Purpose of the Project

The major goal of this project was the development of a highly parallel DNA sequencing method for improving the analysis of genetic mixtures. Many forensic DNA samples are composed of mixed DNA mixtures from different individuals. **Identifying each person who contributed to a DNA sample is very difficult if not impossible with current forensic methods for many samples obtained during a criminal investigation.** To accomplish this goal, we developed a new sequencing technology called **STR-Seq.** Based on a next generation sequencing **(NGS)** approach, our technology enables the parallel analysis of thousands of microsatellites, also known as sequence tandem repeats **(STRs)** for any given sample. This approach has numerous advantages compared to the current standard used by most forensic work; notably the current method involves using a molecular amplification process known as PCR which is analyzed with a system called capillary electrophoresis **(CE)**. In comparison STR-Seq does not use this older technology, but rather involves a highly specialized technology for DNA sequencing of forensic samples. As a result of multiple improvements, STR-Seq delineates multi-component complex DNA mixtures at low fractional representations (i.e. 1% of the total mixture). We completed the following: **(1)** Used the STR-Seq technology to conduct a large population studies of over 1,000 individuals; **(2)** Characterized 1,915 new microsatelites loci that included the discovery of a new class of STRs that are highly polymorphic, much more so than what is currently used by the FBI and others; **(3)** Developed a new method of measuring specific STR genotypes that are represented in a mixture; **(4)** Improved on the sequencing of low DNA quantities. **Overall, we completed all of the objectives of the project which included the publications of our results in the 2017 issue of Nature Communications.** We have included this article with our final report as part of the Appendix.

Overview

Microsatellites are multi-allelic genetic polymorphisms. They are composed of short tandem repeats **(STRs)** with individual sequence motifs that consist of mononucleotides, dinucleotides or larger units up to hexamers. To genotype microsatellites, NGS approaches and other STR assays rely on a limited number of PCR amplicons, typically in the tens. For this project, we developed and demonstrated the high performance of STR-Seq, a technology that analyzes over 2,000 STRs in parallel, and provides the accurate genotyping of microsatellites from DNA mixtures. An added feature that has not been described extensively is the use of microsatellite – single nucleotide polymorphism (SNP) haplotypes. STR-Seq provides can detect both of these genetic variations in tandem arrangement – this feature prove very useful for delineating genetic mixtures.

STR-Seq has multiple features that enable the analysis of complex genetic mixtures:

(1) STR-Seq employs *in vitro* CRISPR-Cas9 targeted fragmentation to produce specific DNA molecules covering the complete microsatellite sequence. By generating DNA molecules that completely cover the intact microsatellite, we dramatically improve the inference of genotypes from STRs.

(2) A PCR amplification-free library preparation provides sequence read data that directly reflect the nucleotide makeup of a single DNA molecule, doing so without other types of complex technologies such as unique molecular barcodes. In addition, we demonstrated that our PCR-free method reduces the errors generated from PCR amplification stutter which affects the current gold standard method used by the FBI and others.

(3) STR-selective primers enable massively parallel, targeted sequencing of large STR sets with much greater capacity than the current forensic genetic method.

Overall, STR-Seq has higher throughput, improved accuracy and provides a greater number of informative microsatellites compared to other methods including CE. **This work regarding STR-Seq is fully described in Nature Communications article [2].** This article is available as part of an open access arrangement, insuring a wide distribution to potential readers and availability for members of the forensic genetic community.

Results

Developing an improved STR-Seq assay with new markers and optimal performance

We **developed a new microsatellite DNA sequencing assay** as described in our Nature Communications article. The technology is show in **Figure 1**. This new and improved assay used those STRs that demonstrated an optimal performance as well as a new set of STR markers. Optimal performance was defined as those STRs that were accurately genotyped even when present in a low fraction of a genetic mixtures.

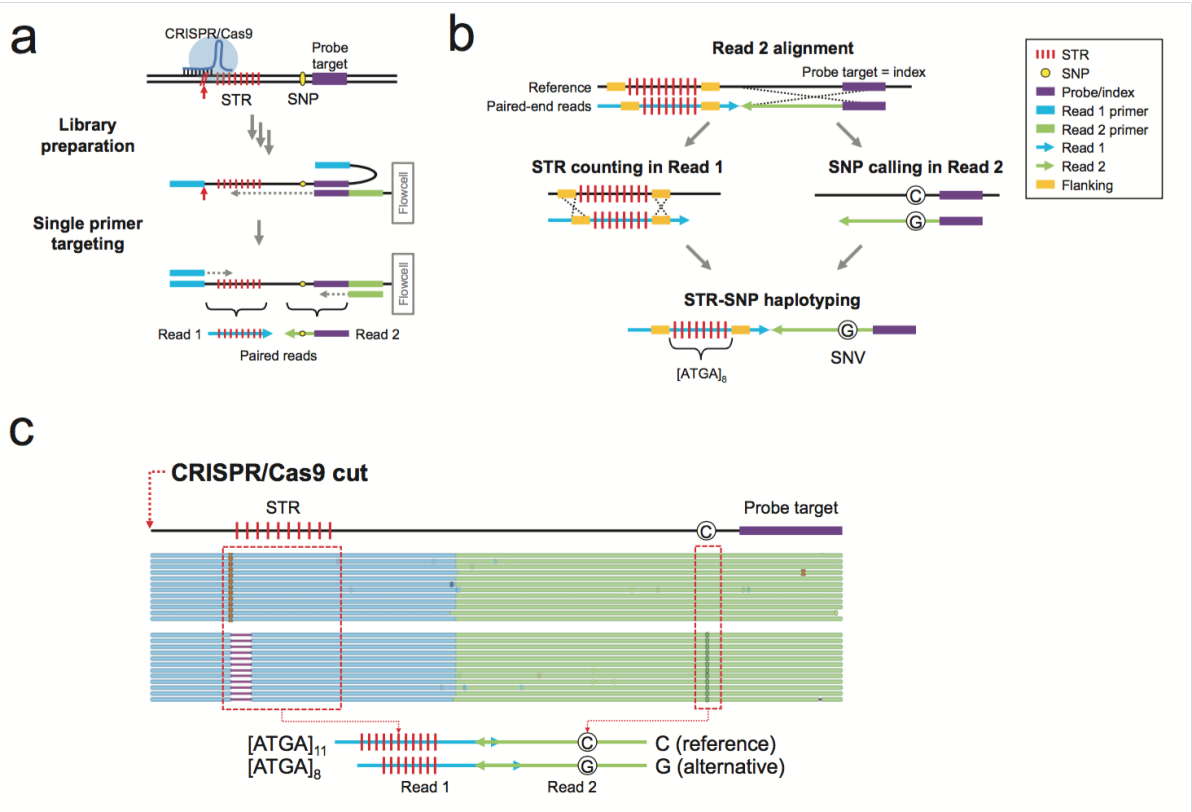


Figure 1. Overview of STR-Seq. ***(a)*** *Guide RNAs and primer probes were designed to target STRs and proximal SNPs. We target both plus and minus strands with only the plus strand targeting is illustrated. In the first step, Cas9 enzyme cleaves upstream of STR. The DNA libraries including the STR and SNP are target sequenced.* ***(b)*** *After initial alignment of Read 2 from any given paired-end set, we use the primer probe sequence derived from Read 2 as an index tag to link the Read 1 microsatellite internal motif and flanking sequences. If the primer probe sequence aligns within 2 bp of the expected primer probe start position, the paired Read 1 was assigned to its specific STR index tag. Based on the human genome reference, we identified the flanking genomic sequences that mark the complete STR segment and then determined the composition (i.e. mononucleotide, dinucleotide, etc.) and overall length of the repeat motif structure. Read 1 sequences that contained both the 5' and 3' flanking sequences with the internal microsatellite were used for genotyping. STR genotypes are called from Read 1. SNPs are phased with the STR genotype to generate haplotypes.* ***(c)*** *As an example of STR-Seq haplotyping, paired end alignments to the reference genome are shown for a STR target (trf747130) for sample NA12878. After the STR genotyping process, 114 and 133 read pairs were identified to have 11 and 8 repeats of a tetranucleotide motif (ATGA) in their Read 1s, respectively. Within each read pair group, all the base calls at the SNP position were identical, being either C (reference) or G (alternative). The site where CRISPR-Cas9 targets is indicated with red arrow, and the two haplotypes are illustrated on the bottom.*

We utilized our bioinformatics pipeline process to design this upgraded assay. First, we developed a bioinformatics process to identify STR markers across the entire human genome. The locations of over **740,000 tandem repeats** were obtained from the UCSC Genome Browser. We identified known STRs with documented polymorphisms and candidate STRs not previously reported to be polymorphic. We limited our selection of STRs to those that could be covered in their entirety within a 150 base pair read produced by an Illumina HiSeq sequencer. To increase the number of potential STR-SNP haplotypes, we identified tandem repeats that were within 100 base pair of a SNP with a high genotype frequency among different populations. Our analysis identified a total of **512,612 tandem repeats** that had at least **one validated SNP within 100 basepairs**. We chose **10,090 tandem repeat loci** that had suitable primers and were in proximity to a SNP position. Overall, this bioinformatics process has provided us with a database from which we can easily develop new and expanded STR-Seq assays to cover many more microsatellites than is currently available.

With this database, candidate primers were identified based on their uniqueness in the human genome reference, requiring at least two edited bases to align in any other location. Targeting primers are positioned on opposing strands; this double strand coverage was particularly useful because a true STR variant should be the same for both the forward and reverse strand reads. This automated bioinformatics pipeline was published and is available as open source software. The STR list included conventional STRs such as Marshfield loci [1], which has been genotyped with the traditional CE method. For this new improved assay, we included 436 Marshfield STRs with established CE genotypes. Also, we included 1,915 new candidates for STRs.

Using this information, we designed a series of STR-Seq tests that evaluate thousands of microsatellites in a single pass. When compared to the PCR-based test, results were found to be highly concordant on small sets of microsatellites.

Genetic mixture tests

We demonstrated that STR-Seq had far superior performance compared to the PCR method used by the FBI for resolving the identity of individual ins a genetic mixture. For example, we generated a six-component mixture. Five DNA samples from unrelated individuals were combined in equimolar ratio and then a minor component DNA (HGDP00924) was added in decreasing ratios ranging from 25% to 0.1%. For HGDP00924's 29 STR-SNP haplotypes, 16 demonstrated a decreasing fraction that correlated with expected mixture ratio. These 16 haplotypes were unique to HGDP00924 compared to the five other samples. Five of the HGDP00924-informative haplotypes were still detectable even at a ratio of 0.1%.

Population analysis of known and candidate STRs

During this project, we completed our microsatellite analysis of 1,004 individuals. The genomic DNA extractions of the individuals were obtained from the Foundation Jean Dausset - Centre d'Etude du Polymorphisme Humain (Paris, France) as shown in **Figure 2**. From these candidates, we determined their genotype heterozygosity and linkage to the flanking SNPs. Ethnic diversity was high, representing multiple geographic regions throughout the world. **One of our major findings of this population analysis is our identification of a new class of STRs that are highly polymorphic as defined by having 20 or more alleles.**

This second generation STR-Seq assay's accuracy was verified using samples with CE-based genotypes that included 436 Marshfield STRs [1]. From the 1,004 individual, we identified a total of 422,659 STR genotypes for the 436 Marshfield STRs. To compare genotypes compared to the CE genotypes, we used a dosage value that is derived from the number of base pairs remaining after subtracting the reference allele. Given that CE genotyping measures differences in amplicon size versus the NGS-based genotyping that counts the number of motifs directly from

a sequence read, the dosage value provides a standardize method for comparing between the two [6]. **STR-Seq accuracy was confirmed by a high correlation between CE and STR-Seq genotype dosages (R^2 = 0.97).**

Discrepancies between STR-Seq and CE-based genotypes were generally related to specific microsatellites having genotypes that exceed the sequence read size (i.e. 150 bases). More importantly, the current PCR-based gold standard method was prone to having specific errors where they could not discriminate the true number of STR repeats based simply on size. Rather, our sequencing approach delineated the exact count of motifs more accurately in these cases, seeing that the context sequence structure was clearly defined.

Using the results from the STR-Seq genotypes obtained from the 1,004 individuals, we determined the heterozygosity of STR markers. The conventional STRs (i.e. CODIS and Marshfield) had relatively high heterozygosity in a narrow range while the new genome-widely screened loci had a wider distribution.

We determined whether our new candidate microsatellites were polymorphic that fulfilled the criteria as being a microsatellite based on their sequence structure [6]. Among the candidate STR loci, we identified a total of 654 that were polymorphic. As noted, these STRs are new and have not been previously described. **Among these new microsatellites, we identified a category of heterozygous STRs (N = 109) that had 20 or more alleles as shown in Figure 2.** Generally, these highly polymorphic microsatellites had dinucleotide repeats.

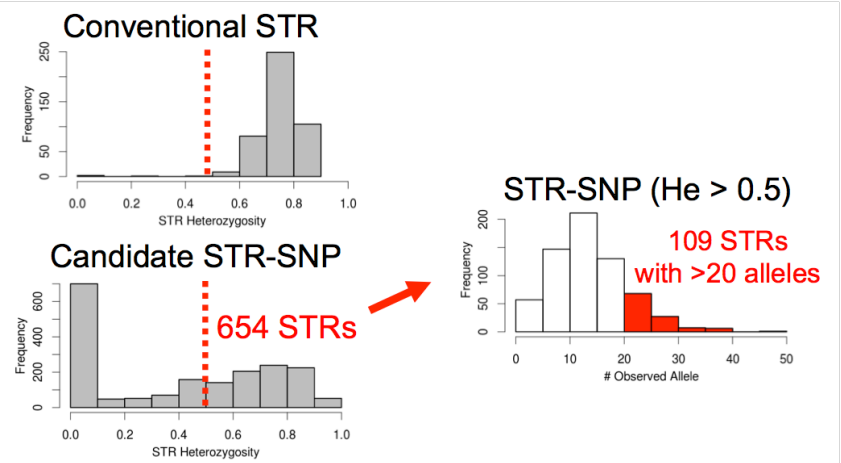


Figure 2. Comparison of heterozygosity comparing conventional STRs versus candidates. *We compared the heterozygosity of STRs among both conventional and established STRs versus our candidates. In this latest iteration of the new STR-Seq assay used for our population study, we identified 653 candidate STRs that were polymorphic. These represented new microsatellites not having been previously described. Subsequently from this group, we identified 109 that were highly polymorphic as denoted by having STRs that had 20 or more alleles.*

STR-SNP haplotype structure

More than 1,000 STRs were phased with at least one flanking SNP having a frequency higher than 0.05. We observed **more STR-SNP haplotypes** from the new candidate loci compared to the known STRs. Approximately 29% of the 1,915 STRs with the reported high frequency flanking SNPs had haplotypes. Among the established, well-characterized STRS (N=453), only 15% had a SNP in close enough proximity to be phased for haplotyping. Overall, the low linkage disequilibrium (LD) was observed (mean $r^2 < 0.1$) considering the short distance between STR and SNP (< 200 bp). Thus, the SNP alleles are not in linkage with STR alleles which is in agreement with recent observations from other studies [6-8]. Greater microsatellite allelic diversity showed a generally weaker linkage with flanking SNPs. When haplotypes are detected in STRs with high allelic diversity, we can expect the haplotype to be highly informative.

Correlation between STR genotype signature and individuals' geographical origin

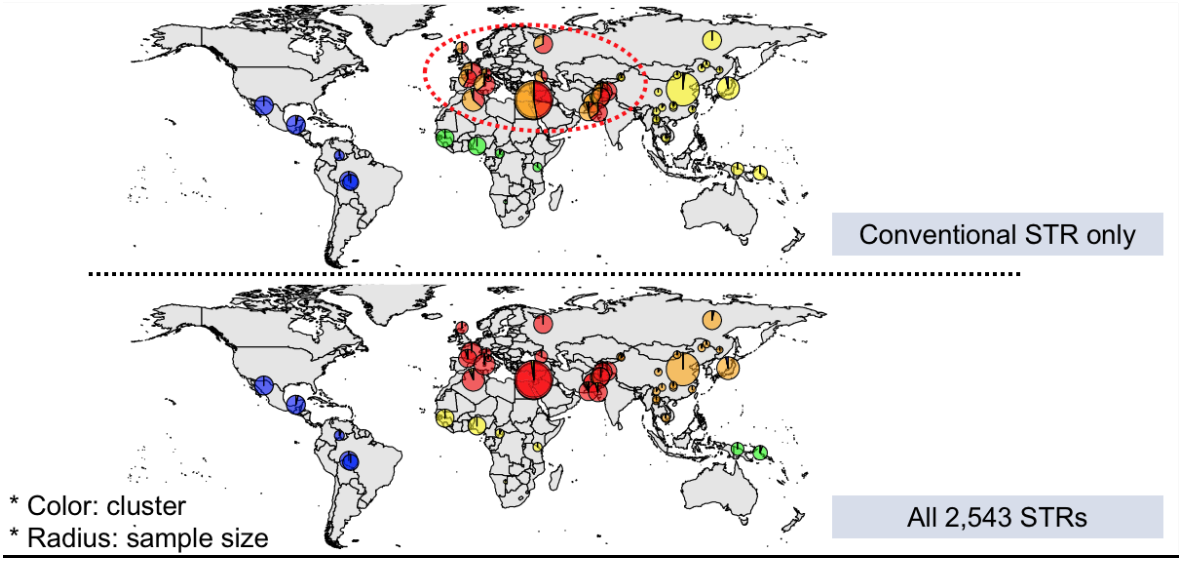
We determined if there was any correlation with an individual's geographical origin and the STR genotype profile derived from over 2,500 STRs per individual. Among our 1,004 samples, there

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

are individuals representing seven regions. This includes the Americas, Oceania, Asia, Europe, Middle East, North Africa, and Subsaharan Africa. We determined the assignment of individuals based on their geographic origin **(Fig. 3)** and compared our analysis with conventional CE-genotypes from the same population. The conventional STRs are from previously reported Marshfield microsatellite markers.

Europe, Middle East, and North Africa clustered in a same group when the number of clusters (K) was set to five. Moreover, populations from western part of Asia (e.g. Sindhi, Pathan, etc.) also clustered together with them, suggesting that their geographical proximity is related to their genetic closeness. For the other regions, when separating Asia into eastern and western parts, the clusters correlated with all the major regional origins. Interestingly, our new marker set defined geography origins better than the conventional STR markers. With STR-Seq results, we were able to distinguish individuals with better accuracy than conventional STRs from CE as denoted below in **Figure 4**.

Figure 4. Linking genographic region to conventional versus expanded STRs. *Conventional STR genotyping could not resolve origin of individuals from Europe and western parts of Asia but with the expanded STR-Seq data, these geographic origins were readily distinguished.*



References

1. Pemberton, T.J., et al., *Sequence determinants of human microsatellite variability.* BMC Genomics, 2009. **10**: p. 612.
2. Shin, G., et al., *CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis.* Nat Commun, 2017. **8**: p. 14291.
3. Hopmans, E.S., et al., *A programmable method for massively parallel targeted sequencing.* Nucleic Acids Res, 2014. **42**(10): p. e88.
4. Myllykangas, S., et al., *Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing.* Nat Biotechnol, 2011. **29**(11): p. 1024-7.
5. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.
6. Willems, T., et al., *The landscape of human STR variation.* Genome Res, 2014. **24**(11): p. 1894-904.
7. Quilez, J., et al., *Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans.* Nucleic Acids Res, 2016. **44**(8): p. 3750-62.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

8. Xu, L., et al., *Systematic Profiling of Short Tandem Repeats in the Cattle Genome.* Genome Biol Evol, 2017. **9**(1): p. 20-31.
9. Carlson, K.D., et al., *MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals.* Genome Res, 2015. **25**(5): p. 750-61.

Scholarly Products

An article describing our new technology, STR-Seq, has been published in **Nature Communications**. As noted previously, the article is made open access and thus, no license or subscription is required to see the article. The article is also highlighted in GenomeWeb (https://www.genomeweb.com/sequencing/stanford-team-develops-targeted-str-sequencing-approach-harnesses-crispr-cas9). Following is the publication information.

Shin, G., et al., *CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis.* **Nat. Commun.**, 2017. **8**:14291. doi:10.1038/ncomms14291

With our second manuscript submission that will detail our population study with STR-Seq, we will release the data publically for general use in forensic genetics. This will be done through the Short Read Archive sponsored by the National Institutes of Health as well as our own website.

Several presentations of this approach were made including meetings at the American Society of Human Genetics in 2017 and National Institute of Justice Forensic Science R&D Symposium at Pittcon 2018.

- Targeted CRISPR/Cas9 DNA fragmentation and selective primer sequencing enables massively parallel microsatellite analysis. 11th Annual Sequencing, Finishing and Analysis in the Future Meeting, Santa Fe, NM, June 2, 2016.
- STR-Seq Enables Massively Parallel Microsatellite Analysis. Roche Diagnostics Seminar, Pleasanton, CA, July 12, 2016.
- Genome-wide sequencing of STRs illuminates the haplotype structure of linked microsatellite-SNPs. Shin, G., American Society of Human Genetics, Vancouver, Canada, October 18–22, 2016. (poster)
- Population Haplotype Analysis of 2,543 STRs and their Flanking SNPs Using a Massively Parallel Next-Generation Sequencing Technology. Shin, G., B. T. Lau, H. P. Ji, H. Lee, M. Kubit, S. M. Grimes, Pittcon, Orlando, FL, February 26–March 1, 2018.

Intellectual Property

Stanford University's Office of Technology Licensing has submitted a patent regarding this technology.

Patent No. 62/175,985. Lau BT, Shin G, Lee H and **Ji HP**. STR-Seq: Highly parallel short tandem repeat analysis method based on programmable target selection. 7/10/2015.

Our results indicate that the technology is working robustly and can be used in a commercial research setting. We are making additional advances that will lead to practitioner application. This includes release of assay designs and genotyping data from our ongoing studies. We have are in discussions with a large biotechnology company about commercialization efforts.

Impact

STR-Seq technology provides a solution for highly parallel analysis across thousands of microsatellites with a genotyping accuracy that is comparable to the traditional CE method. The scale of STR-Seq is 10 to 100 times higher than current microsatellite analysis approaches including other next generation sequencing methods. The analysis of thousands of microsatellites in parallel proves extremely useful for a number of applications. Most importantly, this technology has direct relevance to solving the issue of DNA genetic mixtures, where it is frequently difficult to resolve the identify of these types of samples.

STR-Seq accurately called informative STR-SNP haplotypes that increase the polymorphic context when examining genotypes. For example, an uninformative homozygous variant once phased with an adjacent heterozygous variant yields informative haplotype. As we demonstrate, haplotype detection is a very powerful feature in the analysis of DNA mixtures and improves STR-Seq's sensitivity to identify a minor component DNA sample at a 0.1% ratio. STR-SNP haplotypes that are closely linked in a short interval are rare. In our analysis, only 10% of the microsatellites have informative haplotypes. Therefore, the analysis of more than 1,000 microsatellites enables: (i) discovery of multiple informative haplotypes; (ii) haplotype-based identification of a specific DNA sample that occurs as a low fraction of a multi-sample DNA mixture.

STR-Seq can be run as a PCR amplification-free assay that enables one to link each sequence read to a single DNA molecule without the use of unique molecular indices **(UMI)**. Other targeted sequencing methods require a post-capture PCR step that increases the frequency of amplification errors. To overcome this issue, some STR sequencing assays such as those using molecular inversion probes **(MIP)** have UMI's composed of random sequences[9]. However, random nucleotide tags are prone to PCR and sequencing errors. In the case of the MIP approach, the genomic DNA insert size is limited to 200 bp that restricts its application for identifying some categories of STR-SNP haplotypes.

To enable the sequencing of DNA molecules containing intact microsatellites, we developed an *in vitro* targeting approach that utilizes target specific fragmentation with CRISPR/Cas9. Incorporation of the targeted fragmentation with sequencing library preparation improves STR-Seq's overall performance and this targeted fragmentation process has potential for many applications beyond targeted sequencing. Thus, we demonstrate that there are critical advantages for maintaining an intact target DNA molecule, particularly for highly repetitive segments of the genome. By eliminating PCR amplification artifacts with CRISPR targeted fragmentation, allelic ambiguity is significantly reduced. Overall, STR-Seq has a wide spectrum of applications for forensics and genetics given its breadth and scope.