Appying machine learning methods to a mitochondrial DNA dataset to guess ethnicity

Vassilis Polychronopoulos

December 11, 2012

Abstract

We use a mitochondrial DNA (mtDNA) dataset provided by the FBI to build classifiers that can guess the ethnicity of the bearer of an mtDNA sample, using a number of different algorithms. Mitochondrial DNA is the genetic material that is found in mitochondria, organelles that reside inside the cells. The DNA found in mitochondria is relatively small in size but is highly informative and is therefore a potential good predictor for ethnicity. The majority of previous work on inferring ethnicity from genetic material has focused on autosomal DNA (the chromosomal DNA excluding the sex chromosomes), however, these methods require large amounts of genetic material, whereas mtDNA is small and abundant. Our results are, in the best case, qualitatively similar to the ones of the most recently reported work in the literature, while we obtain a slight improvement in macro-accuracy using SVMs without dimensionality reduction.

Introduction

The problem we address is that of using samples of the mitochondrial DNA of a person to guess the ethnicity to which the person belongs. Efficiently solving this problem can have numerous applications, such as in forensics. Knowledge gained from such studies can also enhance our understanding of the mechanisms of human population genetics and illuminate the patterns of variation in human genetic material.
There are several properties that make mitochondrial DNA, denoted mtDNA, unique and valuable to genetic studies. Mitochondial DNA is contained in mitochondria, organelles that are found in cells and play a crucial role in the production of cellular energy. Each person carries the mtDNA of his or her mother, and when there are no mutations, the mtDNA of a

person is identical to the female parent's mtDNA as it was inherited through the ovum. The male parent's mtDNA is not in any way transmitted to offsprings. This unique property of matrilineal inheritance gives it a special role in human population genetics and has opened up new horizons in our understanding of migrations of prehistoric human populations.
Population groups that share a specific mutation in their mtDNA, are said to belong to the same *haplogroup* and are therefore matrilineally descended from the same prehistoric woman. Tens of different haplogroups have been identified, corresponding to specific mutations in the mtDNA, and the approximate time of the mutations has been estimated. This method can categorize each individual into a haplogroup, however, an individual may bear a number of other mutations apart from the defining mutations of his or her haplogroup. Ultimately, as explained in [9], all members of the human species can trace back their matrilineal lineage to a specific prehistoric woman, called Mitochondrial Eve, that lived around 200,000 years BP (Before Present).
Mitochondrial DNA has a relatively small length but is very informative. In particular, two regions of mtDNA are mainly used in population genetics analyses, Hyper Variable Region 1 and 2 (abbreviated as HVR1 and HVR2), which are highly polymorphic. Genetic information in these two regions appears to have little or no operational role to the functionalities of the organism, and therefore the mutations accumulate there in a constant pace through the generations without affecting the individuals' survival. Knowing the average rate of change in that area, it is easy to estimate the time of separation of two different samples.
Among other achievements, human mitochondrial genetics has proved the out-of-Africa theory for the spread of the human species on the planet, as opposed to the multi-regional hypothesis. The theory states that the human species Homo sapiens, to which all living humans belong, evolved in its current anatomical form in Africa and spread from there to all continents in a number of migrations, with no or minimal interbreeding with other human species that co-existed with Homo sapiens in that period.
Unsurprisingly, specific haplogroups are found in different percentages in different parts of the planet. Variation in human mtDNA tends to increase in indigenous populations, while it decreases when a sample from an indigenous population emigrates. Thus, the entire line of migrations can be traced back all the way to east Africa for all populations.
Even though haplogroups are strongly associated with specific regions, haplogroups alone cannot be used to infer ethnicity. The reason for that is that most of the haplogroups can be found in different ethnicities, albeit with varying percentages. This is due to the fact that most haplogroups are significantly old (some having estimated age of more than 150,000 years) and a number of prehistoric and historic migrations that have followed those mutations have introduced the haplogroups to most populations of the planet. Therefore, the coarse haplogroup categorization alone is not reliable to guess ethnicity in most cases.
Apart from that, though matrilineal inheritance is unique and useful, in environments where a lot of individuals have a diverse family background, with ancestors from different parts of the planet, as is the norm in the USA, matrilineal inheritance of mtDNA may be an obstacle in using it as a reliable predictor. Assigning a specific ethnicity to an individual is mainly phenotypical, that is, based on the appearance and traits of the person. For example, a

person whose ancestors are all Caucasians save for an Asian maternal grandmother, will carry the grandmother's mtDNA which is most likely to be typically Asian, and therefore a classifier would most likely classify this person as Asian, despite the person being of Caucasian phenotype.
These facts could render the whole effort of building classifiers solely based on mtDNA futile, however, as shown in previous work [6], traditional machine learning methods can be used to build classifiers using solely mtDNA with satisfactory results.
We attacked the problem using the same dataset as in [6], evaluating some algorithms that had not been tested before, namely, Naive Bayes, Decision Trees and kNN. We excluded LDA and QDA as they had already been evaluated in [6] with poor results. Evaluation of the SVMs in the dataset without using dimensionality reduction proved computationally feasible, so we also evaluated SVMs without dimensionality reduction which gave qualitatively similar results to the previous work, but with a slight increase in macro-accuracy.

Previous work

We build on already existing work in the literature, and apply machine learning techniques to solve the problem of inferring ethnicity from mitochondrial DNA sequences, as opposed to approaches that rely on autosomal DNA (the normal chromosomal nuclear DNA, excluding the sex chromosomes). Studies [8] [7] have demonstrated that autosomal DNA can be used to infer ethnicity with excellent accuracy, however, a very large amount of genetic material is required for reliable predictions. A *locus* is the location of a gene in a chromosome, usually having a large length. Tens of autosomal loci are needed to achieve satisfactory predictions. Shriver et al. in [8] determine specific autosomal loci and their correlation to specific ethnic groups. They apply a log-likelihood analysis and use it to provide ethnicity estimation. In [7] Philipps et al. used polymorphisms in the autosomal DNA sequence and built classifiers that can predict a sample's membership into one of three coarse ethnicities, Caucasian, African and East Asian.
While these models are able to predict ethnicity with excellent accuracy, the main problem of using autosomal DNA for prediction is that in many cases it is not possible to recover long sequences of autosomal DNA, because of small samples. Autosomal DNA is contained in a single copy in the nucleus of each cell, while mtDNA is contained in multiple copies in each cell. Furthermore, the autosomal loci tend to survive less and degrade faster. On the contrary, mtDNA is easily recoverable even from minute samples due to its high-copy number. Thus, there have been attempts in the literature to infer ethnicity using solely mtDNA samples.
In [4], quadratic discriminant analysis is used to classify samples of mtDNA. However, this study is limited in scope while the results are successful only in classifying Caucasian and South East Africans. The main reference of our work is the study in [6] where a large dataset is used to build models using various traditional machine learning methods to guess ethnicity. In [6], the authors report results using Support Vector Machines (SVM), 1NN, linear and

quadratic discriminant analysis (LDA and QDA). For all methods except 1NN, the authors use principal component analysis (PCA) for dimensionality reduction. The method reduces the n features to k, choosing the k features with the largest sample variance. 1NN, though used without dimensionality reduction, generally performs better than both PCA-LDA and PCA-QDA, while PCA-SVM outperforms all the rest. However, for the African, Asian and Hispanic labels, the accuracy is less than 90% in the cross-validation analysis, even with the best method. In particular, the prediction for Hispanics is always lower than 75%. This may be due to the lack of data, or the use of inefficient algorithms, or it may have to do with the inherent limitations of the task. Hispanics are not a well defined ethnicity, generally regarded as a mixture of Caucasians, Africans and Native Americans, and this is perhaps the reason that the rate of accurate prediction is consistently lower that the other ethnicities.
Our project's goal is to examine alternative machine learning algoriths and report their performance, as well as variations of the algorithms that are used in [6]. We tested kNN for k=3, k=5 and k=21 allowing us to make some interesting observations, and evaluated Naive Bayes and Decision Trees. We verified the results using dimensionality reduction for SVMs, and evaluated it without dimensionality reduction as it proved computationally feasible.

Methodology

The dataset of our project has the same source as the one that Lee et al. use in [6]. It is derived from the forensic and published mtDNA samples from the mtDNA population database of FBI [5], available to the public for research purposes.
The samples of the *forensic* dataset come from the investigation of cases, while the *published* dataset comes from other sources. The dataset contains labels for a large number of ethnic groups. We consider only the samples that are members of four coarse ethnic groups, namely, Caucasian, African, Asian and Hispanic. For that, we need to filter the dataset accordingly. These four coarse ethnicities will be the four labels of our classification problem.
The forensic dataset contains 1,674 Caucasian, 1,305 African, 761 Asian and 686 Hispanic samples, while the published dataset comprises 2,807 Caucasian, 254 African and 915 Asian samples (no Hispanics). As in [6] we first built and evaluated our models using cross-validation analysis of the forensic dataset, and later used the published dataset as an independent test set for evaluation.
A DNA sequence is a sequence of nucleotides and is written using the alphabet $\{C, T, G, A\}$, where each letter of the alphabet represents the nucleotides cytosine, thymine, guanine and adenine respectively, also called *bases*. The database contains the mtDNA samples in a codified way, which is the standard way of representing mtDNA sequences. In particular, the samples are given as a list of polymorphic changes compared to the revised Cambridge Reference Sequence (rCRS)[2]. The rCRS is the complete mtDNA sequence of a Caucasian individual that was sampled in the 70s and is generally used as a point of reference. Thus, the rCRS itself is represented by an empty polymorphism list. A difference with respect to the rCRS may be either a substitution, an insertion or a deletion of one or more nucleotides.

To give an example, a mtDNA sequence that substitutes position 16298 of rCRS with base C (cytosine) will contain polymorphism notation '16298C' in its polymorphism list. An insertion of 4 thymine bases at position 309 is notated '309.4T' while a deletion of the rCRS nucleotide at position 291 is denoted '291D'. Thus, these 3 forms of differences with respect to the rCRS can uniquely identify any mtDNA sequence.
The database is in Microsoft Access form and is provided together with a software tool that allows the user to make a class of queries to the database. In particular, the user can give a polymorphism and search for its existence in the database, or can make searches for a specific sample. This is quite restrictive and that tool did not prove very useful in deriving the feature vectors from the database.
We accessed the .mdb file of the database for each of the two sets (forensic, published) using a software tool provided free for the Ubuntu operating system, the *MDB Viewer*. We extracted the polymorphism lists as comma seperated values, activating the relevant option to create a text file. A typical line of the text file we extracted from the mdb file is the following:

```
"GRC.CAU.000005","4","1","Blood","07/31/2001 12:00:00 AM",
"07/04/1776 12:00:00 AM","RS16024","RF16569","RS1","RF576","249G","263G",
"309.1C","315.1C","","","","","","","","","","","","","","","","","","",""
,"","","","","","","","","","","","","","","","","","","","","","","",""
```

The first field is the id of the sample, specifying that it is a sample of a Caucasian from Greece. The two other fields are information on the database indicating it comes from the forensic dataset and the 'Blood' field indicates that the sequence derived from a blood sample. The first date is the date of addition to the database, while the other date is the date of modification (set to the 4th of July 1776 when there are no mofidications, obviously to honor the US independence day).
The two other fields indicate the start and stop point of the segment of Hyper Variable Region 1 that was sequenced. The full HVR1 spans from position 16024 to position 16569, so the specific sample contains the entire sequence of HVR1.
Likewise, the next two fields indicate the start and stop point of the segment of Hyper Variable Region 2 that was sequenced. HVR2 spans from position 1 to position 576, so again this sample contains the entire HVR2 sequence.
The rest of the fields contain all the polymorphisms of the sample with respect to the rCRS sequence.
One way to derive feature vectors from a set of mtDNA samples presented with the use of polymorphisms lists, is to identify all distinct polymorphisms that are present in the dataset, say n in number, and then create an n-dimensional binary vector. If a sample contains a polymorphism then the corresponding dimension in the binary vector is set to 1, otherwise it is set to 0. The feature vectors can be used immediately to train classification models, or one can apply dimensionality reduction first.
We wrote code in Java that creates the ARFF files suitable for use by the Weka packages. First, we created a program that creates an ARFF file for the samples in the forensic dataset

that contain the entire sequences of HVR1 and HVR2, to obtain the *full-length forensic dataset* as explained in [6]. This subset is significantly smaller than the entire forensic dataset containing 222 Caucasian, 415 Asian, 820 African and 447 Hispanic samples, that is, a total of 1904 samples.

The *trimmed forensic dataset* contains only polymorphisms and samples sequenced in the range 16024-16365 of HVR1. For the forensic dataset, it has the same cardinality as the entire dataset, as all sequences include the said range. The published dataset contains 2540 samples that belong to one of the four coarse ethnicities and whose sequences include said range. This dataset contains 1956 Caucasian, 134 African and 450 Asian samples. We observe that the trimmed published dataset, just as the entire published dataset, lacks Hispanic samples while Caucasians are over-represented and Africans are under-represented. For this reason, it may not be a very suitable test set, however, there are no other samples of this size available in the database. The Genographic Project [1] database claims to have more than 70,000 samples but it is not available publicly so we could not use it.

The forensic and published datasets as extracted from the MDB Viewer, and the Java programs that create the ARFF files can be found in this URL:
http://users.soe.ucsc.edu/∼vassilis/MLproject/.

We include a program that randomly permutes a file with the raw data. In the beginning it was not certain that we will use Weka that automatically randomizes the dataset for cross-validation, so we had created a program that creates random permutations of the initial text file with the raw data. All the other programs for obtaining the full-length and trimmed datasets are there. A user that wishes to use other files (of the same format) as input can change the string variables *trainfile* and *nameOfFile*. The first defines the file that is used to derive the polymorphisms, and the second the file that is used as input to create the final ARFF file. For the forensic dataset both these variables are set to 'foren' to read from that file, while for the published dataset variable *nameOfFile* must be set to be 'published'.

The ARFF files were initially created using numeric attributes where features take values 0 and 1 as described above. While running the experiments, we noticed that the Naive Bayes implementation assuming a real value computed the sample mean and variance to create a real value distribution of the features values. Apart from the fact that this is a mistaken approach, since the values are actually discrete, many sample means were precisely zero as some rare polymorphisms and the fact that the accuracy did not exceed 3 decimal digits made the sample mean to appear as zero. Thus, we made an extra pair of programs to create the ARFF files using discreet values 'yes' and 'no' for the feature vectors. However, for other methods, such as kNN, the features of the vectors should remain numeric.

As said above, one of the techniques that we tried is Naive Bayes. The Naive Bayesian assumption generally does not hold, since some polymorphisms are likely to be present together in members of particular ethnicities. However, the number of polymorphisms is large and some correlation among a minority of the features may not affect the performance significantly. Testing Naive Bayes is definitely interesting, since the naive assumption, though it strictly does not hold, may appromixate reality in a satisfactory manner.

For evaluating the nearest neighbor method, the authors of [6] use the number of mismatch

positions (i.e. the Hamming distance) as the distance metric. Weka did not have a ready implementation for that, but for the case of our feature vectors that can take values only in $\{0, 1\}$, the euclidean distance yields precisely the same results, as it is the square root of the Hamming distance, and the square root is a strictly monotonous function. We evaluated the k-nearest neighbor for k=3, k=5 and k=21.

Finally, decision tree methods are totally absent from the analysis in [6]. We tested the performance of Decision Trees using the traditional algorithm C4.5, as well as a different approach that involved boosting using the LogitBoost strategy (Weka package LADTree).

We used the ready libraries of weka to perform the dimensionality reduction with principal component analysis. However, k-NN was evaluated without dimensionality reduction as any reduction in dimensions would result in loss of information and change in distances. Moreover, decision trees by using the information entropy perform dimensionality reduction implicitly by splitting only based on the dimensions with the highest information gain. Also, SVMs (that are tested using dimensionality reduction in [6]) could be trained using the entire dataset in a reasonable time of computation so we tested SVMs in the unreduced dataset.

We used micro- and macro-accuracy to measure the overall performance of our classification algorithms. Micro-accuracy is defined as $\frac{\Sigma_{i=1}^{K} C_i}{\Sigma_{i=1}^{K} N_i}$ and macro-accuracy is defined as $\Sigma_{i=1}^{K} \frac{C_i}{N_i}$, where K is the number of classes in the dataset, N_i the number of samples in the dataset that belong to class i and C_i is the number of correctly classified samples of class i. Macro-accuracy is a measure of interest because we have imbalanced class sizes, and micro-accuracy tends to over-emphasize the performance of larger classes. In some cases, we also examined the confusion matrix to reach conclusions on the mostly confused ethnicities in the case of mistaken predictions.

Results

k-Nearest Neighbor

On the full-length forensic dataset, we evaluated 1NN. The said subset is rather small and it yielded a macro-accuracy of 79%, lower than all previously reported results save for PCA-QDA. We therefore abandoned the full-length forensic dataset in the rest of our study and used the trimmed forensic dataset, which contains less polymorphisms (417 instead of 780) but has a significantly larger number of instances. On the trimmed forensic dataset, 5-fold cross-validation gave the results that we see in Table 1 for kNN.

We observe that as k increases, both the micro- and macro-accuracy are lower. The accuracy for Hispanics increases and reaches good levels for k=5 but the accuracy for Asians falls very rapidly to reach 54.1% for k=21. This is most likely due to two reasons. First, as k increases the results tend to favor the bigger class, Caucasians, because as the neighbots become more it is more likely to encounter neighbors of the larger class. Secondly, the confusion may have to do with the inherent similarity of Asian samples to other samples that makes it harder to distinguish as the neighbors increase.

Ethnicity	1NN	3NN	5NN	21NN
Caucasian	93.9	94.7	94.7	94.8
Asian	83.3	76.6	70.2	54.1
African	86.0	86.2	85.7	86.4
Hispanic	72.9	74.9	75.2	70.1
Micro-Accuracy	86.49	86.01	84.79	81.51
Macro-Accuracy	84.02	83.10	81.43	76.37

Table 1: 5-fold cross validation for kNN on the trimmed forensic dataset

In Table 2 we can see the confusion matrix which indicates that the Asians are primarily confused with Caucasians and Hispanics. Africans distinguish better, but there is a degree of confusion with Caucasians. For all ethnicities except Asians the results are satisfactory, but the results overall indicate that there is not a benefit either in micro- nor macro-accuracy by using kNN instead of NN, and both these measures decrease in all cases. We also tested

Classified as	Caucasian	Asian	African	Hispanic
Caucasian	1587	19	31	37
Asian	193	412	30	126
African	130	27	1127	20
Hispanic	114	11	80	481

Table 2: Confusion matrix for k=21

kNN in the test set. Sometimes the cross-validation tends to overestimate performance, so it is useful to test using an independent dataset. The published dataset has the drawbacks that we highlighted in the methodology section, nevertheless, it is meaningful to test the algorithm on it. We observe the same pattern of accuracy decrease as k increases. As we

Ethnicity	1NN	3NN	5NN	21NN
Caucasian	94.1	93.8	93.7	94.8
Asian	60.4	54.9	52.9	37.8
African	87.3	87.3	85.1	79.1
Micro-Accuracy	87.79	86.57	85.98	83.85
Macro-Accuracy	80.62	78.67	77.2	70.56

Table 3: kNN on test set

see in Table 3, the results are good for Caucasians and Africans but bad for Asians. As explained in [6] this consistently lower performance for Asians using the test set has to do probably with the fact that the samples in the test set come from different areas of Asia

than those in the train set. In particular, a large percentage comes from Kazakhstan and Kyrgizstan, regions that have no representatives in the training set.

Naive Bayes

We evaluated Naive Bayes on the trimmed forensic dataset using 5-fold cross-validation. The results are shown in Table 4. We had poor results when using Naive Bayes, and we made the observation described in the methodology section. We thus tried to use a changed version of the dataset where the feature vectors are discrete values 'yes' and 'no'. The results are better but still poor with respect to results using other methods. We therefore did not evaluate Naive Bayes any further as it appears that the Naive Bayes assumption does not approximate the reality of the dataset in a satisfactory way.

Ethnicity	Naive Bayes(numeric)	Naive Bayes (discrete)
Caucasian	85.3	90.3
Asian	62.3	65
African	82.8	85.7
Hispanic	59.9	61.7
Micro-Accuracy	76.65	80.18
Macro-Accuracy	72.57	75.69

Table 4: 5-fold cross-validation with Naive-Bayes

Decision Trees

The results of 5-fold cross-validation on the training set (trimmed forensic dataset) using decision trees can be found in Table 5.
Algorithm C4.5 with pruning has good results. Training the decision tree using the LDA-Tree package, which applies the LogitBoost strategy, for 70 and 200 iterations proved very expensive computationally and it consumed more than 24 hours for the 200 iterations to terminate. Nevertheless, the results for the LogitBoost trees are less accurate than those using the traditional C4.5 algorithm.

In Table 6 we show the results for the test set, which are similar to the ones for the cross-validation. Asian predictions are again much lower, while the C4.5 with pruning achieves the best accuracy.

Ethnicity	C4.5 (pruned)	C4.5 (unpruned)	LogitBoost(70 it.)	LogitBoost (200 it.)
Caucasian	93.8	93.8	93	91.9
Asian	78.4	80.4	81.5	83
African	88	86.1	88.4	86.6
Hispanic	71.7	72.9	67.6	71.6
Micro-Accuracy	86.03	85.98	85.74	85.63
Macro-Accuracy	83.00	83.33	82.64	83.25

Table 5: Cross-validation for J48(C4.5) and LDATree

Ethnicity	C4.5 (pruned)	C4.5 (unpruned)	LogitBoost(70 it.)	LogitBoost (200 it.)
Caucasian	91.5	91.1	82.6	84.3
Asian	66.9	62.4	63.1	64.2
African	83.6	84.3	84.3	79.9
Micro-Accuracy	86.73	85.66	79.25	80.47
Macro-Accuracy	80.66	79.29	76.68	76.10

Table 6: Decision Trees on test set

SVMs

In [6], SVMs are declared the winner. We repeated the training and obtained the same results for cross-validation. We also tested SVMs without dimensionality reduction. We used radial basis kernel $K(x_1, x_2) = e^{-\gamma|x_1 - x_2|^2}$ and tuned it for $\gamma = 0.49$ and penalty constant $C = 1$. Cross validation on trimmed dataset, and evaluation on test set for SVMs, yielded the results we can see together in Table 7. SVMs are undoubtedly the winner. Omitting dimensionality

Ethnicity	PCA-SVM	SVM	SVM(test set)
Caucasian	94.62	94.3	92.4
Asian	84.76	84.9	66.4
African	89.81	89.9	90.3
Hispanic	72.59	72.9	-
Micro-Accuracy	88.10	88.06	87.67
Macro-Accuracy	85.45	85.49	83.04

Table 7: SVMs for both train and test set

reduction, yields the same results with a very slight increase in macroaccuracy, and the best

results for the test set compared to all other methods. To reproduce the results with Weka, one has to download the LibSVM library [3] and place the libsvm jar file in the classpath.

Conclusions

The results confirm the conclusions of [6], and declare SVMs to be the winners in predicting ethnicity based on mtDNA sequences. The macro-accuracy achieved for the cross-validation is larger than 85%, for the specific datasets, and corresponds to the accuracies achieved using more than 60 autosomal loci, that is, a substantially larger amount of autosomal genetic material. The results confirm the suitability of mtDNA as a reliable predictor of ethnicity.
Using kNN cannot improve results and leads to a small deterioration that increases with k. while Naive Bayes has low performance. Decision trees using the C4.5 algorithm achieve satisfactory results.
Future work could explore different datasets, because the datasets we used are relatively small compared to the size of existing mtDNA databases [1] but other datasets are not available to the public. Thus, access to abundant and more diverse data seems to be the biggest obstacle. The test set in particular is small and contains no Hispanic samples. Also, future studies can change the way the samples are divided into the 4 coarse ethnicities, as the Asian ethnicity tends to be confused with Caucasians and Hispanics. This is because Asia as a continent encompasses a diversity of phenotypes. We observed that samples from India in the training set are labelled as Caucasians, which can be problematic. An increase in the number of samples of the training set, along with the break-up of the Asian ethnicity into smaller regional sub-groups could refine the predictions.

References

[1] The genographic project. https://genographic.nationalgeographic.com/.

[2] R. M. Andrews, I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. Reanalysis and revision of the cambridge reference sequence for human mitochondrial dna. *Nat Genet*, 23(2):147, Oct 1999.

[3] L. Chang. Libsvm: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/#download/.

[4] T. Egeland, H. Bøvelstad, G. Storvik, and A. Salas. Inferring the most likely geographical origin of mtdna sequence profiles. *Ann Hum Genet*, 68(Pt 5):461–71, 2004.

[5] M. et al. The mtdna population database: An integrated software and database resource for forensic comparison. *Forensic Science Communications*, 4(2), 2002.

[6] C. Lee, I. Mandoiu, and C. Nelson. Inferring ethnicity from mitochondrial dna sequence. *BMC Proceedings*, 5(Suppl 2):S11, 2011.

[7] C. Phillips, A. Salas, J. J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez Dios, M. Calaza, M. C. de Cal, D. Ballard, M. V. Lareu, and Carracedo. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics*, 1(3-4):273–280, Dec. 2007.

[8] M. Shriver, M. Smith, L. Jin, A. Marcini, J. Akey, R. Deka, and R. Ferrell. Ethnic-affiliation estimation by use of population-specific dna markers. *American Journal of Human Genetics*, 60(4):957, 1997.

[9] B. Sykes. *Seven Daughters of Eve*. Corgi, Sept. 2004.