www.sciencemag.org/cgi/content/full/339/6117/321/DC1

Supplementary Materials for

Identifying Personal Genomes by Surname Inference

Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, Yaniv Erlich*

*To whom correspondence should be addressed. E-mail: yaniv@wi.mit.edu

Published 18 January 2013, *Science* **339**, 321 (2013)
DOI: 10.1126/science.1229566

This PDF file includes:

Supplementary Text
Figs. S1 to S6
Tables S1, S2, S5, and S7
Captions for Tables S3, S4, and S6
References

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/339/6117/321/DC1)

Tables S3, S4, and S6 as zipped archives:
S3, Surname haplotype pairs used to challenge Ysearch and SMGF.
S4, Results of database queries using Ysearch and SMGF haplotypes.
S6, Y-STR haplotypes profiled from sequencing datasets.

Table of Contents

Supplementary Text

Supplementary Figures

Supplementary Tables

Supplementary Text

1. Evaluating the general risk of surname recovery

Downloading Ysearch data

The Ysearch website belongs to FamilyTreeDNA (FTDNA), a Texas-based genetic genealogy company. The website allows users, regardless of their testing service, to voluntarily post their Y-STR genotyping results along with their ancestral information and contact details. Based on the data posted on the website, approximately 85% of Ysearch's users were tested with FamilyTreeDNA and the other 15% were tested with other genetic genealogy services. Users from other services are advised to post their results using FamilyTreeDNA nomenclature, and the website offers a conversion table between popular genetic genealogy services and FamilyTreeDNA nomenclature.

With permission from FamilyTreeDNA, we scraped the entire Ysearch database in May 2011. Some areas are protected by reCaptcha and were accessed manually. After parsing and merging the HTML files, we obtained 95,000 surname-haplotype entries, each of which contained: Ysearch userID, surname, ancestral location, and Y-STR results.

Access to the SMGF database

The SMGF website belongs to the Sorenson Molecular Genealogy Foundation, a Utah-based non-profit genetic genealogy organization that was recently acquired by Ancestry.com. The website allows users to query the SMGF database but not to create new records, and all records are from the SMGF program. Unlike the Ysearch database, we could not download the database records to our server. With permission from SMGF, we conducted queries of their database using an automatic script. The webpages that contained the top 10 results based on the SMGF matching algorithm were downloaded and parsed to identify the matches.

Concordance between genealogical databases and the US population

The surname distribution in the general US population was estimated using the Census 2000 study that is based on 270 million records (http://www.census.gov/genealogy/www/data/2000surnames/index.html). The Census study lists 151,671 surnames along with their relative prevalence in the general population and ethnic composition in sorted order. To protect the privacy of the participants and due to sample size limitations, the Census data stops when the cumulative frequency of the surnames reaches 90%, and does not include surnames that are found in less than 100 individuals each.

We compared the surname distribution in Ysearch and SMGF to the distribution in the general US population in order to evaluate the completeness of the databases. We defined the census coverage probability, denoted by *c*, as the chance that the surname of an individual drawn at random from the US population has at least a single haplotype record in one of these databases, and found that *c*=68.5%. The correlation between the US population and the genealogical records was evaluated by a permutation test with 10,000 repetitions. We obtained the following statistics: $E[SSE_{permutations}]=9.01*10^{6}$, $\sigma(SSE_{permutations})=2437$. The hypothesis SSE was $1.99*10^{6}$. The p-value was calculated using one-sided Chebyshev bound.

A mathematical model for the probability of surname recovery

Search method

Our database search method relied on finding a record that shares the closest Time to Most Recent Common Ancestor (TMRCA) with the queried haplotype. The rationale behind this strategy is that close patrilineal relatives have a higher probability of sharing the same surname. For instance, one can imagine that monozygotic twins have a high probability of sharing the same surname, whereas a pair of Y chromosomes whose MRCA lived before the formation of the surname system would have a low probability of sharing the same surname.

Walsh (*1*) has proposed several Bayesian models for estimating the distribution of the TMRCA in non-recombining haplotypes. We used his 'infinite alleles model with differential mutation rates'. Consider two Y chromosome haplotypes with *n* STR loci denoted by $\vec{v} = (v_1, v_2, \dots, v_n)$ and $\vec{u} = (u_1, u_2, \dots, u_n)$, with vector elements corresponding to the allele lengths. Let $\vec{x} = (x_1, x_2, \dots, x_n)$ be a binary vector with $x_i = 1$ for a match at the *i*-th locus of $\vec{v}$ and $\vec{u}$, and $x_i = 0$ otherwise, and let $\vec{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$ be a vector whose elements denote the probability of a mutation per meiosis in each marker. According to Walsh's model, the probability distribution function (PDF) of the TMRCA between the two haplotypes is:

$$P(t|\vec{x},\vec{\mu},N_e) = \frac{e^{-t(\frac{1}{N_e}+2\sum_{i=1}^{n}\mu_i x_i)}\prod_{i=1}^{n}(1-e^{-2t\mu_i})^{1-x_i}}{I(\vec{x},\vec{\mu},N_e)} \tag{1a}$$

where N_e is the effective male population size, and I is a normalization factor to ensure that $\sum_{t=0}^{\infty} P(t|\vec{x},\vec{\mu},N_e) = 1$. Following Thomson et al. (*2*) , N_e was set to 10,000 males. The mutation rates were obtained from the extensive study of Ballantyne, *et al (3).*

The expected TMRCA is denoted by τ and is given by:

$$\tau = \sum_{t=0}^{\infty} t_i \, P(t_i|\vec{x},\vec{\mu},N_e) \tag{1b}$$

The recovered surname was selected according to the record that has the minimal τ to the searched haplotype. Due to technical constraints with the web queries to SMGF and in order to reduce the amount of calculations, we did not determine τ for each of the hundreds of thousands of users in the databases. Instead, we employed the following procedure: (i) Ysearch - identify a set of candidate records that have the maximal number of matching markers to the queried haplotype (ii) SMGF – use the native SMGF search tool to identify the top 10 candidates according to the website's proprietary algorithm (iii) Both – calculate τ for top candidates in Ysearch and SMGF using Eq. 1, and select the record with the minimal τ of the searched haplotype.

Retrieval confidence score

The retrieval confidence score determined the probability that the TMRCA of the retrieved record is indeed shorter than that of (i) a record with a distinct surname that has the second to shortest TMRCA and (ii) a random person from the population. Let P_1 and P_2 be the TMRCA PDFs of the best record and second best record according to Eq.1, and let P_3 be the PDF of coalescent in a Fisher-Wright population: $P_3(t|N_e) = N_e^{-1}e^{-N_e t}$. In addition, let F_i be the cumulative probability distribution function of P_i. The retrieval confidence score, δ, is given by:

$$\begin{aligned} \delta(P_1,P_2,P_3) &= \sum_{j_1=1}^{T} P_1(j_1)\left(\sum_{j_2>j_1}^{T} P_2(j_2)\left(\sum_{j_3>j_1}^{T} P_3(j_3)\right)\right) \\ &= \sum_{j=1}^{T} P_1(j)(1-F_2(j))(1-F_3(j)) \end{aligned} \tag{2}$$

T is the number of generations that is practical for the patrilineal surname system and was set to 20 generations, corresponding to ~1400 AD. P_2 was obtained by scanning records in the list that was generated in step (iii); candidate records with less than 20 markers were excluded as well as records with surnames that matched the top hit.

Surname inference

We set a threshold, δ_0, which denotes the minimal accepted quality for valid surname recovery. If the retrieval passed the confidence threshold, the algorithm inferred that the record's surname is the surname of the input haplotype. Otherwise, the algorithm rejected the inference and returned 'Unknown'. 1.8% of the searches returned records with an empty surname field or with strings that are not found in the surname list of the US census such as 'AshkenaziJewishModal'. The algorithm reported these cases as 'Unknown' as well. Finally, TMRCA ties between two or more records with distinct surnames were also treated as 'Unknown'.

A surname inference resulted in one of the following outcomes: success – the recovered surname is concordant with the true surname, wrong – the recovered surname does not match the true surname, unknown – below confidence threshold, non-valid surnames, and ties.

Following previous record linkage studies (*5, 6*), successful recoveries included a small number of cases where the returned surname displayed a minute spelling variant from the true one, such as Abernathy and Abernethy. These cases can still direct the adversary in tracing back the target at the price of searching for a larger number of individuals. We adopted a stringent approach to detect spelling variants that required that the first letter of both surnames be identical *and* that the Jaro-Winkler string distance (*7*) of the surnames be at least 0.9. This relies on the observation that the suffix of a surname is more prone to mutate than the prefix (*7*). Two percent of the queries showed spelling variants using this approach and they are summarized in the following table:

True surname	**Retrieved surname**	**Jaro-Winkler distance**
ABERNATHY	ABERNETHY	0.977
AYRES	AYERS	0.96
BAIRD	BEARD	0.933
BRALLEY	BRAWLEY	0.947
BRITTON	BRITTAIN	0.944
CHRISTIE	CHRISTISON	0.94

CLARK	CLARKE	0.967
COLLISON	CULLISON	0.964
DENNEY	DENNY	0.967
DUFF	DUFFEL	0.933
FLICKINGER	FLUCKIGER	0.93
MCMURTRY	MCMURTREY	0.984
MILLICAN	MILLIKEN	0.937
PALLETT	PARLETTE	0.919
PARLET	PARLETTE	0.956
SAYRE	SAYER	0.961
SEELYE	SEELY	0.967
WETHERINGTON	WITHERINGTON	0.961

Manual inspection of the genealogical records showed that in a large number of cases the users indicated the spelling variant as an alternative ancestral surname.

Modeling the expected outcomes from a surname recovery

The probability of surname inference from personal genomes is dictated by three factors: the prior distribution of surnames in personal genomes datasets, the distribution of haplotypes within a surname, and the ability to successfully retrieve the surname from the database using the haplotype. For simplicity, we assumed that the distribution of surnames of personal genomes is similar to the distribution of surnames in the population.

Let $I_x(h,s)$ be an indicator function that returns 1 if querying the database with the combination of haplotype *h* and surname *s* returns the outcome *x,* where *x* is either: ‘success’, ‘wrong’, or ‘unknown’. Let f_s be the frequency of a surname and $\alpha(h,s)$ be the frequency of haplotype *h* in the surname *s*. Define $\beta_x(s) \triangleq \sum_{h\in\mathbb{H}(s)} \alpha(h,s)\, I_x(h,s)$, where $\mathbb{H}(s)$ is the set of haplotypes that are associated with the surname *s*. The probability of the surname recovery outcome *x* for a given population is:

$$P(x) = \frac{\sum_{s\in\mathbb{S}} f_s \beta_x(s)}{\sum_{s\in\mathbb{S}} f_s} \tag{3}$$

Where $\mathbb{S}$ is the set of all surnames in the population.

The probability in Eq. 3 can be assessed by sampling individuals from the population using the following estimator:

$$\hat{P}(x) = \frac{\sum_{s\in S} \hat{f}_s \hat{\beta}_x(s)}{\sum_{s\in S} \hat{f}_s} c + \frac{\sum_{s\in \bar{S}} \hat{f}_s \hat{\beta}_x(s)}{\sum_{s\in \bar{S}} \hat{f}_s} (1-c) \tag{4}$$

where S is the set of surnames in the sample that are known to be present in the tested databases and $\bar{S}$ is the set of surnames in the sample that are known to be absent from the tested databases. $\hat{f}_s$ is the estimated frequency of the surname based on the Census data, $\hat{\beta}_x(s) \triangleq \sum_{h \in \mathbb{H}(s)} \hat{\alpha}(h,s)\, I_x(h,s)$, and $\hat{\alpha}(h,s)$ is the frequency of the haplotype-surname combination in the sample, and *c* is the census coverage probability that was determined above. Eq.4 models the outcome rates as a weighted sum of sampling individuals from two distinct strata: those whose surname is found in the databases and those who do not. The two weights mitigate potential ascertainment biases in the sample and increase the confidence that the results reflect the target population.

Estimating the probability of surname recovery by inter-database comparisons

Our input sample relied on a cohort of individuals from the YBase database. This database was maintained by DNA Heritage and was acquired by FamilyTreeDNA in April 2011. FamilyTreeDNA provided us with surname-haplotype records from the database, without other identifiers that can expose the identity of the database users. The YBase and SMGF entries are completely distinct because the SMGF database lists only SMGF users. We took the following steps to remove potential duplicate records between Ysearch and Ybase: first, we asked FamilyTreeDNA to exclude YBase entries whose email addresses appear in Ysearch as well as entries without email addresses. Second, we removed from the downloaded copy of Ysearch all ~900 users that were tested with DNA Heritage. Third, we excluded any YBase user whose haplotype did not show a combination of markers that are typical to the DNA Heritage test panel. Thus, the input cohort was tested with a different company (DNA Heritage) than the database users. This reduces the chance of ascertainment biases due to oversampling of close relatives of the database participants.

Genetic genealogy databases are subject to nomenclature heterogeneity that can confound the analysis. This is especially problematic for DNA Heritage test panels that were subject to five nomenclature changes between 2003 to 2009 (see: http://web.archive.org/web/20100307032155/http://www.dnaheritage.com/helpfiles/DNA_Heritage_nomenclature_changes.pdf). For each input haplotype, we inspected the allelic ranges for markers that underwent significant nomenclature changes, such as DYS452, to decipher the nomenclature stratum and to standardize the haplotype according to the NIST recommended nomenclature. In addition, we set a tolerable genotype range for each marker that is equal to the marker mean value in Ysearch±3std. Entries outside of this range have a high likelihood of nomenclature differences and typos of users. This step filtered approximately 5% of YBase haplotypes. Finally, we selected only YBase haplotypes that have full genotyping results for a set of 34 STR markers (**table S2**) and whose surnames are in the US census. At the end of this process, we retained 911 YBase records (**table S3**).

We used a series of Perl scripts to challenge Ysearch and SMGF with the YBase haplotypes and to compare the returned surnames to the true ones (**table S4**). SMGF searches were conducted with the NIST nomenclature and Ysearch searches were conducted with FamilyTreeDNA nomenclature. The standard deviation was calculated by 30 iterations of re-sampling with replacement participants from the input cohort and repeating the analysis process.

The results of the 911 queries exhibited distinct patterns between the TMRCA of records that exactly match the true surname, records with a spelling variant, and records that returned the wrong surnames (**fig. S1**). The mean TMRCA was 10.3 generations for exact matches, 15.6 generations for a spelling variant, and 24.3 generations for wrong surnames. The TMRCA distribution of exact matches appeared to follow a geometric distribution trend. The TRMCA of records with spelling variants was almost never more recent than 10 generations and was quite different from the distribution of wrong matches. This provides another support for our spelling variations detection algorithm. **fig. S2** shows the final results after processing the results according to Eq. 4.

2. From Surnames to Individuals

The frequency distribution of recovered surnames

We determined the frequency distribution of recovered surnames from the YBase simulations using the following equation:

$$P(s \in \mathbb{S}_i | x = \ success, \delta) = \frac{P(x = success | s \in \mathbb{S}_i, \delta) P(s \in \mathbb{S}_i)}{P(x = success | \delta)} \qquad (5)$$

Where $\mathbb{S}_i$ is a subset of surnames whose frequencies fall in the *i*-th bin out of *j* possible bins. Specifically, we used the following bins:

Bin (i)	Frequency boundaries	Example of surnames in bin
1	>1:400	Smith, Johnson
2	1:400 – 1:4,000	Turner, Collins
3	1:4,000 – 1:40,000	Gates, Sloan
4	1:40,000 – 1:400,000	Bjork, Reach
5	<1:400,000	Kellog, Venter

The term $P(s \in \mathbb{S}_i)$ in Eq. 5 is given by the census data. The other numerator term can be approximated using a slight modification to Eq. 4:

$$\hat{P}(x = success|s \in \mathbb{S}_i, \delta) = \frac{\sum_{s\in\mathbb{S}} \hat{f}_s \hat{\beta}_x(s)}{\sum_{s\in S} \hat{f}_s} c_i + \frac{\sum_{s\in\overline{\mathbb{S}}} \hat{f}_s \hat{\beta}_x(s)}{\sum_{s\in\bar{S}} \hat{f}_s} (1 - c_i) \qquad (6)$$

Where c_i is a normalization factor that denotes the probability that a random person from the US population whose surname is in the *i-th* bin has at least a single entry in Ysearch and SMGF. c_i was determined by intersecting the census data with the list of Ysearch and SMGF. We used $\delta = 0.82$.

The recovered surnames are mostly found in the intermediate bin with a frequency of 1:4,000-1:40,000. Extremely rare surnames have the lowest relative risk for recovery due to the absence of records in Ysearch and SMGF. However, if these databases have even a single record for an extremely rare surname, then there is a 43% chance that the surname will be exposed (**fig. S3**). This phenomenon is potentially due to the small number of male lineages in extremely rare surnames.

Combining surnames with demographic identifiers

The joint probabilities of sex, age, and state were obtained from the US Census Population Estimates Program (www.census.gov/popest/states/asrh/files/SC-EST2009-AGESEX-RES.csv). The data is based on Census 2000 and contains a projection of residents to 2009, which was used in the simulation. Similar to HIPAA, ages that are over 85 were grouped in a single category.

The simulation ran 100,000 times. In each round, a combination of state and age was selected according to their probability in the joint distribution. For instance, there are 287,000 males in California who are 25 years old and 3,500 males in Idaho who are 75 years old. Accordingly, the probability of selecting "California, 25" was 82 times higher than selecting "Idaho, 75". Next, a bin of a recovred surname was selected according to its probability in Eq. 6 and a surname was selected according to its frequency in the bin. For instance, in the case of selecting the 1st bin (≥1:400), Smith had 1.28 higher probability of being sampled than Johnson. Finally, the simulation randomly selected between the return of a spelling variant or exact match, where the former had a probability 11.11%, based on our empirical findings in the Ybase simulations. In case of no spelling variant, the surname frequency was set to the census frequency; otherwise, the surname frequency was selected to be the sum of frequencies of all surnames that can be spelling variants of the original surname according to our spelling variant definition above. The last step portrays a scenario in which the adversary first looks for the target with the returned surname and if he cannot trace the target back, he tries all spelling variants. The number of expected individuals was found by multiplying the surname frequency by the number of males with the selected age and geographical location.

We validated the results of the simulation by comparing them to real datasets of US residents from PeopleFinders (www.peoplefinders.com). These datasets are based on extensive mining of public records, such as voter and drivers license registries, and can be searched by a combination of surname, age, and state. We selected 30 random simulation rounds that passed two criteria: (a) the ages were restricted to 25-35 years to avoid potential confounding due to underrepresentation of minors in public records and conflicting records from deceased individuals (b) the expected number of individuals should be 10-100 to avoid overloading the website. In most cases the lists in PeopleFinders were smaller than expected from simulations. Although we cannot rule out incompleteness of the website, the results also suggest that any underestimation of the list size - if it exists at all - is not significant.

3. Profiling Y-STRs from sequencing data

lobSTR usage

Unless otherwise specified, lobSTR v2.0.0 was used to profile Y-STRs from raw whole-genome sequencing data (*8*). In brief, lobSTR acts in three steps: detecting reads with repetitive elements that are flanked with non-repetitive regions, aligning the flanking regions to a reference, and measuring the repeat length for each STR.

Improved Y-STR reference

We modified lobSTR's standard STR reference to include the genomic locations and nomenclatures of genealogical Y-STRs. These locations were found by conducting *in silico* PCR on the UCSC genome browser using published Y-STR primers (*9-17*) and by searching the FamilyTreeDNA Y chromosome browser (ymap.ftdna.com). Several STR markers reside in duplicated regions of the Y chromosome. For instance, DYS385 has two distinct alleles in a single individual. Since lobSTR filters multi-mappers, we kept only one entry of these markers in the modified reference. Markers DYS448 and DYS449 consist of two STR regions separated by a non-repetitive region. For these, a separate reference entry was created for each region and the final genotype was determined by adding the alleles profiled at each of the two STR regions.

We did not include eight genealogical markers in the reference due to various technical reasons: markers GAAT1B07 and DYS724a/b (also known as CDYa/b) were excluded because their corresponding genomic coordinates could not be determined despite extensive literature searches. DYS726 was excluded because the genetic genealogy nomenclature could not be determined. DYS425 is one of the four repetitive loci of DYF371 (*17*), and using short reads we could not uniquely determine which locus a read originated from. DXYS156-Y was excluded because it is not specific to the Y-chromosome. Marker DYS19b was not included in because it is present in 0.2% of the population (*18*). Marker DYS640 was incorrectly annotated in our original reference and discarded from further analysis. Marker DYS464a-d was excluded because in most cases we typed fewer than four alleles and could not accurately assign typed alleles to forms a-d. In summary, our reference included 34 out of the 36 markers used by the SMGF panel and 79 out of the 87 markers in the most comprehensive test panel of FamilyTreeDNA. The genomic coordinates and conventions used for each Y-STR are given in **table S5**. All coordinates reported in this study follow the hg19 human reference build.

Processing lobSTR calls

lobSTR returns base pair length differences from the UCSC genome reference. Genetic genealogy services use an STR nomenclature that follows the PCR product sizes according to arbitrary primers (*19*). Whenever available we used the NIST nomenclature to translate lobSTR results (http://www.cstl.nist.gov/strbase/ystr_fact.htm). For searches in the Ysearch database results were converted to FamilyTreeDNA nomenclature using a conversion table available from SMGF (http://www.smgf.org/ychromosome/marker_standards.jspx).

For Y-STRs with a single genomic location, the allele with the modal number of supporting reads was used. Y-STR alleles that showed a non-integer number of repeat copies were discarded. We manually inspected a small number of calls where the modal allele was supported by less than 60% of reads aligned to the locus and enhanced the call by removing reads likely to be erroneous, such as reads that contain a high number of sequence mismatches, reads in which the STR resides towards the end of the read, or reads supporting alleles outside the normal range. Importantly, this procedure was executed completely blind to the true allele if it was known. For bi-mapper markers, such as DYS413a/b, the shortest repeat length was assigned to allele "a" and the next to allele "b".

Comparing lobSTR to the HGDP Y-STR panel

General approach

Sequence data for the HGPD panel were downloaded from the NCBI Short Read Archive from experiment SRP009145, sample SRS269343, runs SRX103805-130812. The sample included 10 HGDP individuals: HGDP00456 (Mbuti Pygmy), HGDP00665 (Sardinian), HGDP01284 (Mandenka), HGDP00542 (Papuan), HGDP00521 (French), HGDP00778 (Han Chinese), HGDP01307 (Dai), HGDP00927 (Yoruba), HGDP01029 (San), HGDP00998 (Karitiana). Samples were sequenced to a depth of 25-34x with paired end 100bp reads. Autosomal coverage was calculated using the samtools (*20*) depth tool and gives the average depth of covered bases based on alignments using BWA (*21*). lobSTR 2.0.0 with the improved Y-STR panel was used for the analysis. Y-STR haplotypes for the ten samples are given in table S6**.**

Genotypes for 76 Y-STRs typed by capillary electrophoresis for the 10 HGDP samples were obtained from the CEPH website (ftp://ftp.cephb.fr/hgdp_supp9/). Forty-seven of these markers overlapped with the lobSTR reference and were used to evaluate lobSTR's ability to type Y-STRs.

lobSTR reports alleles as the length difference from the UCSC, whereas the capillary genotypes are reported as the number of repeat copies at each locus. To convert lobSTR output to the same format, we used for following equation: $r + l/p$, where *r* is the number of base pairs of the STR of the lobSTR reference, *l* is the reported lobSTR allele in base-pairs, and *p* is the period of the Y-STR. For all individuals in which lobSTR recovered a genotype for DYS385a/b, only a single allele was returned. If the returned allele matched either the “a” or “b” form reported by the capillary platform, it was considered as correct. This follows our search strategy with the personal genomes, where these partial calls of multi-allelic markers were used to exclude matches not containing the lobSTR call for either allele.

We noticed that the lobSTR calls for all six individuals typed for DYS481 and all three individuals typed for DYS594 are exactly one repeat away from the results in the CEPH study. There is known nomenclature heterogeneity for these markers and some test kits report them with one shorter repeat than as reported by the NIST standard (*22*). Concordantly, we converted lobSTR calls to the shorter allele nomenclature to match that reported by CEPH.

Number of markers profiled at different sequencing coverage levels

Based on our previous experience with lobSTR, we assumed that STR coverage is linearly related to autosomal coverage. For each genome, we used the Picard (http://picard.sourceforge.net) DownsampleSam tool to randomly down-sample reads from the lobSTR alignment file to simulate coverage levels corresponding to autosomal coverage ranging from 1x to 25x. For each coverage level, we repeated the lobSTR allelotyping step to call the Y-STRs. The best-fit saturation curve was found using nonlinear least squares to fit a hyperbolic curve and was extended to predict haplotype lengths for up to 50x coverage.

Further investigation of wrong Y-STR calls

In our previous studies, we found that PCR stutter noise is a major source of error in calling STR alleles. This type of noise usually adds or subtracts a single repeat unit from the true allele. We noticed that the erroneous calls in DYS490 and DYS572 are several repeats away from the true allele, reducing the probability that these errors stem from stutter noise. Further analysis found that these two markers have X chromosome homologs, and that the calling errors can be attributed to misalignment of the X chromosome STRs. We also noticed that these markers were occasionally detected in the female genomes of the CEU panel, which provides further support for this hypothesis. Future algorithm improvements can use the homolog calls from the X chromosome to detect these errors.

4. Cases of Surname Inference from Personal Genomes

Querying genealogical databases

In all surname recovery experiments from personal genomes, database queries utilized the native search interfaces of the websites.

Ysearch was queried using the haplotype matching tool available at http://www.ysearch.org/search_search.asp?uid=&freeentry=true. Online searches were conducted with the default parameters and using the FamilyTreeDNA nomenclature. SMGF was queried using the tool at http://www.smgf.org/ychromosome/search.jspx with the options "Search by Match(%) = 85%" using the NIST nomenclature.

The US male sample from our lab collection

The sequencing experiment was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). To comply with the COUHES approval, we cannot share the specific Y-STR results. As an alternative, we provide summary statistics of the length distribution of the detected Y-STR makers.

Four Catch-All buccal swabs (Epicentre, QEC89100) were used to collect the sample according to the manufacturer's protocol. Genomic DNA was obtained by QuickExtract (Epicentre), followed by phenol-chloroform purification and ethanol precipitation. Library preparation was performed according to the standard Illumina protocol. Three runs of 101bp paired-end reads were generated with a GAIIx platform, generating 740 million reads. Autosomal coverage of 13x (after removing PCR duplicates) was measured using a conventional alignment pipeline as previously described (*23*). **fig. S5A** shows the overlap between the markers that were detected by Illumina versus the genealogical profile from Sorenson Genomics. **fig. S5B** shows the number of STRs that were detected using Illumina and Sorenson as a function of their lengths.

Database retrieval

We created a Ysearch record for the US male using the Ysearch.org website that does not disclose the true surname of the sample and consists of the Y-STR makers that are shared between Sorenson Genomics and Ysearch. Again, a search with the default website interface returned our sample as the top match.

Analyzing Michael Snyder's genome

Raw reads for the blood-derived and saliva-derived DNA of Michael Snyder's genome were downloaded from the NCBI Sequence Read Archive with accessions SRX097307 and SRX097312, respectively. lobSTR 1.0.6 with the native lobSTR reference was used to process both datasets using 20 processors on a server with four 12-core AMD Opteron™ 6100 Series. Forty-eight Y-STR calls were generated. All Y-STR calls were concordant between the blood-derived and the saliva-derived samples. The recovered Y-STR haplotype is given in **table S6.**

Ysearch link to search this haplotype:

http://www.ysearch.org/search_results.asp?uid=&freeentry=true&L1=14&L2=0&L3=16&L4=0&L5=10&L6=0&L7=0&L8=11&L9=13&L10=0&L11=0&L12=12&L13=0&L14=15&L15=0&L16=0&L17=11&L18=11&L19=0&L20=0&L21=0&L22=0&L23=0&L24=0&L25=0&L26=0&L27=0&L28=0&L29=0&L30=0&L31=0&L32=0&L33=0&L34=14&L35=18&L36=16&L37=19&L38=0&L39=0&L40=12&L41=10&L54=11&L55=8&L56=0&L57=0&L58=8&L59=11&L60=10&L61=8&L62=10&L63=0&L42=0&L64=22&L65=0&L66=0&L67=11&L68=12&L69=12&L70=0&L71=0&L49=13&L72=26&L73=0&L51=0&L74=13&L75=11&L76=12&L77=0&L78=9&L79=12&L80=11&L43=0&L44=12&L45=12&L46=0&L47=0&L48=13&L50=10&L52=0&L53=0&L81=9&L82=11&L83=14&L84=9&L85=15&L86=12&L87=0&L88=0&L89=0&L90=11&L91=10&L92=11&L93=0&L94=10&L95=11&L96=0&L97=0&L98=0&L99=0&L100=0&min_markers=8&mismatch_type=absolute&mismatches_max=0&mismatches_sliding_starting_marker=8&recaptcha_challenge_field=03AHJ_VutYkPmq2enCrhZuu94gU9-tcPRX33GpxRzVYZGBmnUWrEecYh8jggsJ0SU37BuJhpK_nMfhB0r8QTNBIe-_lpzJtyC3IRZ6SXlIn1Tnwb9vfGNo5ZojEQ8_8OlQgtCuVj5rTLfLLEXi4vr0-uFyo7upKwcsOFnxGg9SkL81vHEnACEx9H8&recaptcha_response_field=Weighthe+resume&haplo=&region=

SMGF link to search this haplotype:

http://www.smgf.org/ychromosome/search_results.jspx?labStandard=NIST&searchType=genetic&matchPercent=match_85&showCountries=on&showMissingData=on&showAllSurnames=on&DYS385_a=None&DYS385_b=None&DYS426=11&DYS447=None&DYS461=None&DYS388=13&DYS437=None&DYS448=None&DYS462=12&DYS389I=None&DYS438=10&DYS449=None&DYS463=None&DYS389B=None&DYS439=None&DYS452=None&DYS464_a=None&DYS464_b=None&DYS390=None&DYS441=14&DYS454=11&DYS464_c=None&DYS464_d=None&DYS391=10&DYS442=17&DYS455=11&GGAAT1B07=None&DYS392=12&DYS444=13&DYS456=14&YCAII_a=None&YCAII_b=None&DYS393=14&DYS445=10&DYS458=15&YGATAA10=14&DYS394=16&DYS446=None&DYS459_a=None&DYS459_b=None&YGATAC4=None&DYS460=None&YGATAH4=None

Analyzing John West's genome

Raw reads for John West genome were downloaded from NCBI Sequence Read Archive with accession SRA018104. lobSTR 1.0.6 with the improved Y-STR index using the same hardware settings for Michael Snyder genome. lobSTR called 58 Y-STR markers. The recovered Y-STR haplotype is given in **table S6.**

Ysearch link to search this haplotype:

http://www.ysearch.org/search_results.asp?uid=&freeentry=true&L1=13&L2=0&L3=14&L4=0&L5=11&L6=11&L7=14&L8=12&L9=12&L10=13&L11=0&L12=13&L13=0&L14=17&L15=0&L16=0&L17=11&L18=10&L19=0&L20=15&L21=0&L22=0&L23=0&L24=0&L25=0&L26=0&L27=0&L28=0&L29=0&L30=11&L31=10&L32=19&L33=23&L34=15&L35=19&L36=17&L37=17&L38=0&L39=0&L40=12&L41=12&L54=11&L55=9&L56=0&L57=0&L58=8&L59=10&L60=10&L61=8&L62=9&L63=10&L42=0&L64=0&L65=0&L66=16&L67=10&L68=12&L69=12&L70=15&L71=0&L49=12&L72=22&L73=0&L51=13&L74=0&L75=11&L76=14&L77=0&L78=0&L79=0&L80=0&L43=12&L44=11&L45=14&L46=0&L47=0&L48=13&L50=13&L52=0&L53=19&L81=9&L82=0&L83=16&L84=9&L85=16&L86=12&L87=11&L88=13&L89=13&L90=11&L91=10&L92=12&L93=0&L94=11&L95=10&L96=0&L97=0&L98=0&L99=0&L100=0&min_markers=8&mismatch_type=absolute&mismatches_max=0&mismatches_sliding_starting_marker=8&recaptcha_challenge_field=03AHJ_VusNldFpOWxRw2dib-HZoXRWEvElRysd8fba2-AEWcvfROt3W2n0f6ARIuHaqcRgZ1JE92e0aXBEDDpPLRfhPpAYpKvyARJb0FqPs1fP_HPkMw8AiwilCMic_tD_ntx119pL-fmM96E18ekPuaxXIu-0Dw0hIg&recaptcha_response_field=Hcacco+and&haplo=&region=

SMGF link to search this haplotype:

http://www.smgf.org/ychromosome/search_results.jspx?labStandard=NIST&searchType=genetic&matchPercent=match_85&showCountries=on&showMissingData=on&showAllSurnames=on&DYS385_a=11&DYS385_b=14&DYS426=12&DYS447=None&DYS461=12&DYS388=12&DYS437=15&DYS448=None&DYS462=11&DYS389I=None&DYS438=12&DYS449=None&DYS463=19&DYS389B=None&DYS439=13&DYS452=None&DYS464_a=None&DYS464_b=None&DYS390=None&DYS441=14&DYS454=11&DYS464_c=None&DYS464_d=None&DYS391=11&D

YS442=17&DYS455=11&GGAAT1B07=None&DYS392=13&DYS444=12&DYS456=15&YCAII_a=19&YCAII_b=23&DYS393=13&DYS445=13&DYS458=17&YGATAA10=16&DYS394=14&DYS446=13&DYS459_a=None&DYS459_b=None&YGATAC4=None&DYS460=11&YGATAH4=11

Surname recovery using the Craig Venter dataset

Sequence reads for the Venter genome were downloaded from TraceDB (Genbank accession ABBA00000000). We trimmed the first 50bp of every read due to the high error rate at the beginning of Sanger sequence reads and discarded reads whose length after trimming was less than 100bp.

At the default settings, lobSTR 2.0.0 with the improved Y-STR index returned 40 Y-STRs after 40 minutes of runtime using the same hardware settings as described above. Markers returning a non-integer number of repeat copies were discarded.

Ysearch link to search this haplotype:

http://www.ysearch.org/search_search.asp?fail=1&uid=&freeentry=true&L1=0&L2=0&L3=0&L4=0&L5=10&L6=0&L7=0&L8=12&L9=12&L10=12&L11=0&L12=13&L13=0&L14=17&L15=9&L16=0&L17=11&L18=11&L19=0&L20=0&L21=0&L22=0&L23=0&L24=0&L25=0&L26=0&L27=0&L28=0&L29=0&L30=0&L31=0&L32=19&L33=23&L34=0&L35=0&L36=0&L37=17&L38=0&L39=0&L40=12&L41=12&L54=12&L55=9&L56=15&L57=16&L58=9&L59=10&L60=10&L61=8&L62=0&L63=0&L42=0&L64=23&L65=0&L66=16&L67=10&L68=12&L69=0&L70=16&L71=8&L49=0&L72=22&L73=0&L51=0&L74=12&L75=11&L76=0&L77=0&L78=0&L79=13&L80=12&L43=12&L44=11&L45=0&L46=0&L47=0&L48=0&L50=0&L52=0&L53=0&L81=0&L82=0&L83=16&L84=9&L85=0&L86=0&L87=0&L88=0&L89=12&L90=11&L91=0&L92=0&L93=12&L94=11&L95=0&L96=25&L97=0&L98=0&L99=0&L100=0&min_markers=8&mismatch_type=absolute&mismatches_max=0&mismatches_sliding_starting_marker=8&recaptcha_challenge_field=03AHJ_VusyS2psJJIgHViP9Prgl35afzMpQdoc1uJYw3a1I3Lob-ycMFtplYmSlwFUE-GDzsh-4mdVv9uutxFV7-2qugmcKl8jVTG3EnVPwKXNihNKdv-TfVxulspdX1RO-5XhOBVpnPWoZhnxE5OVRCTnXF7fVgXO7TAa-0c-ycvvN9Zp-JDq_Io&recaptcha_response_field=tsshora+infinite&haplo=&region=

SMGF link to search this haplotype:

http://www.smgf.org/ychromosome/search_results.jspx?labStandard=NIST&searchType=genetic&matchPercent=match_85&showCountries=on&showMissingData=on&showAllSurnames=on&DYS385_a=None&DYS385_b=None&DYS426=12&DYS447=None&DYS461=12&DYS388=12&DYS437=None&DYS448=None&DYS462=11&DYS389I=None&DYS438=12&DYS449=None&DYS463=None&DYS389B=None&DYS439=12&DYS452=None&DYS464_a=None&DYS464_b=None&DYS390=None&DYS441=None&DYS454=11&DYS464_c=None&DYS464_d=None&DYS391=10&DYS442=17&DYS455=11&GGAAT1B07=None&DYS392=13&DYS444=None&DYS456=None&YCAII_a=19&YCAII_b=23&DYS393=None&DYS445=None&DYS458=17&YGATAA10=None&DYS394=None&DYS446=None&DYS459_a=9&DYS459_b=None&YGATAC4=None&DYS460=None&YGATAH4=None

Querying Ysearch as described above returned the entry VPBT4 with surname “Venter” as the top hit. The results, including the trace numbers of supporting reads, are summarized in **table S6** and reported in **table S7.** Concordant with Craig Venter’s paternal roots, the top match was the only Venter record in Ysearch with a UK ancestor.

Demographic profiling was conducted using PeopleFinders and USSearch (www.ussearch.com). Female names and users that did not exactly match year of birth=1946 were discarded.

CEU genomes

The CEU male datasets were accessed through the 1000Genomes publicly available Amazon S3 bucket and the European Nucleotide Archive. In cases of father-son pairs, we selected the father for further analysis. All datasets were first processed with lobSTR 2.0.0 with the native STR reference. We reran the 18 CEU genomes that returned the largest number of markers with the improved Y-STR panel. Overall, these genomes had longer read lengths of 76-100bp compared to 36-51bp and were therefore more amenable to STR calling. To validate calls in the low coverage genomes, Y-STRs typed using capillary electrophoresis for 16 Y-STR markers for 10 of the 17 individuals were obtained from He, *et al*. (*24*). In 41/43 comparable markers the genotypes were concordant. The two incorrect cases were off by a single repeat unit and covered only by a single read. All searches were first performed using only the markers typed using lobSTR. Four genomes were supplemented with the markers from He, *et al*. since their searches returned a large number of poorly matching records due to low number of calls in popular markers. Autosomal coverages were measured as reported for the HGDP samples.

Determining the probability of random matches

We determined the probability that at least one household would randomly match the surname and demographic characteristics of the CEU pedigrees. Let *n* be the number of households that hold the recovered surname in the geographical region, *p* the probability that a household matches additional metadata available for the sample, and f_1 *and* f_2 the frequencies of the recovered surname of the paternal and maternal grandfathers. If only one surname was recovered, $f_2=1$. The probability of at least a single random match is:

$$P(\geq 1\, match) = 1 - (1-p)^n \tag{7}$$

In our case, *n* is the number of married households in Utah with the recovered surname. We approximate $n \cong \lceil n_{utah} f_1 \rceil$, where n_{utah} is the total number of married households in Utah, which according to the 2002 census matches to 443,210.

For *p*, we accounted for the additional metadata regarding the number of children, male/female order of the children, and knowledge of the surname of the other set of grandparents. We set *p* to:

$$p = f_2 p_c \frac{1}{2^k} \tag{8}$$

where p_c is the probability that a household has the given number of children, *k* is the number of children in the pedigree and $\frac{1}{2^k}$ is the probability that the male/female order of the children matches that in the pedigree. The upper bound of p_c is 3.5%, which corresponds to the percentage of households in Utah with 5 or more children as determined by the 2000 US Census using the search tool at factfinder2.census.gov. We used this number because data on larger households were not available. This gave the probability of finding at least one random match as:

$$P(\geq 1\,match) = \ 1 - (1 - f_2 p_c \frac{1}{2^k})^{n_{utah} f_1} \quad (9)$$

We note that the order in which surnames are assigned to surnames 1 and 2 does not significantly change this probability as, *1-(1-p)^n* converges to *np* for small *p*, and therefore:

$$P(\geq 1\,match) \approx np = \ n_{utah} f_P f_M p_c \frac{1}{2^k} \quad (10)$$

which also gives the expected number of households that give random matches to the desired characteristics.

One limitation in our analysis is the $n \cong \lceil n_{utah} f_1 \rceil$ approximation that implies that the surname distribution in Utah is very close the surname distribution in the entire US. These two distributions are expected to be relatively close for highly prevalent surnames, but extremely rare surnames can be quite localized. This case was only of a concern for pedigree 3, where its surname is found in only a few hundred individuals in the US. To test the robustness of our analysis, we re-calculated the probability of a random match for this pedigree as if all individuals in the US with this surname live in Utah and each individual is a member of a distinct household. In this scenario, the probability of a random match was 0.3%, which is still significantly low. Notice that this analysis is extremely conservative. The assumption that each of the hundreds of individuals reside in a distinct household is not realistic. In addition, we did not take into account additional metadata, such as the probability to find the exact number of children and the fact that all grandparents were alive during the last year of CEU sample collection, which should further drive down the probability of a random match.

5. Y-STR masking and imputation

One potential solution to surname inference is to mask the Y-STR loci. However, genetic masking is sensitive to imputation strategies. A striking example of this limitation was the ability to recover Jim Watson's masked ApoE status from adjacent SNPs in linkage disequilibrium (*25*), raising the possibility of also bypassing Y-STR masking.

Theoretically, it seems possible to impute genealogical Y-STR haplotypes from Y-SNPs. The rate of SNPs is $3*10^{-8}$ mutations per bp per generation, which translates to a rate of 0.5 *de novo* mutations in the euchromatic region of the Y chromosome per generation. On the other hand, Y-STR variations occur at a smaller rate of ~0.1 mutations per haplotype of 30 markers per generation. This rate difference has been recently demonstrated by deep sequencing the Y chromosomes of two individuals that were separated by 13 meiosis events (*26*). The two individuals had identical Y-STR haplotypes but differed at four Y-SNPs. The excess of *de novo* SNPs over STRs implies that Y-STR haplotypes can be uniquely tagged by Y-SNP haplotypes.

Y chromosome imputation has different properties imputation in autosomal regions. In the autosomes, recombination divides the chromosome into segments with distinct genealogies. The task of autosomal imputation algorithms is to detect segment transitions and match the corresponding ancestral haplotype block from the reference panel (*27, 28*). Y-STRs reside on one long chromosome block. The divide and conquer approach cannot work and the entire Y chromosome block must be imputed in a single step. On one hand, this drastically reduces the computation time needed for imputation. On the other hand, a necessary condition for accurate imputation is that the reference panel must include the Y-STR alleles as a single haplotype block. Accurate imputation will not work if the masked STR alleles are scattered across a collection of reference chromosomes. For instance, if the masked Y-STR haplotype is 14-15-20-11, and the reference has four chromosomes: 14-X-X-X, X-15-X-X, X-X-20-X, and X-X-X-11, where X indicates a mismatch to the masked haplotype, imputation will not return an accurate result. Given that condition, every imputed Y-STR *haplotype* (as opposed to alleles in the autosome) must be documented in the reference panel.

We evaluated the dependency between the reference panel size and the success rates. We focused on Ysearch since SMGF does not list the raw Y-STR haplotypes. Ysearch contains approximately 34,000 unique haplotypes of 30 popular STR markers. These haplotypes cover 34.5% of the haplotypes that segregate in the population according to the Good-Turing frequency estimation procedure (*29*). The reference panels were constructed by re-sampling Ysearch haplotypes using a two-stage procedure: (a) with a probability of 100%-34.5=65.5%, a mock haplotype was sampled. This denotes a haplotype in the reference panel that is not in Ysearch. Otherwise, the procedure continued to the next stage (b) a Ysearch haplotype was sampled according to its frequency in the database. This two-stage procedure was run *N* times, where *N* was the size of the reference panel. Simulating Y-SNPs was not necessary because we assumed that given the size of the haplotype block, imputation always correctly recovers the Y-STR haplotype from the Y-SNP, as long as the former is in the panel. We then conducted surname recovery experiments with YBase using the Ysearch database and the simulated reference panel. If a YBase haplotype was not part of the reference panel, then surname recovery automatically failed and was categorized under the 'unknown' state.

Our results show that with large reference panels of 50,000 male genomes from the US population, the surname recovery success rate is 5% (**fig. S6**). This suggests that imputation is not an immediate threat to masking, but can be problematic as a long term solution.

In addition, we noticed that some community efforts, such as Y Chromosome Genome Comparison (daver.info/ysub), have started linking between Y-SNPs and surnames. These efforts might also enable the bypassing of Y-STR masking.

Supplementary Figures

Figure S1:

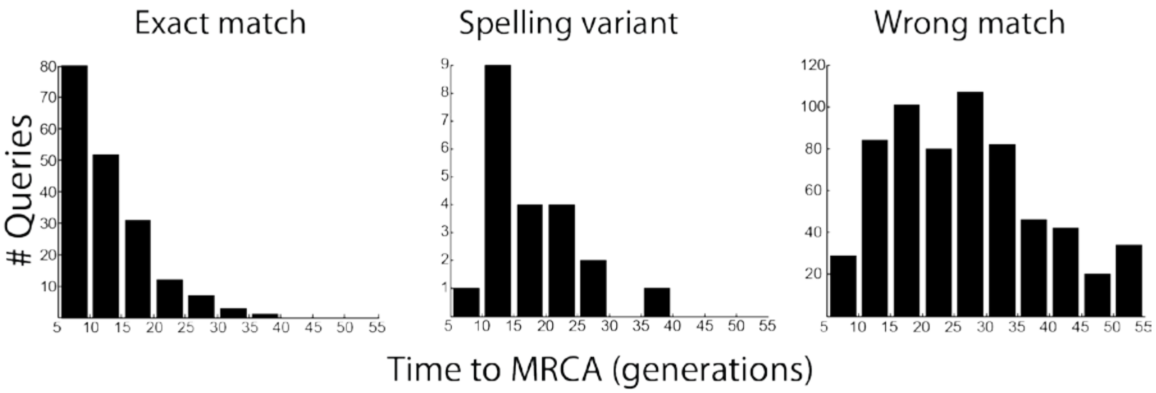

Figure S1: The TMRCA profiles of haplotype queries. Records that matched exactly the input surname (left) showed a geometric-like distribution. For most records with a minute spelling variant from the original surname (center) the MRCA was 10-15 generations ago. Wrong matches (right) mainly showed an ancient MRCA.

Figure S2:

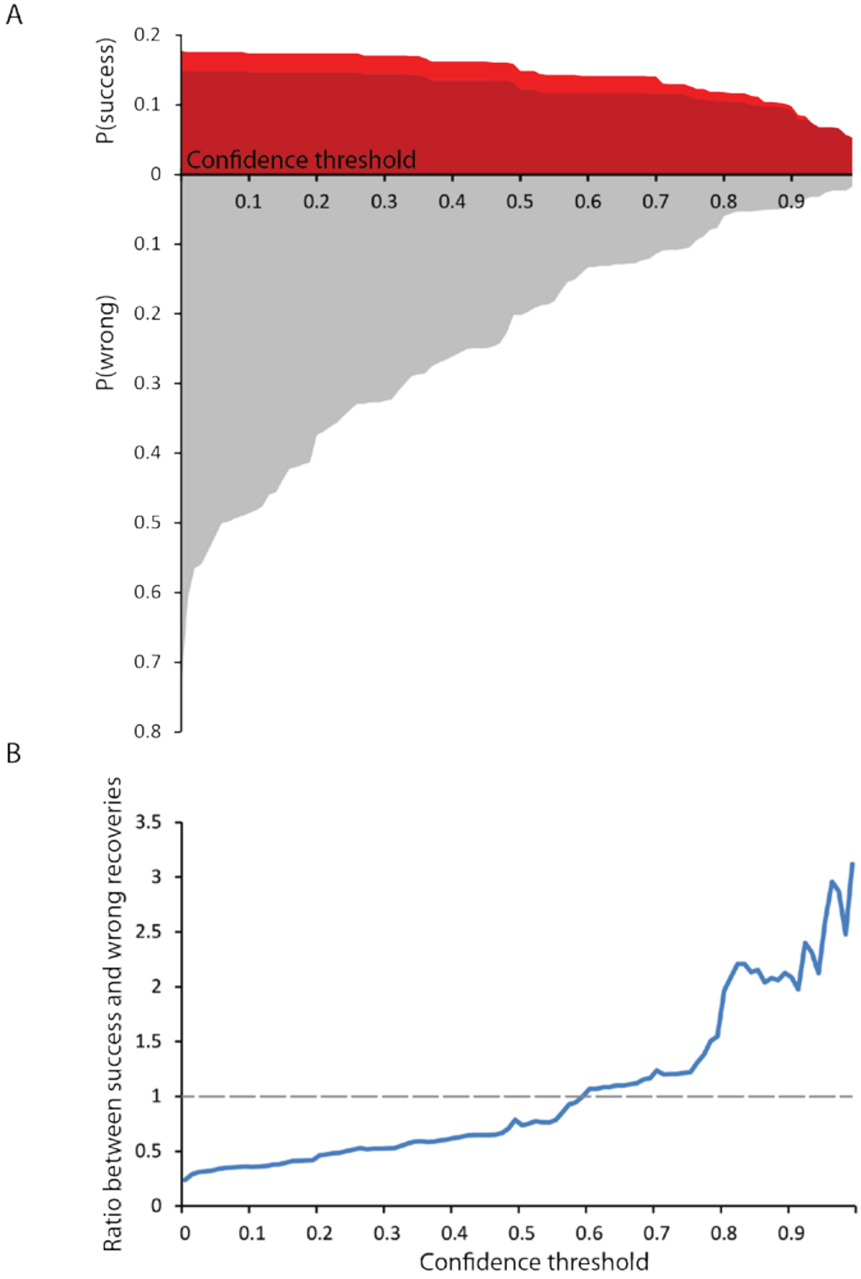

Figure S2: Performance of surname recovery at different confidence thresholds. (**A**) The rate of successful recovery with exact matches (dark red) and spelling variants (light red) versus the wrong recovery rate (gray) as a function of confidence threshold level. (**B**) The ratio between successful recoveries to wrong recoveries.

Figure S3:

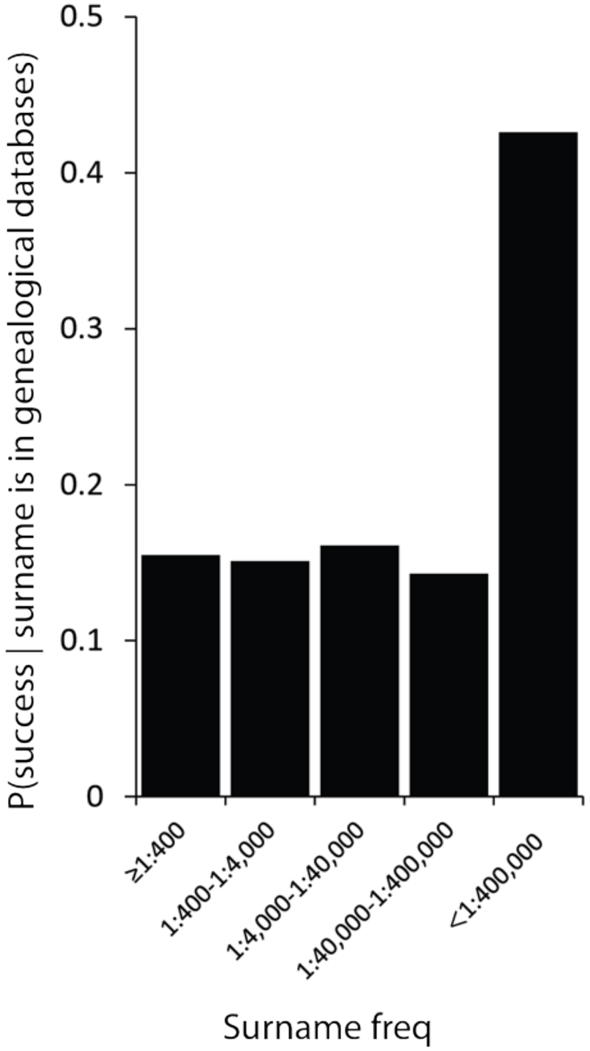

Figure S3: The probability of successful recovery given that the surname has at least one record in Ysearch or SMGF as a function of the surname frequency.

Figure S4:

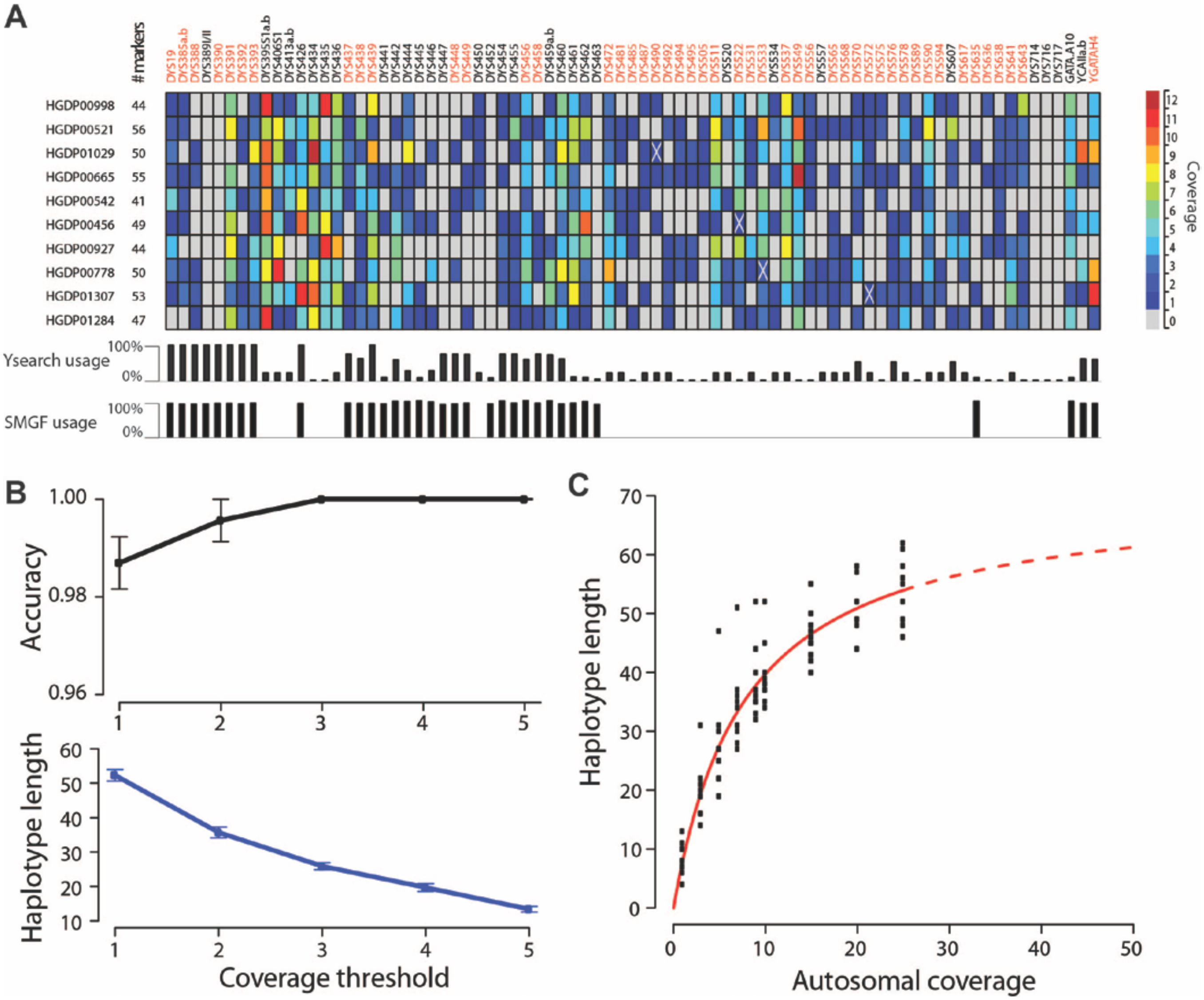

Figure S4: (**A**) lobSTR calling performance on Y-STR haplotypes from ten male genomes. The length of the Y-STR haplotype for each genome is reported on the left. The heatmap denotes the number of reads aligned by lobSTR for each marker. Forty-seven markers (red) were genotyped with capillary electrophoresis. An "X" symbol denotes a discordant allele compared to the electrophoresis calls. Bar plots show the percentage of users in each database that were tested for each marker. (**B**) Expected lobSTR accuracy and Y-STR haplotype length at increasing coverage thresholds. Error bars denote standard error. (**C**) The expected number of alleles in Y-STR haplotypes at different sequencing coverage levels. Different coverage levels were simulated by down sampling from lobSTR aligned reads for the 10 HGDP samples. Black – the number of Y-STR calls for each genome after down sampling. Red – best fit saturation curve.

Figure S5:

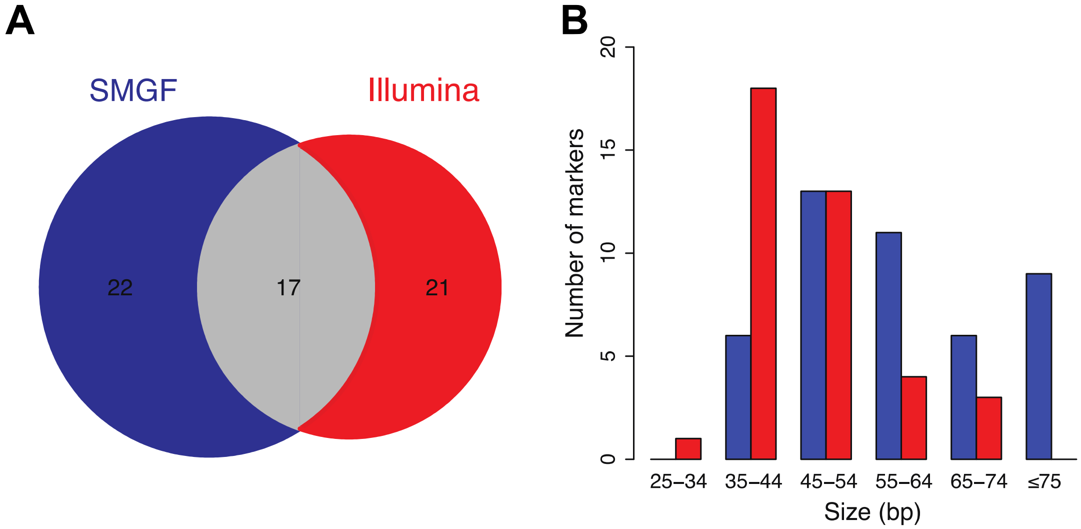

Figure S5: Comparison between Illumina Y-STR profiling and the Sorenson Genomics genetic genealogy service. (A) Illumina profiling returned the results of 38 Y-STR markers. The genetic genealogy service uses a panel of 49 markers, 39 of which are included in lobSTR's Y-STR reference. The results of all 17 markers that were profiled by both strategies were identical. **(B)** The distribution of total STR region lengths is shown for the markers typed by Sorenson (blue) versus markers typed by lobSTR (red).

Figure S6:

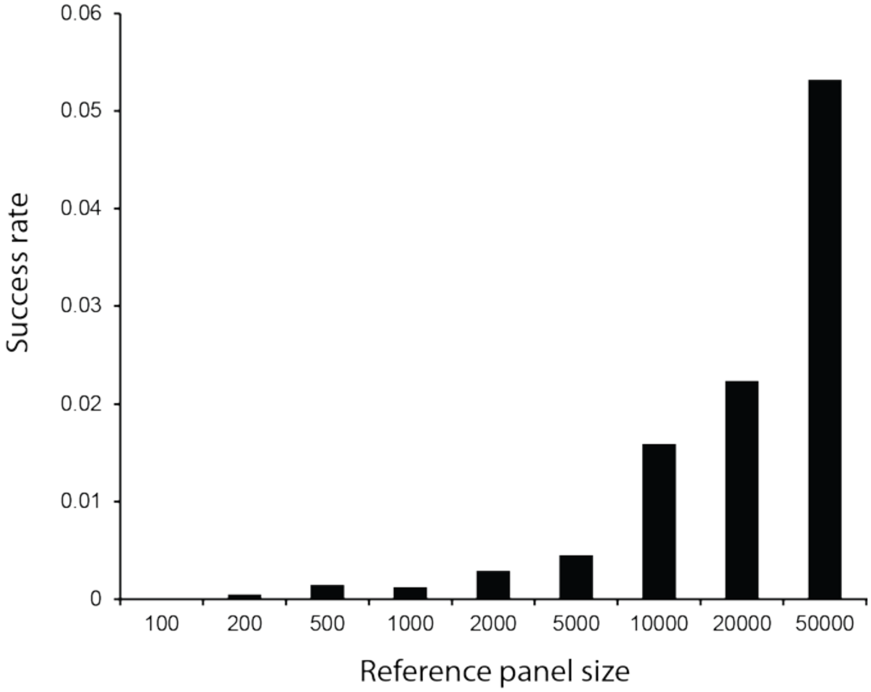

Figure S6: The estimated success rate for surname recovery after imputation as a function of the imputation panel size.

Supplementary Tables

Table S1

	Site	Estimated number of Records	Maintained by:	Availability	Search Interface
Databases	Ancestry DNA (dna.ancestry.com)	50,000	Ancestry.com	Public (fee required)	Search by STR Haplotype
	Family Tree DNA (familytreedna.com)	250,000	Family Tree DNA	Closed	Not searchable
	Oxford Ancestors (www.oxfordancestors.com/)	?	Oxford Ancestors	Closed	?
	SMGF (smgf.org/pages/ydatabase.jspx)	38,000	Sorenson Molecular Genealogy Foundation[1]	Public (free account required)	Search by surname or STR Haplotype
	WorldFamilies (worldfamilies.net/surnames)	150,000[3]	Collection of admins of surname projects.	Public	Search by surname
	Ybase	13,000	DNA Heritage[2]	Previously public. Discontinued.	Discontinued
	Y Chromosome Genome Comparison (daver.info/ysub)	1,000	Volunteers	Public	Download raw SNP data
	Ymatch (dna-fingerprint.com)	1,300	Family Tree DNA	Public	Search by STR and SNP haplotypes
	Ysearch (ysearch.org)	105,000	Family Tree DNA	Public	Search by surname or STR Haplotype
Examples of surname projects	Brown DNA Study (http://brownsociety.org/browndna/results.htm)	800+	Brown members	Public	Table of Y-STR haplotypes
	Clan Donald USA (http://dna-project.clan-donald-usa.org/)	1000+	Donald clan members	Public	Table of Y-STR haplotypes
	McDuffie DNA Surname Project (http://www.mcduffiedna.com/)	150+	McDuffie members	Public	Table of Y-STR haplotypes
	SmithConnections DNA Project (http://www.smithconnections.com/)	500+	Smith members	Public	Table of Y-STR haplotypes
	Williams DNA Project (http://williams.genealogy.fm/)	800+	Williams members	Public	Table of Y-STR haplotypes

List of major genetic genealogy sites that display Y chromosome and surname information. The top section lists genetic genealogy databases. The bottom section lists examples of privately maintained websites that are dedicated to a single surname.

[1] SMGF was recently acquired by Ancestry.com

[2] DNA Heritage was acquired by FamilyTreeDNA in 2011

[3] Includes only users whose surnames are present in the 2000 US Census

Table S2

Count	Marker	Expected mutation rate	Mean	σ
1	DYS19	0.00437	14.34	0.8045
2	DYS385a	0.00208	12.0869	1.6522
	DYS385b	0.00414	14.5464	1.449
3	DYS388	0.000425	12.5142	1.0753
4	DYS389a	0.00551	12.9668	0.6644
	DYS389b	0.00383	29.326	1.0418
5	DYS390	0.00152	23.6032	1.0229
6	DYS391	0.00323	10.4858	0.6104
7	DYS392	0.00097	12.3413	1.1069
8	DYS393	0.00211	13.0752	0.6025
9	DYS426	0.000398	11.6459	0.5198
10	DYS437	0.00153	14.9094	0.6931
11	DYS438	0.000956	11.2206	1.0643
12	DYS439	0.00384	11.66	0.8567
13	DYS442	0.00978	17.2273	1.3301
14	DYS444	0.00545	12.3666	0.892
15	DYS445	0.00216	11.6015	0.9401
16	DYS446	0.00267	13.1767	1.372
17	DYS447	0.00212	24.6396	1.2057
18	DYS448	0.000394	19.3437	0.8748
19	DYS449	0.0122	29.5472	1.6474
20	DYS452	0.00402	30.1854	1.1041
21	DYS454	0.000475	11.0484	0.3744
22	DYS455	0.000426	10.648	0.9704
23	DYS456	0.00494	15.4571	1.1065
24	DYS458	0.00836	16.6389	1.2634
25	DYS459a	0.00013	8.753	0.5017
	DYS459b	0.00013	9.601	0.5422
26	DYS460	0.00622	10.6976	0.639
27	DYS461	0.000989	11.882	0.6914
28	DYS462	0.00265	11.3571	0.6266
29	DYS464a	0.00018	13.8555	1.4488
	DYS464b	0.00018	14.7374	1.0564
	DYS464c	0.00018	15.8236	1.124
	DYS464d	0.00018	16.5742	1.1157
30	DYS635	0.00385	22.6604	1.1601
31	GATA-A10	0.00332	15.5234	1.2242
32	GATA-H4	0.00322	10.7333	0.7801
33	GGAAT1B07	0.0024	10.2854	0.7397
34	YCAIIa	0.002	19.0997	0.905
	YCAIIb	0.002	22.136	1.2624

List of markers used to challenge Ysearch and SMGF. Mutation rates are based on Ballantyne et al. *(3).* YCAII was absent from this study and set to 0.002 according to Walsh (*1*). Mean and standard deviations for marker values are calculated using Ysearch with NIST nomenclature.

Table S3: Surname haplotype pairs used to challenge Ysearch and SMGF. The original data was kindly provided by FamilyTreeDNA based on user-generated content in the discontinued Ybase database. When applicable, the presented haplotypes were subject to NIST nomenclature standardization to reduce nomenclature heterogeneity. The order of the columns follows the Ysearch native order. This table is provided as a separate Excel document.

Table S4: Results of database queries using Ysearch and SMGF haplotypes. This table is provided as a separate Excel document.

Table S5

	Marker	Start (chrY)	End (chrY)	Alt. locations	Ref. allele	Motif structure
1	DYS394/19	9521989	9522052		15	[TAGA]3TAGG[TAGA]n
2	DYS385a/b	20842518	20842573	chrY:19260956-19261212	14	[GAAA]n
3	DYS388	14747535	14747570		12	[ATT]n
4	DYS389I	14612191	14612238		12	[TCTG]m[TCTA]n
	DYS389B	14612338	14612405		29	[TCTG]m[TCTA]n
5	DYS390	17274947	17275042		24	[TCTG]n[TCTA]m[TCTG]p[TCTA]q
6	DYS391	14102795	14102838		11	[TCTA]n
7	DYS392	22633873	22633911		13	[TAT]n
8	DYS393	3131152	3131199		12	[AGAT]n
9	DYS406S1	23843595	23843634		10	[TATC]n
10	DYS413a/b	16099088	16099133	chrY:14676647-14676820	23	[TG]n
11	DYS426	19134850	19134885		12	[GTT]n
12	DYS434	14466533	14466568		9	TAAT[CTAT]n
13	DYS435	14496298	14496333		9	[TGGA]n
14	DYS436	15203862	15203897		12	[GTT]n
15	DYS437	14466994	14467057		16	[TCTA]n[TCTG]2[TCTA]4
16	DYS438	14937824	14937873		10	[TTTTC]n
17	DYS439	14515312	14515363		13	[GATA]n
18	DYS441	14981831	14981908		16	[TTCC]n
19	DYS442	14761103	14761168		17	[TATC]2[TGTC]3[TATC]n
20	DYS444	19226192	19226247		14	[TAGA]n
21	DYS445	22092602	22092649		12	[TTTA]n
22	DYS446	3131458	3131527		14	[TCTCT]n
23	DYS447	15278740	15278854		23	[TAATA]n[TAAAA]1[TAATA]m[TAAAA]1 [TAATA]p

24*	DYS448_1	24365070	24365136		11	[AGAGAT]n
	DYS448_2	24365178	24365225		8	[AGAGAT]n
25*	DYS449_1	8218014	8218074		13	[TTTC]n
	DYS449_2	8218124	8218179		14	[TTTC]n
26	DYS450	8126300	8126344		8	[ATTTT]n
27	DYS452	21620478	21620632		31	[TATAC]m[TGTAC]n[TATAC]p[CATAC][TATAC][CATAC][TATAC]q[CATAC]r[TATAC]s[CATAC][TATAC]t
28	DYS454	8224156	8224199		11	[AAAT]n
29	DYS455	6911569	6911612		11	[AAAT]n
30	DYS456	4270960	4271019		15	[AGAT]n
31	DYS458	7867880	7867943		16	[GAAA]n
32	DYS459a/b	26078851	26078890	chrY:26292857-26293004	10	[TAAA]n
33	DYS460	21050842	21050881		10	[ATAG]n
34	DYS461	21050690	21050737		12	[TAGA]n[CAGA]
35	DYS462	21317047	21317090		11	[TATG]n
36	DYS463	7643509	7643628		24	[AAAGG]m [AAGGG]n [AAGGA]p
37	DYS472	16508484	16508507		8	[AAT]n
38	DYS481	8426378	8426443		22	[CTT]n
39	DYS485	22099634	22099681		16	[TTA]n
40	DYS487	8914174	8914212		13	[TTA]n
41	DYS490	3443765	3443800		12	[TTA]n
42	DYS492	17414337	17414369		12	[ATT]n
43	DYS494	21386168	21386197		10	[TTA]n
44	DYS495	15011300	15011346		15	[AAT]n
45	DYS505	3640831	3640878		12	[TCCT]n
46	DYS511	17304923	17304958		10	[GATA]n

47	DYS520	7730432	7730511		20	[ATAG]n[ATAC]n
48	DYS522	7415625	7415664		10	[GATA]n
49	DYS531	8466195	8466238		11	[AAAT]n
50	DYS533	18393226	18393273		12	[ATCT]n
51	DYS634	18392976	18393035		15	[CTTT]n
52	DYS537	19358850	19358889		10	[TCTA]n
53	DYS549	21520224	21520275		13	[GATA]n
54	DYS556	22601453	22601496		11	[AATA]n
55	DYS557	23234712	23234775		16	[TTTC]n
56	DYS565	16526732	16526775		12	[ATAA]n
57	DYS568	8822555	8822594		11	[AAAT]n
58	DYS570	6861231	6861298		17	[TTTC]n
59	DYS572	3679660	3679699		10	[AAAT]n
60	DYS575	7436257	7436296		10	[AAAT]n
61	DYS576	7053359	7053426		16	[AAAG]n
62	DYS578	22562564	22562599		9	[AAAT]n
63	DYS589	24485693	24485757		12	[TTTTA]n
64	DYS590	8555980	8556019		8	[TTTTG]n
65	DYS594	21656837	21656886		10	[AAATA]n
66	DYS607	18414382	18414457		19	[GAAG]n[GAAA][GAAG][GAAA][GAAG]
67	DYS617	19081518	19081553		12	[TTAn]
68	DYS635	14379564	14379655		23	[TCTA]4[TGTA]2[TCTA]2[TGTA]2[TCTA]2[TGTA]m[TCTA]n
69	DYS636	22634857	22634900		12	[ATTT]n
70	DYS638	17645491	17645534		11	[TTTA]n
71	DYS641	16134296	16134335		10	[TAAA]n
72	DYS643	17426012	17426066		11	[CTTTT]n

73	DYS714	22147731	22147865		27	[TTTCT]m[CTTCT]n[TTTCT]p[CTTCT]q [TTTCT]r
74	DYS717	17313245	17313324		16	[GTACT]m [GTATT]n
75	GATA-A10	18718879	18718938		15	[TCCA]2 [TATC]n
76	GATA-H4	18743553	18743600		12	[TAGA]n
77	YCAIIa/b	19622111	19622156	chrY:19016986-19017135	23	[CA]n
78	DYS395S1a/b	19739341	19739381	chrY:18899736-18899977	15	[AAC]n
79	DYS716	13140129	13140274		28	[ACTCGC][ACTCC]m[ATTCC]n[TATTCTATTGA][ACTCC][ATTCC][ACTCC]2[ATTCA][ATTCC]2[ACTTC][ATTCC]

Y-STR genomic locations and conventions. All coordinates are given for human genome build hg19. Conventions follow NIST guidelines whenever available.

*The values for DYS448 and DYS449 were determined by adding the alleles typed at DYS448_1/DYS448_2 and DYS449_1/DYS449_2. The complete repeat structures for DYS448 and DYS449 are [AGAGAT]mN42[AGAGAT]n and [TTTC]m [N]50 [TTTC]n, respectively.

Table S6: Y-STR haplotypes profiled from sequencing datasets. The allele reported by lobSTR for each marker is given for the Snyder, West, Venter, and HGDP genomes. NA indicates lobSTR did not type that marker. A "-" for markers with multiple forms indicates that not all alleles at that marker could be confidently typed. This table is provided as a separate Excel document.

Table S7

<table>
<tr><th></th><th>Marker</th><th>Craig Venter</th><th>Best Ysearch hit (user VPBT4)</th><th>Supporting reads (Genbank numbers)</th></tr>
<tr><td>1</td><td>DYS388</td><td>12</td><td>12</td><td>gnl|ti|1743110387 1094789366005</td></tr>
<tr><td>2</td><td>DYS391</td><td>10</td><td>10</td><td>gnl|ti|1745937715 1094791529859</td></tr>
<tr><td>3</td><td>DYS392</td><td>13</td><td>13</td><td>gnl|ti|1737227188 1098315434560
gnl|ti|1746572651 1094837174504</td></tr>
<tr><td>4</td><td>DYS395S1a</td><td>15</td><td>15</td><td>gnl|ti|1737262859 1098315213789
gnl|ti|1738241462 1099476577665</td></tr>
<tr><td></td><td>DYS395S1b</td><td>16</td><td>16</td><td>gnl|ti|1733762175 1099341809387</td></tr>
<tr><td>5</td><td>DYS413a</td><td>23</td><td>23</td><td>gnl|ti|1747767205 1094853399058</td></tr>
<tr><td>6</td><td>DYS426</td><td>12</td><td>12</td><td>gnl|ti|1748257622 1094907283222</td></tr>
<tr><td>7</td><td>DYS436</td><td>12</td><td>12</td><td>gnl|ti|1734288742 1099289773803
gnl|ti|1738919653 1099519453685
gnl|ti|1742785287 1094374813149</td></tr>
<tr><td>8</td><td>DYS438</td><td>12</td><td>12</td><td>gnl|ti|1735017296 1099742076569</td></tr>
<tr><td>9</td><td>DYS439</td><td>12</td><td>12</td><td>gnl|ti|1748422688 1094793240651</td></tr>
<tr><td>10</td><td>DYS442</td><td>12</td><td>12</td><td>gnl|ti|1733501518 1099268225126
gnl|ti|1748825315 1094791033063</td></tr>
<tr><td>11</td><td>DYS450</td><td>8</td><td>8</td><td>gnl|ti|1737745507 1098448293867
gnl|ti|1744526565 1094853896084
gnl|ti|1748661445 1094794256244</td></tr>
<tr><td>12</td><td>DYS454</td><td>11</td><td>11</td><td>gnl|ti|1735498386 1095526452861
gnl|ti|1736140170 1096761752422
gnl|ti|1741533528 1094373163843
gnl|ti|1742561346 1094393585481
gnl|ti|1743978858 1094833443046</td></tr>
<tr><td>13</td><td>DYS455</td><td>11</td><td>11</td><td>gnl|ti|1743501959 1094791261520</td></tr>
<tr><td>14</td><td>DYS458</td><td>17</td><td>17</td><td>gnl|ti|1736337474 1098397393123
gnl|ti|1736469939 1098486213789
gnl|ti|1745708415 1094789108897</td></tr>
<tr><td>15</td><td>DYS459a</td><td>9</td><td>9</td><td>gnl|ti|1732454582 1099728652167</td></tr>
<tr><td>16</td><td>DYS461</td><td>12</td><td></td><td>gnl|ti|1737607637 1098448234940</td></tr>
<tr><td>17</td><td>DYS462</td><td>11</td><td></td><td>gnl|ti|1741392194 1094392137431</td></tr>
<tr><td>18</td><td>DYS472</td><td>8</td><td>8</td><td>gnl|ti|1736155657 1096705993089
gnl|ti|1738690968 1099519064932
gnl|ti|1742852160 1094851178370
gnl|ti|1743155673 1094787462520</td></tr>
<tr><td>19</td><td>DYS481</td><td>22</td><td>22</td><td>gnl|ti|1734102710 1099289451847
gnl|ti|1736259266 1098329736933
gnl|ti|1746552196 1094837095559</td></tr>
<tr><td>20</td><td>DYS485</td><td>16</td><td></td><td>gnl|ti|1733291770 1099435048368
gnl|ti|1733792155 1099476054942</td></tr>
<tr><td>21</td><td>DYS492</td><td>13</td><td>13</td><td>gnl|ti|1739143690 1099524067934</td></tr>
<tr><td>22</td><td>DYS494</td><td>9</td><td></td><td>gnl|ti|1746195164 1094829593950
gnl|ti|1748442641 1094793322431</td></tr>
<tr><td>23</td><td>DYS531</td><td>12</td><td>12</td><td>gnl|ti|1734977449 1099549734703
gnl|ti|1736361586 1098415330285
gnl|ti|1744191663 1094846620984</td></tr>
<tr><td>24</td><td>DYS534</td><td>16</td><td>16</td><td>gnl|ti|1745370916 1094780147297</td></tr>
<tr><td>25</td><td>DYS537</td><td>10</td><td>10</td><td>gnl|ti|1735464928 1095527376552</td></tr>
<tr><td>26</td><td>DYS549</td><td>12</td><td></td><td>gnl|ti|1746731253 1094838014991</td></tr>
<tr><td>27</td><td>DYS556</td><td>11</td><td></td><td>gnl|ti|1742128430 1094374558361</td></tr>
<tr><td>28</td><td>DYS557</td><td>16</td><td>16</td><td>gnl|ti|1743535107 1094782120366</td></tr>
<tr><td>29</td><td>DYS565</td><td>12</td><td>12</td><td>gnl|ti|1742310596 1094392588605</td></tr>
<tr><td>30</td><td>DYS568</td><td>11</td><td>11</td><td>gnl|ti|1746854566 1094840623589</td></tr>
</table>

				gnl|ti|1748479999 1094793473108
31	DYS570	17	17	gnl|ti|1742167467 1094374639437 gnl|ti|1747824503 1094853609576
32	DYS578	9	9	gnl|ti|1735984841 1096705446919 gnl|ti|1744305622 1094846063102
33	DYS590	9	9	gnl|ti|1733883306 1099288170508 gnl|ti|1737180065 1098315100193 gnl|ti|1747537076 1094851078820
34	DYS594	10	10	gnl|ti|1742932348 1094912234510 gnl|ti|1746395272 1094833246840 gnl|ti|1746951187 1094846001467
35	DYS617	12	12	gnl|ti|1732482858 1099728740338 gnl|ti|1748787351 1094794051101
36	DYS636	12		gnl|ti|1743887805 1094837801595
37	DYS638	11		gnl|ti|1746043466 1094824088133
38	DYS641	10	10	gnl|ti|1738755005 1099519888704
39	DYS714	25	25	gnl|ti|1743246213 1094916447922
40	YCAIIa	19	19	gnl|ti|1739074839 1099519832477 gnl|ti|1741752326 1094373621762 gnl|ti|1742253410 1093043029666
	YCAIIb	23	23	gnl|ti|1741696014 1094373503777 gnl|ti|1742726858 1094374161659 gnl|ti|1744639237 1094851389510

Craig Venter's haplotype from his personal genome versus the best Ysearch match. Only Ysearch markers with corresponding sequencing results are shown. All alleles are reported using FamilyTreeDNA nomenclature to match the Ysearch convention.

References

1. B. Walsh, *Genetics* **158**, 897 (Jun, 2001).
2. R. Thomson, J. K. Pritchard, P. Shen, P. J. Oefner, M. W. Feldman, *Proc Natl Acad Sci U S A* **97**, 7360 (Jun 20, 2000).
3. K. N. Ballantyne *et al.*, *American journal of human genetics* **87**, 341 (Sep 10, 2010).
4. K. N. Ballantyne *et al.*, *Forensic Sci Int Genet* **6**, 208 (Mar, 2012).
5. S. J. Grannis, J. M. Overhage, C. McDonald, *Medinfo 2004: Proceedings of the 11th World Congress on Medical Informatics, Pt 1 and 2* **107**, 43 (2004).
6. W. E. Winkler, *Wiley S Pro*, 355 (1995).
7. W. E. Winkler, *Proceedings of the Section on Survey Research Methods* 354 (1990).
8. M. Gymrek, D. Golan, S. Rosset, Y. Erlich, *Genome Research*, (Apr 20, 2012).
9. M. Kayser *et al.*, *American journal of human genetics* **74**, 1183 (Jun, 2004).
10. E. Bosch *et al.*, *Forensic Sci Int* **125**, 42 (Jan 24, 2002).
11. J. M. Butler, A. E. Decker, P. M. Vallone, M. C. Kline, *Forensic Sci Int* **156**, 250 (Jan 27, 2006).
12. E. K. Hanson, P. N. Berdos, J. Ballantyne, *J Forensic Sci* **51**, 1298 (Nov, 2006).
13. S. K. Lim, Y. Xue, E. J. Parkin, C. Tyler-Smith, *Int J Legal Med* **121**, 124 (Mar, 2007).
14. N. Leat, L. Ehrenreich, M. Benjeddou, K. Cloete, S. Davison, *Forensic Sci Int* **168**, 154 (May 24, 2007).
15. P. Malaspina *et al.*, *J Mol Evol* **44**, 652 (Jun, 1997).
16. R. Schoske, P. M. Vallone, M. C. Kline, J. W. Redman, J. M. Butler, *Forensic Sci Int* **139**, 107 (Jan 28, 2004).
17. M. A. Jobling *et al.*, *Hum Mol Genet* **5**, 1767 (Nov, 1996).
18. P. Balaresque *et al.*, *Int J Legal Med* **123**, 15 (Jan, 2009).
19. L. Gusmao *et al.*, *Int J Legal Med* **120**, 191 (Jul, 2006).
20. H. Li *et al.*, *Bioinformatics* **25**, 2078 (Aug 15, 2009).
21. H. Li, R. Durbin, *Bioinformatics* **25**, 1754 (Jul 15, 2009).
22. W. Athey, *Journal of Genetic Genealogy* **3**, (2007).
23. Y. Erlich *et al.*, *Genome Research* **21**, 658 (May, 2011).
24. M. He *et al.*, *PLoS One* **4**, e4684 (2009).
25. D. R. Nyholt, C. E. Yu, P. M. Visscher, *European journal of human genetics : EJHG* **17**, 147 (Feb, 2009).
26. Y. Xue *et al.*, *Curr Biol* **19**, 1453 (Sep 15, 2009).
27. E. Halperin, D. A. Stephan, *Nat Biotechnol* **27**, 349 (Apr, 2009).
28. J. Marchini, B. Howie, *Nature reviews. Genetics* **11**, 499 (Jul, 2010).
29. T. Egeland, A. Salas, *PLoS One* **3**, e3988 (2008).
30. K. B. Jacobs *et al.*, *Nat Genet* **41**, 1253 (Nov, 2009).
31. United States. General Accounting Office., United States. (U.S. General Accounting Office, Washington, D.C., 2002).