Patrilineage Clustering within Surname

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. Or in other words, that they belong to the same genealogical patrilineage.

The degree of clustering by surname patrilineage is of some importance to the administrators of Y-chromosome (ySTR) DNA surname projects, like the 8,000+ that have been organized under the auspices of Family Tree DNA. This is because in classifying members of these projects by patrilineage, or in estimating the TMRCAs for patrilineages (or just of haplotype pairs as FTDNA's Tip calculator does), the analyst needs some feel for how much to weight the factor of the shared surname—a factor which calculators do not, and cannot, properly take into account.

The King and Jobling Study

In 2009, a pair of British researchers published the first study on the prevalence of patrilineage clustering within surname.[1] K&J chose 40 English surnames more or less arbitrarily, and sampled the haplotypes of at least 40 bearers of each of those surnames, chosen at random, on 17 ySTR markers. Since the 17 markers they chose (unlike the FTDNA 37-marker panel) were inadequate for definitively sorting people into genealogical patrilineages (most of whose members typically have a surname in common), the authors were obliged to devise a set of crude rules for inferring patrilineage, which somewhat fuzzes their quantitative results

Nonetheless, the study clearly showed, at least for the relatively uncommon British surnames that it favored, that two males with the same surname chosen at random have a significant chance of belonging to the same genealogical patrilineage.

I've characterized the K&J surnames as "uncommon" because only three them even place in the top 2000 surnames that comprise the 1964 U.S. Social Security data.[2] For surnames of this class, K&J found that about 62% of the surname samples constituted patrilineage clusters of two or more haplotypes, with the mean size of the largest cluster representing 41% of the total.

Extending the K&J results to the U.S., and to more prevalent surnames

Being curious about how this degree of patrilineage clustering would compare with that of some of the more common surnames in the American-centric Family Tree DNA surname project, I made my own surname classifications for several of these larger projects, and created the following summary table for comparison with the K&J results.

In this table, 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 Allen, Perkins, Phillips, and Walker, were derived from my analyses of the corresponding FTDNA surname projects,[3] as does the McFarlane data, although for that case I've relied on the patrilineage classifications of the project administrator, and have only reported on the clustering percentage of the largest McFarlane cluster. The other surnames in the table are provided as reference points.

[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.

[2] According to Elsdon C. Smith, *American Surnames* (1969; reprint GPC, 1994)

[3] All of these analyses were made years ago, and the projects have no doubt grown significantly since, though presumably without deviating much in their proportions.

The Data

Surname	Rank	Estimated #Persons			% in Clusters		
	(US)	in 1964/in project			of any	largest	lesser
		in 1000s		Ratio	size	size	sizes (2nd,3rd…)
Smith	1	2,238			15.5	15.5	
Walker	20	451	544	1.21	24	6.9	5.8,4.4,4.1 ...
Allen	23	427	304	.71	75	10.8	10.5,5.6,3.9 ...
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	
McFarlane	494	51	377	7.39	-	27	
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 surname frequency data, and the estimated number of persons bearing the surname comes from *American Surnames*, cited above. The "Ratio" is that of the number of yDNA tested project members to the approximate number of person bearing that surname in the general population as of 1964, expressed in 1,000s.

I derived my clustering percentages from the subset of haplotypes what had tested to at least 37 markers.

Discussion

The lumpy data for the K&J surnames (0% for Bray; 15.5% for the ultra-common surname King; the anomalously small largest cluster for Jefferson, and here is the full table of K&J data) illustrate, I think, the inadequacy of the 17 marker haplotype tested by K&J. The much larger data bases data for the ultra-common surnames, Walker and Phillips (tested to at least 37 markers) provide a much better approximation to the true state of affairs for ultra-common surnames, at least for their American incarnations. And they strongly suggest that **for surnames beyond the very most common, the clustering percentage of the largest cluster can be expect to rise rapidly from the 5-10% for the most common surnames to 20-30% for moderately common surnames, and level off there.** This falls well short of the 40% average for the K&J largest cluster, but then, the K&J surnames are nearly all beyond the 2000 most frequent, and it may be that the American data would climb gradually to that average as surname frequencies decrease.

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. And there are many other factors that differentiate the American population from the British. It is unclear at this point, how, or for that matter whether, there are significant differences between patrilineage clustering within surname in Britain, and in America.