

If two males have the same surname, there’s a chance that they are related—descended from the first male of their line to adopt the surname and pass it on to his children. Until last year, though, there has been no published data that allowed us to estimate the chances of such a relationship. Not surprisingly, the chances vary with the frequency of the surname in the general population, and the probability that two males bearing the surname are thus genetically related may be expressed in terms of the percentages of yDNA-tested males with a common surname who fall into cousin clusters.

Surname	Rank (US)	Est #Persons			% in Clusters	Largest and Lesser Cluster%	
		Total (in 1000s)	Total (in proj)	= Ratio			
Smith	1	2,238			15.5	15.5	
Walker	20	451	544	1.21	24	6.9	5.8, 4.4, 4.1 ...
King	25	405			8.3	8.3	
Phillips	37	337	429	1.27	76	5.7	3.0, 2.2, 1.9 ...
Hayes	100	173					
Perkins	184	120	178	1.48	75	22	19.0, 7.4, 4.4 ...
Harrison	200	113	131	1.16			
Jefferson	475	52			64.3	9.5	
Goldstein	500	51					
Bray	914	30			0	0	
Tuttle	1000	28					
Stanford	1500	19					
Ricks	2000	14					
Stead	>2000	??			76.1	28.3	
Clare	>2000	??			60.6	24.3	
Wadsworth	>2000	??			63.5	32.7	

The above table, which I may add to from time to time, was inspired by a recently published King-Jobling study^[1] (hereinafter “K/J”), which seeks to shed light on the question of how likely it is that two males with the same surname are related genetically. To that end, the authors picked 40 English surnames, and sampled the haplotypes of at least 40 bearers of those surnames, chosen at random, on 17 ySTR markers. Since the 17 markers they chose, unlike the FTDNA 37-marker panel, were inadequate to definitively sort people into [patrilineage](#) groups, they were obliged to devise a set of crude rules for doing this, which diminishes the quantitative significance of their results.

Nonetheless, the study clearly showed, at least for uncommon surnames, that the common surname itself creates a high likelihood of a genetic relationship. All but a very few of the surnames they tested, in fact, were what we in America would call uncommon; only three them even place in the top 2000 in the 1964 U.S. Social Security data. And for surnames of this class, K/J found that about 62% of the surname samples sorted into patrilineage clusters, with the mean size of the largest cluster representing 41% of the total sample (I derived this number myself from their data).

This latter number is extremely significant for our interpretation of yDNA surname project test results, because the calculations for Genetic Distance (GD) and Time of Most Recent Common Ancestor (TMRCA) do not, and cannot take the surname factor into account. They cannot, at least, without consulting surname frequency statistics, and without analyzing enough data to allow the derivation of a function showing the relationship between surname frequency and clustering.

¹ Turi E. King and Mark A. Jobling, “[Founders, drift and infidelity: the relationship between Y chromosome diversity and patrilineal surnames](#)”, *Mol Biol Evol.* 2009 May; 26(5): 1093–1102

It is with that need in mind that I have undertaken a very modest extension of the K/J experiment, using the far more definitive data to be found in the FTDNA surname projects.

The frequency data, and estimated number of persons bearing the surname comes from figures worked up by the Social Security Administration from their 1964 data, as reported in Elsdon C. Smith, *American Surnames* (1969; reprint Genealogical Publishing Co, 1994).

The clustering data for the surnames Smith, King, Stead, Clare, and Jefferson (listed in order of frequency in England) are taken from the K/J study. Those for Perkins, Phillips, and Walker, were derived from the corresponding FTDNA surname projects. And the other surnames are included to provide some surname frequency context.

I derived my clustering numbers by selecting from the most prevalent haplogroups found in the project, all the FTDNA 37- or 67-marker haplotypes, truncating the 67's to 37. I ran these data through Dean McGee's Y-Utility, following my [standard procedure](#).

The clustering numbers themselves were derived as percentages of the total subgroup I examined, which was sorted by my procedure into patrilineage clusters and singletons.

The ratios I've calculated between the total estimated numbers of people bearing the instant surnames in 1964 to the number of project members included in my study indicate (as we should expect) that project size is roughly proportional to surname frequency.

Discussion

Both the 0% clustering of the Bray data, and the 15.5% number for the most common surname, Smith, show, I think, the inadequacy of the mutation-rate-challenged K/J study, at least for these ultra-common surnames. The data for the ultra-common surnames, Walker and Phillips, provide, I think, a much better approximation to the true state of affairs for ultra-common surnames, at least for their American incarnations. Emigration to America constituted a kind of funneling process: selection of a small subset of the British patrilineages for each surname, followed, in many cases, by proliferation of those surnames in America, with the degree of proliferation a function of the date of immigration. Practically, this means that British migrants to New England, nearly all of whom were in place by 1650, dominate the American surname frequency spectrum.

Although the actual surname frequencies can thus be expected to vary considerably between Britain and the U.S., I believe that the very limited data above are already beginning to suggest a simple function (with, probably, a smooth curve with a single slope) relating surname frequency to largest cluster size. No doubt individual surname values are going to fall all around this curve, but the existence of such a function (if it does exist) should provide some basis for American-rooted surname yDNA projects to estimate the likelihood that members of their most common patrilineages are, in fact, related, *independent of any testing*. And where this likelihood is high, it should encourage a spirit of inclusiveness when sorting tested members into patrilineage groups.

I also expect that another function will emerge allowing one to estimate roughly the number of significantly sized patrilineage groups to expect for each surname, and I expect that number to be less than 10, and in most cases less than 5 for most surnames. This is clearly indicated by the K/J data, and it is doubtful that any amount of further study will contravene these general results.

My work with much better articulated FTDNA data (even though this data figures to be contaminated somewhat by selection bias) just as clearly shows the uselessness, and indeed the meaninglessness of the overall clustering percentage. A cluster, by definition, consists of more than one member, but whether sampling turns up one or two people of the same patrilineage, is largely a function of the sample size, and is thus irrelevant to the function relating surname frequency to

genetic relationship. Nevertheless, I shall continue, for now, to include the overall clustering % in my table, just to retain this additional link to the K/J study.

Conclusions

The major finding of my data so far is that **the largest cluster size for ultra-common surnames starts at about 7% and rises rapidly (by the time of Perkins, the 184th most common) to 22%**. Obviously, many more such surname analyses are needed to fill in what I think will be a fairly regular curve relating surname frequency to largest cluster size, and to gain a corresponding feel for the typical distribution of the next largest clusters.

I believe that the best additional candidates for analysis from amongst the FTDNA surname projects, would be those in the range of 30th to 500th or so, which are not actively managed or promoted. I would stick to the larger projects (and more frequent surnames) to minimize the moderate self-selection bias that can be expected to surface through the heightened recruitment from well-worked out, and well-known lines.