Removing Sequencer and PCR Artifacts for Forensic DNA Analysis on Massively Parallel Sequencing Platforms

Scott R Kennedy PhD[1*], Michael J Hipp BS[1]

[1]Department of Pathology, University of Washington, Seattle WA, 98195
*Correspondence should be sent to Scott R Kennedy, scottrk@uw.edu

INTRODUCTION

Because STR analysis depends on length variations in short poly-nucleotide repeats that are amplified using PCR, there are inherent limitations in this technology, such as spurious background peaks resulting from PCR stutter, co-migration, signal oversaturation, and machine noise[1,2]. Of particular concern are PCR derived stuttering artifacts, which arise from slippage of the DNA polymerase of the DNA template[3,4]. Damaged or degraded DNA is particularly prone to this form of error due to the prevalence of DNA adducts that cause erroneous base pairings and enzyme stalling[5]. While there are techniques that allow for the statistical exclusion of stutter peaks, DNA mixture samples, especially combined with DNA damage and template degradation, present a significant challenge[6].

The use of MPS in forensic DNA analysis offers numerous advantages over PCR-CE. However, the technology is not without its disadvantages. The most notable is that MPS protocols often use PCR during library construction, which, as with PCR-CE, has an associated stutter and base misincorporation rate[4], thus giving the appearance of a MAF in a putative DNA mixture. Furthermore, the ability to practically detect MAFs is limited to about 1-2% due to sequencing errors associated with various sequence contexts and base miscalls[7,8]. Damaged DNA is known to worsen this background[9]. While MPS offers numerous improvements over current methods, the field of forensic DNA analysis has profound consequences for both the victim and the accused, therefore it is imperative that the occurrence of false MAFs be eliminated.

A number of approaches have been employed to improve the accuracy of MPS. Removal of DNA damage with the addition of *in vitro* repair kits has been shown to reduce the number of false variant calls in PCR-CE[10,11]. Similar approaches have been shown to be effective in MPS[12]. However, not all mutagenic lesions are recognized by these enzymes, nor is the fidelity of repair perfect. Another approach that has gained significant traction is to take advantage of PCR duplicates arising from individual DNA fragments to form a consensus. Termed “molecular barcoding”, reads sharing unique random shear points or exogenously introduced random DNA sequences before or during PCR are grouped and the most prevalent sequence kept[13-15]. This approach allows for the removal of false MAFs introduced as a base miscall during sequencing. However, the majority of these approaches only barcode a single DNA strand. Thus, misincorporation errors occurring during the first round of PCR will be propagated to all daughter molecules and are unable to be removed by this approach. This issue is especially important for STR genotyping due to the extremely high rate of PCR errors at these types of loci[4]. The only sequencing technology currently capable of identifying PCR errors with high

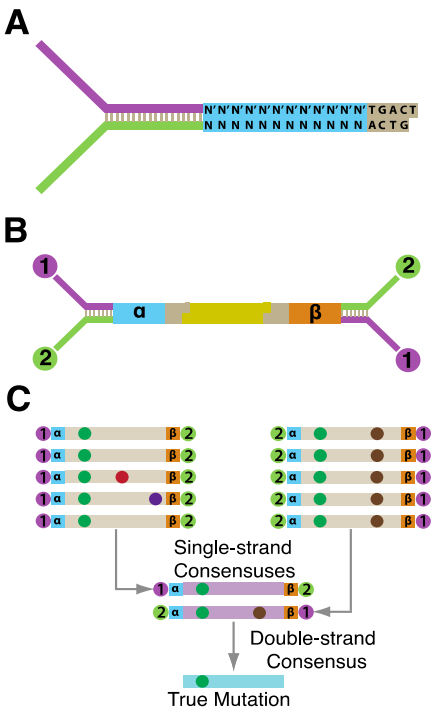

Fig. 1. Duplex Sequencing. (A) Adapter design containing the degenerate double-stranded barcode. **(B)** Ligation of adapters to sheared DNA (*yellow*) generates unique tags on each end (α and β). **(C)**PCR of the two strands produces two related but distinct products. Reads sharing a unique α and β are grouped into families. Mutations are of three types: sequencing mistakes or late arising PCR error (*blue or purple spots*); first round PCR errors (*brown spots*); true mutations (*green spots*).

confidence is Duplex Sequencing[16]. Duplex Sequencing extends the idea of molecular barcoding by using double-strand molecular barcodes to take advantage of the fact that the two strands of DNA contain complementary information. The double-stranded barcode allows for the comparison of both strands of a DNA molecule, whereby a variant is scored only when it is present in both strands. Briefly, sheared duplex DNA is ligated with a random, yet complementary, double- stranded nucleotide sequence (*i.e.* molecular barcode) (Fig. 1A). Following ligation, the individually labeled strands are PCR amplified such that there will be many duplicate "families" that share a common barcode sequence derived from each single parental strand of DNA (Fig. 1B). After sequencing, reads sharing the same barcode sequence are grouped together, and a consensus sequence for each position in the read is calculated for each family to create a "single-strand consensus sequence" (SSCS), with each SSCS being derived from an individual strand of DNA (Fig. 1C). This step filters out random sequencing or late arising PCR errors. Importantly, the SSCS does not filter out base misincorporations and stutter events that occur during the first round of PCR. To remove these errors, the complementary tags derived from the same duplex DNA among the SSCS reads are compared to each other (Fig. 1C). The base identity at each position in a read is kept in the final consensus if the two strands match perfectly at that position. Apparent mutations occurring in only one of the SSCS reads will be filtered out. Upon remapping of the "duplex consensus sequence" (DCS) reads back to the reference genome, any deviations from the reference genome are considered true mutations. Duplex Sequencing has been shown to be highly successful at removing both sequencer and PCR derived artifacts in mitochondrial and nuclear DNA[16-18]. However, these prior studies have focused on the detection of somatic point mutations and small (<5bp) insertions and deletions. We have recently tested the ability of DS to remove PCR stutter. Using our published protocols[17,19], we find that DS can essentially eliminate PCR stutter at STR loci.

METHODS

DNA Samples

DNA used in our studies was obtained from the Coriell archive of the 1000 Genome Project.

Duplex Sequencing

Duplex Sequencing was performed using a modified form of previously published protocols[17,19]. Briefly, we performed targeted genome fragmentation of human genomic DNA to selectively excise the CODIS20+PentaD and PentaE loci from the genome using the *S. pyogenes* Cas9 nuclease. The required guide RNAs (gRNA) (Integrated DNA Technologies) were designed to be specific for flanking regions close (<50bp) to each locus. 30nM gRNAs were complexed with the 30nM Cas9 nuclease following the manufacturer's recommended protocol and then incubated with 10ng of nuclear DNA overnight at 37°C. The reaction was then heat inactivated at 70°C for 10min.

After heat inactivation, AMPure XP Beads (Beckman Coulter, Brea, CA, USA) were used to remove off-target, un-digested high molecular weight DNA by combining the heat inactivated Cas9 digestion with a 0.5x volume ratio of beads. The beads were then separated from the solution with a magnet and the supernatant containing the targeted DNA fragment length was transferred into a new tube. This was followed by a standard AMPure 1.8x volume bead purification eluted into 50 µL of TE_{low} to exchange the buffer and remove small DNA contaminants. The fragmented DNA was then A-tailed and ligated using the NEBNext Ultra II DNA Library Prep Kit (NEB, Ipswich, MA) according to manufacturer's protocol. Duplex Sequencing adapters, described in Kennedy *et al.* [17], were obtained as a commercial adapter prototype synthesized externally through an arrangement with TwinStrand Biosciences. After ligation, adapter ligation and reaction reagents were removed by a 0.8X ratio AMPure Bead purification and eluted into 23 µL of nuclease free water.

PCR copies of every DNA strand in the sequencing library was amplified using KAPA KAPA HiFi HotStart Real-time PCR Master Mix with 2µM MWS13 and MWS20 PCR primers[17,20] following the manufacturer's recommended protocol. A 0.8X ratio AMPure Bead wash was performed to purify the amplified fragments and then eluted into 40µL of nuclease free water.

The post-PCR library consists of DNA fragments from all regions of the genome. Therefore, to enrich for the STR loci of interest, we performed targeted DNA hybridization capture using biotinylated IDT xGen Lockdown Probes (Integrated DNA Technologies, Coralville, IA) specific for the 120bp regions flanking both sides of each STR locus (*i.e.* two probes per locus). Hybridization capture was performed according to the IDT protocol, except for 3 modifications. First, we used blockers MWS60 and MSW61, which are specific to DS adapters, as described elsewhere[17]. Second, we used 75µl of Dynabeads M-270 Streptavidin beads instead of 100µl. Third, the post-capture PCR was performed with the KAPA Hi-Fi HotStart PCR kit (KAPA Biosystems, Woburn, MA, USA) using MWS13 and indexed primer MWS21 at a final concentration of 0.8 µM[17]. The PCR product was purified with a 0.8X AMPure Bead wash.

Samples were quantified using the Qubit dsDNA HS Assay Kit, diluted, and pooled for sequencing. The library was sequenced on the MiSeq Illumina platform using a v3 600 cycle kit (Illumina, San Diego, CA, USA) as specified by the manufacturer. For each sample, we allocated ~7-10% of a lane corresponding to ~2 million reads.

Data Processing

Data were processed using custom designed software, and genotype calls were

performed with a modified form of the HipSTR that is compatible with Duplex Sequencing data[21]. Alpha release of the software is available at https://github.com/fulcrumgenomics/fgstr. Stutter was quantified as described in [22].

ForenSeq Genotyping

1000 Genomes Project DNA samples were processed for the MiSeq FGx platform using Illumina's ForenSeq DNA Signature Prep Kit according to their protocols. Samples were sequenced on an Illumina FGx platform according to Illumina's recommend protocol. Sample genotypes were called using the Illumina ForenSeq UAS package with default settings.

PCR-CE Genotyping

1000 Genomes Project DNA samples were genotyped by PCR-CE by the Defense Forensic Science Center using the Promega 6C kit according to the manufacturer's recommended protocol with a ABI 3130 capillary electrophoresis instrument. Stutter was quantified as described in Brookes *et al.*[23]

Results

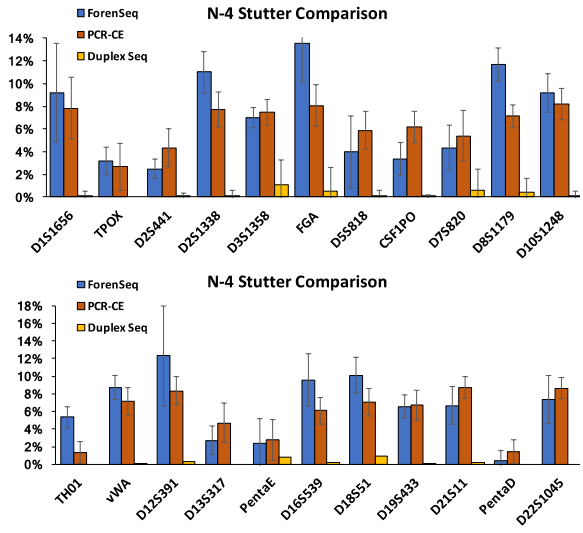

Fig. 2. Loci specific comparison of stutter. Comparison of N-4bp stutter frequency of the CODIS20+PentaD & PentaE loci using PCR-CE(*orange*), the Illumina ForenSeq platform(*blue*), and Duplex Sequecing(*yellow*). Duplex Sequencing (*black*) exhibits dramatically reduced stutter at all loci. (n=12 samples)

Previous reports have quantified the amount of stutter in both PCR-CE and MPS platforms [22-25]. The frequency of stutter typically ranges from between 5-10%, but can be >30%, depending on the STR length involved which prevents the detection of minor contributors below this threshold[24,25]. Because Duplex Sequencing is a MPS-based method designed to eliminate sequencer and PCR artifacts for low frequency point mutations, we wanted to determine if this method could provide similar improvements for STR loci.

We performed a modified Duplex Sequencing protocol that selectively excises the CODIS20+PentaD and PentaE loci from genomic DNA using CRISPR/Cas9 technology [19]. The use of CRISPR/Cas9 allows for the ability to create DNA fragments such that the entire sequencing read is able to traverse the STR repeat, thus ensuring that all reads are informative (Fig. 2)[26]. We performed this protocol on 10ng of genomic DNA samples from the 1000 Genomes Project and sequenced the resulting libraries on an Illumina MiSeq platform and analyzed our data using custom genotyping software. We observe an average stutter frequency of 0.19±0.80%, which is significantly below stutter frequencies reported for both PCR-CE and the Illumina ForenSeq platform[22]. To

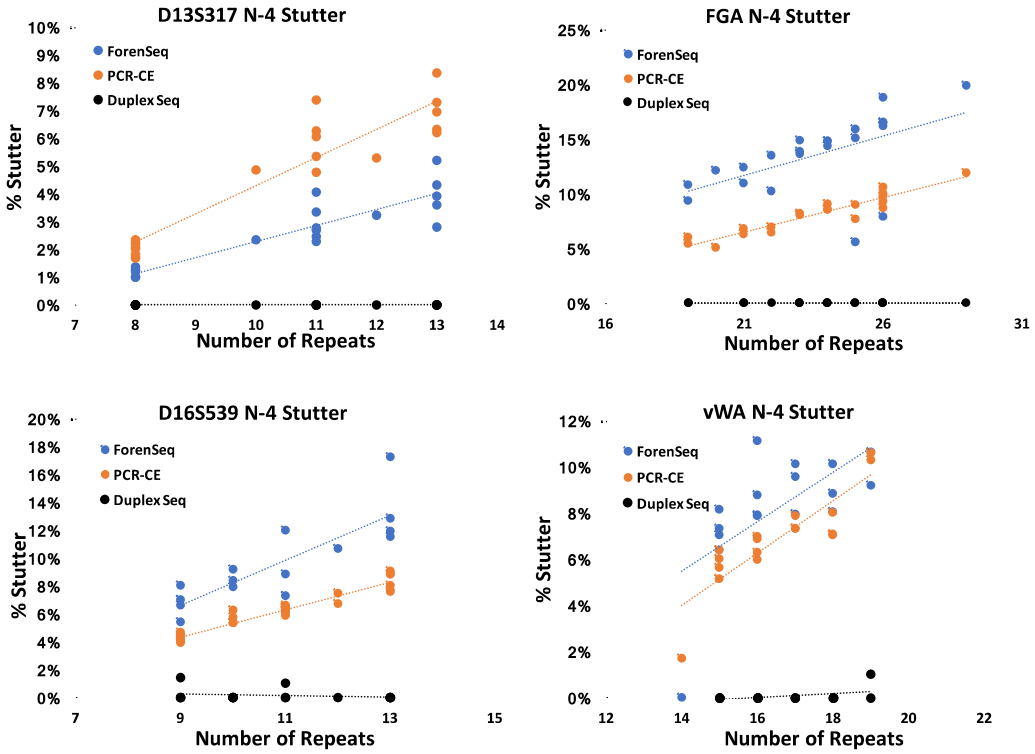

Fig. 3. N-4 stutter percentages as a function of STR length of four representative loci. PCR-CE (*orange*) and the Illumina ForenSeq platform (*blue*) exhibit increased levels of stutter with longer repeat length. Duplex Sequencing (*black*) shows no significant correlation. Each point represents the percent of N-4 signal from a single sample. Only whole repeat units are shown (n=12 samples).

confirm these prior findings, we performed genotyping analysis using PCR-CE and the Illumina ForenSeq kit on the same samples. Consistent with prior reports, we observed a stutter frequency of 5.86±2.84% and 7.21±4.36% for PCR-CE and the ForenSeq kit, respectively. These results show that Duplex Sequencing can effectively remove PCR stutter artifacts.

STR loci differ from each other in both length and sequence. These two factors are known to influence PCR stutter[23]. Therefore, we wanted to determine if different loci exhibited different error rates. As shown in Fig. 2, the stutter rates between the individual loci exhibited only small differences between them, none of which were significant. In contrast, both PCR-CE and the ForenSeq platform exhibited substantial variability between different loci. Additionally, with the exception of the PentaD and PentaE loci, the stutter frequency of Duplex Sequencing was significantly lower than both PCR-CE and the ForenSeq platform.

We next compared the frequency of stutter as a function of the number of repeat units in each genotyped locus. While the overall stutter frequencies varied between loci, PCR-CE and the Illumina ForenSeq platform exhibited a significant linear increase in the percentage of stutter events with STR length, consistent with previous results[23](Fig. 3). In contrast, the frequency of stutter events in Duplex Sequencing did not correspond to STR length, regardless of locus examined (Fig. 3 and data not shown). Together, our data show that Duplex Sequencing is able to substantially reduce PCR stutter by upwards of ~37-fold and that this reduction is independent of the locus examined or the length of the repeat.

Discussion

Current approaches to forensic DNA analysis almost entirely rely on capillary electrophoretic separation of PCR amplicons to identify length polymorphisms in short tandem repeat sequences. This type of analysis has proven to be extremely valuable since its introduction in the early 1990's[27]. Since that time, thousands of publications have introduced standardized protocols and validated their use in laboratories worldwide (Reviewed in Butler, JM[28]). While this approach has proven to be extremely successful, the technology has a number of drawbacks that limit its utility, mainly resulting from background signal arising from PCR stutter. This issue is especially important in samples with more than one contributor due to the difficulty in distinguishing the stutter alleles from genuine alleles[6].

The introduction of MPS systems has the potential to address several challenging issues in forensics analysis. For example, these platforms offer unparalleled capacity to allow for the simultaneous analysis of STRs and SNPs in nuclear and mtDNA. Furthermore, unlike PCR-CE, which simply reports the average genotype of an aggregate population of molecules, MPS technology digitally tabulates the full nucleotide sequence of many individual DNA fragments, thus offering the unique ability to detect MAFs within a heterogeneous DNA mixture[29]. Because forensic specimens comprising two or more contributors remains one of the most problematic issues in forensics, the impact of MPS on the field of forensics could be enormous.

While current MPS platforms offer a number of advantages over conventional PCR-CE approaches, current MPS sample preparation protocols rely on performing a multiplex PCR to enrich and isolate the forensic loci of interest. However, the act of performing the PCR enrichment introduces stutter artifacts which are then sequenced. Consequently, the ability to detect minor contributors is limited to the same extent as PCR-CE. Indeed, both published data, as well as our data reported here, show that stutter frequencies are similar between PCR-CE and Illumina based MPS workflows (Fig. 2)[22-24].

Here, we report the application of an ultra-accurate sequencing method, termed Duplex Sequencing, on forensically relevant STR loci. This method capitalizes on the biochemical redundancy of DNA to greatly lower the error rate of sequencing by allowing for the comparison of PCR duplicates derived from each strand of an original double-stranded DNA molecule. Consequently, DS is capable of removing artifacts arising during both sequencing and PCR steps performed during library preparation. Importantly, DS does not prevent the occurrence of PCR artifacts, such as stutter, from occurring. Instead, unique to DS, the method is able to detect when stutter is likely to have occurred and remove the read as artifactual.

Our data demonstrate that DS is able to dramatically reduce the level of PCR stutter compared to conventional PCR-CE and MPS based approaches. Furthermore, DS exhibits no significant correlation between stutter levels and STR length. Based on our data, the detection of MAFs is at or below 1%. This dramatic reduction in background levels opens up the possibility of detecting minor contributors in more challenging samples, such as is commonly encountered in sexual assault cases where the victim's DNA derived from vaginal epithelial cells is frequently at much higher levels than the perpetrator's.

The current Duplex Sequencing workflow can be employed with only modest

deviations from the normal Illumina library preparation workflow. Moreover, the concept of Duplex Sequencing could be generalized to essentially any sequencing platform, such as the Thermo-Fisher Ion Torrent. In sum, the compatibility with existing workflows and platforms, along with the ability of Duplex Sequencing to radically lower the stutter rates, offers a powerful MPS-based tool for use in the emerging field of forensic genomics.

References

1. Pompanon F, Bonin A, Bellemain E, Taberlet P. Genotyping errors: Causes, consequences and solutions. Nat Genet. 2005;6: 847–859.

2. Butler JM, Buel E, Crivellente F, McCord BR. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. Electrophoresis. 2004;25: 1397–1412.

3. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 1992;20: 211–215.

4. Shinde D, Lai Y, Sun F, Arnheim N. *Taq* DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$ microsatellites. 2003;31: 974–980.

5. Quach N, Goodman MF, Shibata D. *In vitro* mutation artifacts after formalin fixation and error prone translesion synthesis during PCR. BMC Clin Pathol. 2004;4: 1.

6. Budowle B, Onorato AJ, Callaghan TF, Manna Della A, Gross AM, Guerrieri RA, et al. Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Foren Sci. 54: 810–821.

7. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011;11: 759–769.

8. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39: e90.

9. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J Mol Diagn. 2013;15: 623–633.

10. Diegoli TM, Farr M, Cromartie C, Coble MD, Bille TW. An optimized protocol for forensic application of the PreCR™ Repair Mix to multiplex STR amplification of UV-damaged DNA. Foren Sci Int Genet. 2012;6: 498–503.

11. Robertson JM, Dineen SM, Scott KA, Lucyshyn J, Saeed M, Murphy DL, et al. Assessing PreCR™ repair enzymes for restoration of STR profiles from artificially degraded DNA for human identification. Foren Sci Int Genet. 2014;12: 168–180.

12. Do H, Dobrovic A. Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil-DNA glycosylase. Oncotarget. 2012;3: 546–558.

13. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J. Parallel, tag-directed assembly of locally derived short sequence reads. Nat Meth. 2010;7: 119–122.

14. Miner BE, Stöger RJ, Burden AF, Laird CD, Hansen RS. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Res. 2004;32: e135–e135.

15. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci USA. 2011;108: 9530–9535.

16. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci USA. 2012;109: 14508–14513.

17. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc. 2014;9: 2586–2606.

18. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet. 2013;9: e1003794.

19. Nachmanson D, Lian S, Schmidt EK, Hipp MJ, Baker KT, Zhang Y, et al. CRISPR-DS: An efficient, low DNA input method for ultra-accurate sequencing. BioRxiv 2017: 1–14. doi:10.1101/207027

20. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, et al. Sequencing small genomic targets with high efficiency and extreme accuracy. Nat Meth. 2015;12: 423–425.

21. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Meth. 2017;14: 590–592.

22. Aponte RA, Gettings KB, Duewer DL, Coble MD, Vallone PM. Sequence-based analysis of stutter at STR loci: Characterization and utility. Foren Sci Int Genet. 2015;5: e456-e458.

23. Brookes C, Bright J-A, Harbison S, Buckleton J. Characterising stutter in forensic STR multiplexes. Foren Sci Int Genet. 2011;6: 58-63.

24. Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, et al. Developmental validation of the MiSeq FGx Forensic Genomics System for targeted next generation sequencing in forensic DNA casework and database laboratories.

Foren Sci Int Genet. 2017;28: 52–70.

25. Zeng X, King J, Hermanson S, Patel J, Storts DR, Budowle B. An evaluation of the PowerSeq™ Auto System: A multiplex short tandem repeat marker kit compatible with massively parallel sequencing. Foren Sci Int Genet. 2015;19: 172–179.

26. Shin G, Grimes SM, Lee H, Lau BT, Xia LC, Ji HP. CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. Nat Commun. 2017;8: 14291.

27. Edwards A, Civitello A, Hammond HA, Caskey CT. DNA typing and genetic-mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet. 1991;49: 746–756.

28. Butler JM. Forensic DNA typing: Biology, technology, and genetics of STR markers. 2nd ed. New York: Elsevier Academic Press; 2005.

29. Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, et al. Quantification of rare allelic variants from pooled genomic DNA. Nat Meth. 2009;6: 263–265.