

# High resolution Y-chromosome analysis of surname association with paternal relatedness

S. R. Woodward<sup>1,2</sup>, N. M. Myres<sup>1</sup>, J.B. Ekins<sup>1</sup>, J.E. Ekins<sup>1</sup>, L. A. D. Hutchison<sup>1</sup>, L. Layton<sup>1</sup>, M. L. Lunt<sup>1</sup>, S. S. Masek<sup>1</sup>, A. A. Nelson<sup>1</sup>, M. E. Nelson<sup>1</sup>, K. L. Pennington<sup>1</sup>, U. A. Prego<sup>1</sup>, J. L. Peterson<sup>1</sup>, K. H. Ritchie<sup>1</sup>, T. Tolley<sup>1</sup>. 1) Sorenson Molecular Genealogy Foundation, Salt Lake City, UT; 2) Department of Molecular Biology, Brigham Young University, Provo, UT. e-mail: scott@smgf.org

## Abstract

The correspondence of surname and relatedness has been used as an indicator of population subdivision in both epidemiological and population migration studies. Recently, a limited number of studies have used the Y-chromosome to assess the validity of using surnames as a measure of paternal relatedness due to its analogous paternal transmission. These studies have reported varied results depending on the particular surname, the resolving power of the marker system, and the population(s) under investigation. To further characterize the correlation of surname and paternal relatedness we have analyzed 13,489 male samples representing 9,444 surnames as 36 Y-chromosome STR loci. We report surname frequencies and classify non-unique surnames as mono or polyphyletic. Paternal genealogies were used to identify potential ancestral and/or geographic origins for each lineage within each of the polyphyletic surname groups. These findings have application to estimating non-paternity rates, compiling forensic databases, population genetic studies, epidemiological studies and ancestry testing.

## Background

Surnames, as currently understood and used in the Western world, have been used in some populations for between 500 and 1000 years. Broadly, most surnames fall into four categories. 1. Surnames derived from first names include Johnson, Williams, and Thompson. Most often they are patronymic, referring to a male ancestor, but occasionally they are matronymic. 2. Occupational surnames refer to the occupation of the bearer. Examples include Smith, Clark, and Wright. 3. Locational or topographic surnames are derived from the place that the bearer lived. Examples include Hill, Woods, and Ford and 4. Surnames derived from nicknames include White, Young, and Long. It would be expected that at the time of adoption of a surname there may not have been any correlation between the selection of a surname and the the Y-chromosome haplotype. In Europe, surnames began to be used in the 12th century, but it took several centuries before the majority of Europeans had one. The primary purpose of the surname was to further distinguish people from one another. In the 13th century about a third of the male population was named William, Richard or John. To uniquely identify them, people began referring to different Williams as William the son of Andrew (leading to Anderson), William the cook (leading to Cook), William from the river (leading to Rivers), William the brown-haired (leading to Brown), and so on. Eventually these surnames became inherited, being passed from parents to children.

Y-chromosome haplotypes have been used for paternal line identification in forensic, paternity and ancestry testing. 36 locus haplotypes allow for high resolution differentiation between paternal lineages.

## Material and Methods

DNA samples of 13,489 males were selected from the Sorenson Molecular Genealogy Foundation (SMGF, [www.smgf.org](http://www.smgf.org)) database. 36 Y-chromosome STR were typed using standard PCR and capillary electrophoresis methods (1). Descriptions of the loci including frequency data are found at [http://smgf.org/marker\\_details.html](http://smgf.org/marker_details.html).

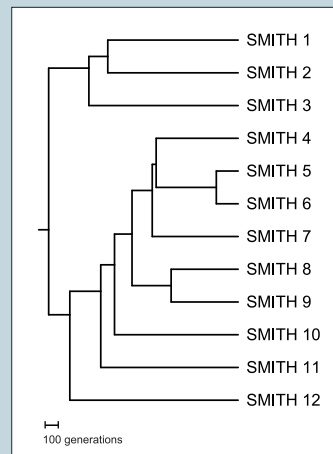
**Surname Frequencies**

Surname	Samples	relative frequency	frequency	US Census freq
SMITH	138	1,000	0,010	1,000
JOHNSON	92	0,667	0,007	0,800
JONES	75	0,543	0,006	0,621
BROWN	72	0,522	0,005	0,621
TAYLOR	62	0,449	0,005	0,311
WILLIAMS	53	0,384	0,004	0,690
MILLER	50	0,362	0,004	0,424
DAVIS	37	0,268	0,003	0,480
WILSON	27	0,196	0,002	0,339
MOORE	19	0,138	0,001	0,312

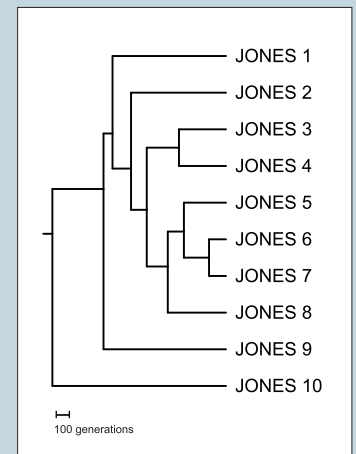
**Table 1.** Top ten non-unique surnames in the SMGF database. Including a frequency normalized on the Smith surname and a comparison of the frequency of the surnames in the 1990 US Census.

## Results

The 13,489 male genotypes in the database represent 9,444 unique surnames resulting in 2,529 non-unique surnames. The top ten non-unique surnames and their number of occurrences represented include: Smith (138), Johnson (92), Anderson (88), Jensen (87), Jones (75), Brown (72), Christensen (71), Hansen (70), Taylor (62). This is not a random sample of surnames throughout the world but skewed towards Scandinavian and United Kingdom surnames, reflecting the demographics of the SMGF database at the time of sampling. When the frequencies of the common SMGF surnames are normalized on the most frequent surname (Smith) the representation of the other surnames is consistent with the frequencies of the 1990 US Census (4). Table 1. From the list of non-unique surnames, test for mono or polyphyletic origins were performed for the Smith, Jones, and Wells groups using methods within BATWING (2) and PHYLIP (3). All groups were shown to be polyphyletic within the time frames of surname usage (30 - 50 generations). Figures 1 and 2. Although there are individuals in the database that share surnames and also share Y-chromosome haplotypes, these individuals can also be shown by traditional genealogical methods to belong to the same lineage.



**Figure 1.** Tree of the Smith surnames in the SMGF database. Scale represents 100 generations. All haplotypes are consistent with a Y-chromosome haplotypes that predate the origin of the Smith surname.



**Figure 2.** Tree of the Jones surnames in the SMGF database. Scale represents 100 generations. Again it is shown that the Y-chromosome haplotypes predate the origin and use of the Jones surname.

## Conclusions

It is not entirely unexpected that groups of males that share non-unique surnames may have polyphyletic origins based on Y-chromosome haplotypes. The non-unique Y-chromosome haplotypes have coalescence times and TMRCA ranging between 750 and 1200 generations which far exceed the time frame for the beginning of the use of surnames. Y-chromosome haplotypes reflect relationships at a much deeper number of generations than traditional western surnames.

## References

1. Ekins, J.B, et al. Inference of Ancestry: Constructing hierarchical human reference population and assigning unknown individuals. Submitted to Human Genomics.
2. Wilson, Weale & Balding 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. Journal of the Royal Statistical Society: Series A (Statistics in Society), 166: 155-188.
3. Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.
4. US Census 1990.

