CSE 494/594 Algorithms in Computational Biology
Project Proposal

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

March 1, 2020

1 Introduction

Topic description, background information and what's been done

- Forensic identification based on DNA
- mtDNA (Amorim et al.), HLA, DIP-STR (Kuffel et al.)

2 Motivation

Why is this project important:

- Need for improvements in the field of forensic identification

3 Methods

- Genome Analysis Toolkit Best Practices Pipeline pdf by Van der Auwera et al. (2013) [4].
- Innovation of our approach is combining mtDNA and HLA (DIP-STR method).

Different statistical approaches exist to determine whether the DNA an individual is present in a genomic DNA mixture (e.g. [1], [2], [5]). 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. We intend to develop a novel framework by combining the power of single nucleotide polymorphisms (SNPs) from mtDNA and HLA, both highly polymorphic. Kuffel et al. [3] conclude that the application of HLA together with any standard short tandem repeat (STR) based analysis (e.g. deletion/insertion polymorphism (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 simulation. (**Include more information or, if not clear yet, add TBD**)

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.

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 [3].

References

[1] RG et al. Cowell. “Analysis of forensic DNA mixtures with artefacts”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.1 (2015), pp. 1–48.

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

[3] Agnieszka Kuffel, Alexander Gray, and Niamh Nic Daeid. “Human Leukocyte Antigen alleles as an aid to STR in complex forensic DNA samples”. In: *Science & Justice* (2019).

[4] Geraldine A. et al. Van der Auwera. “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline”. In: *Current Protocols in Bioinformatics* (2013). DOI: 10.1002/0471250953.bi1110s43.

[5] Samuel H. et al. Vohr. “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.