Inferential Genotyping of Y Chromosomes in Latter-Day Saints Founders and Comparison to Utah Samples in the HapMap Project

Jane Gitschier[1,*]

One concern in human genetics research is maintaining the privacy of study participants. The growth in genealogical registries may contribute to loss of privacy, given that genotypic information is accessible online to facilitate discovery of genetic relationships. Through iterative use of two such web archives, FamilySearch and Sorenson Molecular Genealogy Foundation, I was able to discern the likely haplotypes for the Y chromosomes of two men, Joseph Smith and Brigham Young, who were instrumental in the founding of the Latter-Day Saints Church. I then determined whether any of the Utahns who contributed to the HapMap project (the "CEU" set) is related to either man, on the basis of haplotype analysis of the Y chromosome. Although none of the CEU contributors appear to be a male-line relative, I discovered that predictions could be made for the surnames of the CEU participants by a similar process. For 20 of the 30 unrelated CEU samples, at least one exact match was revealed, and for 17 of these, a potential ancestor from Utah or a neighboring state could be identified. For the remaining ten samples, a match was nearly perfect, typically deviating by only one marker repeat unit. The same query performed in two other large databases revealed fewer individual matches and helped to clarify which surname predictions are more likely to be correct. Because large data sets of genotypes from both consenting research subjects and individuals pursuing genetic genealogy will be accessible online, this type of triangulation between databases may compromise the privacy of research subjects.

Genotypic data can provide powerful insights into human evolution, migration, and history. Analysis of the Y chromosome, which is inherited largely intact via male descendancy, has proven to be particularly effective in tracking lineages of global significance, such as in the population of Asia by male-line relatives of Genghis Khan.[1] This method has also allowed authentication of claims of ancestry, such as those of the Lemba, an African tribe that practices Jewish rituals and claims Jewish lineage.[2]

Looking closer to home, I considered whether any genetic "dynasties" might be similarly revealed in the United States by genotypic analysis. The population of the Latter-day Saints ([LDS] Mormon), by virtue of its historical polygamy, manifested fecundity and rapid expansion and seemed promising for investigation. Moreover, because genealogical record keeping is a key activity in the LDS faith, both genotypic data and pedigree information are accessible via online repositories.

I determined whether it might be possible to trace the Y chromosomes of founders of the Mormon population and whether male descendants of those founders might be represented in the CEPH (*C*entre d'*E*tude du *P*olymorphisme *H*umain) samples that were originally collected from multigenerational families in Utah[3–5] and that now comprise the CEU (Utah residents with ancestry from northern and western Europe) set in the HapMap project.[6] I chose to begin by investigating the records of the two most well-known early leaders of the LDS: Joseph Smith, Jr., founder of the LDS Church, and Brigham Young, who took the helm of the Church after Smith's demise in 1844 and led the group from Illinois to Utah.

I used three resources for this investigation: (1) FamilySearch, a genealogical registry run by the LDS; (2) Sorenson Molecular Genealogy Foundation (SMGF), a nonprofit organization that displays genotypic and pedigree information, provided with informed consent, on the web; and (3) Y chromosome genotyping of the CEU samples, conducted during my sabbatical in the laboratory of Chris Tyler-Smith at the Wellcome Trust Sanger Institute.

Using FamilySearch.org, I tracked the ancestors, descendants, and male lineages parallel to those of Joseph Smith, who was born on December 23, 1805 in Sharon, Vermont, and died on June 27, 1844 in Carthage, Illinois. According to FamilySearch data, Smith, who received the Golden Tablets and whose revelation of polygamy launched that practice in the LDS, took 24 wives, yet fathered children only with his first wife, Emma Hale. Of the ten progeny, only five lived past infancy, and four of these were male.

By accessing and analyzing the data in SMGF and FamilySearch databases, I inferred the haplotype of Joseph Smith's Y chromosome by a two-step process as follows: First, I searched the Y chromosome database in SMGF under the surname "Smith" and then leafed through the associated pedigree information until a connection was found with the Joseph Smith of interest. This was readily accomplished, given that the FamilySearch pedigree information is linked with the SMGF genotyping database, but it involved my generating a large family tree to verify who was related to whom. Second, I employed a "guess-and-check" approach to discerning the Y haplotype of the individual of who contributed DNA within the branch of the pedigree of interest. This laborious process was

[1]Department of Medicine and Pediatrics and Institute for Human Genetics, University of California, San Francisco, CA 94143, USA
*Correspondence: jane.gitschier@ucsf.edu
DOI 10.1016/j.ajhg.2009.01.018. ©2009 by The American Society of Human Genetics. All rights reserved.

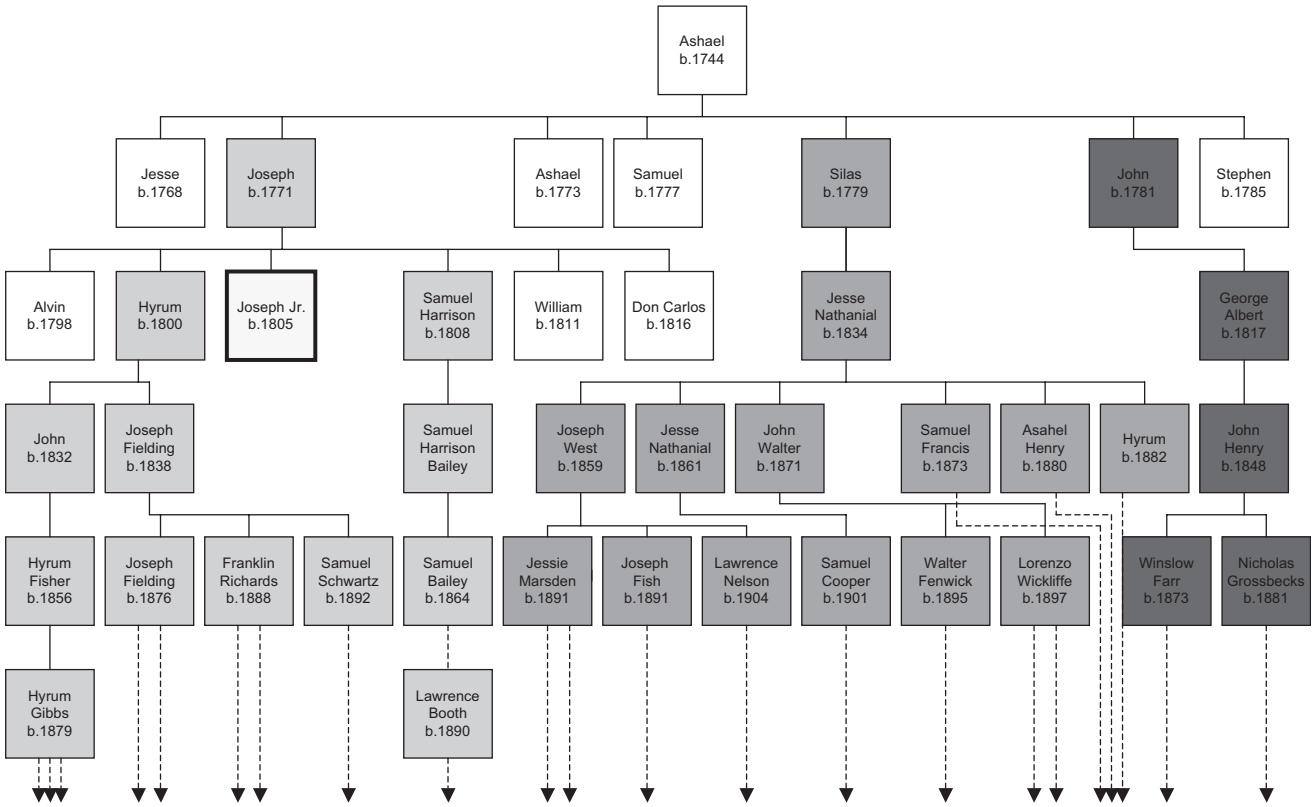

Figure 1. Reconstructed Pedigree Related to Joseph Smith, Jr., Founder of the LDS Church
Each arrow indicates a direct descendant with genotyping information found in the SMGF Y-chromosome database. Branches of the family are shaded in gray tones to be consistent with haplotype information presented in Table 1. Joseph Smith, Jr. is indicated by a symbol with a bold outline.

necessary because the SMGF database does not provide alleles for individuals directly but rather forces the user to "guess" an allele at a particular marker for a particular individual. The database signals a correct guess by changing the color of the query box from dark to light blue. I used the SMGF-generated table of allele frequencies for the markers to make informed guesses for alleles at each marker in the haplotype and worked my way through alleles, starting with the most common allele and iteratively researching the database until the color change signaled the correct allele assignment at a particular marker. Once the complete haplotype for the individual who contributed DNA was discerned, I used it as the query instead of the surname to search the SMGF Y chromosome database again, thus unmasking a series of either identical or closely related haplotypes from related individuals.

In the case of Joseph Smith, I did not find any direct descendants who contributed DNA to the SMGF project, but I did find evidence for contributions from descendants of two of his brothers (Hyrum, b. 1800 and Samuel Harrison, b. 1808) as well as from descendants of two of his paternal cousins (Jesse Nathanial, b. 1834 and George Albert, b. 1817). As illustrated in Figure 1 and presented in Table 1, Y chromosome haplotypes from a total of 22 descendants of Ashael Smith (b. 1744), Joseph Smith's grandfather, were generated from ~40 short tandem repeat (STR) markers and deposited. By parsimony, an ancestral consensus Y haplotype for Ashael Smith, and by extension for Joseph Smith, Jr., can be proposed, as presented in Table 1. Of particular note, during revision of this manuscript, I was informed by Scott Woodward and Ugo Perego of SMGF that they had previously reported a haplotype, involving a subset of the markers described herein, for Joseph Smith in a Mormon historical journal;[7] the haplotype they reported is identical to the consensus prediction herein.

The 22 haplotypes (Table 1) comprised 22 of the 23 best hits in the Sorenson database for the consensus query sequence combined with the surname "Smith." One haplotype, derived from a descendant of Bernard Culbert Smith, is also part of the cluster, suggesting that this individual's ancestor was also closely related to the Ashael Smith clan, but I could find no genealogical records within FamilySearch to support this contention. Table 1 shows that when an allele deviates from the consensus sequence among the Smith relations, it does so by a single repeat unit, consistent with a stepwise mutational model previously observed for Y STR marker allele changes.[8] The actual haplotype of Joseph Smith's Y chromosome could, in fact, have deviated from the consensus by the gain or loss of a repeat at one or a few markers.

Table 1. Compilation and Prediction of Y Chromosome Haplotypes in Joseph Smith, Jr., and Brigham Young Pedigrees

SMGF PARTICIPANT'S MOST RECENT NAMED ANCESTOR	DYS385	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS394/19	DYS426	DYS437	DYS438	DYS439	DYS441	DYS442	DYS444	DYS445	DYS446	DYS447	DYS448	DYS449	DYS452	DYS454	DYS455	DYS456	DYS458	DYS459	DYS460	DYS461	DYS462	DYS463	DYS464	GGAAT1B07	YCAII	YGATAA10	YGATAC4	YGATAH4.1
Hyrum Gibbs Smith, b.1879	11,13	12	14	30	24	11	14	13	14	12	[16]	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Hyrum Gibbs Smith, b.1879	11,13	12	14	30	24	11	14	13	14	12	[16]	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Hyrum Gibbs Smith, b.1879	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	■	18	30	30	11	11	17	17	9,11	■	12	11	■	15,16,16,17	10	19,23	15	23	21
Joseph Fielding Smith, b.1876	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,12	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Joseph Fielding Smith, b.1876	11,13	12	14	30	24	[10]	14	13	14	12	15	12	12	14	[18]	12	12	13	■	■	30	30	11	11	17	17	9,13	11	12	11	■	15,16,16,17	10	19,23	15	23	21
Franklin Richards Smith, b.1888	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	■	18	30	30	11	11	17	17	9,14	11	12	11	■	15,16,16,17	10	19,23	15	23	21
Franklin Richards Smith, b.1888	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,15	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Samuel Schwartz Smith, b.1892	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,16	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Lawrence Booth Smith, b.1890	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,17	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Jesse Marsden Smith, b.1891	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Jesse Marsden Smith, b.1891	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	[18]	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Joseph Fish Smith, b. 1891	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Lawrence Nelson Smith, b. 1904	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Samuel Cooper Smith, b.1901	11,13	12	14	30	24	[10]	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Lorenzo Wickliffe Smith, b.1897	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Walter Fenwick Smith, b. 1895	11,13	12	14	30	24	11	14	13	14	12	[14]	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Lorenzo Wickliffe Smith, b.1897	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	[29]	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Samuel Francis Smith, b.1873	[11,14]	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	[19]	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Asahel Henry Smith, b.1880	11,13	12	14	30	[25]	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Hyrum Smith, b.1882	11,13	12	14	30	24	11	14	13	14	12	15	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Winslow Farr Smith, b.1873	11,13	12	14	30	24	11	14	13	14	12	[14]	12	[13]	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
Nicholas Grossbecks Smith, b.1881	11,13	12	14	30	24	11	14	13	14	12	[14]	12	12	14	17	12	12	13	25	18	30	30	11	11	17	17	9,10	11	12	11	24	15,16,16,17	10	19,23	15	23	21
JOSEPH SMITH, JR. (PREDICTED)	**11,13**	**12**	**14**	**30**	**24**	**11**	**14**	**13**	**14**	**12**	**15**	**12**	**12**	**14**	**17**	**12**	**12**	**13**	**25**	**18**	**30**	**30**	**11**	**11**	**17**	**17**	**9,10**	**11**	**12**	**11**	**24**	**15,16,16,17**	**10**	**19,23**	**15**	**23**	**21**
Clifford Earl Young, b. 1883	■	■	■	■	■	■	■	■	■	■	14	■	■	■	16	12	12	13	■	■	■	30	11	■	16	■	10,10	■	■	11	■	■	10	■	15	23	■
Seymour La Von Young, b. 1903	11,15	12	12	28	24	11	13	13	15	12	14	12	12	14	16	12	12	13	25	19	29	30	11	11	16	17	10,10	11	12	11	24	15,15,17,17	10	19,23	15	23	21
Samuel Claridge Young, b. 1877	11,15	12	12	28	24	11	13	13	15	12	14	12	12	14	16	12	12	13	25	19	29	30	11	11	16	17	10,10	11	12	11	24	15,15,17,17	10	19,23	15	23	21
Samuel Claridge Young, b. 1877	11,15	12	12	28	24	11	13	13	15	12	14	12	12	14	16	12	12	13	25	19	29	■	11	11	16	17	10,10	11	12	11	24	15,15,17,17	10	19,23	15	23	21
Lyle Legrand Young, b. 1886	11,15	12	12	28	24	11	13	■	15	12	14	12	12	■	16	12	■	13	25	19	■	30	■	■	■	■	■	11	■	11	■	15,15,17,17	10	19,23	15	23	21
BRIGHAM YOUNG (PREDICTED)	**11,15**	**12**	**12**	**28**	**24**	**11**	**13**	**13**	**15**	**12**	**14**	**12**	**12**	**14**	**16**	**12**	**12**	**13**	**25**	**19**	**29**	**30**	**11**	**11**	**16**	**17**	**10,10**	**11**	**12**	**11**	**24**	**15,15,17,17**	**10**	**19,23**	**15**	**23**	**21**

■ indicates missing data
[ ] indicates allele changes compared to the consensus sequence

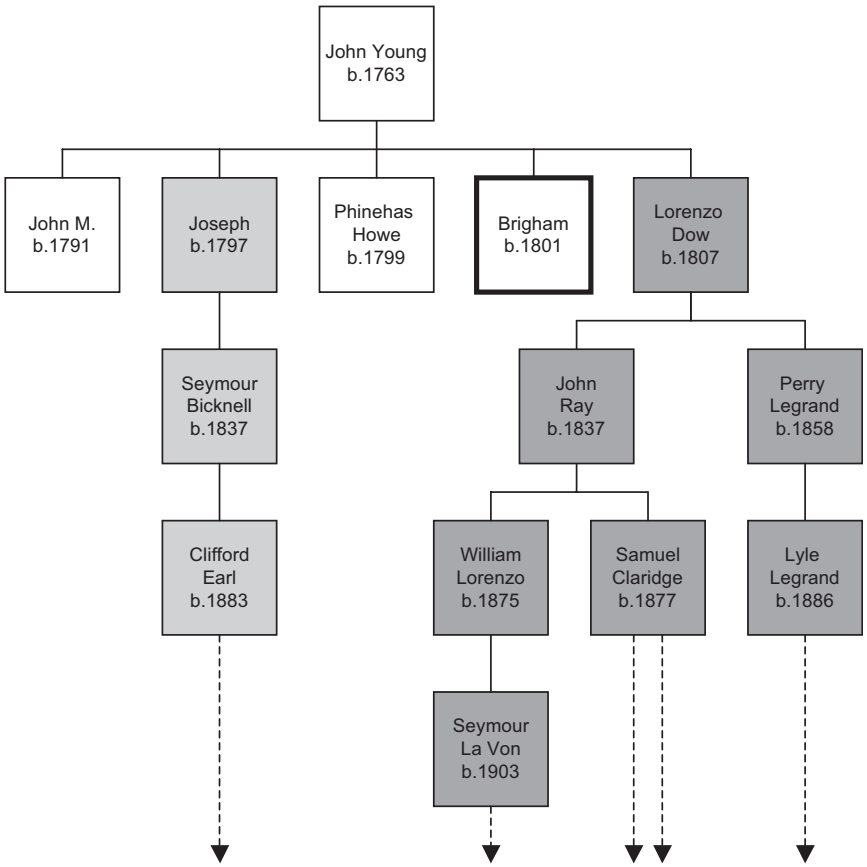

Figure 2. Reconstructed Pedigree Related to Brigham Young
Each arrow indicates a direct descendant with genotyping information found in the SMGF Y-chromosome database. Branches of the family are shaded in gray tones to be consistent with the haplotype information presented in Table 1. Brigham Young is indicated by a symbol with a bold outline.

Performing a search with the Smith consensus haplotype query, but without regard to last name, pulled up nearly the same set of individuals in the top 23 hits, with individuals having McCall and Clair as surnames displacing only two Smiths. Additional surnames emerged within the first-50 best hits, including McClellan, Robertson, Murray, Loftus, White, Wilson, Douglass, and Lockhart, with only modest changes (two mismatches; data not shown). The haplotypes for these individuals are more similar to the Smith ancestral genotype of interest than to the genotypes of other Smith clans. Whether the genotypes are shared by identity by descent (through adoption out of the family or misattributed paternity) or are fortuitously similar could not be determined because I was unable to link the pedigrees using FamilySearch information.

For Brigham Young (b.1801), investigation of FamilySearch pedigree records indicates 37 wives, with 58 offspring, 22 of whom were male. Yet, no direct descendants of Brigham Young appear to have contributed DNA to the SMGF collection. Connections could be made only to his older brother Joseph, b. 1797, and to four descendants of his younger brother Lorenzo Dow Young, b. 1807 (Figure 2; Table 1). Although the genotyping data from Joseph's relatives is very limited, these five samples provide consistent evidence for a consensus haplotype that may be attributable to Brigham Young himself.

Using the Young surname to screen the SMGF database indicated additional individuals who are clearly related to one another but differ from the Youngs of interest in at least eight to ten marker calls. I could not find any genealogical record to connect these other Youngs to that of Brigham Young's family. By genotype query, the closest match to the Youngs of interest are a collection of individuals with the surname Fuller, who differ at three to five markers, but I failed to establish any genealogical connections to the family of Brigham Young. Thus, it appears that the male relatives of Brigham Young may not have made a very large contribution to the LDS gene pool, at least as evidenced by the SMGF repository, with the obvious caveat that SMGF database does not provide an unbiased sampling of that population.

DNA samples from all unrelated male individuals who constituted the CEU set of the HapMap collection were contributed with informed consent and were obtained from Coriell Institute by the Wellcome Trust Sanger Institute. Samples were genotyped with the YFiler kit (Applied Biosystems), which consists of 17 highly informative short tandem repeat (STR) markers. In the case of each father-son pair, I chose to analyze the father. The resulting genotypes for each of the 30 samples will be reported elsewhere[9] but are additionally displayed in Table 2 for ease of the present discussion. All repeat-size measurements are made according to the International Society of Forensic Genetics (ISFG) guidelines.[10] For consistency with the SMGF nomenclature, the two independently variable alleles generated at marker DYS385 are listed together (separated by a comma), and the allele size for marker DYS389II is the sum of the allele for marker DYS389I as well as an independently varying additional allele, whose call can be determined by subtraction. Table 2 summarizes the resulting haplotype data for these 30 samples.

Although each genotype in this CEU set of samples is unique, two sets of samples may be derived from paternally related individuals. Samples 11839 and 12872 differ by only one repeat at a single marker (DYS390), suggesting that their respective pedigrees, 1349 and 1459, could be distantly related. Similarly, samples 11881 (from pedigree

Table 2. Y STR Haplotypes for CEU Samples and Summary of Predicted Surnames in SMGF Database

CEU Sample ID	DYS385	DYS389I[a]	DYS389II[a]	DYS390	DYS391	DYS392	DYS393	DYS394/19	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	YGATAC4	YGATAH4.1	Number of Different SMGF Surnames with Exact Marker Match	Predicted State of Origin[b]	Matches Per Assayed Markers
6993	11,14	13	29	24	10	13	13	14	15	12	12	19	15	19	23	22	1	Utah	17/17
6994	13,14	12	28	22	10	11	13	15	16	10	11	20	14	16	22	20	5	Utah (3)	17/17
7022	13,15	12	28	22	10	11	13	14	16	10	11	20	16	14	21	20	0		
	13,14	*12*	*28*	*22*	*10*	*11*	*13*	*14*	*16*	*10*	*11*	*20*	*15*	*14*	*21*	*20*		*Utah*	*15/17*
	13,14	*12*	*28*	*22*	*10*	*11*	*13*	*14*	*16*	*10*	*11*	*20*	*14*	*14*	*21*	*20*		*Utah*	*15/17*
7034	11,14	14	30	23	10	13	13	14	15	12	12	19	17	17	23	20	0		
	11,14	*13*	*29*	*23*	*10*	*13*	*13*	*14*	*15*	*12*	*12*	*19*	*17*	*17*	*23*	*20*		*Arizona*	*16/17*
7357	12,14	12	28	24	10	13	13	14	15	12	12	19	16	17	23	21	1	Canada	17/17
	12,14	*13*	*29*	*24*	*10*	*13*	*13*	*14*	*15*	*12*	*12*	*19*	*16*	*17*	*23*	*21*		*Utah*	*16/17*
	12,14	*12*	*28*	*24*	*10*	*13*	*14*	*14*	*15*	*12*	*12*	*19*	*16*	*17*	*23*	*21*		*Idaho*	*16/17*
	12,14	*12*	*28*	*24*	*10*	*13*	*13*	*14*	*15*	*12*	*12*	*19*	*15*	*17*	*23*	*21*		*Utah*	*16/17*
11829	14,15	12	28	23	10	11	13	14	16	10	11	20	14	17	21	21	0		
	14,15	*12*	*28*	*23*	*10*	*11*	*13*	*14*	*16*	*10*	*11*	*20*	*14*	*16*	*21*	*20*		*Utah*	*15/17*
	14,15	*12*	*28*	*23*	*10*	*11*	*13*	*14*	*16*	*10*	*11*	*20*	*14*	*14*	*22*	*21*		*Utah*	*15/17*
	14,15	*12*	*28*	*23*	*10*	*11*	*13*	*14*	*16*	*10*	*11*	*20*	*14*	*15*	*21*	*20*		*Utah*	*15/17*
11831	11,14	14	30	24	11	13	13	14	15	12	12	19	16	17	24	21	2	Utah (1)	17/17
11839	11,15	13	29	25	11	13	13	14	15	12	12	19	15	16	23	21	1	Australia	17/17
	11,15	*13*	*29*	*25*	*11*	*13*	*13*	*14*	*15*	*12*	*12*	*19*	*15*	*17*	*23*	*21*		*Utah*	*16/17*
	11,15	*13*	*29*	*24*	*11*	*13*	*13*	*14*	*15*	*12*	*12*	*19*	*15*	*16*	*23*	*21*		*Utah*	*16/17*
	11,15	*13*	*29*	*24*	*11*	*13*	*13*	*14*	*15*	*12*	*12*	*19*	*15*	*16*	*23*	*21*		*Idaho*	*16/17*
	11,14	*13*	*29*	*25*	*11*	*13*	*13*	*14*	*15*	*12*	*12*	*19*	*15*	*16*	*23*	*21*		*Utah*	*16/17*
11881	13,14	12	28	22	10	11	13	14	16	10	12	20	14	14	21	20	1	Brazil	17/17
	13,14	*12*	*28*	*22*	*10*	*11*	*13*	*14*	*16*	*10*	*12*	*20*	*14*	*15*	*21*	*20*		*Utah*	*16/17*
	13,14	*12*	*28*	*22*	*10*	*11*	*13*	*14*	*16*	*10*	*11*	*20*	*14*	*14*	*21*	*20*		*Utah*	*16/17*
11992	14,14	13	30	22	10	11	13	14	16	10	12	20	13	15	21	21	1	Utah	17/17
11994	11,14	12	28	24	11	13	13	13	15	12	12	19	15	15	23	22	2	Utah, Texas	17/17
12003	14,15	14	32	23	10	12	14	15	14	10	11	20	13	14	21	19	0		
	14,15	*14*	*32*	*23*	*10*	*12*	*14*	*15*	*14*	*10*	*11*	*20*	*14*	*15*	*21*	*19*		*Utah*	*15/17*
12005	11,14	13	29	24	11	13	12	14	15	12	12	19	15	17	23	21	3	Utah (1)	17/17
12043	11,14	12	27	24	10	13	13	14	15	12	11	19		19	23	20	1	Utah	16/16
12056	14,15	13	29	24	10	13	13	14	15	12	11	19	16	15	23	21	1	Utah	17/17
12144	11,14	13	28	24	11	13	13	15	14	12	12	20	15	17	23	21	0		
	11,14	*13*	*30*	*24*	*11*	*13*	*13*	*14*	*14*	*12*	*12*	*20*	*15*	*17*	*23*	*21*		*Utah*	*15/17*
12146	12,15	13	29	22	10	11	13	15	14	10	11	17	14	18	21	20	1	Utah	17/17
12154	11,14	13	29	24	13	14	14	14	15	12	11	19	17	17	23	21	1	Utah	17/17
12155	11,14	14	31	25	10	11	13	15	14	11	10	19	15	15	23	21	4	Utah (2)	17/17
12248	11,14	14	30	25	11	13	13	14	15	12	12	19	15	18[c]	23	20	1	Utah	17/17[c]
12264	11,13	13	29	23	10	13	13	14	16	12	12	19	16	17	23	21	0		
	11,13	*13*	*29*	*24*	*10*	*13*	*13*	*14*	*16*	*12*	*12*	*19*	*16*	*17*	*23*	*21*		*Utah*	*16/17*
12716	11,14	12	28	24	11	13	13	14	15	12	12	19	16	17	23	20	5	Utah (1) Idaho (1)	17/17
12750	13,14	12	28	22	10	11	13	14	16	10	12	20	12	14	21	20	0		
	13,14	*12*	*28*	*22*	*10*	*11*	*13*	*14*	*16*	*10*	*12*	*20*	*14*	*15*	*21*	*20*		*Utah*	*15/17*
	13,14	*12*	*28*	*22*	*10*	*11*	*13*	*14*	*16*	*10*	*11*	*20*	*14*	*14*	*21*	*20*		*Utah*	*15/17*
12760	11,13	13	29	23	11	13	13	14	15	12	12	18	17	17	23	21	1	Utah	17/17
12762	11,14	14	30	24	11	13	13	14	15	12	12	18	16	17	23	21	12	Utah/Idaho (1)	17/17
12812	12,14	13	29	24	10	14	14	13	15	12	12	19	16	16	23	21	0		
	12,14	*13*	*29*	*24*	*10*	*13*	*13*	*13*	*15*	*12*	*12*	*19*	*16*	*16*	*23*	*21*		*Utah*	*15/17*
	12,14	*13*	*29*	*24*	*10*	*13*	*13*	*13*	*15*	*12*	*12*	*19*	*16*	*16*	*23*	*21*		*Utah*	*15/17*
12814	11,14	13	30	24	11	13	12	15	15	12	12	20	15	17	23	21	0		
	11,14	*13*	*30*	*24*	*11*	*13*		*15*	*15*	*12*	*12*	*20*			*23*	*21*		*Utah*	*14/14*
	11,14	*13*	*30*	*24*	*11*	*13*	*13*	*14*	*15*	*12*	*12*	*20*	*15*	*17*	*23*	*21*		*Utah*	*15/17*
12872	11,15	13	29	24	11	13	13	14	15	12	12	19	15	16	23	21	3	Idaho (1)	17/17
12874	11,14	13	28	24	11	13	13	14	14	12	12	18	15	17	24	20	0		
	11,14	*13*	*28*	*24*	*11*	*13*	*13*	*14*	*14*	*12*	*12*	*18*	*15*	*17*	*23*	*20*		*Texas*	*16/17*
12891	13,14	12	28	23	10	11	13	14	16	10	11	20	14	15	21	20	7	Wyoming (1)	17/17
Joseph Smith	**11,13**	**14**	**30**	**24**	**11**	**14**	**13**	**14**	**15**	**12**	**12**	**18**	**17**	**17**	**23**	**21**			
Brigham Young	**11,15**	**12**	**28**	**24**	**11**	**13**	**13**	**15**	**14**	**12**	**12**	**19**	**16**	**17**	**23**	**21**			

indicates haplotype in SMGF with close match to CEU haplotype above it
indicates allele in SMGF sample deviating from CEU allele
indicates missing information in SMGF sample

a A change in the allele size for DYS389I will cause a change in allel size for DYS389II, as described in text.
b The number in paranetheses indicates the number of surnames corresponding to that state having the genotype.
c The allele size for this CEU sample appears to be 1 nucleotide shy of the full 18 tetrarepeats

1347) and 12750 (from pedigree 1444) also differ at a single marker (DXYS456), albeit by two repeat units, making the degree of their genetic relationship more tenuous.

Comparison of the haplotypes from these 30 individuals with the predicted haplotypes for Joseph Smith and Brigham Young for this same set of 17 markers (indicated at the bottom of Table 2) indicates that none of the HapMap contributors appears to be descended from either LDS founding family.

Because the SMGF proved so effective in the case of tracking the Smiths and Youngs, I became curious to know whether any of the 30 independent CEU Y haplotypes are represented in the SMGF Y chromosome repository, which comprises data contributed by over 23,000

men with a repertoire of 13,164 unique surnames. Because this resource is enriched for samples taken from the Utah population under discussion, it could provide a source of identity information, although the names of the contributor and his most recent (presumably, living) ancestors are typically masked. According to the SMGF website, approximately half of the samples appear to have been contributed by individuals in Utah; the other half appear to have been contributed from individuals throughout the remaining parts of the United States and the world.

As shown in Table 2, 20 of the 30 CEU Y chromosome haplotypes exactly match that of at least one individual in the SMGF database. In three of these cases (7357, 11839, and 11881), a single perfect match was made to an individual whose most recently named ancestor resided outside the United States, in Canada, Australia, and Brazil, respectively. For the remaining 17 haplotypes, at least one of the perfectly matching genotypes correspond to individuals whose most recently named ancestor resided in a state with a substantial LDS population, namely Utah, Idaho, Wyoming, and Texas.

For the remaining ten CEU haplotypes, SMGF repository genotypes were identified with either one or two mismatches, and some of the contributors of these were also derived from Utah or nearby states. In almost every case, the mismatches deviated from the query haplotype by a single repeat unit at a given marker, which would be expected for a close relative. The retrieved mismatch genotypes, indicated by shading and italics, are also shown in Table 2.

Each of the matching SMGF haplotypes is associated with a surname, and the obvious question emerges as to whether these surnames indeed correspond to the surnames of the CEU contributors themselves. This direct question is unanswerable. In consultation with investigators at the University of Utah, where the samples were collected, we jointly concluded that confirmation of the predicted surnames would violate the ethical constraints of informed consent obtained during the collection of these samples because the names of the subjects would be used in the analysis. Moreover, in deference to the privacy of those who contributed the CEU samples I have not included the predicted surnames in Table 2. Instead, I have attempted to assess the accuracy of the predictions with several simulations, as follows:

To challenge the power of using a collection of only 17 STR markers to accurately screen the SMGF database, I submitted the 17-marker subset of alleles, shown in Table 2, corresponding to the consensus haplotypes for both Joseph Smith and Brigham Young. In response to each of these queries, conducted without regard to surname, I retrieved largely the same set of individuals that I had uncovered with the larger set of ~40 markers. This very limited test suggests that a reasonable guess as to male ancestors may be made by a relatively small set of highly informative markers, at least within this targeted database.

Another measure of the specificity of a 17-marker haplotype lies in asking how many samples one would expect to find with the given haplotype within a particular database. In the case of the Y chromosome, the expected frequency of a given STR haplotype cannot be generated simply by assessing the products of the allele frequencies at each locus;[11] indeed such calculations, based on the SMGF marker allele frequencies, would predict matches for each of the 30 CEU samples of between 1 in 10 billion and 1 in 10 trillion unrelated individuals. Because the Y chromosome is inherited as an intact unit without undergoing recombination, there are strong associations between pairs of alleles. These associations are somewhat counterbalanced by the mutability of the STR markers over time. Consequently, the frequency of a particular Y STR haplotype within a population cannot be assessed a priori but demands an empiric estimate from actual data sets. Thus, to assess haplotype frequency, I used the 30 CEU haplotypes as queries in two additional large databases, as follows:

First, I searched a collection of 10,254 Y haplotypes derived for the identical set of 17 STR markers and deposited into the Y Haplotype Reference Database (YHRD), a compendium that is specifically designed to assess the frequency of Y STR haplotypes in world-wide populations, including the United States. Only two of the 30 haplotypes were found to have an exact match in this data set: an identical match for sample 11839 was observed in one individual of Portuguese ancestry (out of 303 in that population), and two identical matches were found for sample 12005, one in an admixed population of 50 individuals from Cordoba, Columbia and one in a European population of 384 from Ravenna, Italy. For most (28 of 30) of the CEU haplotypes, this analysis indicates a very conservative estimate of finding a given haplotype in fewer than 1 in 10,254 individuals in the worldwide population.

Second, Family Tree DNA kindly agreed to help me by querying a subset of their private Y haplotype database of 55,000 individuals. This database is enriched for Americans of Western and Northern European ancestry and in this regard may provide a better comparison for the SMGF database. However, this collection of individuals was genotyped with only 16 of the 17 Y STR markers. Because marker Y_GATAC4 is not routinely included in the Family Tree DNA marker set, this analysis provides an overestimate of the prevalence of a particular genotyped by roughly 2- to 10-fold, depending on the frequency of the actual Y_GATAC4 allele. Moreover, because this Family Tree DNA cohort is 2.4 times as large as that of SMGF, one would expect a greater number of matches if the distribution of haplotypes in the two sample sets were equivalent. Yet, of the 30 CEU haplotypes, 14 (47%) failed to match a single 16-marker haplotype in this Family Tree DNA database and only eight (26%) detected one to three individuals with perfectly matched 16-marker haplotypes. The remaining 8 matched between 9 and 84 individuals, with the number of unique surnames ranging from 6 to 60.

Indeed the two CEU haplotypes (from samples 12762 and 12891) that matched the largest number of Family Tree DNA haplotypes also did so in the SMGF database. Although these more-common haplotypes would be unreliable predictors of surname in the SMGF database, most of the 17-marker haplotypes generated for the CEU samples would be expected by chance to be found in fewer than 1 in 55,000 individuals, and the surname predictions made in the SMGF database would probably be more accurate for these.

The above queries demonstrate that a set of possible surnames can be unmasked for the CEU families, and I posit that a fair number of these predictions are likely to be correct, assuming my genotyping was performed correctly and given the population involved, the database searched, the comparisons with other databases, and the effectiveness of this approach for the Smith and Young pedigrees.

One could imagine that scrutiny of the HapMap archive, which now consists of over 3.1 million genotypes per individual,[6] might predict some physical, health, and behavioral attributes associated with particular alleles, which, in combination with a surname, might lead to a further embellishment of identity. A reading of the HapMap consent form, available online, shows that subjects who consented for the HapMap project were informed that their genotypic data would be extensive, that it would be posted on the internet, and that their cell lines would be widely distributed to enable genetic research beyond the HapMap project itself. The threat to privacy through this type of cross-database triangulation was anticipated by the HapMap project[12] and was included as a potential risk on the HapMap consent form.

One of the great concerns in human genetics is maintaining the privacy of individuals who contribute samples for research purposes. Although this concern is raised typically in the context of private medical information,[13] I would argue that the biggest risk to loss of anonymity lies with genealogical investigations. Indeed, it is the very nature of genealogical research to seek out connections, and use of DNA information tremendously augments this ability. Although currently these quests are limited to mitochondrial and Y DNA markers, in the future whole-genome genotypes, and indeed entire genome sequences, will probably be posted online by individuals who are eager to make connections with relatives outside of exclusively matrilineal and patrilineal ancestry.

By contributing samples and associated genealogical information to repositories specializing in genetic genealogy, individuals make important contributions to our collective knowledge, but they do so at the risk of unmasking personal information for unwitting relatives who may have contributed DNA in anonymity for research purposes. This problem will be exacerbated in the near future, as larger numbers of subjects are engaged for genetic research, more individuals seek their genetic heritage, further deposit of DNA sequences in shared databases is demanded by public funding, genome sequences proliferate as technology becomes faster and cheaper, algorithms to query them improve, and computers increase in speed and capacity.

These observations may prove to stimulate ideas for improving informed consent and refining public access to detailed genotyping of human subjects. I proffer some thoughts on these issues, as follows:

First, if a research study plans to provide open access to genetic data, it is imperative that study investigators clearly inform the subjects that their genomic data will be accessible online and that it may be possible for others to make inferences about their identity through comparisons with genomic data deposited into other online databases. These genomic data may include that derived from closely related individuals, who may have deposited publicly accessible data without the knowledge of the subject. As this report illustrates, of particular concern are databases designed for genealogical research because DNA information in them is often linked to names.

Second, the enthusiasm for shared genetic data, especially those generated through public funding, must be tempered by privacy concerns for the participants, given that DNA itself is the ultimate "identifier." Researchers instead should consider limited but facilitated access to DNA databases, as outlined in NIH Data Sharing Policy and Implementation Guidance.

Third, I propose establishment of a secure, password-protected comprehensive human genetic database, analogous to GenBank and which I provisionally refer to as "GenomeBank." This archive would be compiled by contributions from individuals themselves and made searchable for the purpose of genealogical investigation. This genotypic information could be contributed to the database directly from a commercial personal genome service or a research laboratory under the authorization of the participating individual, who may choose to add additional identifying information for genealogical research. By instituting and applying consistent numbering and nomenclature for all data, it should be possible to compare genotypes and/or sequences among samples. Under this model, if two genotypes or sequences are found to have sufficient similarity to suggest recent shared ancestry, individuals contributing the information would be notified of potential relationship and would have the opportunity to communicate with each other if both parties agree, an algorithm that is commonly used in online social networking or dating services. In this manner, genealogical research can make full use of the advances in the genotyping and sequence technology and genetic connections can be made without restriction to patrilineal or matrilineal ancestry. Moreover, individuals will not be restricted to finding connections within a small cohort of individuals who happen to have used the same personal genome service, and research subjects will have value added to their participation in the form of genealogical research. With time, such an enterprise could be useful for individuals who have been displaced from their blood relatives by

adoption, war, or migration, and it could form the framework for the delineation of a "world-wide pedigree."

Acknowledgments

I am very grateful to Chris Tyler-Smith for giving me the opportunity to work in his laboratory, and I thank members of his group, especially Ya-Li Xue and Tatiana Zerjal, for their patience and instruction, and the Howard Hughes Medical Institute for its support. I thank Mark Leppert, Jeff Botkin, and Dennis Drayna for stimulating discussions and advice, Natalie Myres of Sorensen Molecular Genealogy Foundation for helpful email discussions, Bennett Greenspan and Eileen Krause of Family Tree DNA for help in accessing their databases, and Beth Theusch for thoughtful reading of this manuscript. Finally, I would like to acknowledge the many Utah families who have made extraordinarily important contributions to human genome mapping since its inception almost three decades ago and who continue to contribute to research and to inspire those of us who conduct it.[5]

Received: September 5, 2008
Revised: October 29, 2008
Accepted: January 13, 2009
Published online: February 12, 2009

Web Resources

The URLs for data presented herein are as follows:

Coriell Institute for Medical Research, http://ccr.coriell.org
FamilySearch, http://familysearch.org
Family Tree DNA, http://familytreedna.com
HapMap project, http://hapmap.org
NIH Data Sharing Policy and Implementation Guidance, http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#archive
Sorenson Molecular Genealogy Foundation, http://www.smgf.org
Y Chromosome Haplotype Reference Data, http://www.yhrd.org

References

1. Zerjal, T., Xue, Y., Bertorelle, G., Wells, R.S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., et al. (2003). The genetic legacy of the Mongols. Am. J. Hum. Genet. *72*, 717–721.
2. Spurdle, A.B., and Jenkins, T. (1996). The origins of the Lemba "Black Jews" of southern Africa: evidence from p12F2 and other Y chromosome markers. Am. J. Hum. Genet. *59*, 1126–1133.
3. White, R., Leppert, M., Bishop, D.T., Barker, D., Berkowitz, J., Brown, C., Callahan, P., Holm, T., and Jerominski, L. (1985). Construction of linkage maps with DNA markers for human chromosomes. Nature *313*, 101–105.
4. Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M., and White, R. (1990). Centre d'etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. Genomics *6*, 575–577.
5. Prescott, S.M., Lalouel, J.M., and Leppert, M. (2008). From linkage maps to quantitative trait loci: The history and science of the Utah Genetic Reference Project. Annu. Rev. Genomics Hum. Genet. *9*, 347–358.
6. The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
7. Perego, U.A., Myres, N.M., and Woodward, S.R. (2005). Reconstructing the Y-Chromosome of Joseph Smith: Genealogical applications. J. Mormon Hist. *31*, 42–60.
8. Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C., Krawczak, M., Nagy, M., Dobosz, T., 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.
9. He, M., Gitschier, J., Aerjal, T., de Knijff, P., Tyler-Smith, C., and Xue, Y. (2009). Geographical affinities of the HapMap samples. PLoS One, in press.
10. Gusmao, L., Butler, J.M., Carracedo, A., Gill, P., Kayser, M., Mayr, W.R., Morling, N., Prinz, M., Loewer, L., Tyler-Smith, C., and Schneider, P.M. (2006). DNA commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis. Forensic Sci. Int. *157*, 187–197.
11. Roewer, L., Kayser, M., de Knijff, P., Anslinger, K., Betz, A., Caglia, A., Corach, D., Furedi, S., Henke, L., Hidding, M., et al. (2000). A New Method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. Forensic Sci. Int. *114*, 31–43.
12. The International HapMap Consortium. (2004). Integrating ethics and science in the International HapMap Project. Nat. Rev. Genet. *5*, 467–475.
13. Lowrance, W.W., and Collins, F.S. (2007). Identifiability in genomic research. Science *317*, 600–602.