CSE 494/594 Algorithms in Computational Biology
Project Proposal

Alexandra Dent, Juan Garcia Mesa, Matthew Huff, Raj Shah

March 1, 2020

1 Introduction

Forensic genetics encompasses the use of genetics in order to settle legal disputes such as criminal investigations. To do so, biological material such as blood, hair, or semen is collected and its DNA compared against the DNA of some reference sample, such as that of a suspect. Certain highly polymorphic regions of DNA enable satisfactory discrimination between individuals. Since the 1980s, the regions of choice have been the loci of short tandem repeat (STR) sequences, which are repetitive regions 50-500 bp in length that are subject to high mutation rates due to replication slippage. The use of this technique for convoluted samples, which contain DNA from different sources, has been improved by pairing STR polymorphism with deletion/insertion polymorphism (DIP-STR) [6].

There exist, however, even more possible accessories for deconvolution. Already, in samples where nuclear DNA is dilute or degraded, the circular and robust mitochondrial DNA (mtDNA) is a popular choice for genetic identification due to single-nucleotide polymorphisms (SNPs) in its highly variable control region (16,023-576 bp) [5, 12]. Furthermore, the human leukocyte antigen (HLA) gene system is the most polymorphic in the human genome and the growth of second-generation sequencing may resolve crucial ambiguities in Sanger-based HLA typing [1].

2 Motivation

Courtrooms demand a high level of confidence for the admission of evidence. Current techniques give the likelihood of falsely matching an innocent suspect, otherwise known as the probability of a man not excluded $P(RMNE)$ between one in a billion to one in a thousand. A $P(RMNE)$ between 10^{-9} and 10^{-6} may affect the ability for evidence to be admissable in court. Isaacson, et. al [11]. Current methods, with the inclusion of a minor allele ratio of 0.01 as the only statistically viable results, yielded a $P(RMNE)$ within the guidelines of courtroom admissability requirements with a sample between 3-5 individuals and between 0 and 10 allele mismatches, and anything outside that was deemed to be nonviable for court admissability.

mtDNA, as a stable maternal line of DNA, reduces some of the stochastic errors of traditional DNA sequencing and allows for more significant analysis assumptions to be made in regards to matching members of the same family line. To the contrary, most HLA class I alleles are very rare, often falling within a single person or family, and the $HLA-A$, $HLA-B$, and $HLA-C$ regions encode highly polymorphic HLA class I molecules. Robinson, et al. [13]. There exists at this time a large HLA sequence database that contains over 10,000 alleles that could be utilized to narrow down matches. However, work by Robinson et. al. shows that by removing SNP and recombinant alleles reduces HLA class I variability to 11 $HLA-A$, 17 $HLA-B$, and 14 $HLA-C$ alleles that hold all significant variation in exons 2 and 3, which leads to the potential of fast detection of matching HLA alleles. The combination of matching both stable mtDNA sources and highly varied HLA alleles may yield to a lower probability of $P(RMNE)$ with higher sample numbers for forensic analysis.

3 Methods

Different statistical approaches and workflows exist to determine whether the DNA an individual is present in a genomic DNA mixture (e.g. [3], [2], [4]). However, there is a common pattern that all methods follow and that we intend to replicate.

First, the development of a robust theoretical framework for detecting the presence of an individual in a mixture sample is needed. Current probabilistic approaches are generally divided into deterministic versus Bayesian analysis to deconvolute a set with multiple contributors [10]. We intend to develop a novel framework by combining the power of SNPs from mtDNA, well established and widely used in the filed of forensic identification, together with HLA, both highly polymorphic genomic regions. Kuffel et al. [1] conclude that the application of HLA together with any standard STR-based analysis (e.g. DIP-STR) can show a significant increase in the probability of positive identification. We do not discard the use of DIP-STR to complement and strengthen our analysis.

Then, the following step is to test the limits of differentiating power of our framework through computational simulations. Hu et al. (2014) [10] provide a thorough review of software, including dynamic and web-based, that have been applied to produce accurate analysis of complex DNA profiles. These include LoComatioN [7], targeting low-copy DNA profiling in mixed DNA samples, an open-source R package developed by Forensim [8] that interprets and weights forensic DNA evidence, and LRmix software [9], which builds upon the previous Forensim package. Other approaches based on coalescent simulation of pairs of alleles have also been done and remain a possibility, although perhaps slightly out of the scope of this project.

Finally, if possible, demonstrate the validity of the simulation results with data from real world samples. Fortunately, there exist many online data bases that provide public data fitted for this validation. An thorough search is yet to be done to determine what database will be used, given all previous milestones are completed successfully and we are to perform validation with real data.

The innovation of the workflow of the project will be combining genomic polymorphism from both mtDNA and HLA. The latter has been understudied and has recently shown promising results when combined with other standard methods [1].

References

[1] Agnieszka Kuffel et al. "Human Leukocyte Antigen alleles as an aid to STR in complex forensic DNA samples". In: *Science & Justice* (2019).

[2] Nils Homer et al. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays". In: *PLoS genetics* 4.8 (2008).

[3] RG Cowell et al. "Analysis of forensic DNA mixtures with artefacts". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.1 (2015), pp. 1–48. DOI: `https://doi.org/10.1111/rssc.12071`.

[4] Samuel H. Vohr et al. "A method for positive forensic identification of samples from extremely low-coverage sequence data". In: *BMC Genomics* 16 (2015). DOI: `10.1186/s12864-015-2241-6`.

[5] António Amorim, Teresa Fernandes, and Nuno Taveira. "Mitochondrial DNA in human identification: a review". In: *PeerJ* 7 (Aug. 2019). 7314[PII], e7314–e7314. ISSN: 2167-8359. DOI: `10.7717/peerj.7314`. URL: `https://pubmed.ncbi.nlm.nih.gov/31428537`.

[6] Angel Carracedo and Paula Sänchez-Diz. "Chapter 20A Forensic genetics: From classical serological genetic markers to DNA polymorphisms analyzed by microarray technology". In: *Forensic Science*. Ed. by Maciej J. Bogusz. Vol. 2. Handbook of Analytical Separations. Elsevier Science B.V., 2000, pp. 695–706. DOI: `https://doi.org/10.1016/S1567-7192(00)80075-3`. URL: `http://www.sciencedirect.com/science/article/pii/S1567719200800753`.

[7] Peter Gill, Amanda Kirkham, and James Curran. "LoComatioN: a software tool for the analysis of low copy number DNA profiles". In: *Forensic Science International* 166.2-3 (2007), pp. 128–138.

[8] Hinda Haned. "Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics". In: *Forensic Science International: Genetics* 5.4 (2011), pp. 265–268.

[9] Hinda Haned and Peter Gill. “Analysis of complex DNA mixtures using the Forensim package”. In: *Forensic Science International: Genetics Supplement Series* 3.1 (2011), e79–e80.

[10] Na Hu et al. “Current developments in forensic interpretation of mixed DNA samples”. In: *Biomedical reports* 2.3 (2014), pp. 309–316.

[11] J. et al. Isaacson. “Robust detection of individual forensic profiles in DNA mixtures”. In: *Forensic Science International: Genetics* 14 (2014).

[12] Mafia Victoria Lareu and Antonio Salas. “Chapter 20B Mitochondrial DNA in forensic genetics”. In: *Forensic Science*. Ed. by Maciej J. Bogusz. Vol. 2. Handbook of Analytical Separations. Elsevier Science B.V., 2000, pp. 707–720. DOI: `https://doi.org/10.1016/S1567-7192(00)80076-5`. URL: `http://www.sciencedirect.com/science/article/pii/S1567719200800765`.

[13] J. et al. Robinson. “Distinguishing functional polymorphism from random variation in the sequences of $>$ 10,000 HLA-A, -B and -C alleles”. In: *PLoS Genetics* (2017).