Challenges in Genomic Privacy: An Analysis of Surname Attacks in the Population of Britain[1]

Sahel Shariati Samani*, Mark Elliot* and Andrew Brass**

* School of Social Sciences University of Manchester, Manchester UK M13 9PL
** Faculty of Biology, Medicine and Health, University of Manchester, Manchester UK M13 9PL

ABSTRACT

In 2013, Gymrek et al. reported that personal genomes can be re-identified through surname inference using patrilineal information inherent in the Y chromosome. They highlighted that the attack is based on freely available resources. This finding has raised significant concerns about the privacy of participants in genomic studies and genomic privacy in general. However, the findings are much less clear cut than the high profile nature of the paper might suggest and the experiments reported in the paper are somewhat *ad hoc*. Therefore, a more thorough analysis of the risk of privacy breaches of genomic data through surname inference is desirable. The current paper analyses this risk in the British population.

Our work demonstrates: (i) that although re-identifying personal genomes by surname inference attack is possible, the risk is relatively low in the population of Britain and crucially dependent on the scale of external resources used to perform the attack; (ii) that many different factors influence the risk and so the risk of re-identifying genomic data via this route is specific to each genomic dataset and hence the risk should be assessed for every dataset individually and (iii) that attaching geo-demographic metadata to genomic data could greatly facilitate re-identification and so we advise that caution should be adopted with such attachments.

Keywords: Genomic privacy, Surname attack, Disclosure control

[1] Please Cite as Samani, S., Elliot, M. J. and Brass, A. (2017) 'Challenges in Genomic Privacy: An Analysis of Surname Attacks in the Population of Britain'. *Cathie Marsh Institute Working Paper* 2017-03

1 BACKGROUND

Human cell nuclei contain two sex chromosomes termed *X* and *Y*. Typically, there are two X-chromosomes in each cell of a female *(XX)*, while male cells include a copy of one of their maternal X-chromosomes and the paternal Y-chromosome *(XY)*. Y chromosomes are therefore necessarily passed from father to son. As a result the Y chromosome can be used to track paternal lineages [1].

This does not mean that the Y chromosome is unchanging. Sex cells (sperm and eggs) undergo a form of DNA replication called meiosis. Occasionally, errors can occur during meiosis which will change the form of the Y chromosome passed from father to son. Such errors can take a number of forms. For example, a single base can be changed in a Single Nucleotide Polymorphism (SNPs). These errors are relatively infrequent and are therefore useful for tracking changes that occur over long periods of time, for example in historical studies of the migration of human populations [2]. A second form of error can be observed in regions of the chromosome that contain short repetitive regions known as Short Tandem Repeats (STRs[2]). In particular, the number of repeats in any given STR can change. Y-chromosomal Short Tandem Repeats (Y-STRs) are the most changeable parts of the Y chromosome and the most likely to differ between generations. This rate of change has been studied for father/son pairs [3] and pedigrees [4]. This high rate of changeability means Y-STRs can be useful in distinguishing between more recent male lines [5]. The Y chromosome therefore provides an insight into male inheritance patterns and serves as a link to other data that might also correlate with patrilineal inheritance.

From the Mediaeval period, in Britain and elsewhere, the addition of the surname to the given names of individuals became common practice. In most societies, surnames are passed from the father to the child. This means that – for male offspring – surnames are usually inherited in parallel with the Y-chromosome and so a culturally inherited feature (surname) is correlated with a genetically inherited element (the Y chromosome) [6, 7]. This leads to the possibility that the Y chromosome of an individual can be used to infer that individual's surname. This has clear implications for privacy.

[2] Sometimes referred to as "microsatellites",

Several studies have looked at the degree of the association between surnames and Y-chromosome sequence variation. In 2000, Sykes and Irven [8] investigated the association between Y-chromosome haplotype[3] and surname for a sample of males sharing the same surname "Sykes". They demonstrated that Sykes males with the same ancestry shared the same Y-chromosome haplotype, and it was reported that there was a significant association between whether males had the surname "Sykes" or not and distribution of the Y-chromosome haplotypes. King and Jobling [9] analysed Y-chromosomal pattern diversity within 40 British surnames using 1,678 samples. Their analysis illustrated a strong relationship between surnames and Y-chromosome haplotypes. This correlation could have many applications in genealogy and forensics [5]. In genealogy in particular, it has sparked interest in the potential of genetic genealogy for enriched understanding of the family trees and origins. As sequencing technologies improve and their costs reduce, the use of Y-chromosome sequencing for these secondary purposes becomes more viable and has triggered the establishment of several databases (and associated project websites) containing Y-STR haplotypes and associated surnames [10]. This in turn has led some researchers to explore whether it is possible to use such genealogical data resources as a tool for determining the identity of unidentified Y chromosome sequence data.

In 2013, Gymrek et al. [11] presented a study that shows that databases that include Y-chromosome haplotypes and the associated surnames pose a threat to the confidentiality of personal genomic data. They demonstrated that they could recover surnames associated with personal genomes by profiling their Y-STR haplotypes and querying genealogical databases[4]. They then showed that combining a surname with other demographic data could lead to re-identification of the target genome leading to heightened concerns about genomic privacy. However, as we argue in this paper, the findings are less clear cut than the high profile nature of the paper might suggest and so it is crucial at this point to conduct more general analyses of the real risk of privacy breaches through surname inference.

1.1 Identifying Personal Genomes by Surname Inference

In 2008, Lunshof et al. [12] introduced the idea that a combination of surnames, genotypes and geographical information is a threat to privacy. Gitschier [7] pursued this idea

[3] The term haplotype refers to a combination of alleles that are inherited together as a block from a single parent.
[4] Strictly these are genetic genealogical databases, for brevity we drop the word genetic here.

experimentally by examining 30 unrelated CEU[5] participants in the HapMap project and reported that the detection of the potential surnames was possible. Nevertheless, these potential surnames correspond to multiple individuals and the study did not itself lead to re-identification of genomics data at the individual level. In 2013, Gymrek et al. [11] followed this to simulate an actual re-identification of personal genome data, which they refer to as *end-to-end re-identification*. They demonstrated that it was possible to recover surnames from personal genomes by profiling Y-STRs and using genetic genealogy databases and that the target could then be re-identified by combining the surname with other types of information, such as age and state of residency.

To recover surnames, Y-STR haplotypes are first required. Gymrek et al. [11] used lobSTR which is an algorithm to profile STRs from raw sequencing reads and produce Y-STR haplotypes. They then used Ysearch (www.ysearch.org) and SMGF (www.smgf.org)[6], as their primary resources to underpin surname inference. These databases have built-in search engines which allow users to query with Y-chromosome STR haplotypes and search for possible matching records based on genetic similarity. These search engines usually retrieve matching surnames with some information related to the paternal line, like pedigrees and geographical locations. The two datasets included about 135,000 records with approximately 39,000 surnames between them and Gymrek et al. [11] claimed that they are representative of the distribution of surname frequencies in the United States[7].

A brief description of Gymrek et al.'s algorithm for inferring the surname of a given Y-STR haplotype follows: First, the database record that has the shortest number of generations to most recent common ancestor is retrieved. Then, a confidence score, generated through comparison with other possible matches, is calculated and compared with a pre-defined threshold. If the score passes the threshold, the recovered surname will be assigned to the input haplotype, otherwise the input haplotype will be categorised as "unknown".

The recovered surnames are then combined with demographic data and, in one experiment, pedigree information to perform the end-to-end re-identification. This type of auxiliary data is associated with the genomic data in the Coriell Cell repository where the 1000 genome

[5] CEU participants are Utah residents with Northern and Western European ancestry whose samples were collected by CEPH (Centre d'Etude du Polymorphisme Humain).
[6] This web site is no longer available.
[7] As evidence for that, they state that the $\log_{10}$ of the number of records per surname is correlated with $\log_{10}$ surname frequencies in the United States with $R^2 = 0.78$. We note that correlation of decimal logs is at best a proxy measure of the equivalence of distributions and that even logged the relationship is non-linear. However, this detail is not central to our argument, so we simply note this in passing.

project database is housed and so this is a reasonable scenario to be exploring. Gymrek et al. [11] actually report two slightly different experiments one which uses surname, year of birth and state of residence (where the target is particular individuals) and the other which adds in pedigree (where the target is entire extended families).

To protect any information against possible privacy threats, systematic analysis and evaluation of plausible attacks is necessary. Based on such a risk assessment, we can better understand and manage the attacks. Thus, in this paper we carry out a more extensive analysis of the risk of re-identifying personal genomes and breaching privacy of genomic data by surname inference in a British context. The contribution of the paper can be summarised as follows:

1. Based on a model of the distribution of surname frequencies, we quantify the probability of *recovering the surname associated with a genome*. By recover we mean correctly linking a surname with a genome.
2. We then quantify the risk of *re-identification* of a personal genome using surname inference attack.
3. We then analyse the impact on the risk of re-identification of combining surnames with metadata on age and location.

Note that in the later simulations the data we use is for England and Wales only; for the sake of brevity we will refer to this by its (now somewhat archaic) name of "Britain" [13].

2 METHOD

2.1 Intruder Model

To analyse attack scenarios and perform a disclosure risk assessment, after understanding the key characteristics and main uses of the data, we need to define the situations in which a disclosure might occur [14]. In this paper, we first consider a scenario where an intruder holds a (single) de-identified personal genome and aims to re-identify that genome via surname inference. For such an attack, the intruder first needs to profile STRs from the genome and produce Y-STR haplotypes. We assume that either the intruder has the genomic knowledge required for this purpose or the genomic data he holds is in the form of Y-chromosome haplotypes. We also assume that the intruder has access to the required resources including a database of linked surnames and Y-STR haplotypes and also has the

necessary expert knowledge and bioinformatics tools needed to recover the matched surname S. Finally we make three simplifying assumptions. The first is that an intruder with a sufficiently compelling motivation to carry out such an attack exists, the second is that that the payload of such an attack is unrelated to its probability and the third is that there is no *data divergence* [14] present which hampers any of the linkages[8]. These simplifying assumptions allow us to treat the *risk* a successful attack and the *probability* that an attack will be successful as being synonymous and make the calculations tractable. This type of simplifying assumption is standard in disclosure control research; the net effect is to produce an upper bound on the risk.

We then expand this scenario and assume that the intruder has access to a genomic database including m personal genomes and he intends to re-identify at least one of the genomes via surname inference. It should be noted that in this work we will always be considering male genomic data as the attack necessarily uses patrilineages.

2.2 The Probability of Surname Inference

To model this attack, we assume that the intruder has access to a database D of n surname-Y-STR haplotype pairs sampled randomly from the male population of Britain including N people. We refer to this database as the *external genealogical database*. He also holds a de-identified personal genome G (known as *target genome*) selected randomly from the whole population. For an intruder to be able to recover the surname using the external genealogical database, at least one male whose surname is S should be in that database. Therefore, the probability of the surname S being recovered for G can be estimated as the probability of having at least one of the F_S males with surname S from the male population in the external genealogical database:

$$P(recovering\ S) = P(\exists S \in D) = 1 - P(\nexists S \in D) = 1 - \frac{\binom{N-F_S}{n}}{\binom{N}{n}}$$

[8] More specifically, the model assumes that there is a one-to-one empirical mapping between haplotypes and surnames in the population. This will rarely be true, particularly for popular surnames, where multiple haplotypes are likely to be associated with a single surname. There are also other sources of error in the mapping: Some individuals have acquired their surname from adopted or step parents, from their biological mother, or from having changed their names in adulthood. There may also be errors in the sequencing process. Gymrek et al. estimate the total error rate to be 5% and call this conservative but give no information as to how they arrived at this figure, which is an unknown empirical quantity. In line with standard disclosure control approaches for this type of scoping exercise and in order to progress the analysis, we make a simplifying assumption here that the error rate is zero. We will return to the issue in the discussion.

To expand the scenario, we assume that the intruder has access to a genomic database GD of m de-identified personal genomes (refered to as *target genomic database*). We assume that each personal genome G_i has a surname S_i and that GD is a simple random sample of the population. The intruder also holds an external genealogical database D including n surname-Y-STR haplotype pairs sampled randomly from the male population of Britain. The probability of the intruder to be able to recover at least one of the m personal genomes can be computed as:

$$P(recovering\ at\ least\ 1\ S) = 1 - P(recovering\ no\ S) = 1 - \prod_{i=1}^{m}\left(\frac{\binom{N-F_{S_i}}{n}}{\binom{N}{n}}\right)$$

2.3 The Impact of Surname Inference on Re-Identification Risk

Since the number of males who share a given surname can vary significantly depending on the popularity of the surname, the recovery of a surname also varies in terms of its impact on the risk of re-identification. For instance, correctly associating a de-identified genome with the surname "Smith" (which occurs over 600,000 times in the UK population) is very different from associating the surname "Austin" or "Rubaduka" (whose occurrences are approximately 28,000 and 10 in the UK population, respectively) [15]. The rarer the surname is in the population, the bigger the impact of its inference is on the probability of re-identification. We can express this quite simply; the impact of inferring surname S, which occurs F_s times in the population is:

$$I(S) = \frac{1}{F_s}$$

Therefore, the probability of re-identifying a single de-identified male genome via surname inference can be estimated as follows[9]:

$$P(reidentification|G) = P(recovering\ S) \times I(S) = \left(1 - \frac{\binom{N-F_s}{n}}{\binom{N}{n}}\right) \times \frac{1}{F_s}$$

[9] Strictly, this is only the risk of re-identification if the intruder's strategy is to attribute the surname and then to draw randomly from the list of people with that surname and "guess" that that is the person to whom the genome belongs. This is the best that the intruder could do if they have no other auxiliary information, but it is probably reasonable to say that no intruder would adopt such a strategy.

If the intruder has access to a target genomic database GD of m de-identified personal genomes, rather than a single genome, then the probability of re-identifying at least one of the genomes in the genomic database via surname inference can be calculated as:

$$P(reidentification\ of\ at\ least\ 1\ G|GD) = 1 - P(reidentification\ of\ no\ G)$$
$$= 1 - \prod_{i=1}^{m} P(not\ reidentifying\ G_i)$$
$$= 1 - \prod_{i=1}^{m} (1 - P(reidentification|G_i)$$
$$= 1 - \prod_{i=1}^{m} (1 - \left(1 - \frac{\binom{N-F_{s_i}}{n}}{\binom{N}{n}}\right) \times \frac{1}{F_{s_i}})$$

2.4 Modelling Frequencies Distribution of Surnames

In order to assess the surname inference attack and the impact of surname inference on the risk of re-identification, we need to model the distribution of the surname frequencies in the population. In the literature, some studies explore the distribution of surname frequencies and a review of these can be found in [16]. For instance, Fox and Lasker [17] demonstrate that the distribution of surname frequencies in the UK population follows a Discrete Pareto Distribution which means that the number of surnames occurring t times in the population is proportional to $t^{-\beta}$, where β is a positive constant. Thus, the number of surnames which occurs t times in the population can be modelled by:

$$F(t) = \alpha . t^{-\beta}$$

where α is a constant and the $f(t)$ sum to 1.

This shows that the distribution of surname frequencies follows a simple linear regression model on logarithmic scale. The dataset used in these experiments is a publicly available one from [18]. It comprises the 250 most common surnames in Britain, published by National Statistics 2002, which have frequencies between 27,000 and 660,000 the UK population.

This data was fitted to a 2nd order polynomial where $R^2 = 0.9991$:

$$y = -0.142 * x^2 - 0.162 * x + 5.639$$

where,

$x = log(rank_s)$

$y = log(F_s)$

$rank_s$: $surname's\ rank\ in\ the\ population$

F_s: $frequency\ of\ surname\ S\ in\ the\ population$

3 RESULTS

In this section, we first evaluate the effectiveness of a surname attack and then given the impact of surname inference on the risk of re-identification, we assess the overall risk.

3.1 Part 1: Genomic Information Only

3.1.1 Surname Inference Attack (Given an Intruder with a Single Genome)

In the first model, we used the model of the distribution of surname frequencies described in section 2.4 and estimated the probability of surname inference given an external genealogical dataset of n surname-Y-chromosome haplotypes as described earlier. The results presented use a range of external database sizes from 1,000 to 500,000. Figure 1 shows the probability of recovering surname S with rank r given the model of surname frequencies distribution and different external genealogical databases.

Figure 1 shows that the probability of recovering a surname decreases markedly as the rank of surname in the population increases. It is also clear that we are likely to be able to infer a common surname even when the sample database includes as few as 1,000 entities[10]. We also see that the probability of recovering a particular surname being recovered is higher for bigger external genealogical databases.

3.1.2 Estimation of Risk of Re-Identification

However, the full re-identification risk must also consider the number of people with a particular surname. Specifically, we used the distribution model to compute the number of the male with surname S and rank r in the whole population and then compute the risk $P(reidentification|G)$. Similar to the above experiment, we modelled four genealogical

[10] This is a theoretical estimate which assumes do data divergence. In practice, it would require a larger number of entries. See section 4 for discussion of this issue.

databases of different sizes to explore the effect of its size on the risk. Figure 2 shows the overall risk of re-identifying a genome associated with an individual with surname S and rank r using the genealogical databases of the four sizes.

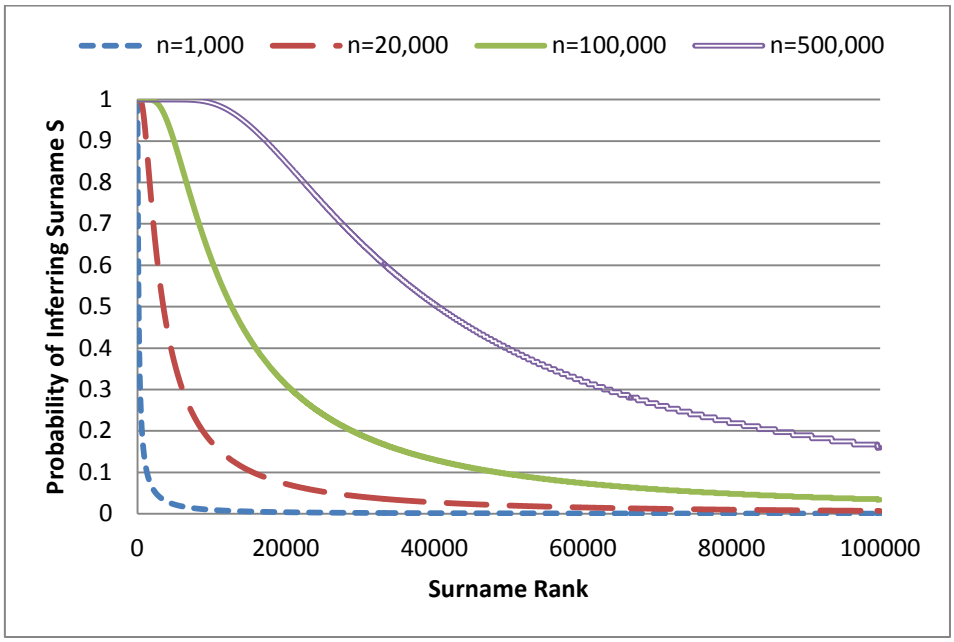

Figure 1- Probability of recovering surname S with rank r

We use external genealogical databases of surname-Y-chromosome haplotypes of different sizes ranging from 1,000 to 500,000.

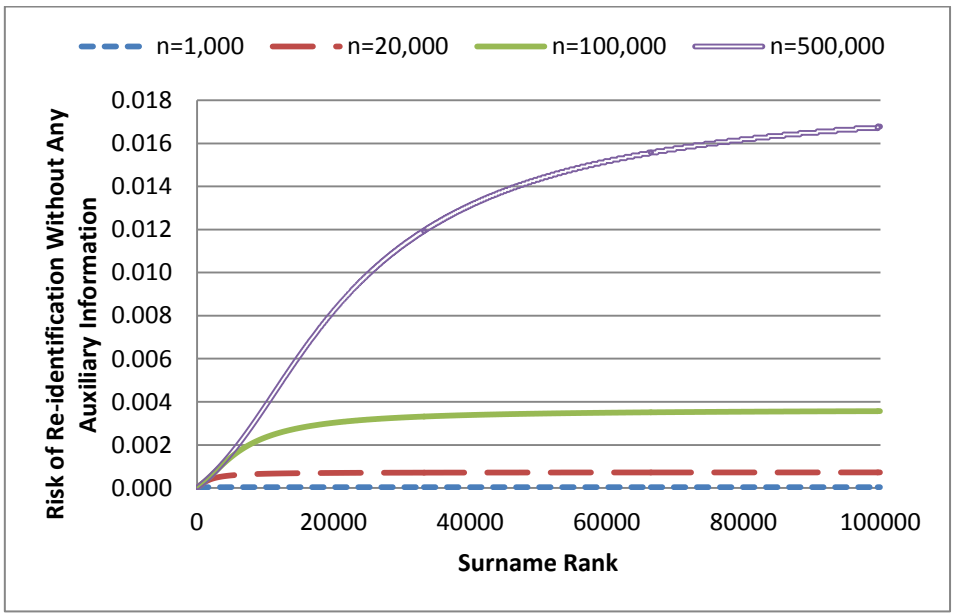

Figure 2- Overall risk of re-identifying a genome associated with an individual with surname S and rank r

We use four external genealogical databases with different sizes n from 1,000 to 500,000.

Figure 2 shows that the overall risk of re-identification increases as the surname's rank rises and the number of males with that surname decreases. This illustrates three things:

1. That the overall risk of re-identification is higher for surnames which are rarer even though they are less likely to be recovered in the first step. Conversely, the overall risk is quite low for common surnames even though the probability of recovering them is high.
2. That the overall risk of re-identifying a genome associated with surname S increases as the external genealogical database grows. It indicates that the risk of re-identifying common surnames is very low even for big n whereas the risk of re-identification of rare surnames increases as n grows.
3. That the maximum risk of re-identification is less than $0.5*10^{-4}$ when n=1,000 where the maximum risk increases to approximately 0.0168 when n rises to 500,000 samples.

3.1.3 Surname Inference Attack (Given an Intruder with a Target Genomic Database of m Records)

For this scenario we first generated a target genomic database of m samples selected randomly from the whole population with the distribution of surname frequencies as described before. Then having the genomic database, we used the above model to quantify the probability that an intruder would recover the surname associated with at least one of the genomes in the genomic database given a genealogical dataset. The external genealogical database includes n surname-Y-chromosome haplotypes sampled randomly from the British population. We performed the simulation using simulated target genomic databases with a range of different sizes and external genealogical databases of four sizes. As above we then considered the re-identification risk given the number of people expected to share that surname.

Figure 3 shows the mean risk of at least one re-identification associated with a target genomic dataset of size m, given four genealogical datasets including different number of samples. As Figure 3 shows, the overall risk of re-identification increases linearly as the size of genomic database (m) grows, but stays low when the external genealogical database includes a few number of samples n=1,000. The risk however increases markedly as n grows (as expected).

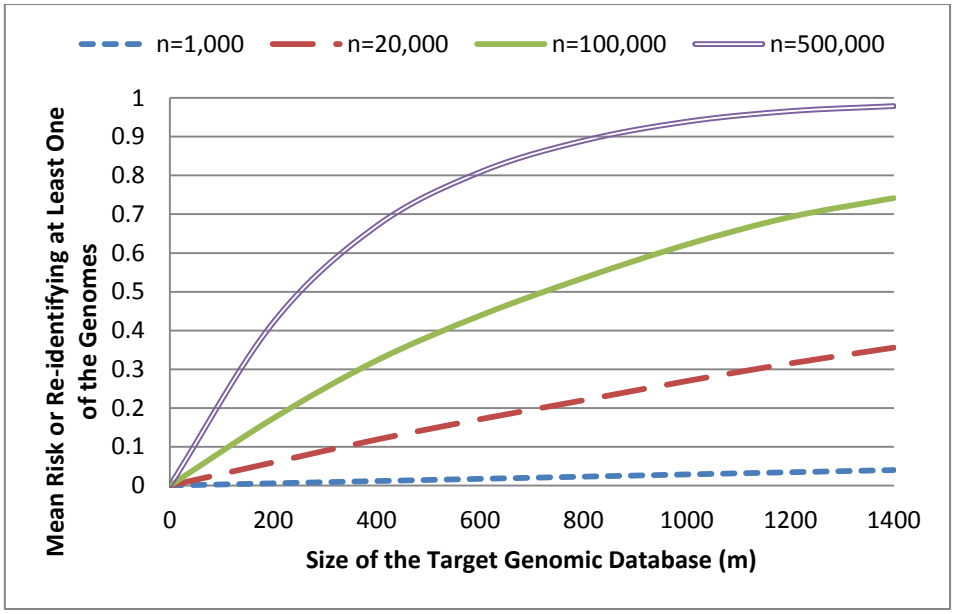

Figure 3- Mean risk of re-identifying at least one of the genomes in the target genomic database of size m

We use an external genealogical dataset including n surname-Y-STR haplotypes pairs of four sizes from ranging from n= 1,000 to 500,000.

3.2 Part 2: Target Genomic Database of m Records with Attached Auxiliary Information

Based on comprehensive simulations with the US census data, Gymrek et al. [11] demonstrate that searching for individuals using the combination of surname, state of residency and year of birth using online resources produced median cross classified frequencies of twelve male records. Therefore, they conclude, such a combination will generate sufficiently few matches that manual investigation is feasible.

Here we simulate this ad hoc study, at scale, with the British population. Barrai et al. [19] show that the US population is highly mobile, people from different origins are spread over the entire area of the US. In contrast, in Great Britain, Anglo-Saxon surnames are often spatially concentrated in the areas where they first became popular [20]. Cheshire et al. [21] demonstrate that in Great Britain there is a strong association between surname distribution and geographical locations and mobility is not as strong a phenomenon in this population. It is therefore a meaningful exercise to make probabilistic inferences about an individual's location of residence from their surname, which is potentially useful to an intruder trying to track somebody whilst being only in possession of their genome. On the other hand, adding the individual's location of residence (at least at the coarse geographical scale that Gymrek et al. use) is likely to be generally less informative to an intruder than in the US case. However,

if an individual's region of residence is not one where people of that surname are concentrated, then a high differentiation of the individual from the rest of the population is provided which may increase the risk of re-identification significantly for such individuals. This is what Elliot and Dale [14] refer to as the risk impact of multivariate skew.

To comprehend the effect of combining geographical region and age – the two additional pieces of information that Gymrek et al. used for an end-to-end re-identification of a genome – with surnames on the overall risk of re-identification of a de-identified genome in a British context, we performed three experiments that will be discussed in the following sections.

3.2.1 Effect of Attaching Geographical Regions to Genomes on the Overall Risk of Re-Identification

The rarer a surname is in a geographical region, the bigger is the impact of attaching this auxiliary information to genomes on the overall risk of re-identification of a genome via surname inference. We simply express the impact as following:

$$I(S|GOR) = \frac{1}{F_{S,GOR}}$$

where $F_{S,GOR}$ represents the frequency of males with surname S in *government office region* GOR (a close approximation, in terms of average size of population, to the state of residence geography used by Gymrek et al).

Therefore, the overall probability of re-identifying a single de-identified genome, via surname inference, given its associated geographical region can be quantified as following:

$$P(reidentification|G, GOR) = P(recovering\ S) \times I(S|GOR) = \left(1 - \frac{\binom{N-F_S}{n}}{\binom{N}{n}}\right) \times \frac{1}{F_{S,GOR}}$$

where N is the total number of males in the population of Britain, n is the number of surname-Y-STR haplotype pairs in the external genealogical database and F_S is the number of males with surname S in the population.

In this experiment, we selected ten arbitrary common surnames from the British male population (from Smith with rank 1 to Heywood with rank 990) [18] and obtained their

frequency distribution in ten different geographical regions of Britain[11]. We then estimated the overall risk of re-identifying a genome related to each of the above surnames for different GORs. We performed the same experiment simulating external genealogical databases of eight different sizes (n) ranging from 1,000 to 1,000,000. Figure 4 illustrates this when n=1,000 and n=20,000, showing that the mean risk increases as the surname's rank in the population rises, emphasising that the rarer a surname is, the higher is the overall risk of re-identification. However, the mean risk is relatively low when n is small (less than 0.5×10^{-3}). It also shows that the mean of the overall risk increases as n becomes larger. It can also be seen that increasing n has more impact on the overall risk of re-identification of genomes associated with rarer surnames compared to the very common ones.

To demonstrate the impact of knowing a genome's geographical region on the overall risk of re-identifying that genome via surname inference, Figure 5 shows the overall risk of re-identifying a genome with no additional information (only based on surnames) and the mean risk of re-identifying a genome associated to surname S with rank r, given the geographical information. We used an external genealogical dataset with 1,000 samples, and the ten surnames that we had their frequency distributions in different geographical regions.

As Figure 5 shows, the mean risk of re-identifying a genome via surname inference increases markedly when geographical information is attached to the genome. For instance, the risk of re-identifying a genome which is related to a male named "Heywood" with rank 990 in the population is about 3.4×10^{-5} when n=1,000 and no geographical information is attached, whereas the mean risk increases to 5.5×10^{-4} if we add geographical information – more than 16 times bigger. Figure 5 also illustrates that adding geographical information has more impact on the risk of re-identification of genomes related to rarer surnames in comparison with the more common ones.

[11] This data was provided by Paul Langley from the UCL Centre for Advanced Spatial Analysis. Led by Paul Langley and Richard Webber, researchers in this centre study the distribution of surnames in the UK and have launched an online tool to map surname concentration by county.

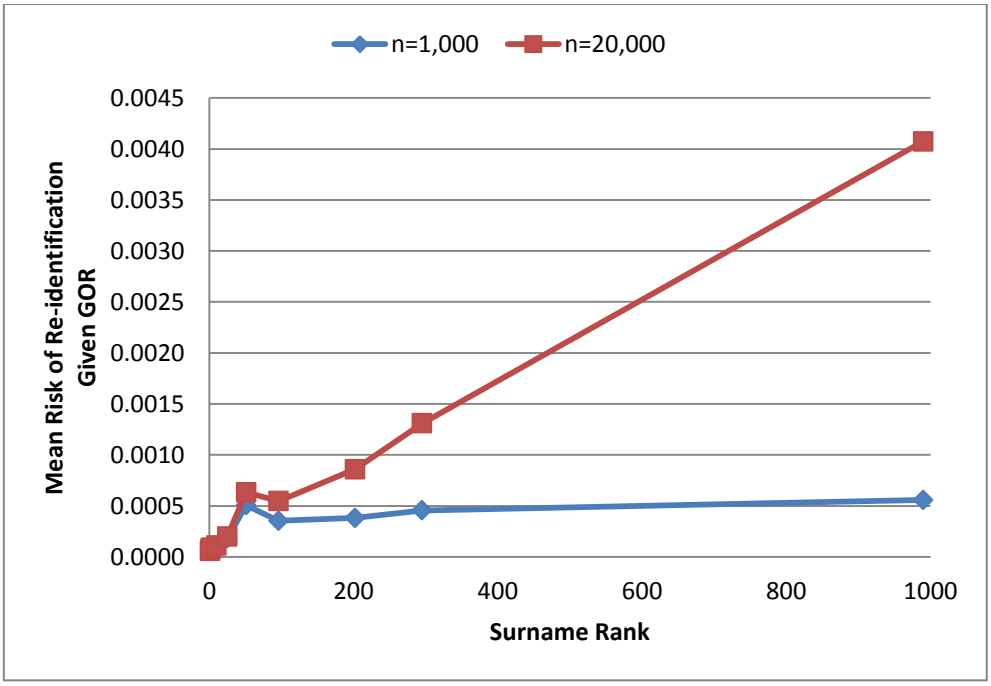

Figure 4- Mean risk of re-identifying a genome related to each surname knowing the geographical region associated with the genome

We use two external genealogical datasets with sizes n=1,000 and n=20,000.

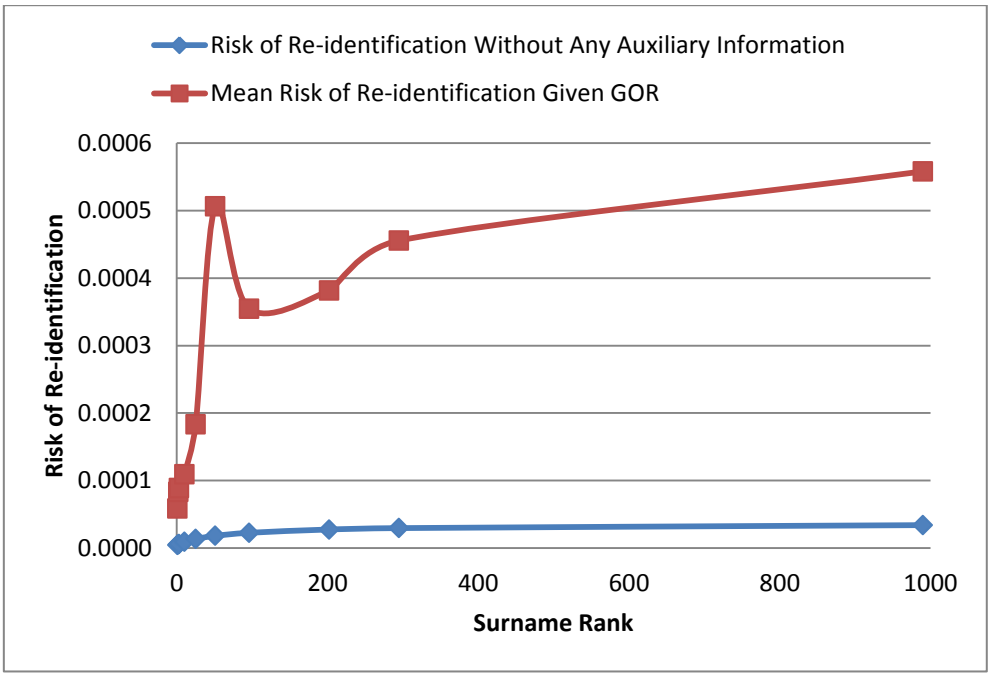

Figure 5- The impact of adding geographical information to the target genome *G* associated to surname *S* with rank *r* on the risk of re-identification

We use an external genealogical dataset with n=1,000.

3.2.2 The Effect of Attaching Geographical Regions and Age to Genomes on the Overall Risk of Re-Identification

Attaching age information to genomes has a slightly different impact on the overall risk of re-identification as attaching geographical information. In principle, the rarer a surname is in a particular age range, the more impact it has on the overall risk. However, it is unlikely that the relationship is as strong as with GOR. Any effects would arise from differential fertility rates by surname and differential migration patterns. Unfortunately we could not obtain data for the distribution of surname dependent upon age. To simplify this whilst still allowing us to take into account the differentiating effect of age on risk we treat the distribution of surname by age as uniform. Given this assumption, the impact of attaching age to genomes, given that the geographical information is also attached, is indicated by the following function:

$$I(S|GOR, AGE) = \frac{1}{F_{S,GOR}} \times \frac{1}{F_{GOR,AGE}}$$

where $F_{S,GOR,}$ represents the frequency of males with surname S in geographical region GOR and $F_{GOR,AGE}$ represents the frequency of males in GOR who have age = AGE.

Therefore, we can quantify the probability of re-identifying a single de-identified genome, via surname inference, given its associated geographical region and age as following:

$$P(reidentification|G, GOR, AGE) = P(recovering\ S) \times I(S|GOR, AGE)$$
$$= \left(1 - \frac{\binom{N-F_S}{n}}{\binom{N}{n}}\right) \times \frac{1}{F_{S,GOR}} \times \frac{1}{F_{GOR,AGE}}$$

where N is the total number of males in the population of Britain, n is the number of surname-Y-STR haplotype pairs in the external genealogical database and F_S is the number of people with surname S in the population.

In the following experiment, the dataset used for the age frequency distribution in different geographical regions, is a publicly available one from [22]. It includes the number of males and females in different age groups in ten geographical regions in Britain obtained from census 2001, published by office for National Statistics. In this dataset, the data is provided for every single age from 0 to 24, then for every age group of 5 years from [25-29] to [85-89], plus 90 and over as the last group. In order to simulate the effects of knowing a subject

age to within a year we assumed that the age profiles were equally split across the 5 year age ranges (i.e. we assumed that the number of people aged 33 could be estimated as a 1/5 of the all people aged between 30 and 34)[12]. From this we could then calculate the risk of re-identification as a function of surname for various external genealogical database sizes when we add age and geographical region data. Figure 6 shows the mean risk of re-identifying a genome related to a male with surname "Smith" or "Heywood" located in each geographical location knowing the age (in the form of year) associated with the genome, when n is equal to 1,000 or 20,000. Note that these means mask variation in risk and in such cases the maximum risk tends to 1. That is there will be some combinations of age, region and surname which are unique in the population.

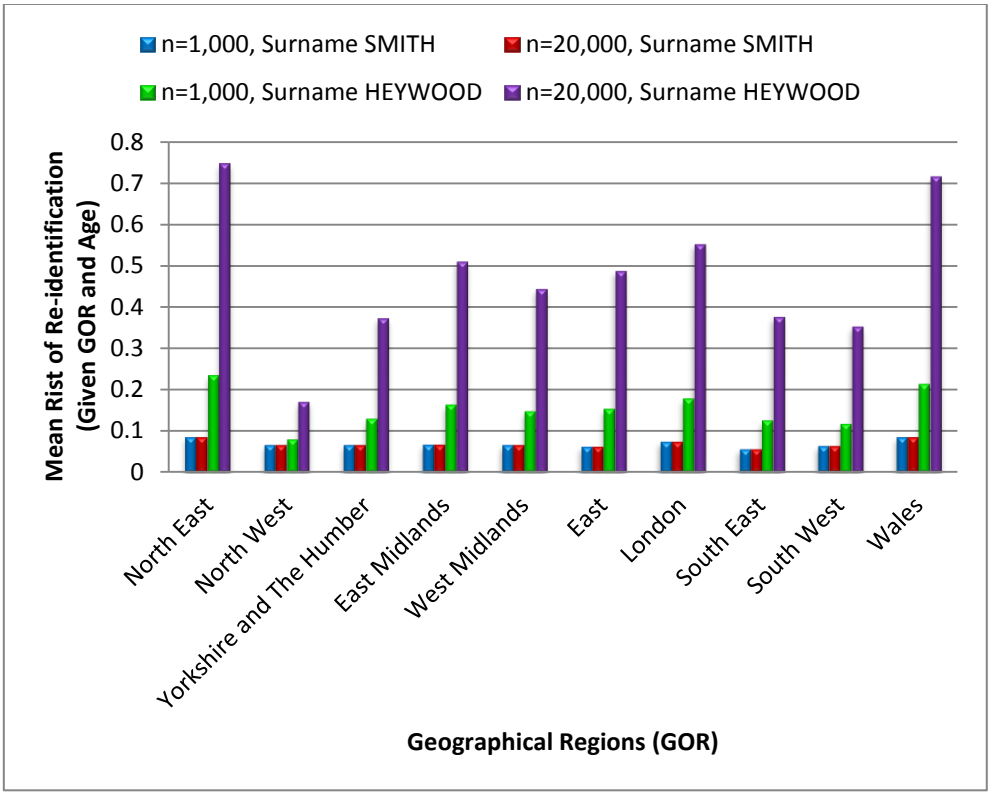

Figure 6- Mean risk of re-identifying a genome related to a male with surnames "SMITH" and "HEYWOOD" located in each geographical location knowing the age in years associated with the genome

We use two external genealogical datasets with sizes n=1,000 and n=20,000.

Figure 6 shows that adding age data to the target genome as well as geographical information increases the mean risk of re-identification considerably. For instance, the mean risk of re-identification of a genome related to surname "Heywood" increases from approximately 0.0005 to 0.2 (400 times bigger) when we add age and geographical region data to the

[12] This is obviously a simplifying assumption but is a close enough approximation to the underlying distribution for our current purposes. The post 100 population is excluded here and for males this tapers off very fast so much so that it is fair to assume that exact age plus region plus surname will be unique pretty much all the time.

genome (where n=1,000). Similar to Figure 5, Figure 6 shows that increasing the size of the external genealogical database n increases the mean risk of re-identification of rarer surnames more noticeably than the common ones. It can also be seen in Figure 6, that some combinations of age, geographical region, and surname which are very likely to be identified in the population. We also performed the same experiment using larger external genealogical databases and our results demonstrated that for n>20,000 the mean risk is similar (within a few percentage points).

4 DISCUSSION

In this paper, we have systematically analysed the risk of an intruder being able to re-identify a male genome within the British population using a surname attack. The headline results are that the probability of an intruder being able to re-identify a single genome with no auxiliary information attached is low even utilising genealogical resources larger in scale than those currently available. With an intruder who has obtained access to a de-identified genomic database, the risk of them being able to correctly re-identify at least one of the genomes in the database increases significantly as the size of the genealogical resource and the de-identified genomic database increase. If auxiliary information is attached to the genome then the risk also increases markedly. On the numbers given here the risk can be said to exceed the principal of negligibility [23] and therefore such data would be classified as personal data. However, the interpretation of those numbers is dependent on the assumptions that we have made in constructing the intruder models. We will now examine those assumptions.

1. **Knowledge assumption.** We assumed that either the intruder has the genomic knowledge (or the genomic data he holds is in the form of Y-chromosome haplotypes) and necessary expert knowledge and bioinformatics tools needed to recover the matched surname.

   To recover surnames, the intruder must have access to the target Y-STR haplotypes. This means that the intruder should either (i) have a good knowledge of genomics to be able to profile STRs from raw sequencing reads and to be able to produce the Y-STR haplotypes and also has access to high coverage raw sequencing reads for the target or (ii) have access to the target's Y-STR haplotypes with a large number of markers. This highlights that the risk associated with genomic data via this route is

dependent on the properties of each particular genomic dataset. For example, genomic data being sequenced for medical purposes usually includes only some particular parts of the individuals' genomes and profiling their Y-STRs is impossible and hence performing such an attack is impractical.

2. **Means Assumption.** We also assume that the intruder has access to the required resources including a database of linked surnames and Y-STR haplotypes.

   To perform such an attack, the existence of databases that contain Y-STR haplotypes and associated surnames is necessary. Such databases should either be directly accessible by the intruder or they can be queried using Y-STR haplotypes. Currently, several databases and surname project websites exist that contain some Y-STRs and associated surnames [10]. Two of these – Ysearch [24] and Ymatch [25] both maintained by Family Tree DNA – contain data from throughout the world and can be searched by STR haplotypes. These databases include approximately 185,000 and 1,300 records respectively.

   In addition, there is one UK specific project, named the Oxford Genetic Atlas Project (OGAP) [26], includes both Y-chromosome and mitochondrial databases involving over ten thousand volunteers from Britain and Ireland. Their services include deduction of the maternal and paternal clan, ancient ancestral mother and father as well as resolving genealogical relationship. However, the OGAP databases are neither publicly accessible nor searchable.

   Considering the approximate number of records in each genetic genealogical database that are currently available, our results show that the maximum risk of re-identifying a single individual's genome by surname inference (assuming no auxiliary information) is less than 0.075% where $n<200{,}000$. Bear in mind that there are various factors which moderate this figure, deriving from the manner in which these databases have been populated. Firstly, these databases are global, so the proportion of records relevant to the British population will be smaller than the total, and that geographical separation is likely to be associated with separation of genomic patrilineage, this indicates that 0.075% represents an upper ceiling of the risk. On the other hand, socioeconomic biases among the customers of genetic genealogy companies may imply that certain groups and their surnames are over or underrepresented, so for example certain particularly rare but economically successful surnames may be

associated with much higher risk. This will cause some error variance in our estimates, but is very difficult if not impossible to control for.

However, this risk increases as the number of records available in genetic genealogical databases grows. Due to the rapid progress in genomic research and people's interest in knowing about their family origins and history, the number of these databases is rising and several companies have plans to develop larger databases. Therefore, it is vital to consider policies governing such databases, their distribution, and the number of entities they can contain as they do significantly affect the re-identifiability of genomic data.

It is worth noting that in their study Gymrek and her colleagues did not recover the surnames by using the publicly available search engines provided by Ysearch or SMGF; they downloaded the Ysearch database records onto their own server with the agreement of Family Tree DNA [27]. They did that in order to facilitate their analyses and carry out informative meta-analyses. This does somewhat contradict their assertion that the technique relies on free publicly available resources as not every public user can download these databases. In practice, this would not prevent a real intruder carrying out exactly the attack that they did but it would make it more difficult for the intruder to be able to verify certain information and therefore would impact on the intruder's confidence in any given match. For example, as they had the complete database, they could confirm that the database was representative of the surname distribution for the US population and so they had higher confidence in the surnames recovered, could measure false positives in the matching and so on.

3. **100% surname to mapping assumption.** We assume that all the males with the same surname have the same Y-STR haplotype with the same ancestral origin. Therefore, to recover their surnames, we just need to have one of the males with that surname in the genealogical database. However, this assumption overestimates risk. We summarise the three main reasons for this, following King and Jobling [9]: First, most surnames, in particular common surnames, had several independent origin families during the period that surname usage became established, and therefore their Y-STR haplotypes are distinct as a consequence. For example, not all the "Smiths" in the population have the same paternal ancestor (at the time of surname establishment), so their Y-STR haplotypes will also vary. This is particularly true for surnames with

an occupational or patronymic origin, which make up the majority of the most common surnames in the British population, rather than regional surnames, which were often specific to small settlements and thus have a much stronger genetic component. Second, there is the concept of *Non-Patrilineal Transmissions (NPT)*, which refers to the introduction of non-paternal descendants into a surname group, for example by adoption, name change, inheriting the mother's surname, or paternity misattribution. NPT has been estimated to occur at a rate of 1-4% per generation. Third, there is the relatively high rate of Y-STR mutations, so even two patrilineally related males who share a surname may have different Y-STR haplotypes. For an intruder to recover the surname of a genome, it would be necessary for the genealogical database to have a Y-STR haplotype which belongs to someone from the same family origin, with the same surname, and without too many mutations. This decreases the probability of inferring surnames compared to an estimation based on our assumptions.

4. **No divergence assumption.** In common with most risk assessments of this sort we assume that data that are supposed to correspond do so. As all data processes carry a risk of errors (for example, contamination, or data entry errors) so all data processes that rely on this assumption therefore necessarily inflate estimates of risk.

5. **Motivation assumption.** We assume an intruder with a sufficiently compelling motivation to carry out such an attack exists. This greatly simplifies the calculation of risk which is otherwise reliant on an equation such as that of Marsh et al. [28]:

$$pr(idnetification) = pr(attempt) \times pr(identification|attempt)$$

where the first element of the right hand side is impossible to measure (but is almost certainly less than 1).

6. **Assumption that the payload is independent of the probability.** This assumption is related to the preceding one. The payload is likely to be related to the intruder's motivation. If it is given that a motivated intruder exists then the impact of the attack is likely to be unrelated to its risk. However if the payload motivates the intruder then it won't be independent.[13]

[13] A related point here is that the equation of risk with event probability is based on an immature model of risk. Mature models of risk consider both the likelihood of an event and its payload/impact.

The net effect of this set of assumptions is to inflate our estimates of risk. However, they do not do so by a fixed or reliable quantity (and indeed the effect is likely to change over time). We would therefore regard that the estimates reported herein are upper bounds.

5 CONCLUSION

Genomic research and its applications are progressing rapidly; whilst producing much valuable knowledge the genomics revolution also raises serious privacy concerns. Recently, it has been shown that male genomes are vulnerable to re-identification by surname inference. By modelling the distribution of surname frequencies in Britain, we have demonstrated that even though it is possible to infer the surname associated with a personal genome, the probability of such an inference is significantly dependant on the size (and availability) of external genealogical databases that include Y-chromosome haplotypes and the associated surnames.

We observe that common surnames are less informative and are thus less likely to lead to re-identification, whilst rare surnames may be very informative. Considering the impact of recovering different surnames on the overall risk of re-identification, we illustrate that the per-record risk of re-identification via surname inference is relatively low when the size of the databases containing Y-chromosome-surname haplotypes is not too large. This emphasises that it is crucial to re-consider policies concerning these databases, their availability, and the number of records they can contain.

Further, this work outlines that there are several other factors that affect the risk of re-identifying personal genomes by surname inference. In particular, additional non-genomic information that may be attached to the genomic data is critical in determining whether an end-to-end re-identification is possible. The quantity and level of detail of such additional data is thus important. We demonstrate that with large databases the simple act of adding age and region starts to move some groups of surnames into high risk categories. We noted that it was particularly risky to include the exact age of an unusually old person (relative to the general population).

Finally, we note that the assumptions that we made in constructing our intruder scenarios cause our risk estimates to be inflated and that therefore the estimates should be regarded as upper bounds. Overall, we argue that in a British context, these upper bounds on the per-

record risk of re-identifying personal genomes by surname inference are – at present – low. However, if personal genomes are to be shared we should be very careful with what metadata are associated with the shared genomes and we must also be alert to the development of large (both in number of entries and of markers) genetic genealogical databases. In general, the paper demonstrates that the risk of a privacy breach of genomic data via this route is strongly dependant on the particular properties of each genomic dataset and therefore should be assessed on a case by case basis.

ACKNOWLEDGEMENTS

I would like to thank the Economic and Social Research Council (ESRC) [ES/J500094/1] for supporting this work, the ESRC's Consumer Data Research Centre [ES/L011840/1] for providing us with some of the data for the work in chapter 6 and Prof. Natalie Shlomo and Dr. Catherine Heeney for their comments on an early draft of that chapter.

REFERENCES

1. KREBS, J. E., LEWIN, B. & KILPATRICK, S. T. 2013. *Lewin's Genes XI*, Jones & Bartlett Learning.
2. HARTL, D. L., CLARK, A. G. & CLARK, A. G. 1997. *Principles of population genetics*, Sinauer associates Sunderland.
3. GUSMAO, L., SÁNCHEZ- DIZ, P., CALAFELL, F., MARTIN, P., ALONSO, C., ÁLVAREZ- FERNÁNDEZ, F., ALVES, C., BORJAS- FAJARDO, L., BOZZO, W. & BRAVO, M. 2005. Mutation rates at Y chromosome specific microsatellites. *Human mutation,* 26**,** 520-528.
4. HEYER, E., PUYMIRAT, J., DIELTJES, P., BAKKER, E. & DE KNIJFF, P. 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human Molecular Genetics,* 6**,** 799-803.
5. KING, T. E. & JOBLING, M. A. 2009. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends in Genetics,* 25**,** 351-360.
6. HEENEY, C., HAWKINS, N., DE VRIES, J., BODDINGTON, P. & KAYE, J. 2011. Assessing the Privacy Risks of Data Sharing in Genomics. *Public Health Genomics,* 14**,** 17-25.
7. GITSCHIER, J. 2009. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *The American Journal of Human Genetics,* 84**,** 251-258.
8. SYKES, B. & IRVEN, C. 2000. Surnames and the Y chromosome. *The American Journal of Human Genetics,* 66**,** 1417-1419.
9. KING, T. E. & JOBLING, M. A. 2009. Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. *Molecular Biology and Evolution,* 26**,** 1093-1102.
10. CONGIU, A., ANAGNOSTOU, P., MILIA, N., CAPOCASA, M., MONTINARO, F. & DESTRO BISOL, G. 2012. Online databases for mtDNA and Y chromosome polymorphisms in human populations. *J Anthropol Sci,* 90**,** 201-215.
11. GYMREK, M., MCGUIRE, A. L., GOLAN, D., HALPERIN, E. & ERLICH, Y. 2013. Identifying personal genomes by surname inference. *Science,* 339**,** 321-324.
12. LUNSHOF, J. E., CHADWICK, R., VORHAUS, D. B. & CHURCH, G. M. 2008. From genetic privacy to open consent. *Nature Reviews Genetics,* 9**,** 406-411.
13. HUNT, J. 2016. *What's the Difference Between Great Britain and the UK?* [Online]. Available: http://mentalfloss.com/article/85686/whats-difference-between-great-britain-and-uk [Accessed 2017].
14. ELLIOT, M. & DALE, A. 1999. Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics,* 14**,** 6-10.
15. JOLY, Y., NGUENG FEZE, I. & SIMARD, J. 2013. Genetic discrimination and life insurance: a systematic review of the evidence. *BMC Medicine,* 11**,** 25.
16. ROSSI, P. 2013. Surname distribution in population genetics and in statistical physics. *Physics of Life Reviews,* 10**,** 395-415.
17. FOX, W. R. & LASKER, G. W. 1983. The distribution of surname frequencies. *International Statistical Review/Revue Internationale de Statistique***,** 81-87.
18. *Surnames of England and Wales - the ONS list - How common (or rare) is your surname?* [Online]. Available: http://www.taliesin-arlein.net/names/search.php [Accessed 2017].
19. BARRAI, I., RODRIGUEZ-LARRALDE, A., MAMOLINI, E., MANNI, F. & SCAPOLI, C. 2001. Isonymy structure of USA population. *American Journal of Physical Anthropology,* 114**,** 109-123.
20. CHESHIRE, J. A., LONGLEY, P. A. & SINGLETON, A. D. 2010. The surname regions of Great Britain. *Journal of Maps,* 6**,** 401-409.
21. CHESHIRE, J. A., MATEOS, P. & LONGLEY, P. A. 2009. Family names as indicators of Britain's changing regional geography.
22. *official labour market statistics* [Online]. Available: https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp?theme=26&subgrp=2001+Census [Accessed 2017].
23. ELLIOT, M., MACKEY, E., O'HARA, K. & TUDOR, C. 2016. *The Anonymisation Decision Making-Framework,* Manchester, UKAN Publications.
24. *ysearch* [Online]. Available: www.ysearch.org [Accessed 2017].
25. *DNA fingerprint* [Online]. Available: http://www.dna-fingerprint.com/ [Accessed 2017].
26. *oxford ancestors* [Online]. Available: http://www.oxfordancestors.com/ [Accessed 2017].
27. GYMREK, M., MCGUIRE, A. L., GOLAN, D., HALPERIN, E. & ERLICH, Y. 2013. Identifying personal genomes by surname inference (supplementary materials). *Science,* 339**,** 321-324.

28. MARSH, C., SKINNER, C., ARBER, S., PENHALE, B., OPENSHAW, S., HOBCRAFT, J., LIEVESLEY, D. & WALFORD, N. 1991. The Case for Samples of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society. Series A (Statistics in Society),* 154**,** 305-340.