypes in surname groups in Colombia demonstrates the powerful male-specific founder effects caused by Spanish and Portuguese colonisation [25].

Most population studies of Y-chromosome diversity categorise donors into local sub-populations on the basis of at least two generations of residence. However, this is compromised by migration in preceding generations. The geographical specificity of surnames suggests surname-based sampling as a means to choose modern Y chromosomes in a way that reflects their past population distributions [26]. This was done in a study of the Viking contributions to the Wirral peninsula and West Lancashire, in northwest England [27]. Historical and other evidence suggests colonisation

by Norse Vikings, beginning in 902 CE. Independent samples were recruited for each place: the 'modern' sample, based simply on two-generations of residence; and the 'medieval' sample, based on a history of residence plus the possession of a surname known from documentary evidence to have been present in the region prior to 1572 CE. The distributions of Y haplotypes in the two sample types were significantly different, and this could be accounted for by a greater Norse contribution to the 'medieval' samples, as judged by admixture analysis. This supports the idea that surname-based ascertainment provides a sample that more closely reflects past populations, prior to immigration from elsewhere.

Several studies of surnames and Y-haplotypes have used the diversity present within surnames to make inferences about the past rates of non-paternity [2-4, 25]. The assumptions and methods vary, but there is agreement that rates are <5% per generation, and in some cases <1% [25]. These rates are therefore consistent with modern estimates where there is no prior suspicion of non-paternity [28], and contradict the oft-quoted 'urban mythical' figure of 10% per generation.

Forensic application

The link between surname and Y-chromosomal haplotype suggests the idea of predicting a surname in forensic investigations [29]. In a case where an autosomal DNA profile yields no matches in a DNA database, a list of surnames with associated Y-STR haplotypes could allow a Y-profile to be matched with one or more surnames. This would provide a means to prioritise a suspect list; the surname prediction would act only as an investigative tool, since autosomal profiling could be used to exclude or

match individuals once they were identified. The validity of this approach has been confirmed in principle [1], but has yet to be used in practice; it might be compromised in the mixed and urban populations commonly encountered in criminal investigations. The link between surname and Y-haplotype is weak for common names (*Figure 2*), and including all rare ones is impractical, so the approach would be most useful for intermediate frequency surnames. In some cases, sharing of common haplotypes across surnames could result in many surnames being returned. In a sample of 1814 men carrying 164 names the commonest 17-STR haplotype was shared across 16 different surnames [30].

While surname prediction might have useful forensic applications, it also has the potential to infringe the privacy of those contributing DNA anonymously for medical research. For example, the surnames of the donors of the European members of the HapMap [31] DNA collection could be guessed at using published genotyping data and public databases of names and haplotypes [32]. In a highly publicised case, a 15-year old boy conceived by anonymous sperm donation traced his biological father by surname prediction through testing of his own Y chromosome, and exploiting public databases together with information on the father's date and place of birth [33].

Genetic genealogy and the rise of recreational genetics

Without doubt the most active area of exploitation of the link between surnames and Y-haplotypes is in the area of genetic genealogy, driven by the massive popular interest in family history, the availability of commercial DNA testing, and the ease of communication afforded by the internet. Many

companies offer Y-chromosome analysis, which is done using DNA extracted from buccal samples received from customers by post. More broadly, genetic genealogy forms part of 'recreational genetics', which includes the use of genome-wide markers to assess personal ancestry, relatedness and disease susceptibility, and this growing activity is also providing useful information for surname studies.

Directed commercial Y testing is usually seen as an adjunct to the traditional methods of genealogical research [34], and can, for example, show that two men with the same surname share a haplotype and therefore a recent common ancestor [35, 36]. Estimates of the time during which that ancestor lived [37] might also be offered, subject to considerable uncertainty. More generally, a group of men sharing a surname can collaborate to have their Y chromosomes analyzed, which can lead to the refinement of family trees, or the inclusion or rejection of branches for further genealogical investigation. Thousands of such 'surname projects' are currently in existence (*Box 3*). The size of Y-chromosome/surname datasets, often made freely available online by customers, is large (*Table 2*), and despite the possibly biased ascertainment of samples these represent a very useful general resource, and give opportunities for collaboration between the academic and amateur communities. One recent example is the characterisation of a set of novel SNPs within the generally rare hg G, which was facilitated by the easy identification and recruitment of DNA donors carrying hg G chromosomes via public genetic genealogy databases [38].

The interpretation of the relationships among customers' Y haplotypes depends on the haplotype resolution. Though some companies offer Y-SNP analysis, most offer only Y-STR typing, since this is highly discriminating and

universally applicable; the number of STRs typed varies from 15 to 67. Notably, 67 STRs is far more than are analysed in most academic studies, which are generally restricted by budgetary considerations to 20 or fewer. Generally speaking, the more markers typed the better (aside from the increasing probability of typing errors), since this reduces ambiguity in the interpretation of shared haplotypes. However, as the number of STRs increases so, too, does the probability of detecting an STR mutation between close relatives [29], and this needs to be taken into account.

Companies offering broader recreational genetics services use microarray-based methods to type up to ~1 million SNPs genome-wide, and return information to customers. The relevance for surname studies is that a proportion of the SNPs typed in these analyses are annotated as Y-linked (for example, 858 SNPs on the commercially typed Illumina 1M chip) and so they provide potential information about Y lineages. However, the SNP validation status and the correspondence with well-studied Y-SNPs [39] is in many cases unclear, and this is being resolved through the sharing of genotypic data from SNP chips among genetic genealogists [40, 41]. For example, haplogroup R1b1b2 is the commonest Y-lineage in western Europe, reaching over 90% in Ireland, and it has been difficult to find SNPs to subdivide it for population studies. The SNP rs34276300, known as S116, has been identified through comparing SNP chip results as a useful marker to subdivide hg R1b1b2, and is now being incorporated into academic studies. This is an area in which closer collaborations between amateurs and academics could prove particularly useful.

Members of the amateur community often display an impressive level of knowledge about aspects of molecular evolution, population genetics and

statistics; some of this is evinced in the quarterly online *Journal of Genetic Genealogy* (www.jogg.info). While it lacks the standard scientific peer-review system of traditional journals, it is nonetheless attracting academic geneticists among its authors, and is an interesting model for public involvement in scientific publication. Other resources for genetic genealogy are listed in *Box 3*. Thanks to the advances in DNA technology and the power of the internet, genetics is now joining astronomy as a science in which amateurs can make useful discoveries.

Genetic genealogy is fun, fascinating, and has much to contribute to academic science, but does it have any drawbacks? One obvious problem is the danger of detecting unexpected past non-paternities, or of having cherished oral histories disproven, both of which happened in the case of a family who believed themselves to be descendants of President Thomas Jefferson [42]. Although the Y chromosome is notoriously lacking in robust disease associations [6], some interstitial Y-chromosomal deletions (with incidences up to ~1 in 4000 males [43]) are certainly associated with male infertility [44], and can be signalled by the absence of specific Y-STRs and SNPs [45]. Beyond the genealogical aspects, the assignment of Y-lineages to particular geographical origins or ethnic groups can be misleading [46, 47]. None of these potential pitfalls seem likely to put off the customers of DNA typing companies, however.

Future developments

Sampling of a wider variety of populations and their surnames will help to alleviate the current geographical bias, and should lead to interesting new insights about social and demographic history. However, most new

advances will arise from exploitation of recent technological developments. Improvements to the methods of analysis of ancient DNA should allow the testing of genealogical links between living individuals and putative patrilineal ancestors, and also among archaeological human remains [48, 49]. High-resolution Y-typing and mitochondrial DNA sequencing together with whole-genome SNP analysis should allow reliable reconstructions of genealogies *de novo*, at least for the past few generations; this will include the establishment of links across the sexes, which cannot be achieved by the analysis of uniparentally inherited markers alone. In terms of relatedness, surname-ascertained cohorts of men who share Y-chromosomal coancestry lie between the traditional pedigree and the population, and application of whole-genome typing to such groups could be useful in understanding the history of recombination [50], and for genetic epidemiological purposes.

Recent application of conventional and 'next-generation' sequencing [51] technologies has revealed a large number of putative Y-SNPs in two named individuals, Craig Venter [52] and James Watson [53]. Such 'celebrity genomics' [54] projects will add further famous names to the webpages of genealogical geneticists, to join the motley crew of Genghis Khan, Thomas Jefferson, Marie Antoinette, Jesse James *et al.* (www.isogg.org/famousdna.htm). As the cost of sequencing continues to fall, private individuals will fund their own genome projects, and it seems inevitable that SNPs will be identified that are specific to particular surnames or their branches, providing powerful resources for genealogical research.

Acknowledgements

Our work is supported by the Wellcome Trust, including a Senior Research Fellowship in Basic Biomedical Science (grant number 057559) to MAJ. We thank Pablo Mateos for making available summary information on surname frequencies and distributions; Steve Archer, Patricia Balaresque, Francesc Calafell, Holly Eckhardt, Dafeng Hui, Andrew Nicoll, and Himla Soodyall for assistance; and three anonymous reviewers for helpful comments.

Figure Legends

Figure 1: Current Y diversity within a surname is influenced by founder numbers, non-paternity, genetic drift and mutation.

In this hypothetical genealogy all males share a patrilineal surname which originated 20 generations ago in two unrelated founding men carrying different Y haplogroups (hgs; see colour key top right), T and R1a, that themselves share common ancestry ~1600 generations [39] ago. Subsequently further diversity was introduced by non-paternity events, adoptions or surname changes (shown by stars, and the different haplogroup colours) or STR mutations (different shades of haplogroup colours). Diversity was reduced by genetic drift: all current hg T chromosomes within the surname descend from the original founder, whereas all current hg R1a chromosomes have a most recent common ancestor (MRCA) only 9 generations ago. In each case, white dots and bold lines indicate genealogical connections between current chromosomes and their MRCA. Current haplogroup diversity within the surname is very different from that in the general population [2] (pie charts to right, with sectors proportional to haplogroup frequency); in particular, hg T is not found in the general population sample, but represents 35% of the chromosomes in the surname sample.

Figure 2: Reduced Y chromosome diversity within surname groups.

Diversity of Y chromosome haplotypes among control males and five surname groups is represented by median joining networks. Circles within the networks represent Y haplotypes, with area proportional to frequency, and coloured according to haplogroup, as shown in the key top right. (a)

Control British males (n=110), all with different surnames, show high diversity and very few non-unique haplotypes. (b) British males sharing the most frequent surname *Smith* (n=58) resemble controls, with high diversity and little haplotype sharing. (c) British males with the medium-frequency surname *Jefferson* (n=85) show lower diversity and examples of shared haplotypes forming many descent clusters (dotted ellipses). (d) British males with the low-frequency surname *Attenborough* (n=31) show very low diversity, with 87% falling into a single descent cluster within hg E1b1b1. (e) Irish males with the common surname *Ryan* (n=62) show low diversity and a major descent cluster. (f) Irish males with the medium-frequency surname *McEvoy* (n=50) show higher diversity than those within *Ryan*.
For explanation of networks and descent clusters, see *Box 2*.

Box 1: Surname derivations

Most heritable surnames derive from a limited number of etymological sources [55, 56]. Here are some genetically relevant examples (British unless otherwise specified):

• Patronyms ('son of…'): *Bateson*, *Jeffreys*, *Watson*

• Clan or group membership: *Haldane* ('half-Dane'); *McKusick* (Irish - 'descendant of Isaac'), *Wallace* ('a Celt')

• Occupations or status: *Fisher* (fisherman), *Wright* (maker of machinery/objects), *Franklin* (feudal status term), *Chakraborty* (Indian – local landlord), *Müller* (German – miller)

• Specific places: *Charlesworth* (Derbyshire, England), *Darlington* (Co. Durham, England), *Crick* (Northamptonshire, England), *Pontecorvo* (near Rome, Italy), *Tsui* (ancient state of Xu, China)

• Landscape features: *Bridges*, *Ford* (river crossing), *Southern, Suzuki* (Japanese – pampas grass)

• Nicknames or characteristics: *Darwin* ('dear friend'), *Hodgkin* (pet form of *Roger*), *Sturtevant* ('hasty individual'), *Klug* (German – 'wise, prudent'), *Ochoa* (Basque – from *otxoa*, 'wolf')

Many surnames have one or more spelling variants; these were generally fixed relatively recently, when spellings were formalised [2, 57].

In Iceland, surnames are not heritable, but patronymic: the surnames of a son or daughter of the father *Stefán*, for example, will be *Stefánsson* and *Stefánsdottir*, and in the next generation the surnames will change again. Many heritable surnames in other countries have evolved from previously non-heritable patronymic systems.

Box 2: Markers for Y-chromosome diversity

Two types of polymorphic marker are commonly used to distinguish Y chromosomes from one another [6]. Binary markers such as single nucleotide polymorphisms (SNPs) have low mutation rates, typically ~10^{-8} per base per generation [58], and mostly represent unique events in human evolution. Short tandem repeats (STRs) are multiallelic markers, new alleles arising largely by single-step mutation at a typical rate of ~10^{-3} per STR per generation [8].

Binary markers are used in combination to define monophyletic haplotypes ('haplogroups'), which are arranged into a maximum parsimony tree [39, 59] containing major clades labelled A through T (*Figure Ia*). Each clade is further subdivided into alphanumerically named subclades (*Figure Ib*), the whole tree currently comprising 586 markers defining 311 haplogroups [39]. Application of new sequencing technologies (www.1000genomes.org) will yield thousands of new markers, and serious nomenclature problems, since the current system will become impossibly unwieldy. Some haplogroups are frequent in particular populations, and therefore provide relatively little discriminatory power.

The majority of widely used Y-STR markers are tri- and tetranucleotide repeats, of which there are >200 on the chromosome [60]. Combinations of Y-STRs (typed in PCR multiplexes) define more informative haplotypes within the haplogroups. Relationships among Y-STR haplotypes are often displayed in median-joining networks [61] (*Figure Ic*), which can also incorporate

haplogroup information. Closely related sets of haplotypes (typically found within surnames) define 'descent clusters', and, given an estimate of average STR mutation rates, time to the most recent common ancestor (TMRCA) for a cluster can be estimated [62].

Typing an STR multiplex is a highly efficient way both to distinguish between Y chromosomes and to indicate haplotype relationships, and can even be used to predict a haplogroup [63]. Each new haplogroup-defining SNP arose on a single chromosome, carrying a single Y-STR haplotype. Over time, mutation led to a limited repertoire of variation among the Y-STR haplotypes within this haplogroup, deriving from the founding haplotype [64]. The power of haplogroup prediction depends on the number of STRs typed, and, in some cases, specific diagnostic STR alleles. Distinguishing between closely related haplogroups is usually difficult, and, indeed, they may share identical Y-STR haplotypes, even when many STRs are typed. In such cases, SNP typing is essential.

Figure legend for Box 2:

Figure I: Y-chromosomal markers.

a) Phylogeny showing major haplogroups (A-T) defined by binary markers [39].
b) Detailed phylogeny of haplogroup I, showing SNPs on branches (not all are included) and alphanumeric names of sub-haplogroups [39].
c) Median-joining network of Y-STR haplotypes within a surname, indicating labelling conventions and examples of descent clusters (after [2])

Box 3: Resources for genetic genealogy studies

Aside from the Y-haplotype databases listed in Table 2, there are many useful resources for those interested in surnames and genetics, for example:

• **The International Society of Genetic Genealogy** (www.isogg.org) advocates the use of genetics as a tool for genealogical research, and provides a support network for genetic genealogists. It hosts the ISOGG Y-haplogroup tree, which has the virtue of being regularly updated.

• The ***Journal of Genetic Genealogy*** (www.jogg.info – and see text) publishes articles on individual surname studies, new methods of analysis, insights into mutation rates, geographic patterns in genetic data, and information that helps to characterise haplogroups.

• **The Guild of One-Name Studies** (www.one-name.org) exchanges and publishes information about one-name studies based on traditional historical and genealogical research, and including DNA information.

• Some DNA typing companies, including **Family Tree DNA** (www.familytreedna.com) and **DNA Heritage** (www.dnaheritage.com), host many 'surname projects'.

• **Wikipedia**'s pages on Y haplogroups (en.wikipedia.org/wiki/Human_Y-chromosome_DNA_haplogroups) provide up-to-date information on specific Y lineages, and useful information on particular SNPs can be found in the wiki-based **SNPedia** (www.snpedia.com). Details of Y-STRs and useful links are in **STRBase** (www.cstl.nist.gov/biotech/strbase/y_strs.htm).

• Information on how DNA information can be used in studying surnames can be found in popular books, including Smolenyak & Turner's *Trace your roots with DNA* [65] and Fitzpatrick's *DNA & genealogy* [66]

Table 1: Surname statistics for selected countries and populations.

Country or population (% of total surveyed)	**Mean no. of carriers / surname**	**Most frequent surname**	**% carrying most frequent surname**	**% carrying 10 most frequent surnames**	**Hereditary surname time-depth (years)**	**Ref.**
Great Britain (75)	28	*Smith*	1.22	5.9	~700 (England); ~300 (Wales)	[57, 67]
Ireland (69)	63	*Murphy*	1.71	10.5	~900	[67, 68]
Netherlands (28)	9	*De Jong*	0.54	3.7	~200	[26, 67]
Germany (35)	23	*Müller*	0.89	3.9	~700	[67, 69]
Norway (74)	29	*Hansen*	1.41	9.3	~100 (most of rural population)	[67, 70]
France (32)	17	*Martin*	0.33	1.6	~500	[67, 71]
Spain (21)	37	*Garcia*	3.66	19.8	~500	[67, 69]
Italy (27)	12	*Rossi*	0.33	1.5	~600	[67, 69]
India (0.3)	19	*Sharma*	2.44	12.8	complex history	[67]
Japan (35)	904	*Sato*	1.44	10.4	~800 (governing classes); ~150 (majority)	[67, 72]
China (22)	72195	*Li*	7.4	>30	~5000	[73]
Australia (37)	630	*Smith*	1.23	5.4	Most names imported from elsewhere	[67]
USA (23)	43	*Smith*	0.9	4.8	Most names imported from elsewhere	[67]
Canada (13)	17	*Smith*	0.31	2.0	Most names imported from elsewhere	[67]
Tristan da Cunha (100)	40	*Green*	0.23	100	Names imported from elsewhere	[74]
Lancaster County Amish (nk)	[a]	*Stoltzfus*	26	82	Names imported from elsewhere	[75]

[a] **:** 27 names in 5,538 households

Table 2: Publicly available Y haplotype databases.

Haplotype database	Description	Markers	Database size (no. of haplotypes)
Ybase (ybase.org)	Maintained by testing company (DNA Heritage); users can add their own data. Contains surnames, plus geographical, and genealogical information.	Up to 49 STRs plus haplogroups	14,462
Ysearch (ysearch.org)	Maintained by testing company (Family Tree DNA); users can add their own data. Contains surnames, plus geographical information.	Up to 100 STRs plus haplogroups	71,919
Sorenson Molecular Genealogy Foundation database (smgf.org)	Run by non-profit organisation carrying out DNA typing. Contains surnames, plus geographical and detailed genealogical information.	Up to 43 STRs	31,706
Y-STR Haplotype Reference Database (ystr.org)	Collaborative academic project run by the International Forensic Y-User Group [76]. Contains geographical population data only, and has global coverage.	Up to 17 STRs	72,055

References

1. King, T.E., *et al.* (2006) Genetic signatures of coancestry within surnames. *Curr. Biol.* 16, 384-388

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

3. McEvoy, B., and Bradley, D.G. (2006) Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. *Hum. Genet.* 119, 212-219

4. Sykes, B., and Irven, C. (2000) Surnames and the Y chromosome. *Am. J. Hum. Genet.* 66, 1417-1419

5. Wolinsky, H. (2008) What's in a name? *EMBO Rep.* 9, 1171-1174

6. Jobling, M.A., and Tyler-Smith, C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* 4, 598-612

7. Heyer, E., *et al.* (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* 6, 799-803

8. Gusmão, L., *et al.* (2005) Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.* 26, 520-528

9. Redmonds, G. (2004) *Names and History: People, Places and Things.* Hambledon and London

10. Fenner, J.N. (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* 128, 415-423

11. Zhivotovsky, L.A., *et al.* (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* 74, 50-61

12. Darwin, G.H. (1875) Marriages between first cousins in England and their effects. *J. Statist. Soc.* 38, 153-184

13. Lasker, G.W. (1985) *Surnames and genetic structure.* Cambridge University Press

14. Rogers, A.R. (1991) Doubts about isonymy. *Hum. Biol.* 63, 663-668

15. Colantonio, S.E., *et al.* (2003) Use of surname models in human population biology: a review of recent developments. *Hum. Biol.* 75, 785-807

16. Mateos, P. (2007) A review of name-based ethnicity classification methods and their potential in population studies. *Popul. Space Place* 13, 243–263

17. Wei, I.I., *et al.* (2006) Using a Spanish surname match to improve identification of Hispanic women in Medicare administrative data. *Health Serv. Res.* 41, 1469-1481

18. Quan, H., *et al.* (2007) Mortality, cause of death and life expectancy of Chinese Canadians in Alberta. *Can. J. Public Health* 98, 500-505

19. Webber, R. (2007) Using names to segment customers by cultural, ethnic or religious origin. *J. Dir. Data Digit. Mark. Pract.* 8, 226–242

20. Hill, E.W., *et al.* (2000) Y chromosomes and Irish origins. *Nature* 404, 351-352

21. Moore, L.T., *et al.* (2006) A Y-chromosome signature of hegemony in Gaelic ireland. *Am. J. Hum. Genet.* 78, 334-338

22. McEvoy, B., *et al.* (2008) Genetic investigation of the patrilineal kinship structure of early medieval Ireland. *Am. J. Phys. Anthropol.* 136, 415-422

23. McEvoy, B., *et al.* (2006) The scale and nature of Viking settlement in Ireland from Y-chromosome admixture analysis. *Eur. J. Hum. Genet.* 14, 1288-1294

24. King, T.E., *et al.* (2007) Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. *Eur. J. Hum. Genet.* 15, 288-293

25. Bedoya, G., *et al.* (2006) Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. *Proc. Natl. Acad. Sci. U. S. A.* 103, 7234-7239

26. Manni, F., *et al.* (2005) New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. *Am. J. Phys. Anthropol.* 126, 214-228

27. Bowden, G.R., *et al.* (2008) Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England. *Mol. Biol. Evol.* 25, 301-309

28. Anderson, K.G. (2006) How well does paternity confidence match actual paternity? Evidence from worldwide nonpaternity rates. *Curr. Anthropol.* 47, 513-520

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

30. King, T.E. (2007) The relationship between British surnames and Y-chromosomal haplotypes. Ph.D. thesis, University of Leicester

31. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437, 1299-1320

32. Gitschier, J. (2009) Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *Am. J. Hum. Genet.* 84, 251-258

33. Motluk, A. (2005) Anonymous sperm donor traced on internet. *New Scientist* 2524, 6

34. Roderick, T.H. (2000) The Y chromosome in genealogical research: "From their Ys a father knows his own son". *Natl. Geneal. Soc. Q.* 88, 122-143

35. Trumme, T., *et al.* (2004) Genetics in genealogical research - reconstruction of a family tree by means of Y-haplotyping. *Anthropol. Anz.* 62, 379-386

36. Kayser, M., *et al.* (2007) Relating two deep-rooted pedigrees from Central Germany by high-resolution Y-STR haplotyping. *Forensic Sci. Int. Genet.* 1, 125-128

37. 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

38. Sims, L.M., *et al.* (2009) Improved resolution haplogroup G phylogeny in the Y chromosome, revealed by a set of newly characterized SNPs. *PLoS ONE* 4, e5792

39. Karafet, T.M., *et al.* (2008) New binary polymorphisms reshape and increase resolution of the human Y-chromosomal haplogroup tree. *Genome Res.* 18, 830-838

40. Turner, A. (2008) SNPs on chips: a new source of data for Y chromosome studies. *J. Genet. Geneal.* 4, iii-iv

41. Athey, T.W., and Wilson, J.F. (2008) Y-SNP rs34134567 defines a large subgroup of haplogroup G2a-P15. *J. Genet. Geneal.* 4, 149-150

42. Williams, S.R. (2005) Genetic genealogy: the Woodson family's experience. *Cult. Med. Psychiatr.* 29, 225–252

43. Kuroda-Kawaguchi, T., *et al.* (2001) The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.* 29, 279-286

44. Vogt, P.H. (2005) AZF deletions and Y chromosomal haplogroups: history and update based on sequence. *Hum. Reprod. Update* 11, 319-336

45. King, T.E., *et al.* (2005) Inadvertent diagnosis of male infertility through genealogical DNA testing. *J. Med. Genet.* 42, 366-368

46. King, T.E., *et al.* (2007) Thomas Jefferson's Y chromosome belongs to a rare European lineage. *Am. J. Phys. Anthropol.* 132, 584–589

47. Bandelt, H.J., *et al.* (2008) The brave new era of human genetic testing. *Bioessays* 30, 1246-1251

48. Gerstenberger, J., *et al.* (1999) Reconstruction of a historical genealogy by means of STR analysis and Y-haplotyping of ancient DNA. *Eur. J. Hum. Genet.* 7, 469-477

49. Keyser-Tracqui, C., *et al.* (2003) Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia. *Am. J. Hum. Genet.* 73, 247-260

50. Coop, G., *et al.* (2008) High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319, 1395-1398

51. Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics. Hum. Genet.* 9, 387-402

52. Levy, S., *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.* 5, e254

53. Wheeler, D.A., *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872-876

54. Check, E. (2007) Celebrity genomes alarm researchers. *Nature* 447, 358-359

55. Hanks, P., and Hodges, F. (1988) *A Dictionary of Surnames*. Oxford University Press

56. Reaney, P.H., and Wilson, R.M. (1997) *A Dictionary of English Surnames*. Oxford University Press

57. McKinley, R.A. (1990) *A history of British surnames*. Longman

58. 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

59. Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* 12, 339-348

60. Kayser, M., *et al.* (2004) A comprehensive survey of human Y-chromosomal microsatellites. *Am. J. Hum. Genet.* 74, 1183-1197

61. Bandelt, H.-J., *et al.* (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37-48

62. Forster, P., *et al.* (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* 59, 935-945

63. Schlecht, J., *et al.* (2008) Machine-learning approaches for classifying haplogroup from Y chromosome STR data. *PLoS Comput. Biol.* 4, e1000093

64. de Knijff, P. (2000) Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am. J. Hum. Genet.* 67, 1055-1061

65. Fitzpatrick, C. (2005) *DNA & genealogy*. Rice Book Press

66. Smolenyak, M., and Turner, A. (2005) *Trace Your Roots with DNA*. Rodale Press

67. Longley, P., *et al.* (2009) publicprofiler/worldnames. http://www.publicprofiler.org/worldnames/

68. MacLysaght, E. (1969) *The surnames of Ireland*. Irish University Press

69. Currer-Briggs, N. (1982) *Worldwide family history*. Routledge and Kegan Paul

70. Morgan, F.E. (1981) *The significance of Norwegian " farm names "*. Norwegian American Museum

71. Morlet, M.-T. (1997) *Dictionnaire étymologique des noms de famille*. Perrin

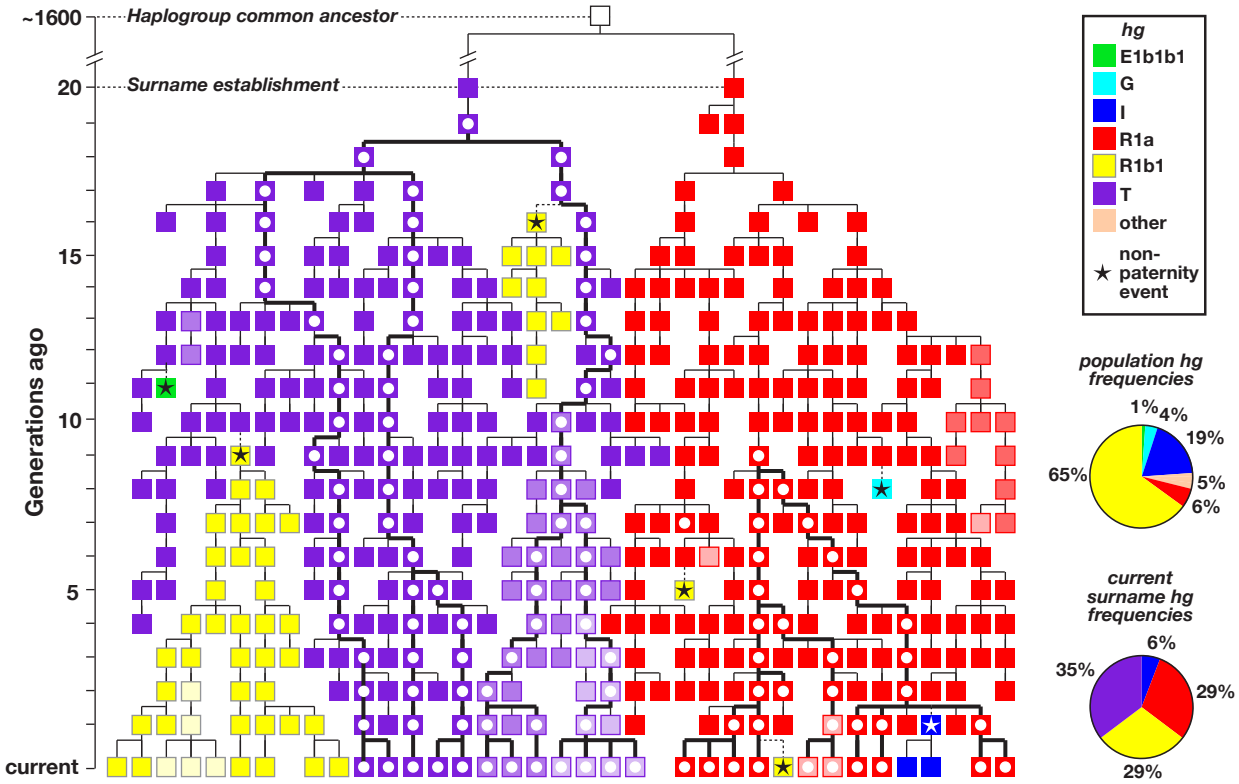
72. Yasuda, N. (1983) Studies of isonymy and inbreeding in Japan. *Hum. Biol.* 55, 263-276

73. Yuan, Y. (2007) Science and culture of surnames. *Chinese Nat. Geog.* 2, 38–39

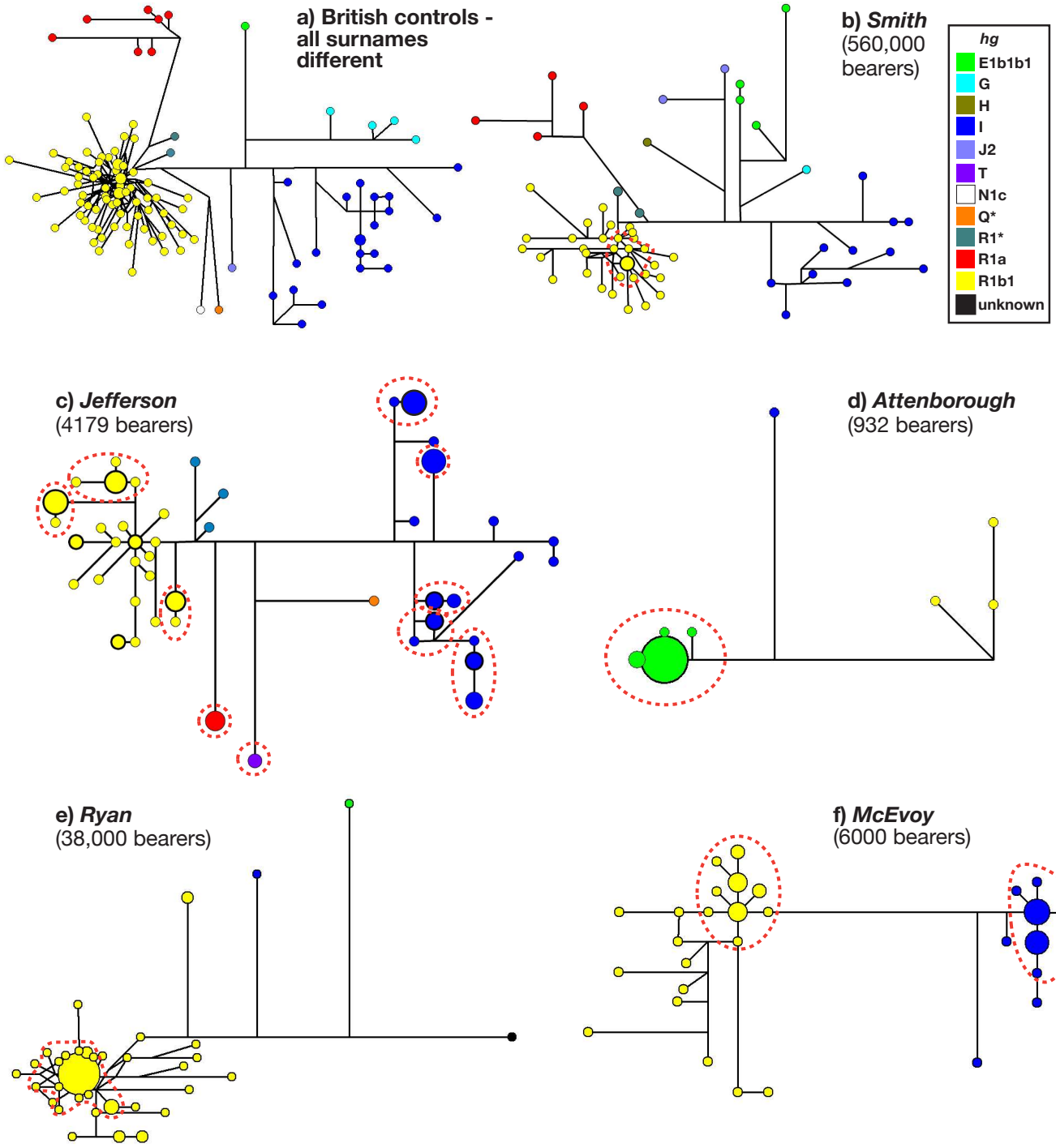
74. Soodyall, H., *et al.* (2003) Genealogy and genes: tracing the founding fathers of Tristan da Cunha. *Eur. J. Hum. Genet.* 11, 705-709

75. Pollin, T.I., *et al.* (2008) Investigations of the Y chromosome, male founder structure and YSTR mutation rates in the Old Order Amish. *Hum. Hered.* 65, 91-104

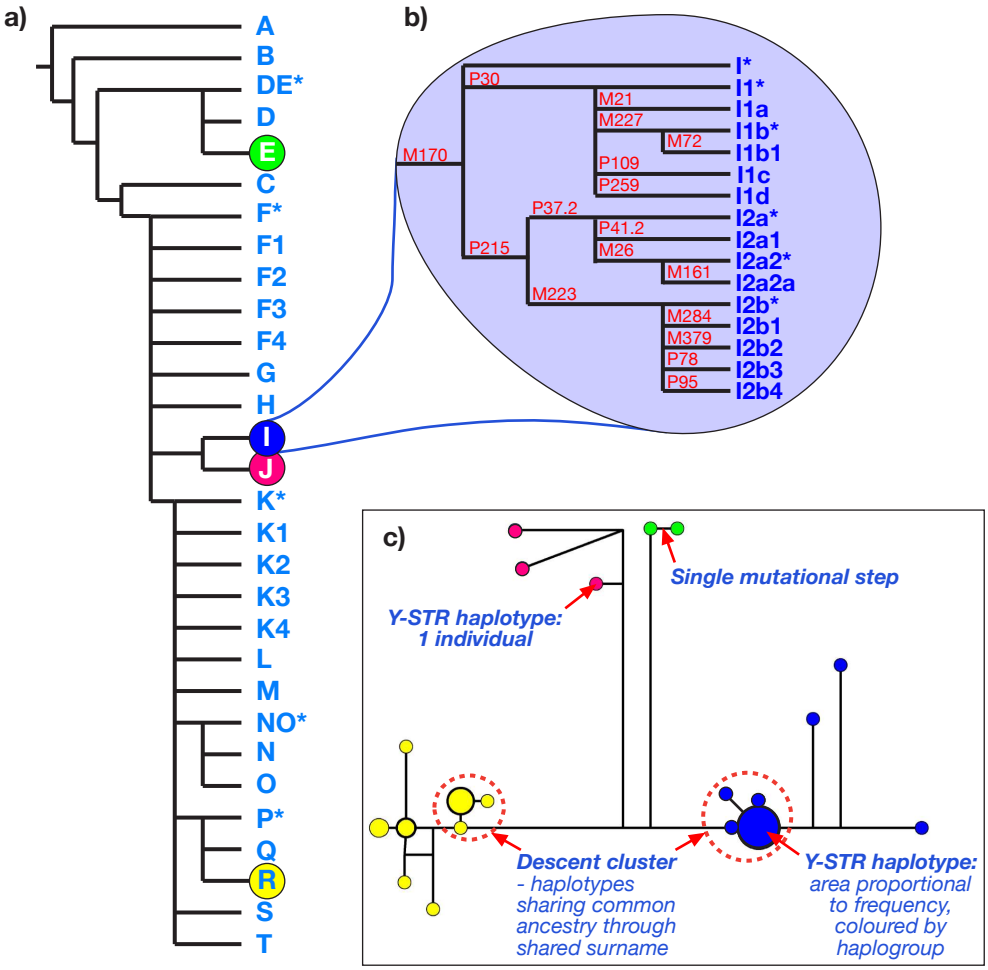
76. Willuweit, S., *et al.* (2007) Y chromosome haplotype reference database (YHRD): update. *Forensic Sci. Int. Genet.* 1, 83-87



King & Jobling, Figure 1



King & Jobling, Figure 2



King & Jobling, Figure for *Box 2*