Ysurnames? The patrilineal Y-chromosome and surname correlation for DNA kinship research

Sofie Claerhout, [1,*] Jennifer Roelens, [2] Michiel Van der Haegen, [1] Paulien Verstraete, [1] Maarten H.D. Larmuseau, [3,4,5] Ronny Decorte, [1,6]

[1] Forensic Biomedical Sciences, Department of Imaging & Pathology, KU Leuven, Leuven 3000, Belgium
[2] Department of Earth and Environmental Sciences, KU Leuven, Leuven 3000, Belgium
[3] Laboratory of Socioecology and Social Evolution, Department of Biology, KU Leuven, Leuven 3000, Belgium
[4] Histories vzw, Mechelen 2800, Belgium
[5] Department of Human Genetics, KU Leuven, Leuven 3000, Belgium
[6] Laboratory of Forensic genetics and Molecular Archaeology, UZ Leuven, Leuven 3000, Belgium

*Corresponding author at the laboratory of Forensic Genetics and Molecular Archaeology, Kapucijnenvoer 33, Leuven, Belgium. E-mail: sofie.claerhout@kuleuven.be

Abstract

The Y-chromosome is a widely studied and useful small part of the genome providing different applications for interdisciplinary research. In many (Western) societies, the Y-chromosome and surnames are paternally co-inherited, suggesting a corresponding Y-haplotype for every namesake. While it has already been observed that this correlation may be disrupted by a false-paternity event, adoption, anonymous sperm donor or the co-founding of surnames, extensive information on the strength of the surname match frequency (SMF) with the Y-chromosome remains rather unknown. For the first time in Belgium and the Netherlands, we were able to study this correlation using 2,401 males genotyped for 46 Y-STRs and 183 Y-SNPs. The SMF was observed to be dependent on the number of Y-STRs analyzed, their mutation rates and the number of Y-STR differences allowed for a kinship. For a perfect match, the Yfiler® Plus and our in-house YForGen kit gave a similar high SMF of 98%, but for non-perfect matches, the latter could overall be identified as the best kit. The SMF generally increased due to less mismatches when encountering (1) deep Y-subhaplogroups, (2) less frequently occurring surnames, and (3) small geographical distances between relatives. This novel information enabled the design of a surname prediction model based on genetic and geographical distances of a kinship. The prediction model has an area under the curve (AUC) of 0.9 and is therefore useable for DNA kinship priority listing in estimation applications like forensic familial searching.

Keywords: Surname; Y-chromosome; Y-STR; Y-SNP; Familial searching

1. Introduction

A last name or surname is the portion of a personal name indicating the family, tribe or community this person belongs to. It can be used as a powerful tool to determine judicial co-ancestry between individuals. In most modern societies, the transmission of surnames is linked with the biological inheritance of the male specific Y-chromosome, from father to son (1). As 95% of the Y-chromosome is defined as a non-recombining region (NRY) with the X-chromosome, it can contribute to the determination of paternal lineages since it is well-preserved through many generations. This makes the Y-chromosome an interesting and widely used marker providing numerous applications for interdisciplinary genetic research, such as population genetics and genetic genealogy (2).

The Y-chromosome has also been of interest for forensic genetics as it provides evidence for a biological paternal kinship and singles out male DNA components in mixtures with high female DNA background. This gives rise to different forensic genetic applications such as paternity tests, identification of suspects and familial searching (3,4). The latter is a perpetrator identification method where investigators purposefully search for a relative with a close match in the national DNA database or via large-scale voluntary DNA mass screening using autosomal or Y-chromosomal DNA analysis (5). As autosomal DNA kinships fade away over generations due to recombination events, Y-chromosome analysis can provide the opportunity to identify patrilineages, and can thus be used to find distant or close biological relatives who share a most recent common ancestor (MRCA) with the perpetrator. A biological kinship can be verified and discriminated using DNA polymorphisms on the Y-chromosome such as single nucleotide polymorphisms (Y-SNPs) and short tandem repeats (Y-STRs). Y-SNPs are bi-allelic markers with a slow mutation rate of approximately 2×10^{-8} mutations per generation (mpg) (6,7) and can therefore be used to determine a person's bio-geographical origin as it divides the entire male human population into 20 evolutionary Y-haplogroups (8). The main Y-haplogroups had their origin long before the practice of surnames in human populations was introduced (9,10). Therefore, when two namesakes do not share their Y-haplogroup, they do not share recent paternal ancestry and are thus not related on a familial level. On the other hand, Y-STRs are fast-mutating DNA polymorphisms with their mutation rate varying between 10^{-4} and 10^{-2} mpg (11–14). A set of Y-STR alleles is referred to as a Y-haplotype and can be used to attribute a person to a familial lineage (11). Provided that their Y-SNP haplogroup also matches, a close Y-haplotype profile between two males could indicate a biological kinship sharing a common paternal ancestor, but this strongly depends on the number of Y-STR markers analyzed in order to decrease false positive kinships (15).

An interesting research question for genetic genealogy involves the strength of the correlation between surnames and the Y-chromosome, which could eventually provide a label of regional and familial relationships. This could be useful for familial searching to identify the surname of the perpetrator from his DNA left at the crime scene (16). By implementing surnames, a priority list of people bearing the same surname in close match with the perpetrator can be extracted, as successfully used in the extensively discussed murder case of Marianne Vaatstra (2). In this case, two close genetic matches with the perpetrator were found in the first cohort of men. After genealogical research, these two were found to be related, but did not share their surname as their common ancestor was traced back before the time surnames were registered. Police investigators further focused on both surnames of these relatives and could therefore eventually identify the true offender.

Research on the strength of this correlation has already received attention in several studies. The link may be disrupted in cases of an adoption, a surname change, a maternal surname transmission or an independently co-founded surname (1). Although these discrepancies can be identified by analyzing the genealogy with archival information, they may remain undetected in case of a hidden adoption, an unprecedented baby exchange, an anonymous sperm donor or an extra-pair paternity (EPP) event (17). However, several population studies already showed rather low EPP frequencies per generation (1-2%) (18–24). Within the Belgian and Dutch population, an EPP rate of 1.44% per generation (95% CI: 1.36-1.51%) could be observed (25–27). Spatiotemporal differentiation analysis identified a lower historical EPP compared to the current rate and a significantly higher EPP rate within genealogical pairs living in rural areas compared to the city (26). Sociological or demographic factors have not yet been studied. The Y-chromosome-surname correlation could also be influenced by surname frequency and/or the specific geographical origin of the surnames. King *et al.* (2006) investigated unrelated males accounting for 150 different British surnames. They observed a correct surname prediction in only 19% of the cases; however for less common surnames, prediction sensitivity increased to 34%. This indicates an inverse correlation between surname frequency and genetic Y-chromosome co-ancestry, pointing to polyphyletism (multiple founders) as the major driver (28). These results were confirmed by a more recent study comparing 37 Spanish surnames (29). In contrast, an Irish study obtained no significant correlation through the analysis of 43 surnames, because probably there was uncertainty about the Irish surname transmission due to the conversion to the English language (30). In 2015, Solé-Morata *et al.* confirmed the positive correlation between Y-diversity and surname frequency through the analysis of 50 Catalan surnames. Furthermore, they predicted surnames by attributing Y-haplotypes to self-defined major descent clusters (MDC). The sensitivity of this prediction was estimated 40.8% and assessed a false discovery rate (FDR)

of 17% (23). Unfortunately, these were just estimations and, additionally, defining major descent clusters can be a time-consuming and highly subjective job. Furthermore, their predictions were only based on 17 Y-STRs, which is not enough to make a distinction between unrelated males within common Y-SNP subhaplogroups as for example 'R1b-M269' (15,31). Through the analysis of more Y-STRs, a higher resolution of discrimination could be obtained, which is preferable for surname prediction in forensic investigations.

Lastly, the surname and Y-chromosome correlation could also be influenced by the age of hereditary surname practices within the area being examined. Surname tradition has taken hold separately in different cultures around the world (23,32). In Europe, surnames aroused during the Roman Empire, but died out as a result of Germanic and Persian influences and re-emerged again during the Middle Ages (33). Paternally inherited surnames became a more common tradition for Belgium in the 14th to 15th century, while in the Netherlands this was introduced in the 16th to 17th century. However, it was not until 1795 and 1811 that respectively Belgium and the Netherlands were obligated to register their surnames at Civil Services (34). This could be important in the finding of a surname match as the number of meioses permitted between a genealogical pair has to lie within a genealogical timescale since hereditary surnames arose, thus having an MRCA born after the age of surname introduction.

As the combination of genetics with genealogy opens up new investigative leads within, for instance, population genetics and forensic familial searching, it is in our interest to gain more knowledge concerning the correlation between the Y-chromosome and surnames and calculate prediction accuracies that can be reached without searching for MDCs. In this study, we examine for the first time the Y-chromosome and surname match frequency (SMF) relation for more than 1,100 surname clusters within a Belgian and Dutch population (the Low Countries in Europe) where no close kinships of less than seven generations are present. Furthermore, we indicate the importance of diverse influencing factors on the SMF: different Y-haplotypes, Y-SNP subhaplogroup typing, age of hereditary surname practices, surname frequency, surname anthroponymy and the geographical distance between relatives. Finally, a surname prediction model including a spatial factor next to the genetic aspect was developed.

2. Materials and methods

2.1. Database

The present study includes previously obtained Y-chromosomal and genealogical data from 2,401 males, whereof the majority are randomly registered men collected through a collaborative research project with the non-profit organization 'Histories vzw'. A small part of our database (28%) are judicially related males collected to study the extra-pair paternity (EPP) rates, haplogroup specific Y-STR mutation rates and parallel Y-STR evolution (14,25,35). These judicially relatives do not necessarily share a common surname, as relatives with a surname change within their paternal lineage were also included. No close kinships of less than seven meioses (or generation steps) are present in the database, which is a stronger threshold compared to the limit of being first-degree cousins within the study of Solé-Morata *et al.* (23). Previously confirmed biological relatives were not excluded from our analysis since this would create a bias towards extra-pair paternities, which would reduce the representativeness of the database in true DNA kinship research. Additionally, genetic kinships are needed for the investigation of our different research parameters (e.g. the influence of genetic and geographical distance as further explained in section 2.3). Through written consents, the permission for DNA analysis and scientific publication of the anonymized results were received. The Ethical Commission of University Hospital Leuven accepted and approved these studies (S54010, S55864, S59085). DNA samples were collected, extracted and genotyped (Y-haplotype and Y-subhaplogroup), as described in Claerhout *et al.* (14). Detailed Y-STR information concerning Y-chromosome position, primers, length, repeat motif and genomic reference sequences are available in Claerhout *et al*. (35). Y-chromosomal data used in this study has been submitted to the open access Y-STR Haplotype Reference Database (YHRD, https://yhrd.org), available under accession numbers YA003651, YA003652, YA003653, YA003739, YA003740, YA003741, YA003742, YA004300 and YA004301. Each included subject has complete information on his surname and municipality of residence, and has his residence located in the Low Countries (Belgium n=1,859; the Netherlands n=544). The spatial distribution of the samples based on their residence is illustrated in **Figure 1**. All surnames with similar denotation or pronunciation were composed into 1,128 surname clusters as spelling (capital letters, truncations and accents) could be altered from generation to generation (36). For example, the surname cluster 'Dhondt' has 12 variants in our database: D'Hondt, D'hondt, D'Hont, D'hont, D'Hond, D'hond, Dhondt, Dhont, dhondt, DenHond, DeHondt and DeHond.

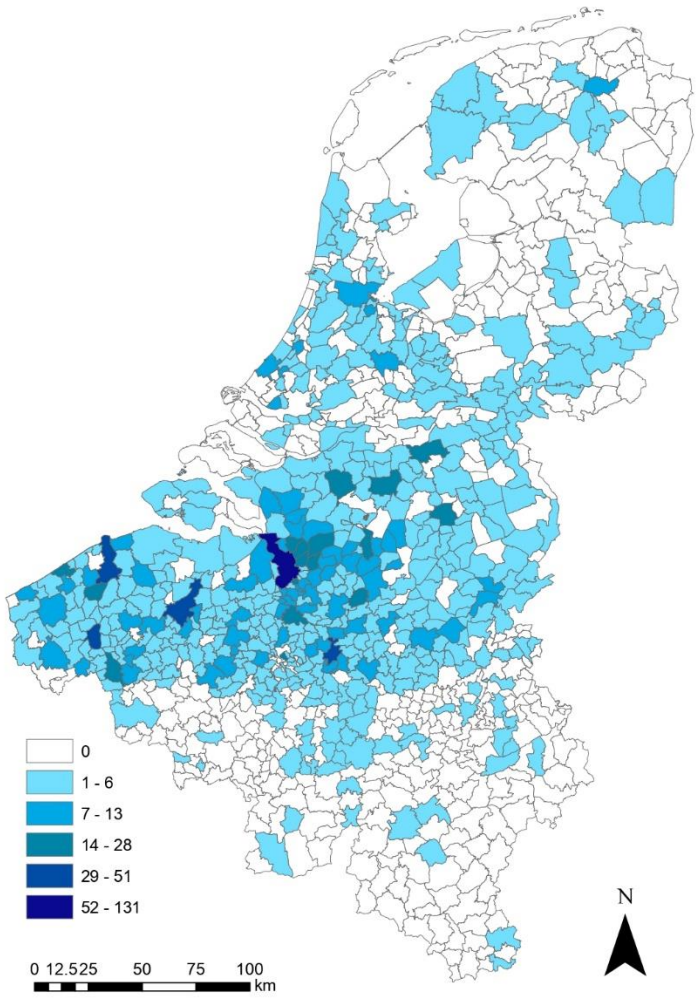

***Figure 1**: The spatial distribution of all 2,401 samples based on their residence in the Low Countries. Belgium n=1,859; the Netherlands n=544.*

2.2. Y-STR haplotypes

All samples were genotyped for one or more Y-haplotype sets (**Table 1**). Three Y-haplotype sets are commercially available PCR amplification kits frequently used for forensic Y-STR genotyping, namely the Yfiler® kit (Thermo Fisher Scientific), the PowerPlex® Y23 System (Promega Corporation) and the Yfiler® Plus kit (Thermo Fisher Scientific), containing respectively 17, 23 and 27 Y-STRs. The fourth Y-haplotype set is our in-house kit, further referred to as YForGen, containing 46 Y-STRs. Additionally, we compared two altered Y-haplotype kits, named YForGen[-] containing all in-house Y-STRs except the rapidly mutating (RM) Y-STRs with a mutation rate of at least 10^{-2} mpg, and the YCommon kit with 38 Y-STRs including all common typed Y-STRs for every individual in the database (13,14). A detailed list

of the genotyped Y-STRs within the six different Y-haplotypes can be found in **Table S1**. An overview of the Y-haplotypes and their characteristics is represented in **Table 1**. All samples were at least genotyped for the YCommon, which includes among others the well-known Y-STRs from the smallest Yfiler® kit. 67% of the samples were genotyped for the Y-STRs of the PowerPlex® Y23 kit, and 44% of the samples were fully typed for our in-house YForGen kit. The latter contains all Y-STR loci included in the Yfiler® Plus kit. The average mutation rates for all Y-haplotypes were derived from the individual Y-STR mutation rates identified in Claerhout *et al.* (14). The maximum number of Y-STR differences allowed between genealogical pairs separated by 20, 30 and 40 meioses were obtained through the online tMRCA calculator of J.D. McDonald, which is based on the infinite allele model (IAM) of Walsh (www.scs.illinois.edu/~mcdonald/tmrca.htm) (37). These values are based on the most probable tMRCA estimations for a range of Y-STR differences per Y-haplotype kit and will also give an indication of the most probable distance between two males for a number of differences. The resulting probability curves for the number of meioses between a genealogical pair are visualized in **Figure S1** according to the observed Y-STR differences per Y-haplotype.

Table 1: *The six investigated Y-haplotype kits with their characteristics and maximum number of Y-STR differences for genealogical pairs separated by 20,30 and 40 meioses. RM: rapidly mutating.*

Y-STR haplotypes	No. Y-STRs	% RM Y-STRs	Average mutation rate ($\times 10^{-3}$)	No. samples	Max. no. Y-STR differences		
					20 meioses	30 meioses	40 meioses
Yfiler®	17	6	3.54	2,401	1	2	3
PowerPlex® Y23	23	13	4.15	1,616	2	3	4
Yfiler® Plus	27	30	5.81	1,063	3	5	7
YForGen	46	22	4.94	1,063	5	7	10
YForGen[-]	37	0	3.15	1,616	2	4	5
YCommon	38	16	4.21	2,401	3	5	7

2.3. Data analysis

To obtain information about the subject-to-subject relationships in the database, a distance (or similarity) matrix was created to compare each unique pair of subjects in the database. With 2,401 subjects, a total number of 2,881,200 unique subject-to-subject relationships had to be analyzed. The similarities or distances calculated between the subjects were the *surname cluster match*, *haplogroup match*, *subhaplogroup match*, *number of Y-STR differences* and *geographical distance*. The first three similarities were defined as dummy variables in which a match is set to '1' and a non-match is set to '0'. The *number of Y-STR*

differences was calculated as the number of differences that occurred between the two subjects for the specific Y-haplotype kit. The *geographical distance* was calculated as the Euclidian distance between the X,Y-coordinates of the centers of the municipalities of residence. For each analysis, the surname match frequency (SMF) was calculated as the proportion of *surname cluster matches*.

2.3.1. Y-haplotypes and Y-SNP genotyping

To investigate the effectiveness of each Y-haplotype kit, SMFs were calculated for different numbers of occurring Y-STR differences. The influence of Y-SNP haplogroup and subhaplogroup typing was also investigated for each Y-haplotype kit. Kinship validation for a *haplogroup match* was limited to the annotation of the main haplogroups (A to T). For a *subhaplogroup match*, the two main Y-haplogroups of West-Europe 'I' and 'R' were further subdivided into I1, I2, R1a, R1b-A and R1b-B (see **Figure S2** for a detailed 'R1b' Y-SNP subdivision phylogenetic tree). Furthermore, meeting the conditions of having a subhaplogroup match, the SMF was also calculated for the number of Y-STR differences allowed in a biological kinship separated by 20, 30 and 40 meioses (see **Table 1** for the allowed number of Y-STR differences). Additionally, in order to examine the influence of the age of hereditary surname practices, the SMF for the Belgian subset was compared to the Dutch subset SMF using the 20-30-40 meioses cut-offs, meeting the conditions of having a Y-SNP subhaplogroup match.

2.3.2. Surname anthroponymy and frequencies

For each surname cluster, the anthroponymy was defined. Seven surname origin groups were identified: ancestry, geography, characteristic, profession, mythology, combi and undefined. A list of the origin categories with the corresponding number of samples and additional explanation, including examples, are given in **Table S2**. To examine the influence of the surname anthroponymy, SMFs were calculated for each surname origin cluster separately. Each subject of the surname origin was then compared to all the other subjects in the database. To investigate the influence of the surname frequency, the frequency of each surname cluster was determined using the data available in the Belgian population register (2008, www.familienaam.be) and the Dutch center for family history 'Nederlandse FamilienamenBank' (2007, www.cbgfamilienamen.nl/nfb). Surname spelling variants were aggregated when frequencies were obtained. SMFs were calculated for each surname cluster including more than one subject. Each subject of the surname cluster was then compared to all the other subjects in the database. This analysis was done with our in-house YForGen kit. To have genetic matches sufficiently strong enough to evaluate the sensitivity of the surname

based on their frequency, the minimal threshold for each surname cluster was defined as at least three genetic matches. A genetic match was here defined as having less or equal to seven Y-STR differences between the subjects, which is the limit for 30 meioses (**Table 1**). Additionally, surname matches and mismatches were visualized for three surname clusters with a high, intermediate and low number of occurrence in the Low Countries including respectively 49, 31 and 25 samples from our database.

2.3.3. Geographical distance

The geographical distance was calculated as the Euclidean distance between the coordinates of the centers of the residence municipalities of two subjects. In order to observe the importance of the geographical distance on the number of genetic matches and the SMF, the geographical distances were grouped in intervals of 5 km up to a total distance of 200 km. This analysis was performed for all Y-haplotypes. The maximum number of Y-STR differences was set on 40 meioses and varied thus per Y-haplotype as previously mentioned (**Table 1**).

2.3.4. Surname prediction model

A surname prediction model was developed for our in-house YForGen kit. A logistic regression model was used to calculate the probability of a positive surname match. The dataset was randomly split into 80% training and 20% validation. The set-up of the model was done using the scikit-learn library for Python (38). The predictor variables used for the logistic regression model were based on the number of Y-STR differences, the individual Y-STR mutation rates and the geographical distance between the two subjects. Based on preliminary results, it was decided to set-up the model for kinships having a Y-SNP subhaplogroup match. Surname match probabilities were calculated as:

$$Surname\ match\ probability = \frac{e^{f(predictor\ variables)}}{1 - e^{f(predictor\ variables)}} \qquad \text{(Formula 1)}$$

A general model evaluation was done by calculating the area under a receiver operating characteristic (ROC) curve (AUC) (39). A ROC curve is a graphical way to illustrate the ability of the model to classify the kinships by varying the cut-off values of the surname match probabilities. This is done by plotting the true positive rate (TPR, sensitivity) against the false positive rate (FPR) for different cut-offs. Only kinships with probabilities above the cut-off value are then classified as a surname match. For the different cut-offs, the TPR, FPR and false detection rate (FDR) were also calculated separately to evaluate the model.

3. Results

3.1. Y-haplotypes and Y-SNP genotyping

Surname match frequencies (SMF) were compared for the three commercially available Y-STR kits, namely Yfiler® (17 Y-STRs), PowerPlex® Y23 System (23 Y-STRs) and Yfiler® Plus (27 Y-STRs), together with our in-house YForGen kit (46 Y-STRs). **Figure 2A** visualizes the SMF per number of Y-STR differences (0-7) for the four previously mentioned Y-STR kits. The surname sensitivity slope varies for the different Y-haplotypes, but generally decreases when the number of Y-STR differences allowed increases. For a perfect match (no Y-STR differences on the Y-haplotype), the Yfiler® Plus and YForGen have a similar SMF of 98%. On the allowance of three Y-STR differences, the SMF for the YForGen kit (97%) remained significantly higher. The SMF drops more rapidly for the Yfiler® than for the PowerPlex® Y23, Yfiler® Plus and the YForGen, piercing the x-axis respectively at 3, 5, 7 and 12 Y-STR differences. Therefore, the YForGen Y-haplotype kit clearly outperforms the three commercially available kits.

Figure 2B compares the SMF per number of Y-STR differences (0-14) for our in-house YForGen kit containing ten RM Y-STR loci, with two altered kits, namely the YForGen[-] (37 Y-STRs) which excludes all RM Y-STRs from our in-house kit, and the YCommon (38 Y-STRs) which includes all common genotyped Y-STRs for every individual in the database. Despite the almost equal number of markers within these two kits, the SMF for the YForGen[-] decreased more rapidly. Additionally, the influence on SMF when excluding the Y-SNP subhaplogroup match as a kinship restriction is visualized by a dotted line for each Y-STR haplotype in **Figure 2A,B**. The SMF curve including the main haplogroup (A to T) as a kinship restriction follows the exact same trend of the curve without any haplogroup restriction, while on the other hand the determination of a deep Y-SNP subhaplogroup clearly has a positive impact on the SMF.

The total number of same-surname pairs in the dataset (surname matches) together with the false negative surname ratio per Y-haplotype are shown in **Figure S3**. A significant lower number of false negative surname matches, and thus higher number of correct surname matches using the Y-haplotype, was observed for every kit when Y-SNP subhaplogroup was included as a boundary condition. For example, the percentage false negative surname matches for our in-house YForGen kit was 70% without Y-SNP analysis and only 29% with subhaplogroup restriction. In other words, this indicates that the surname matches found in the database increased significantly from 30 to 71% when encountering Y-SNP subhaplogroups. When comparing Y-haplotypes, the YForGen and Yfiler® Plus both had a

significant lower number of false negative mistakes compared to the PowerPlex® Y23 ($p = 0.04$), YForGen[-] ($p = 0.04$) and YCommon ($p = 0.03$). As shown in **Figure S3** (bar plots), there must be emphasized again that the number of surname matches per kit diverge as the number of typed samples for each Y-haplotype kit was not identical (**Table 1**). However, noticeable for every Y-haplotype kit is that the number of surname matches reduces with 50% when subhaplogroups are taken into account. Additionally, the normalized difference between the number of surname matches from YForGen and Yfiler® Plus compared to the others is only significant without encountering subhaplogroups.

In **Figure 2C**, the SMF is visualized for the six Y-haplotypes according to the permitted number of Y-STR differences for genealogical pairs separated by 20, 30 and 40 meioses (or generation steps). The SMF for the YForGen kit was significantly higher compared to all the other Y-haplotypes for kinships separated by 30 and 40 meioses. For kinships comprising of 20 meioses, the SMF of the YForGen kit remained the highest, but the differences between the SMF of the Yfiler® Plus and the YForGen[-] kit were not of any significance. The Y-SNP genotyping influence on SMF is also visualized in **Figure 2C** for the six Y-haplotypes (striped). When the subhaplogroup boundary condition was considered, the SMF was generally observed to be higher. A significant difference was observed within our YForGen kit for kinships separated by 40 meioses. Finally, in **Figure 2D**, the SMF for all Y-haplotype kits with Y-SNP subhaplogroup restriction on the Belgian subset of the database was compared to the SMF of the Dutch subset. Overall, the Dutch SMF was lower compared to the Belgian SMF, but these differences were not observed to be significant. The highest SMF for the three time intervals (20-30-40 meioses) for both subsets were observed using our in-house YForGen kit: Belgium (94%-89%-60%) and the Netherlands (90%-89%-55%).

published as DOI:10.1016/j.fsigen.2019.102204

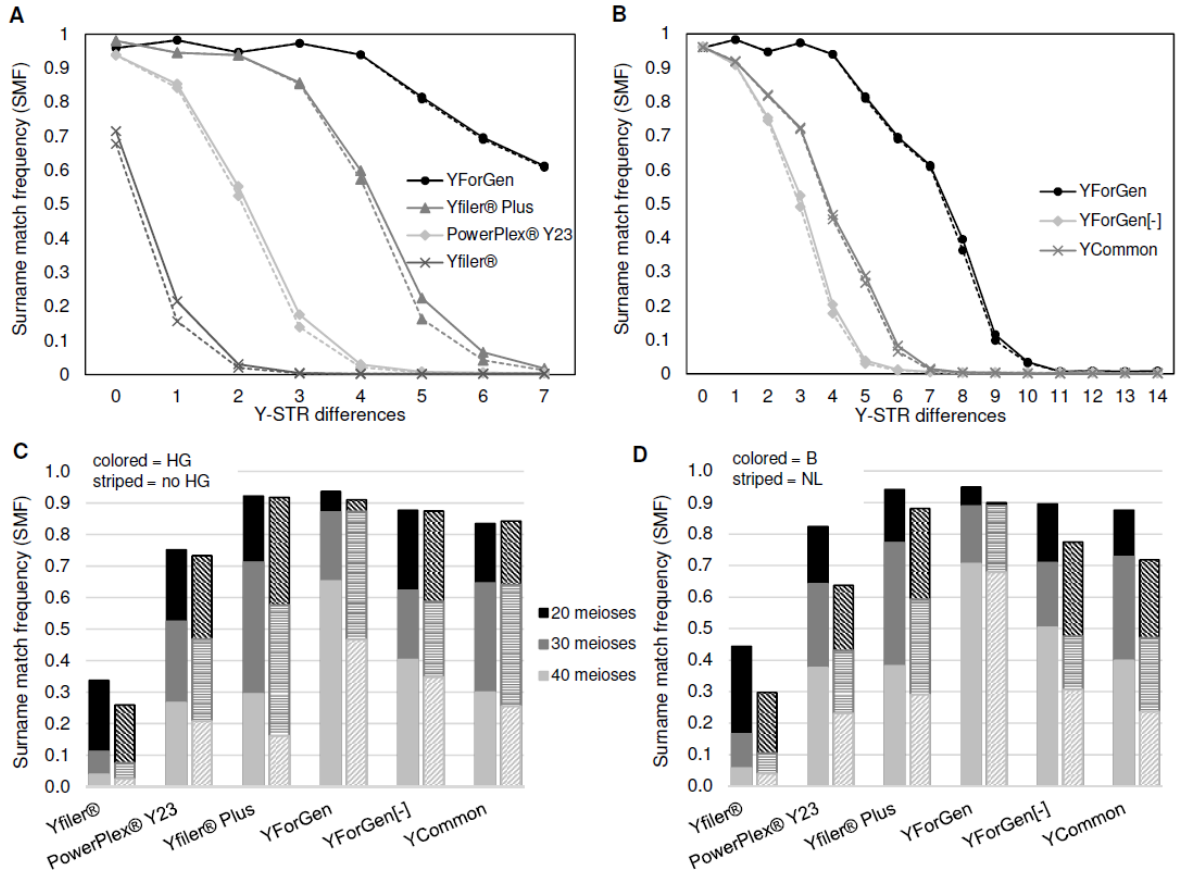

***Figure 2**: Surname matching frequency (SMF) for different Y-haplotypes with and without Y-SNP (sub)haplogroup kinship restriction. Solid line/colored: Y-SNP subhaplogroup restriction; striped: no restriction (A) SMF per number of Y-STR differences (0-7) for three commercially available Y-haplotypes, compared to our in-house kit YForGen. (B) SMF per number of Y-STR differences (0-14) for the YForGen kit compared to two altered in-house Y-haplotypes: YForGen[-] and YCommon. (C) Y-SNP genotyping influence on SMF according to the permitted number of Y-STR differences for a genealogical pair separated by 20, 30 and 40 meioses. (D) SMF for Belgium (B, colored) versus the Netherlands (NL, striped).*

To estimate the added value of RM Y-STRs, the genetic match ratio of the YForGen[-] kit (max. 5 Y-STR differences) was compared to our in-house YForGen kit (max. 10 Y-STR differences) containing ten RM Y-STRs (22%). Without subhaplogroup restriction, one genealogical pair belonging to two different subhaplogroups within 'R1b' was incorrectly accepted using the YForGen[-] kit due to only four Y-STR differences, but correctly excluded with our YForGen kit due to 12 Y-STR differences. In general, we observed that no kinships were falsely excluded within the YForGen kit despite the incorporation of RM Y-STRs. When the subhaplogroup restriction was encountered and the full database was examined with the YForGen kit, 65% of the genetic matches contained a surname match, while for the YForGen[-] kit this percentage significantly decreased to 57% ($p = 0.01$). After additionally analyzing the 65% YForGen surname matches with the YForGen[-], 19 genetic kinships (1.9%) were falsely excluded as they contained six Y-STR differences which exceeds the maximum number of

five for this Y-haplotype. The majority of the Y-STR differences (35%) between these 19 kinships were observed within the *DYS464* loci (4.10×10^{-3} mpg), followed by the second marker *DYS389I* (4.24×10^{-3} mpg) containing 5% of all the differences. Without the *DYS464* marker, only three genealogical pairs would have received a false negative kinship mistake.

3.2. Surname anthroponymy and frequencies

Surnames were divided into seven origin groups, listed and explained in **Table S2**. Three origin groups were excluded for further analysis due to the low number of samples (mythology) or the lack of a clear origin group (combi and undefined) which would give rise to aberrant ratios. The influence on surname match frequencies (SMF) for the other four origins (ancestry, characteristic, profession and geography) was studied by comparing all subjects within the origin group to the entire database. **Figure 3** visualizes the SMF per number of Y-STR differences (0-7) using the commercially available Yfiler® Plus kit. For completeness, the surname sensitivity for the five other Y-haplotype kits can be found in **Figure S4**. For three Y-STR differences, the ancestry group has a significant higher match ratio (0.87) in comparison to the other origin groups: characteristic ($p = 0.02\times10^{-4}$), profession ($p = 0.46\times10^{-3}$) and geography ($p = 0.04$). However, no clear SMF differences between the four surname origins can be observed. Additionally, the surname matches and mismatches using our YForGen kit for each origin group were plotted per number of Y-STR differences (0-10) against the geographical distance of residence (km) (**Figure S5**). The ancestry group demonstrates again a significant higher percentage of correct surname matches compared to the other origin groups: characteristic ($p = 0.68\times10^{-3}$), profession ($p = 0.70\times10^{-2}$) and geography ($p = 0.02$).

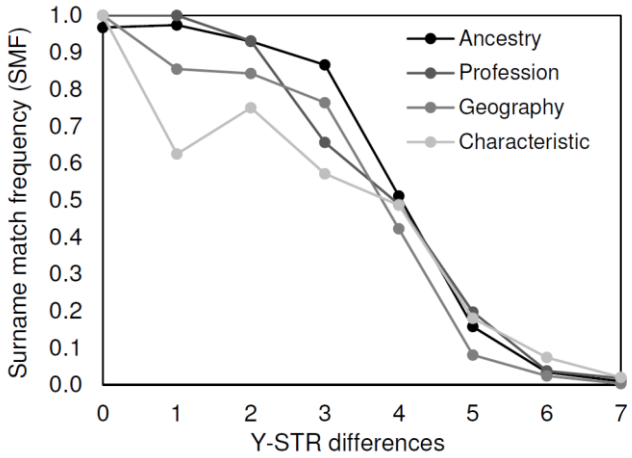

Figure 3: *Influence of surname anthroponymy on SMF. SMF per number of Y-STR differences (0-7) for the four main surname origin groups.*

To investigate the influence of surname frequency, a subset of three Dutch surname clusters was collected with a high, intermediate and low number of occurrence in the Low Countries, respectively covering 0.12%, 0.02% and 0.01% of the population and 49, 31 and 25 samples from our database. The surname matches and mismatches for the three surname clusters compared to the entire database were plotted per number of Y-STR difference (0-10) against the distance of residence (km) (**Figure 4A-C**). In general, the surname with the lowest frequency has a significant (Chi-square test, $p < .0001$) higher percentage of correct surname matches (77%) compared to the other two surnames with intermediate (49%) and high (41%) frequencies. Additionally, the surname containing the lowest frequency (**Figure 4C**) has no surname mismatches (x) up to and including seven Y-STR differences. For the intermediate (**Figure 4B**) and high (**Figure 4A**) frequency surname clusters, this could be observed up to respectively four and two Y-STR differences. The higher the surname frequency, the more surname mismatch pollution will be identified and the more important the distance will be. Further, the SMF was plotted against surname frequency for the full database with a Y-haplotype kinship threshold of seven Y-STR differences and a genetic match threshold of at least three matches (**Figure 4D**). A non-significant negative trend ($p = 0.06$) was observed between the SMF and the surname frequency with a low coefficient of determination (R^2) of 0.018. For this analysis, we lack equally spread data points, as the distribution of surname frequencies in our database (94.9% low, 3.5% intermediate and 1.5% high surname frequency) strings along with the surname division in the population.

 published as DOI:10.1016/j.fsigen.2019.102204

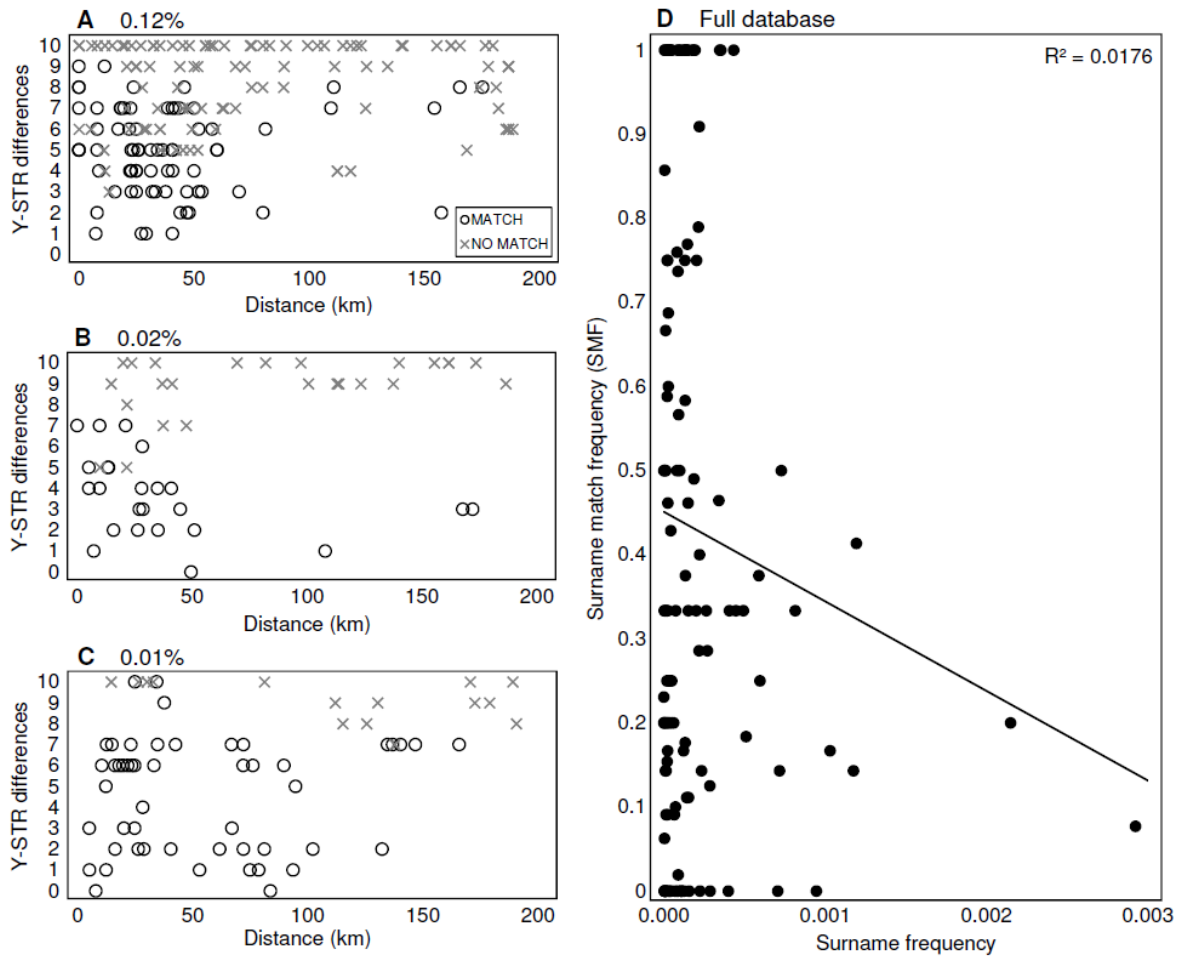

Figure 4: *Influence of surname frequencies on SMF studied with subhaplogroup restriction for the in-house YForGen kit. (A-C) Three surname clusters with a difference in the number of occurrences in the Low Countries, resp. 0.12%, 0.02% and 0.01% of the population. Y-STR differences from 0-10 plotted against distance of residence (km). circle: surname match; cross: no match. (D) SMF plotted against surname frequency for the entire database. Boundary conditions are seven Y-STR differences and at least three genetic matches (p = 0.06).*

3.3. Geographical distance

The surname match frequency (SMF) was visualized in function of the geographical distance of the biological relatives for the in-house YForGen kit (**Figure 5, scatter plot**). Y-STR differences up to 40 meioses were allowed, but a match in subhaplogroup was required. The SMF starts in the first 5 km bin with a value of 0.9. A high coefficient of determination (R^2) of 0.945 was obtained, meaning that a clear exponential decrease in SMF can be observed when the geographical distance between relatives increases. When only the number of genetic matches is considered, and thus ignoring the sharing of a surname (**Figure 5, bar plot**), it can also be observed that more genetic kinships occur within closer distances (R^2 = 0.8941). This indicates that males living in close proximity (up to 40 km) have more similarities in their Y-haplotype. The number of genetic matches decreases until a search radius of 60 km is reached. From then on, the number of genetic matches is more distributed.

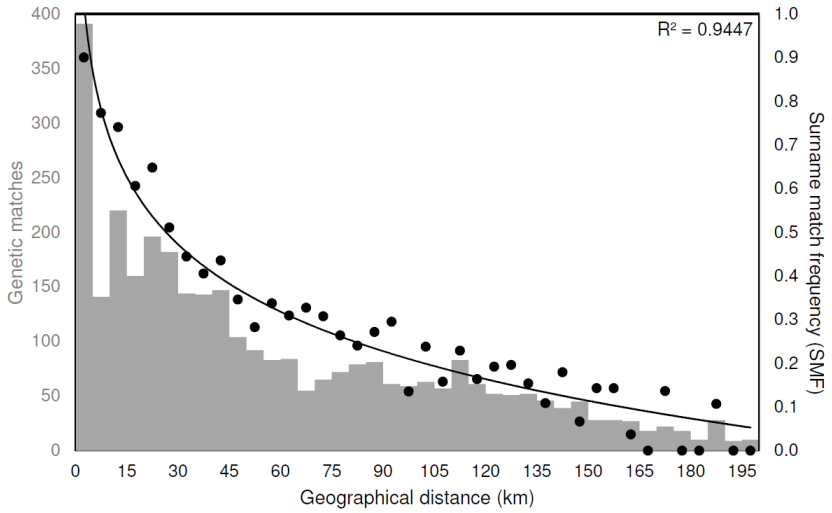

Figure 5: *Influence of geographical distance on SMF studied with subhaplogroup restriction for the in-house YForGen haplotype. Geographical distances were binned in intervals of 5 km with the number of genetic matches on the primary y-axis (bar plot) and the SMF on the secondary y-axis (scatter plot).*

3.4. Surname prediction model

The surname prediction model was developed using a logistic regression model. Based on previous results, it was decided to set-up the model for kinships having a subhaplogroup match. This resulted in a dataset of 106,409 unique kinships, in which only 1.34% of the kinships had a surname match. Four significant predictor variables were selected and resulted in a statistically significant logistic regression model ($p < .0001$). The coefficients, p-values and relative importance for the predictor variables are represented in **Table 2**. The negative logistic regression coefficients for the number of Y-STR differences and the geographical distances - thus the higher, the lower the probability on a surname match - are in line with the results in section 3.1. Also essential to the prediction model are the variables linked to the different individual Y-STR mutation rates, where the average mutation rate of the changed Y-STRs is calculated. Therefore, equal kinships are prioritized on the defined average mutation rate which takes a correction for rapidly mutating Y-STRs into account.

Table 2*: Coefficients, p-values and relative importance of the predictor variables of the logistic regression model for predicting surname matches.*

Variable	Coefficient	P-value	Relative importance
Number of Y-STR differences	-0.4489	<.0001	0.853
Geographical distance (km)	-0.0059	<.0001	0.146
Average mutation rate	0.0342	<.0001	1.4×10^{-5}
Range of mutation rates	0.0131	<.0001	1.1×10^{-5}
Intercept	3.0249		

The model was evaluated using the validation dataset containing 20% of the individuals randomly selected from the database. The ROC-curve of the model is represented in **Figure 6A.** The area under the ROC curve (AUC) is 0.891, indicating a very good model where the ranking of the surname match probabilities can be trusted. In **Figure 6B**, the TPR, FPR and FDR were calculated for each possible cut-off, in which only kinships with probabilities above the cut-off value are classified as having a surname match. It is clear that there is a trade-off between the sensitivity (or TPR) and the FDR of the model. Although the probabilities are the most important to rank kinships for a surname match, we suggest to only use kinships with probabilities larger than 0.3, keeping the FDR lower than 15% and the model sensitivity (TPR) at 70%.

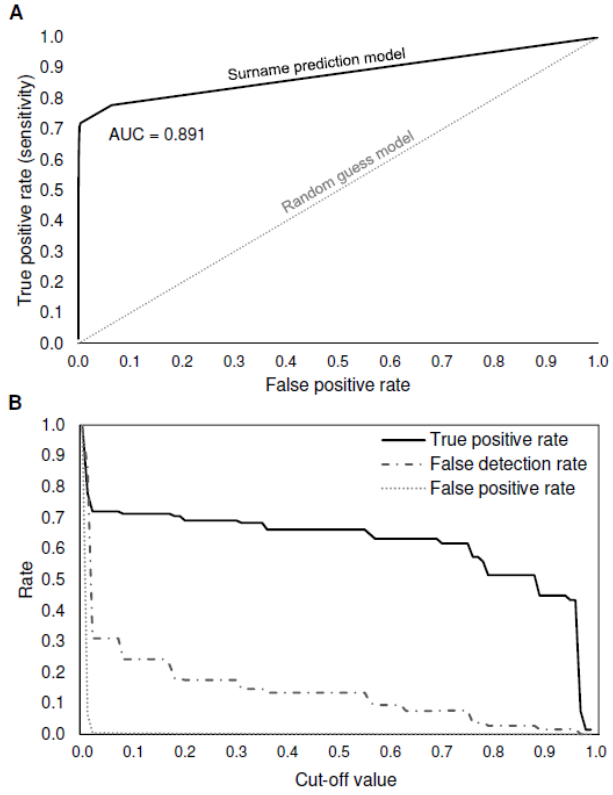

Figure 6: *Surname prediction model. (A) ROC-curve of the surname prediction model with an area under the curve (AUC) of 0.891, (B) True positive rate, false detection rate and false positive rate for the different possible surname match probability cut-offs.*

4. Discussion

In most populations, the Y-chromosome and surnames are co-inherited from father to son. This can serve as an interesting feature for interdisciplinary genetic kinship research, e.g. population genetics, genetic genealogy and forensic genetics. This correlation, however, may be disrupted in cases of an extra-pair paternity event, a maternal surname transmission, a (hidden) adoption, an anonymous sperm donor, a (hidden) baby exchange, a surname change, or an independently co-founded surname (23,25,40). In this study, we investigate for the first time the strength of the Y-chromosome-surname correlation for a population with residence in Belgium and the Netherlands (the Low Countries in Europe). By combining in-depth Y-chromosomal data with surnames, we were able to study its paternal inheritance correlation in more detail. With a database comprising 2,401 males within the Low Countries, the surname match frequency (SMF) was analyzed. Additionally, we examined the influence on the SMF of various factors, namely Y-haplotype, inclusion of Y-SNP subhaplogroup analysis, surname frequency, surname anthroponymy and the geographical distance between relatives.

4.1. Y-haplotype comparison

Surname match frequencies (SMF) were compared for six Y-haplotypes to investigate the influence of the number of markers and their individual mutation rates. Previous research in 2014 indicated a higher number of uninformative genetic matches when samples were analyzed with the smaller Yfiler® kit (17 Y-STRs) and compared to the PowerPlex® Y23 kit (23 Y-STRs) (41). In our study, this statement can be confirmed and expanded by including the larger Yfiler® Plus kit (27 Y-STRs) and our in-house YForGen kit (46 Y-STRs). As expected, when a Y-haplotype is more expanded, the chance of a false positive kinship match due to a coincidence of high Y-STR resemblance decreases (15,31), leading to a higher probability for a surname match. The SMF for every Y-haplotype is thus inversely correlated with the allowed Y-STR differences. For a perfect Y-haplotype match, the Yfiler® Plus and YForGen have a similar SMF of 98% (**Figure 2**). The 2% false positive surname matches could be explained by a possible interruption in their surname patrilineage due to a surname change or an extra-pair paternity (EPP) event of their common ancestor. This means that two males with a similar Y-haplotype can share a male ancestor, but not their surname. This confirms the previously obtained EPP rate for the Belgian and Dutch population of 1.44% per generation (95% CI: 1.36-1.51%) (26,27). The overall lower SMF for kinships separated by 40 meioses compared to 20 and 30 meioses in **Figure 2C** can therefore also be explained by the higher chance on interrupting events in the patrilineage.

Overall, the most extended Y-haplotype (our in-house YForGen kit including 46 Y-STR loci) clearly outperforms the SMF from the other kits. This suggests that a more expanded kit increases surname match frequencies as the chance on a random match causing a false positive surname mistake decreases. Although, despite the almost equal number of markers within our in-house kits YForGen[-] (37 Y-STRs) and YCommon (38 Y-STRs), the SMF for the YForGen[-] decreased more rapidly. This is probably a result of the difference in the number of rapidly mutating (RM) Y-STRs included. By the inclusion of more RM Y-STR markers (YCommon = 16%, YForGen[-] = 0%), false positive matches are decreased due to an increase of the differences especially with the RM Y-STR markers and therefore causing an overall higher SMF. The SMF is thus not only dependent on the number of markers included, but also on their individual mutation rates, which is also confirmed by the significance of the variables based on the mutation rates in the surname prediction model. RM Y-STRs are therefore not only interesting to increase the discrimination power between closely related individuals but also to exclude potential distant paternal relationships. Additionally, without the inclusion of RM Y-STRs, an increase in false negative surname matches was observed, pointing to *DYS464* as the most discriminative Y-STR. This may be due to the fact that *DYS464* is a highly polymorphic multi-copy Y-STR located on palindrome 1, containing four loci. This makes it hard to assign the capillary electrophoresis allele output to the correct loci, leading to an interpretation bias (14). However, it must be noted that the inclusion of RM Y-STRs causes an increase in the Y-haplotype average mutation rate, meaning a higher allowed number of Y-STR differences between a genealogical pair. For this reason, it is necessary to include individual Y-STR mutation rates rather than average mutation rates for kinship analysis purposes as false negative mistakes could be due to the wrong derivation of the maximum allowed number of Y-STR differences per Y-haplotype.

4.2. Y-SNP genotyping

Furthermore, the influence of Y-SNP haplogroup and subhaplogroup genotyping on the surname match frequency was investigated. As main Y-haplogroups had their origin long before the practice of surnames in human populations was introduced, biologically related males with corresponding surnames but different Y-SNP (sub)haplogroups do not share common paternal ancestry (9,28). In this case, there must have been either multiple founding fathers for the same surname or a co-founding factor (false-paternity, maternal surname transmission, etc.). In literature, it has already been described that (deep) Y-SNP determination might be useful to study genetic kinships rather than Y-STR genotyping alone (15). Y-haplogroup analysis decreases the amount of false positive kinships made based on high Y-STR haplotype resemblance, which could lead to an increase in surname matches. However, we observed no influence on the SMF when only the main Y-SNP haplogroup was

encountered as a boundary condition. Nevertheless, when the more detailed Y-SNP subhaplogroup was considered, the SMF generally increased. This trend was most observed with the comparison of males separated by 30 and 40 meioses, as the deep subhaplogroup decreased the chance on a random match with a high number of Y-STR differences (**Figure 2C**). Additionally, a significant higher number of correct surname matches was observed for every Y-haplotype kit when the Y-SNP subhaplogroup was included as a boundary condition (**Figure S3**). Based on these results, the Y-SNP subhaplogroup boundary condition was implemented in the surname prediction model.

4.3. Age of hereditary surname practices: Belgium vs. the Netherlands

The influence on SMF of a country and therefore the time when surnames were introduced was observed through the comparison of the Belgian subset (n=1,859) with the Dutch subset (n=544). Paternally inherited surname practices in Belgium were introduced from the 14th until the 15th century, while in the Netherlands this lasted until the 16th and 17th century. But it was not until 1795 and 1811 that respectively Belgium and the Netherlands were obligated to register their surnames at Civil Services (34). The highest SMF for the three time intervals (20-30-40 meioses) for both subsets were once again observed for samples analyzed with our in-house YForGen kit (**Figure 2D**). An overall lower SMF has been identified for the Dutch subset which could be due to the fact that the Dutch hereditary surname traditions came into use later than in Belgium, meaning that biological kinships have their MRCA before surnames were introduced. However, when a generation time of 30 years is assumed, relatives separated by 20-30-40 meioses have an MRCA who respectively lived around the year 1700-1550-1400 (14). This means that only the Dutch genealogical pairs separated for theoretically 40 meioses share an MRCA who lived before the surname introduction (16th and 17th century). Though, despite the low Dutch SMF for the Yfiler® (4%), the SMF for the other kits is at least 25%. This could be explained by human migration events or a false generation estimation including a wrong generation time interval. The distraction of the number of generations between a biological kinship is based on the online tMRCA calculator using the formula of Walsh (37). Unfortunately, this only relies on the number of changed Y-STRs and the average Y-haplotype mutation rate as previously mentioned. To further improve tMRCA estimations, the formula needs to be adapted especially when RM Y-STRs are incorporated. This can be optimized by the incorporation of individual Y-STR mutation rates together with Y-STR mutation characteristics, such as multi-step and hidden parallel modifications (14,35). The number of generations could therefore be overestimated.

4.4. Surname anthroponymy and frequencies

Surname match frequencies (SMF) were compared for the four main surname origin groups: ancestry, profession, geography and characteristic. In general, all four origin groups composed high SMF (97-100%) for an exact Y-haplotype match (zero Y-STR differences) and thus low false negative surname mistakes. However, on the allowance of up to three Y-STR differences, the SMF of geography and especially characteristic surnames decreased more rapidly in which significant differences were observed compared to the ancestry origin group. These findings contradict a previous study on Spanish surname types where no differences between surname types were observed (29). Nevertheless, it can be concluded that a larger sample size per origin group is needed to increase the power of this comparison which would result in a more well-grounded conclusion.

The surname frequency was examined as an independent factor influencing the SMF when examining a genetic match for familial searching. Since this was already observed in other countries (Spain and Great Britain), surname frequency seems to be correlated with Y-STR diversity as more frequent surnames are believed to have multiple (unrelated) founders (23,28,29). Within the Low Countries, we can confirm this previously stated correlation through the observation that the SMF decreases when the surname frequency increases. However, it is important to note that 95% of the surnames have low frequencies, and despite the fact that we have enough males included in the database bearing a surname with a high frequency, they were clustered as only a few data points. For a stronger correlation, more surnames with high frequencies are needed. Yet, the correlation could still be confirmed since a significant higher percentage of correct surname matches was found when comparing surname clusters containing a low, intermediate and high frequency in the Low Countries (**Figure 4**). A low SMF for surnames with a high number of carriers in a population can be explained by diverse factors. With more surname bearers, the chances on a Y-chromosome lineage interruption increases. Furthermore, these surnames are probably founded by multiple founders (polyphyletic surnames) which increases the number of unique Y-haplotypes per surname cluster. Haplotype frequency and diversity can also be influenced by genetic drift and migration.

4.5. Geographical distance

The last factor with a possible influence on surname match frequency was the geographical distance between relatives. With genetic kinships widely covered across Belgium and the Netherlands, it can be questioned if paternally related males living in close vicinity have a higher chance to be bearers of the same surname than relatives living distantly. To repeat, no

close kinships of less than seven generations are present in the database. Furthermore, 33% of the genetic kinships are crossing the national border, meaning that one relative lives in Belgium and the other one in the Netherlands. Approximately 10% of these crossing-border kinships eventually also share their surname. For every Y-haplotype, we observed a strong negative exponential trend of a surname match when the geographical distance between biologically related males increased (up to 200 km). This indicates the added value to implement the residence of a male in a database when used for surname prediction in familial searching. However, as addresses are apt to change with the opportunity of migration, caution is advised. Paternal ancestors of around the 1850s, before the Industrial Revolution took place, did not have much migration opportunities, and can thus be used as a proxy. The occurrence of foreign surnames in Flanders or names of Flemish origin in the neighboring countries can be caused by migration events in history. The migration from France to Flanders at the end of the 16th century has previously been found to have left a genetic mark on the Y-chromosome of males currently living in Flanders (42). Surname and Y-chromosome correlations found in other countries may therefore also be of interest to the Flemish population and vice versa. Especially for populations sharing a language e.g. Dutch in Flanders and the Netherlands.

4.6. Surname prediction model

When DNA kinship is based upon a Y-chromosome match between two relatives, genealogical information like surnames can be useful due to their paternally co-inheritance. This can be used as an interesting identification tool for forensic investigations involving the possibility to predict the surname of the perpetrator from his DNA left at the crime scene. Taking the various factors influencing the surname match frequency into account, a surname prediction model could be developed. Based on previous results, only kinships with corresponding subhaplogroups were considered. Furthermore, when building the model using the geographical distance between the places of residence, it needs to be assumed that perpetrators commit the crime in the residence they live in. Research on journey-to-crime literature revealed that serial rapists on average commit their crime within a radius of 5 km (43). For serial killers, the median home to crime distance travelled was 15 km in the USA, 9 km in the UK and 11 km in Germany (44,45). Based on this data, we are confident using the geographical distance in the prediction model.

The predictor variables used for the model are the number of Y-STR differences, the geographical distance and the average and range of mutation rates of the changed Y-STRs. The logistic regression model had a very good AUC-value of 0.89, which indicates that it is possible to rank the possible surname matches with their calculated probabilities using the

trained prediction model. The number of Y-STR differences and geographical distance variables contribute for more than 99% to the prediction model. The variables of the Y-STR mutation rates contributing only for a very small part to the model are however significant and can make the difference between a match and non-match.

The study of Solé-Morata *et al.* estimated a surname prediction sensitivity of 40.8% and assessed a false discovery rate (FDR) of 17% (23). With an equal FDR, the surname prediction sensitivity of our model would be 72%. This increase can be explained by the analysis of more Y-STR markers (46 instead of 17 Y-STRs), and by considering the individual Y-STR mutation rates, Y-SNP subhaplogroups and the geographical distance. Another factor could be the difference in sampling since they collected males bearing 50 preselected surnames with a different kinship threshold as previously described in section 2.1. The advantage of our model is that it can be implemented in forensic identification investigations knowing the Y-chromosome profile of the still unknown perpetrator and the place of crime. This information can then be matched to a database and a probability of surname match can be calculated using the logistic regression formulas and the trained variable coefficients in **Table 2**. Finally, this newly obtained information can also be of interest for population genetics and genetic genealogy concerning the correlation strength between the co-inherited Y-chromosome and surnames.

4.7. Ysurnames in a forensic case?

The Y-haplotype and Y-subhaplogroup derived from a biological sample after forensic DNA analysis could in practice be used to generate a priority list of possible surnames, with compatible predictor variables in a Y-chromosomal DNA database, using the previously described prediction model. The compiled priority suspect list could serve the investigators as a potential tool to identify the assailant in a forensic case (16). This approach of identifying a person using his Y-chromosome as a search tool has already been put into practice successfully by children looking for their biological fathers after anonymous sperm donations, and subsequently has potential applications in forensic casework as shown in the solved murder case of Marianne Vaatstra (2,46). The latter investigation stimulated two national law adaptations in the Netherlands concerning forensic use of Y-SNP bio-geographic ancestry and regarding familial searching where DNA databases can be used to indirectly identify a DNA donor through his relatives. The legally approved familial searching includes additional Y-chromosome testing of samples stored in the national DNA database, and the allowance of a large-scale, voluntary DNA mass screenings under certain circumstances. This is only allowed for serious crime leading to many years of imprisonment, and is particularly meant as

last resort to solve cold cases where all other attempts have already failed (2). Unfortunately, for most countries (including Belgium), there is no legal margin for familial searching, nor has there any database containing Y-chromosomal profiles been established.

Ultimately, the success of a Y-chromosome workflow is establishing national or local Y-chromosome databases. In Austria, they are consistently including Y-haplotypes in their national DNA databases with primary objective to identify sexual offences (47). Although the Y-chromosome Haplotype Reference Database (YHRD, www.yhrd.org) contains a very large collection of Y-haplotype information worldwide (48), this does not include surname information and can among other reasons not be used for forensic familial searching. Resolving sexual assaults and cold cases will continue to improve by the utilization of Y-chromosome analysis and by the growth of the national DNA database (49). To set up a database for surname prediction or to draw a strategy for mass screening in a geographical limited area, we observed that it is both important to know which surnames to include as well as which Y-haplotype kit to use for genotyping (see sections 4.1 and 4.4). It has been proposed that including intermediate frequency surnames may be the best approach, since common surnames have a lower SMF and including all rare ones would be impractical (16,40,42). Nevertheless, surnames with high and low frequency can still provide valuable indications, as well as region-specific surnames present in the area where the crime was committed. Concerning Y-chromosomal genotyping, the most valuable analysis includes Y-SNP subhaplogrouping together with extended Y-STR genotyping using the best Y-haplotype kit, which was observed to be our in-house YForGen kit (46 Y-STRs), or at least the commercially available Yfiler® Plus kit (27 Y-STRs). Although these kits provide a lower amount of false negative surname matches and exclude more false positive surname mistakes, we still have to handle surname information in a forensic context with caution as this can cause ethical and social concerns. These issues may have an impact on family and personal privacy, such as the revealing of hidden genetic relationships (unknown paternity or adoption) or designating an innocent individual as a family member of the suspect based on their genetic material (50). Hereby a list of important policy recommendations regarding the use of surnames in forensic familial searching has to be described in order to reduce the personal harm and to provide the best social protection for the relatives and inhabitants.

5. Conclusion

For the first time in Belgium and the Netherlands (the Low Countries), we were able to study the surname match frequency (SMF) using Y-chromosomal and genealogical data from 2,401 males. As expected, the SMF was observed to be inversely correlated with the allowed number of Y-STR differences. Additionally, when a Y-haplotype was more expanded including rapidly mutating Y-STRs, the chance of false positive kinships decreased leading to an overall higher SMF. For a perfect match, a high SMF of 98% could be observed, whereas 2% false positive surname matches could be explained by an interruption in the surname patrilineage due to an extra-pair paternity (EPP) event, a maternal surname transfer, a (hidden) adoption, an anonymous sperm donor or a baby exchange. Our in-house YForGen kit, which genotypes 46 Y-STR loci, could overall be identified as currently the best kit encountering the highest SMF. When deep Y-subhaplogroups were considered, the SMF generally increased. Moreover, we identified a lower SMF within more frequent surnames, as they are believed to have multiple (unrelated) founders and thus more mismatches. In addition, we observed a strong negative exponential correlation between the geographical distance and the SMF. This novel information enabled the design of a surname prediction model with an area under the curve (AUC) of 0.9 and can therefore be used for DNA kinship priority listing in for example forensic familial searching.

Supplemental Data

Supplemental Data include five figures and two tables and are available on the website of Forensic Science International: Genetics.

Declaration of Interests

The authors declare no competing interests.

Author Contributions

Sofie Claerhout: Conceptualization, Data Curation, Formal Analysis, Investigation, Project Administration, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing.

Jennifer Roelens: Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Review & Editing.

Michiel Van der Haegen: Investigation.

Paulien Verstraete: Investigation.

Maarten H.D. Larmuseau: Funding Acquisition, Resources.

Ronny Decorte: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing.

Acknowledgments

We are grateful for all DNA donors of the genetic genealogy project and the non-profit organization 'Histories vzw' for their participation. We also want to thank ir. Simon Vanpaemel and Karel Wanzeele for their valuable input, together with Margot Debunne and Drew Glatczak for reading and improving the manuscript. Funding was provided by KU Leuven [BOF-C1 grant number C12/15/013].

Accession Numbers

The accession numbers for the Y-chromosomal data reported in this paper from the Y-STR Haplotype Reference Database (YHRD, https://yhrd.org) are YA003651, YA003652, YA003653, YA003739, YA003740, YA003741, YA003742, YA004300 and YA004301.

References

1. Jobling MA. In the name of the father: Surnames and genetics. Trends Genet. 2001;17(6):353–7.
2. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. Springer Berlin Heidelberg; 2017;136(5):621–35.
3. Andersen MM, Balding DJ. How convincing is a matching Y-chromosome profile? PLoS Genet. 2017;13(11):e1007028.
4. Calafell F, Larmuseau MHD. The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. Hum Genet. Springer Berlin Heidelberg; 2017;136(5):559–73.
5. Bieber FR, Brenner CH, Lazer D. Finding criminals through DNA of their relatives. Science (80-). 2006;312(5778):1315–6.
6. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000;
7. Balanovsky O. Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. Hum Genet. Springer Berlin Heidelberg; 2017;136(5):575–90.
8. Van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. Seeing the wood for the trees: A minimal reference phylogeny for the human Y chromosome. Hum Mutat. 2014;35(2):187–91.
9. Jobling MA, Tyler-Smith C. The human Y chromosome: An evolutionary marker comes of age. Nat Rev Genet. 2003;4(8):598–612.

10. Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, et al. Y chromosome sequence variation and the history of human populations. Nat Genet. 2000;26(november):358–61.

11. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, et al. Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet. 2010;87(3):341–53.

12. Burgarella C, Navascués M. Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data. Eur J Hum Genet. 2011;19(1):70–5.

13. Ballantyne KN, Ralf A, Aboukhalid R, Achakzai NM, Anjos MJ, Ayub Q, et al. Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats. Hum Mutat. 2014;35(8):1021–32.

14. Claerhout S, Vandenbosch M, Nivelle K, Gruyters L, Peeters A, Larmuseau MHD, et al. Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences. Forensic Sci Int Genet. 2018;34:1–10.

15. Larmuseau MHD, Vanderheyden N, Van Geystelen A, Van Oven M, De Knijff P, Decorte R. Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. Ann Hum Genet. 2014;78(2):92–103.

16. King TE, Jobling MA. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet. 2009;25(8):351–60.

17. Larmuseau MHD, Vanoverbeke J, Van Geystelen A, Defraene G, Vanderheyden N, Matthys K, et al. Low historical rates of cuckoldry in a Western European human population traced by Y-chromosome and genealogical data. Proc R Soc B Biol Sci [Internet]. 2013;280(1772):20132400. Available from: http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2013.2400

18. Anderson KG, Tower DH, A US. How Well Does Paternity Confidence Evidence from Worldwide Nonpaternity Rates. Curr Anthropol. 2006;47(3):513–20.

19. Wolf M, Musch J, Enczmann J, Fischer J. Estimating the Prevalence of Nonpaternity in Germany. Hum Nat. 2012;23(2):208–17.

20. Sasse G, Müller H, Chakrabory R, Ott J. Estimating the Frequency of Nonpaternity in Switzerland. Hum Hered. 1994;44(6):337–43.

21. Greeff JM, Erasmus JC. Three hundred years of low non-paternity in a human population. Heredity (Edinb) [Internet]. Nature Publishing Group; 2015;115(5):396–404. Available from: http://dx.doi.org/10.1038/hdy.2015.36

22. Boattini A, Sarno S, Pedrini P, Medoro C, Carta M, Tucci S, et al. Traces of medieval migrations in a socially stratified population from Northern Italy. Evidence from uniparental markers and deep-rooted pedigrees. Heredity (Edinb). 2015;114(2):155–62.

23. Solé-Morata N, Bertranpetit J, Comas D, Calafell F. Y-chromosome diversity in Catalan suRNAme samples: Insights into suRNAme origin and frequency. Eur J Hum Genet. 2015;23(11):1549–57.

24. Larmuseau MHD, Matthijs K, Wenseleers T. Cuckolded Fathers Rare in Human Populations. Trends Ecol Evol. 2016;31(5):327–9.

25. Larmuseau MHD, Claerhout S, Gruyters L, Nivelle K, Vandenbosch M, Peeters A, et

al. Genetic-genealogy approach reveals low rate of extrapair paternity in historical Dutch populations. Am J Hum Biol. 2017;29(6):1–9.

26. Claerhout S. Genetisch-genealogisch verwantschapsonderzoek in de Lage Landen op basis van Y-chromosomale variatie. Master dissertation (KU Leuven, Belgium); 2016.

27. Gruyters L, Claerhout S. Spatio-temporele analyse van de menselijke koekoeksgraad in de Lage Landen. Master dissertation (KU Leuven, Belgium); 2017.

28. King TE, Ballereau SJ, Schürer KE, Jobling MA. Genetic Signatures of Coancestry within Surnames. Curr Biol. 2006;16:384–8.

29. Martinez-Cadenas C, Blanco-Verea A, Hernando B, Busby GBJ, Brion M, Carracedo A, et al. The relationship between surname frequency and Y chromosome variation in Spain. Eur J Hum Genet [Internet]. Nature Publishing Group; 2016;24(1):120–8. Available from: http://dx.doi.org/10.1038/ejhg.2015.75

30. McEvoy B, Bradley DG. Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. Hum Genet. 2006;119(1–2):212–9.

31. Solé-Morata N, Bertranpetit J, Comas D, Calafell F. Recent radiation of R-M269 and high Y-STR haplotype resemblance confirmed. Ann Hum Genet. 2014;78(4):253–4.

32. King TE, Ballereau SJ, Sch??rer KE, Jobling MA. Genetic signatures of coancestry within surnames. Curr Biol. 2006;16(4):384–8.

33. Manni F, Toupance B, Sabbagh A, Heyer E. New Method for Surname Studies of Ancient Patrilineal Population Structures , and Possible Application to Improvement of Y-Chromosome Sampling. Am J Phys Anthropol. 2005;126:214–28.

34. Marynissen A. Alles over familienamen. Retrieved from https://familienaam.be/alles-over-familienamen [Internet]. https://familienaam.be/alles-over-familienamen; Available from: https://familienaam.be/alles-over-familienamen

35. Claerhout S, Van Der Haegen M, Vangeel L, Larmuseau MHD, Decorte R. A game of hide and seq : Identification of parallel Y-STR evolution in deep-rooting pedigrees. Eur J Hum Genet. 2018;27:637–46.

36. Debrabandere F, De Baets P. Woordenboek van de familienamen in België en Noord-Frankrijk. Amsterdam: Veen; 2003.

37. Walsh B. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics. 2001;158(2):897–912.

38. Pedregosa F, Weiss R, Brucher M. Scikit-learn : Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

39. Hanley J, Mcneil B. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology. 1982;143:29–36.

40. King TE, Jobling MA. Founders, drift, and infidelity: The relationship between y chromosome diversity and patrilineal surnames. Mol Biol Evol. 2009;26(5):1093–102.

41. Larmuseau MHD, Vanderheyden N, Van Geystelen A, Decorte R. A substantially lower frequency of uninformative matches between 23 versus 17 Y-STR haplotypes in north Western Europe. Forensic Sci Int Genet [Internet]. Elsevier Ireland Ltd;

2014;11(1):214–9. Available from: http://dx.doi.org/10.1016/j.fsigen.2014.04.002

42. Larmuseau MHD, Vanoverbeke J, Gielis G, Vanderheyden N, Larmuseau HFM, Decorte R. In the name of the migrant father: Analysis of surname origins identifies genetic admixture events undetectable from genealogical records. Heredity (Edinb) [Internet]. Nature Publishing Group; 2012;109(2):90–5. Available from: http://dx.doi.org/10.1038/hdy.2012.17

43. Warren J, Reboussin R, Hazelwood RR, Cummings A, Gibbs N, Trumbetta S. Crime Scene and Distance Correlates of Serial Rape. J Quant Criminol. 1998;14(I):35–60.

44. Lundrigan S, Canter D. Spatial patterns of serial murder: an analysis of disposal site location choice. Behav Sci Law. 2001;19(4):595–610.

45. Lundrigan S, Canter D. A multivariate analysis of serial murderers' disposal site location choice. J Environ Psychol. 2001;21:423–32.

46. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science (80-). 2013;

47. Neuhuber F, Klausriegler E, Kreindl G, Zahrer W, Dunkelmann B, Pickrahn I, et al. The efficiency of Y-chromosome markers in forensic trace analysis and their inclusion in the Austrian National DNA Database. Forensic Sci Int Genet Suppl Ser. 2013;

48. Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A, et al. Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int. 2001;118(2–3):106–13.

49. Amy Jeanguenat. Y-STR Testing: Enhancing Sexual Assault and Cold Case Workflows. SAKI, Sex Assault Kit Initiat. 2018;2.

50. Kim J, Mammo D, Siegel MB, Katsanis SH. Policy implications for familial searching. Investigative Genetics. 2011.