Forensic Science International: Genetics Supplement Series 4 (2013) e334–e335

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Deep Sequencing Provides Comprehensive Multiplex Capabilities

B. Budowle [a,b,*], D.H. Warshauer [a], S.B. Seo [a], J.L. King [a], C. Davis [a], B. LaRue [a]

[a] Institute of Applied Genetics, Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Fort Worth, TX, United States
[b] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form 4 October 2013
Accepted 5 October 2013

Keywords:
Sequencing, STR, SNP, mtDNA, Database, Alignment, Software

ABSTRACT

Massively parallel sequencing offers the potential to type a large battery of forensically, relevant markers of a large number of individuals simultaneously. This summary describes progress, on development of STR, SNP, and mtDNA marker typing systems and demonstrates that transition to this technology is promising.

1. Introduction

For the past few decades, sequencing has been performed primarily by Sanger sequencing [1]. However, the methodology is labor-intensive, has a relatively low throughput, and is costly on a per nucleotide basis. In contrast, massively parallel sequencing (MPS) provides a much higher throughput of specified targets at a substantial reduction in cost. The technology makes possible simultaneous analysis of a large battery of genetic markers, far exceeding the current capacity of 15–24 STRs of a fluorescent multiplex capillary electrophoresis (CE) system [2]. All forensically relevant STRs and human identity SNPs (comprising between 400 and 500 markers and much more) can be typed simultaneously. Barcoding enables a number of samples to be sequenced at the same time. In addition, sequencing of the entire mtDNA genome could be carried out in a single analysis.

A significant outcome of a large battery of markers is that more searches can be generated when comparing a profile with those housed in a DNA database. A standard set of core-STRs is essential for sharing data within and between databases. However, an unintended consequence has arisen. There can be a tendency to type evidence first with the core-loci, even if the evidence would be analyzed better with a different set of markers. Ideally, the state of the evidence should dictate what markers would be best suited for analysis. This constraint can be alleviated if reference samples were typed more comprehensively for autosomal, Y and X STRs, SNPs and mtDNA (or a reasonable subset of these markers). Then, forensic scientists could be more judicious in addressing the needs of an analysis. The ultimate outcome would be that more investigative leads could be developed and the value of national databases would be enhanced.

With its economies of scale, MPS can provide a system such that reference samples can be typed economically for a large battery of markers and eventually, if commercialized, could exceed a cost benefit compared with current costs for typing a modicum of autosomal STRs. This paper describes some initial work and findings with MPS and genetic marker typing.

2. Initial STR and SNP design and testing

A panel of forensically relevant markers was selected for multiplexing using the GAIIx™ and MiSeq™ platforms and associated chemistries (Illumina Inc., San Diego, CA). The set of markers comprised 31 autosomal STRs, 26 X-STRs, 29 Y-STRs, and 378 SNPs. This number of markers is modest by throughput capacity of the MiSeq (~8 billion bases) but large by multiplex CE standards. Library preparation of DNA was performed using the Illumina® TruSeq™ Custom Enrichment protocol (Illumina). The TruSeq™ library preparation chemistry was selected initially because there is no PCR amplification required. Custom probes were designed using the DesignStudio software (Illumina). Since this initial panel is designed for reference sample typing, the amount of template DNA (~1 μg) is not particularly limiting. However, for applications requiring greater sensitivity of detection, a PCR-based enrichment method should be sought (see below for example with mtDNA). After capture, paired-end sequencing was carried out on the GAIIx™ and MiSeq™ sequencing platforms (2 × 148 and 2 × 251, respectively).

All SNPs were typable and a concordance study with an alternate sequencing chemistry demonstrated correct typing with

* Corresponding author at: Institute of Applied Genetics, Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Fort Worth, TX, United States. Tel.: +1 8177352979.
E-mail address: bruce.budowle@unthsc.edu (B. Budowle).

1875-1768/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigss.2013.10.170

almost all common loci [3]. For evaluation of STR typing, success was based on comparison with results from standard CE typing. Of the 22 autosomal and 22 Y STRs analyzed using on the GAIIx™ platform, 21 and 19 were typed, respectively. There were no discordant types [4]. However, the D21S11, DYS389II, DYS439, and DYS448 loci did not yield results. The negative results were due to the shorter read length of the GAIIx™ platform not allowing for full coverage of the repeat region at these loci. Analysis on the MiSeq™ platform, which has a longer read length capability, provided typable and concordant results. The autosomal STR results are similar to those described by Bornman et al. [5]. Overall the initial results demonstrated that a large panel of markers can be typed by MPS, and the results will enable analysis for many more markers than can be analyzed simultaneously by CE.

To facilitate STR allele calling of MPS data and to be able to compare nominal allele results from CE-based STR typing, the software STRait Razor [4] was developed. The software is a Linux-based Perl script that designates alleles at STR loci based initially by identifying flanking sequences and then determining the length of the repeat sequence within these flanking regions. This software can interpret both single-end and paired-end FASTQ sequence data, handle repeat motifs ranging from simple to complex, and does not require a reference composed of extensive allelic sequence data. This software offers an improvement over existing MPS STR-calling software packages [6–8].

STRait Razor successfully called alleles at all STR loci that were fully sequenced. Moreover, the software provided additional value by allowing assessment of the impact of library preparation methods and read length on allele detection. STR alleles can only be detected in reads that encompass the complete repeat region of an allele, and the enrichment and library preparation method can impact allele detection. For example, the HaloPlex™ chemistry (Agilent, Santa Clara, CA) for library preparation relies on restriction enzyme cleavage [9] which, in turn, creates fragments with consistent start and end points. The limitation of the method is that, depending on the length of the allele in question and the position of the repeat region within the resulting cleaved fragment(s), it is possible for sequencing reads to be produced that only partially span the repeat region and its associated flanking sequences or the site may not be captured by the designed probes. If a repeat region is situated toward the beginning of a HaloPlex™ fragment, the allele is likely to be detected in one direction of a paired-end analysis. However, when the reads are too short and the fragment is sequenced from the opposite direction, then the repeat region is oriented toward the end of the read and may not be completely covered. This situation was observed in loci such as D7S820 and vWA, where the alleles were detected only in one set of paired-end reads and not the other. Some library preparation redesign and increased read length may overcome the truncated repeat region reads. Therefore, with HaloPlex a marker or two may be incompatible. Given that many more STRs can be typed by MPS and library preparation is critical for performance, one may consider giving up a couple of current "core" STR loci; a decision of little consequence, as many more STRs can be typed than are practically possible with CE technology. The overall practicality of laboratory flow should be a criterion for long-term functionality.

The TruSeq™ chemistry is less prone to HaloPlex™-specific cleavage site issues because DNA is fragmented randomly for a much more varied positioning of repeat regions within the resulting fragments. Therefore, it is likely that at least some reads will encompass the entire repeat region of an STR in the panel. However, the non-enzymatic random fragmentation employed by the TruSeq™ chemistry results in lower read counts for some alleles in comparison with HaloPlex™. This limitation may be the cause for undetected alleles in a sample at loci DYS439 and DYS448 following TruSeq™ preparation and GAIIx™ sequencing.

3. Initial mtDNA design and testing

To date, forensic analyses of mtDNA were restricted to the hypervariable regions residing in the D-loop because of time, cost and labor constraints. With the capacity of MPS, whole mitochondrial genome sequencing was attempted. Target enrichment was performed using long-PCR as described by Gunnarsdóttir et al. [10] and was followed by library preparation using the Nextera® XT library preparation method. This method employs tagmentation, requires as little as 1 ng of DNA, is less laborious, allowing preparation of a number of samples in a microplate format, and can be performed in a shorter time than that required for the TruSeq™ approach. Sequencing was performed on a MiSeq™. Data analyses were performed on FASTQ data via BWA [11], SAMtools [12], and GATK [13], and calls were confirmed utilizing IGV [14], and HaploGrep [15].

Up to 96 samples were prepared and analyzed simultaneously in a single run. Overall, sequence results were concordant with extant sequence data derived by alternate chemistries, demonstrating that sequencing whole mitochondrial genomes is no longer an obstacle for a non-genome center laboratory. However, to improve on workflow and make MPS of mtDNA practical for the application-oriented laboratory further emphasis will be necessary in the areas of heteroplasmy, strand bias, coverage, and particularly alignment issues.

While the MPS described herein can provide extensive data, available software tools were rather limited. Different software packages did not always provide the same results. Most of the discordant results were due to alignment and different conventions from what are common within the forensic science community (e.g. 5′ vs 3′ alignment conventions). To effectively determine the mtDNA sequences, several software packages (described above) and manual processing were used. Without well-developed software analysis, sequencing will continue to be tedious and time-consuming.

Conflict of interest

None.

Acknowledgements

This work was supported in part by award no. 2012-DN-BXK033, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice. The authors also would like to thank Illumina, Inc. for the support of this study. C.D. is employed by Illumina.

References

[1] F. Sanger, et al. Proc Natl. Acad. Sci. U. S. A. 74 (12) (1977) 5463–5467.
[2] D.R. Hares, Forensic Sci. Int. Genet. 6 (1) (2012) e52–e54.
[3] S. Seo, et al. Int. J. Leg. Med. (2013) (in press) PMID: 23736940.
[4] D.H. Warshauer, et al. Forensic Sci. Int. Genet. 7 (2013) 409–417.
[5] D.M. Bornman, et al. Biotechniques: Rapid Dispatches (2012) 1–6.
[6] M. Gymrek, et al. Genome Res. 22 (2012) 1154–1162.
[7] B. Langmead, et al. Genome Biol. 10 (2009) R25.
[8] S.L. Fordyce, et al. Biotechniques 51 (2011) 127–133.
[9] Agilent Technologies® HaloPlex™ Specifications: http://www.genomics.agilent.com/GenericB.aspx?pagetype=Custom&subpagetype=Custom&pageid=3081.
[10] E.D. Gunnarsdóttir, et al. Genome Res. 21 (1) (2011) 1–11.
[11] H. Li, R. Durbin, Bioinformatics 26 (5) (2010) 589–595.
[12] H. Li, et al. Bioinformatics 25 (16) (2009) 2078–2079.
[13] A. McKenna, et al. Genome Res. 20 (9) (2010) 1297–1303.
[14] J.T. Robinson, et al. Nat. Biotechnol. 29 (2011) 24–26.
[15] A. Kloss-Brandstätter, et al. Hum. Mutat. 32 (1) (2011) 25–32.