Forensic Science International: Genetics 30 (2017) 93–105

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures

Samuel H. Vohr[a,*], Rachel Gordon[b], Jordan M. Eizenga[a], Henry A. Erlich[b], Cassandra D. Calloway[b,c], Richard E. Green[a]

[a] Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High St., Santa Cruz, CA 95064, USA
[b] Center for Genetics, Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr Way, Oakland, CA 94609, USA
[c] Forensic Science Graduate Program, University of California, Davis, 1 Shields Ave, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 25 January 2017
Received in revised form 5 May 2017
Accepted 26 May 2017
Available online 29 May 2017

Keywords:
Forensics
Mixtures
Mitochondrial DNA
Massively parallel sequencing
Next-generation sequencing
Haplogroups
Deconvolution
Mixemt

ABSTRACT

Massively parallel (next-generation) sequencing provides a powerful method to analyze DNA from many different sources, including degraded and trace samples. A common challenge, however, is that many forensic samples are often known or suspected mixtures of DNA from multiple individuals. Haploid lineage markers, such as mitochondrial (mt) DNA, are useful for analysis of mixtures because, unlike nuclear genetic markers, each individual contributes a single sequence to the mixture. Deconvolution of these mixtures into the constituent mitochondrial haplotypes is challenging as typical sequence read lengths are too short to reconstruct the distinct haplotypes completely. We present a powerful computational approach for determining the constituent haplotypes in massively parallel sequencing data from potentially mixed samples. At the heart of our approach is an expectation maximization based algorithm that co-estimates the overall mixture proportions and the source haplogroup for each read individually. This approach, implemented in the software package mixemt, correctly identifies haplogroups from mixed samples across a range of mixture proportions. Furthermore, our method can separate fragments in a mixed sample by the most likely originating contributor and generate reconstructions of the constituent haplotypes based on known patterns of mtDNA diversity.

1. Introduction

Mixtures of DNA from more than one individual are commonly found in forensic samples [1]. For example, a stain found at a crime scene may contain DNA from both a victim and a perpetrator. Mixed samples can also be the result of factors not directly related to a forensics case. A sample believed to be from a single individual may contain contaminating DNA from the surface from which it was collected or from handling prior to or during laboratory work. As many forensics samples represent a limited and invaluable resource, mixed samples, although notoriously difficult to interpret, often contain information essential to an investigation. Therefore, accurate interpretation of mixed DNA samples is a critical need as well as challenge for forensic analysis.

The interpretation of DNA from mixed samples has been the subject of research since the early days of DNA-based identification [2,3]. Many distinct approaches have been proposed for different genotyping technologies using various genomic loci. For short-tandem repeat profiles (STRs), methods have been proposed that examine electropherogram peaks associated with PCR amplification products of different lengths [4,5]. While these methods can detect whether a sample of DNA represents a mixture, it is not always possible to reconstruct the individual contributors from a mixed profile alone, as no linkage information exists between the genetic marker loci, which have been chosen to be statistically independent. In addition, determining the number of contributors to a mixture can be difficult and is often complicated by allelic dropout, especially in degraded or low-template DNA samples [6–8].

Mitochondrial DNA (mtDNA) represents a promising target locus to aid in the interpretation of mixed DNA samples. Unlike the nuclear genome, where each cell carries two distinct copies inherited from each parent, the mitochondrial genome, a haploid lineage marker, is only passed from mother to child. As a result, an individual is expected to contribute only a single haplotype to a mixture sample. Each cell contains hundreds or thousands of identical or nearly identical copies of the mitochondrial genome, making it a more tractable target for recovery and amplification in

* Corresponding author.
E-mail address: svohr@ucsc.edu (S.H. Vohr).

http://dx.doi.org/10.1016/j.fsigen.2017.05.007

trace and degraded samples [9]. The mitochondrial genome also does not undergo recombination during meiosis. As a result, patterns of variation persist in populations across time and geography. Human genetic variation within mtDNA is extremely well sampled and catalogued [10–12]. These data have yielded a comprehensive human mtDNA phylogeny with reconstructed mutation history and a detailed and standardized haplogroup nomenclature [13].

The highly polymorphic mtDNA hypervariable regions (HVRI, HVRII) have been used to analyze forensic and missing person case samples through PCR amplification and Sanger sequencing [14,15]. Through this approach, it is possible to observe the variants that are present in mixed samples. However, estimating individual contributions is difficult as trace electropherogram peak areas or heights do not necessarily correlate with mixture proportions [4,16,17]. Moreover, Sanger sequence data from PCR amplification of mixtures alone cannot be used to reconstruct mtDNA haplotypes from each contributor as the primary signal describes the bases present at each reference position but no linkage information between the variants.

Recently, studies have focused on applying massively parallel sequencing technologies, also known as next-generation sequencing, to sequence PCR amplification products from mtDNA hypervariable regions in a high-throughput, clonal approach [18–20]. Unlike Sanger sequencing, massively parallel sequencing generates a survey of individual DNA molecules in the sample including the linkage information they contain. Therefore, novel and low-frequency variants can be easily detected, along with the haplotype background on which they appear. Additionally, it is possible to estimate quantitatively the contribution of each mtDNA haplotype that is present in a mixture sample by counting sequence reads. In addition to the sequencing of PCR products, mitochondrial DNA can be sequenced from shotgun libraries and assembled into a complete mtDNA sequence. This approach allows for the molecules to be observed in a way that more closely resembles how they appear in the input sample including short fragments that are not easily recovered by amplification. Shotgun approaches have been used to sequence and reconstruct the mitochondrial genomes of Neandertals [21], ancient humans [22], and other species [23] from highly-fragmented and damaged DNA [24]. Target-capture enrichment strategies have also been employed with this approach to recover mtDNA from such samples more efficiently [25–29]. The use of whole mitochondrial genome sequencing has been explored for forensic applications with a focus on heavily-degraded samples where DNA fragments may be too short to be amplified efficiently [29–32].

Direct sequencing of the mtDNA molecules in a sample poses a unique set of data interpretation challenges that are not present when examining PCR amplified sequences from a single locus. When examining amplification products through massively parallel sequencing, each sequence represents the same genomic region and, in the case of mixed samples, each distinct haplotype is attributable to a unique contributor, with the exceptions of cases of heteroplasmy, PCR artifacts, sequencing errors, and matrilineally-related contributors [20]. In contrast, sequences obtained from a shotgun library are from random segments of the mtDNA genome and each may carry a different set of variant positions. As a result, shotgun fragments may not be directly comparable with each other and determining the contributing haplotypes is difficult. Complete mtDNA sequences can easily be reconstructed from shotgun sequences when samples contain DNA exclusively from a single individual. Mixed samples, however, are not easily interpreted. The human mtDNA locus is over 16 kilobases long while sequence read lengths of current massively parallel sequencers are limited to a few hundred base pairs. Moreover, the DNA recovered from many forensics samples is degraded and consists of short fragments. Since extended regions of the contributing haplotypes may be identical and long-range information is not available to link variants to the same haplotype, standard genome assembly strategies cannot reconstruct the complete constituent mtDNA genomes from mixed sequence data. Alternatively, the constituent haplotypes can often be reconstructed by detecting variable sites within the mixture and assigning variants to the major and minor haplotypes based on their frequency in the sample [33,34]. However, this approach can only be applied to mixtures of exactly two individuals with sufficiently distinct mixture proportions (e.g., 75% and 25%). Phylogenetic information, i.e., variants and haplotypes known from surveys of mitochondrial diversity, has been shown to be useful for detecting and interpreting mixed samples [35,36].

We present a fast and powerful approach for interpreting mtDNA sequence data from mixed or unmixed samples, implemented in the software package mixemt. We address two fundamental questions in mixture interpretation. First, how many individuals (contributors) are present in a potentially mixed sample and at what relative proportions? Second, what are the variants associated with each contributing haplotype? To answer these questions, we make use of the large catalog of defined mitochondrial haplogroups to identify the distinct haplotypes that contribute to the mixture. We then reconstruct the haplotypes of each contributor by assigning reads to the haplogroup from which it most likely originated. Through this strategy, reads carrying novel variants, which provide the most power to discriminate between individuals, are partitioned by contributor using common, well-described variants. We demonstrate with *in silico* and *in vitro* mixture samples that our method can reliably detect the haplogroups present in mixed samples of two and three individuals and estimate their relative mixture proportions. We also show that our method can recover variants, including novel mutations, for each contributor in mixtures of two individuals. While we describe and evaluate our method using shotgun sequence data and expect our approach to be most useful in the analysis of highly-degraded samples where PCR-based methods often fail, our approach can be applied to massively parallel sequencing of amplification products as well.

2. Materials and methods

2.1. Description of phylogenetic interpretation approach

In a mixture sample, each sequenced DNA fragment, i.e., a single or paired-end read, represents a partial observation of a complete mtDNA haplotype. Our goal is to infer from these data the haplogroups that are present in a mixture, their relative proportions and, ultimately, the sequence for each contributor to the extent possible. Exhaustively searching the space of all possible 2, 3, 4, etc. contributor mixtures and proportions to fit the observed data is computationally infeasible. A confounding factor is that individual fragments contain little information on their own. Many fragments overlap conserved regions where few variants occur in the population and often the variants that fall within a fragment are shared by several haplogroups. A robust approach must consider each fragment in the context of all other fragments in a mixture. We propose an approach in which a sample of sequenced fragments is treated as a mixture of all known mtDNA haplogroups. We employ the expectation maximization algorithm to co-estimate mixture proportions and the probabilities of each fragment originating from each haplogroup. We then reduce the pool of possible contributors by applying heuristic filters to identify the most likely haplogroup contributors and remove any spurious haplogroup signals. Finally, the fragments are assigned to