

Short Tandem Repeat Profiling via Next Generation Sequencing for Cell Line Authentication

Yi-Hsien Chen^{1#}, Jon P. Connelly^{2, 3#}, Colin Florian¹, Xiaoxia Cui^{1*}, Shondra M. Pruett-Miller^{2,3*}

¹Genome Engineering & Stem Cell Center (GESC@MGI), Department of Genetics, Washington University School of Medicine, St. Louis, 63110, USA

²St. Jude Children's Research Hospital, Department of Cell & Molecular Biology, Memphis 38105, USA

³Center for Advanced Genome Engineering

#These authors contributed equally to this work.

Address correspondence to:

*x.cui@wustl.edu

*shondra.miller@stjude.org

Key words: next generation sequencing, NGS, short tandem repeat, STR, targeted deep sequencing, cell line authentication, capillary electrophoresis, cell identity

Running title: STR profiling via NGS for cell line authentication

Abstract

Cell lines are indispensable models for modern biomedical research. In the era of CRISPR gene editing, they serve as versatile tools for preclinical studies, allowing patient specific mutations to be modeled or corrected and the resulting phenotypic outcomes studied. A large part of their usefulness derives from the ability of a cell line to proliferate over multiple passages (often indefinitely) allowing multiple experiments to be performed. However, over time, the cell line identity and purity can be compromised by human errors. Both cross contamination from other cell lines and even complete misidentification are possible. Routine cell line authentication is a necessary preventive measure and has become a requirement for many funding applications and publications. Short tandem repeat (STR) profiling is the most common method for cell line authentication and is usually carried out using standard polymerase chain reaction (PCR)-capillary electrophoresis (CE) analysis (STR-CE). Here we evaluated next generation sequencing (NGS)-based STR profiling of human and mouse cell lines at 18 and 15 loci, respectively, in a high-throughput format. Using the program STRight written in Python, we demonstrate that NGS-based analysis (STR-NGS) is superior to standard STR-CE in terms of the ability to report the sequence context of repeat motifs, sensitivity, and flexible multiplexing capability. STR-NGS is a valuable alternative for cell line authentication.

Introduction

Since the introduction of short tandem repeats (STRs) as polymorphic DNA signatures (Puers et al., 1993), STR profiling has become the gold standard for identity confirmation in contemporary forensic science (Butler, 2007). STRs, also known as microsatellites or simple sequence repeats, are DNA segments containing core repeat units of 2-6 nucleotides that are scattered throughout the genome (Ellegren, 2004). In addition to the original 13 core STR loci (Bruce Budowle, 1997), 7 more loci were included in the Combined DNA Index System (CODIS) that is used for forensics in the United States (Hares, 2015). These STR loci are highly polymorphic, genetically unlinked, and offer powerful and accurate individual identification.

Human and mouse cell lines are important research models for mechanistic studies, target identification and therapeutic development. However, cell cultures are at risk for misidentification due to human errors and cross-contamination from other cell lines. Examples of misidentified cell lines jeopardizing scientific research continue to grow in number (Nardone, 2007), demonstrating the urgent need for frequent cell line authentication. As a result, cell line authentication is now required by an increasing number of journals prior to publication as well as for grant applications (Almeida et al., 2016). A method that is sensitive, high-throughput and economical is highly desirable.

The framework of using STRs for the authentication of human cell lines was first introduced in 2010 (Barallon et al., 2010). To date, many thousands of human cell line STR profiles are available, reflecting the unique donors from whom they were originally derived (Novroski et al., 2016). The American Tissue Culture Collection (ATCC) has published standard guidelines which recommend the use of at least eight STR loci

(TH01, TPOX, vWA, CSF1PO, D16S539, D7S820, D13S317 and D5S8181 plus Amelogenin for gender identification) for human cell line authentication (ASN-0002, 2011). Moreover, recent studies report several additional STR loci that may be used to authenticate mouse cell lines (Almeida et al., 2019; Almeida et al., 2014).

Currently, STR profiling is predominately performed by resolving multiplexed, fluorescently-labeled PCRs using capillary electrophoresis (Deforce et al., 1998). However, loci of the same size but with different sequences cannot be distinguished using this conventional method. The full sequences and nucleotide variations found in STR loci provide additional data that aid in identification. The conventional STR method also requires access to a specialized instrument, a Genetic Analyzer, which has limited potential for further improving sensitivity or throughput. With continuous technical improvements, NGS has become an attractive alternative for STR profiling. Different NGS platforms such as Roche/454, Ion torrent, and Illumina systems, have proven capable of sequencing the majority of STR loci in forensic science (Fordyce et al., 2015; Mikkelsen et al., 2014; Van Neste et al., 2014). NGS-based STR (STR-NGS) profiling has several advantages over the conventional method, including high-throughput, low cost when running many samples, flexibility with which STR loci to include, quantitative measurements for mixed samples, and high resolution for single nucleotide polymorphisms (SNPs) (Bornman et al., 2012; Shin et al., 2017). This method has yet to be applied to cell line authentication.

In the present study, we evaluate the accuracy and sensitivity of the Illumina MiSeq platform for STR profiling of human and mouse cell lines and demonstrate the method is

valid and scalable for routine quality control of human and mouse lines used in biomedical research.

Materials and methods

Control samples and Cell lines

Two human iPSC lines, BJFF.6 from a male donor, and AN1.1 from a female donor were generated at the Genome Engineering & Stem Cell Center (GESC) at Washington University in St. Louis. iPSCs were maintained on Matrigel (Corning) coated plates in Stem-Flex medium (Thermo Fisher). In addition, four mouse cell lines, NIH3T3, MC3T3-E1, CT26 and 4T1 were obtained from the American Type Culture Collection (ATCC) and cultured in the ATCC-recommended media. All cell lines were grown in a humidified incubator set at 5% CO₂ and 37°C.

Extraction and quantification of DNA

DNA was extracted using a crude DNA extraction buffer (10mM Tris pH8.0, 2mM EDTA, 0.2% Triton X-100 and 200 µg/ml proteinase K) or DNA Blood Mini kit (Qiagen) following the manufacturer's instructions. DNA was quantified using NanoDrop One spectrophotometer (Thermo Fisher).

STR loci and primer design

Eighteen STR loci (CSF1PO, D13S317, D16S539, D18S51, D19S433, D2S1338, D21S11, D3S1358, D5S818, D7S820, D8S1179, FGA, PentaD, PentaE, TH01, TPOX, vWA and Amelogenin) recommended from the ATCC and the Combined DNA Index

System (CODIS) were used for human cell line profiling. Fifteen STR loci (18-3, 4-2, 5-5, 6-7, 9-2, 12-1, 15-3, X-1, 1-1, 2-1, 3-2, 8-1, 11-2, 17-2 and 19-2) were used for mouse cell line STR profiling (Almeida et al., 2019; Almeida et al., 2014). The Primer Blast design tool from NCBI was used to design PCR primers flanking STR regions based on the reference sequences from STRBase (Ruitberg et al., 2001) (<https://strbase.nist.gov/>) for human loci and two recent studies (Almeida et al., 2019; Almeida et al., 2014) for mouse loci, respectively. To increase PCR specificity, nested forward primers were used for the PentaE locus and only the internal forward primer has the adapter sequence required for NGS. All primer sequences are listed in Table S1 and S2.

Optimized PCR amplification and processing using Illumina chemistry

A two-step PCR strategy is used to amplify each STR locus for NGS. PCR1 amplifies the individual STR locus and adds a partial Illumina adapter. A portion of the PCR1 amplicon is then used as template for PCR2 and adds a unique index and the remaining Illumina adaptor to each sample. PCR conditions for STR loci (PCR1) have been optimized. Results for detailed optimizations are found in Supplemental Figure 1. Tetramethylammonium (TMA) oxalate solution was made by mixing a 2 to 1 molar ratio of aqueous TMA hydroxide (Sigma Aldrich) and ammonium oxalate monohydrate (Thermo Fisher), respectively (Markowitz, 1957). PCR reactions contain 0.5 mM of each primer, 70 ng of genomic DNA, 2 mM TMA oxalate, and 1X Platinum SuperFi PCR Master Mix (Thermo Fisher). The samples were amplified using a Veriti thermal cycler (Thermo Fisher) as follows: 98°C for 2 min; 28 cycles of 98°C for 15 s, 60°C for 15 s,

and then 68°C for 30 s; followed by an extension for 2 min at 72°C. For multiplex PCRs, samples were amplified using the following reaction mixture: 0.2 μM final concentration of primer sets for each locus (6 sets in each reaction), 140 ng of genomic DNA in the reaction, and 1X Platinum SuperFi PCR Master Mix (Thermo Fisher). Multiplex PCRs were carried out in a Veriti thermal cycler (Thermo Fisher) with the following cycle conditions: 98°C for 2 min; 30 cycles of 98°C for 15 s, 60°C for 45 s, and then 68°C for 1 min; followed by an extension for 5 min at 72°C. Products of PCR1 were subjected to an additional 5 cycles with non-overlapping dual-indexing primers without TMA oxalate, and products of PCR2 were pooled and purified using 0.6X SPRI beads. The purified library was run on a BioAnalyzer using a high sensitivity chip to confirm quality and quantity. It was then denatured with 0.1nM NaOH, followed by neutralization with 0.1M HCl. The denatured pool was diluted to 20pM following Illumina guidelines. It was sequenced on a MiSeqv2 using a MiSeq Reagent Kit 500v2 kit with the following inputs: 250 bp Read1, 10 bp Index1, 8 bp Index2 and 250 bp Read2. Samples were demultiplexed using the index sequences, FASTQ files were generated using the Illumina bcl2fastq software, allowing 1 bp mismatch.

STR data analysis

Human STR-CE profiling was performed by Cell Line Genetics (WI, USA). STR-NGS was analyzed using the python script STRight. This program is a modified version of CRIS.py (Connelly and Pruett-Miller, 2019) and takes a csv file containing target STR data as input. The csv files (STR_human.csv and STR_mouse.csv) contain 9 columns: (1) STR name, (2) Reference sequence, (3) Start target sequence, (4) End target

sequence, (5) repeat size, (6) bp_modifier, (7) the number of repeats in a reference sequence found on strbase.nist.gov (this can be used as a quality control check if novel STRs are being added), (8) SNP_modifier, and (9) notes. Briefly, the program runs through each line of the fastq files and tests whether the start and end target sequences of each STR can be matched to the read. If both sequences are found, the distance between the two sequences is measured in base pairs. Because SNPs could occur in the flank region outside the STR repeat region, the start and end sequences for STRight analysis do not always land precisely at the beginning and end of each locus. Instead, they may be placed several base pairs upstream or downstream of the repeat to avoid known SNPs. In addition, some STR loci have additional nucleotides inside the repeats (i.e., DS21S11 has 13 additional nucleotides among “TCTA” and “TCTG” repeats). To take these extra base pairs into account, the bp_modifier value is used to subtract any extra base pairs from the total length (i.e., if the start sequence lands 5bp upstream from the start of the STR repeat region, then a bp_modifier of 5 would subtract 5 base pairs from the measured distance). The repeat number of a locus is determined by dividing the calculated distance by the size of each repeat, for example, 4 for an “AAAG” repeat. Finally, there are special cases where SNPs in a given locus create an extra repeat. For example, the STR D13S317 is a “TATC” repeat whose end usually terminates right before the sequence “AATC”. Some individuals however have an A>T variant creating a termination sequence of “TATC”, and which also adds an additional repeat. Typical STR-CE utilizes a primer mixture containing two reverse primers which have the 5’ nucleotide landing on either the “A” or “T” of the SNP to detect that variant. STRight contains a SNP_modifier which checks whether the A/T SNP is present in the

read. If the indicated SNP (in this case “T”) is present, then the SNP_modifier value (in this case 1), would add an extra repeat to the total number to account for the SNP. STRight reports the top repeat lengths for each STR along with sequence information. STRight is free to use and source code is available at <https://github.com/patrickc01/STRight>

Analysis of mixed samples

Genomic DNA was extracted from diploid iPSC cell lines, BJFF.6 and AN1.1 cells, and added to individual reactions with a final concentration of 140 ng of total DNA at BJFF.6/AN1.1 ratios of 1:1, 5:1, 10:1, 20:1, 100:1, 200:1, and 1000:1. PCR amplification is described above, and analysis was performed in the same way as the single source samples.

Results

NGS optimization for human STR profiles. We developed an NGS-based STR profiling (STR-NGS) method that amplifies and analyzes common human and mouse STR loci for repeat length. Primers designed to each STR locus amplify a region of interest (ROI) containing the STR repeats flanked by left and right consensus sequences (Figure 1A). Our NGS library construction is a two-step process. PCR1 amplifies the region of interest and adds partial Illumina adaptors to the amplicons. PCR2 is the indexing PCR and adds unique identifiers to each sample. To improve NGS read quality, we optimized the PCR1 reaction at each locus. First, different primer sets were tested for amplification efficiency and specificity, and the final primer pairs for

the 17 CODIS STR loci and sex (AMEL) locus in this study are listed in Table S1.

Second, we evaluated different DNA polymerases, including several high-fidelity DNA polymerases- AccuPrime Taq DNA polymerase high fidelity from Thermo Fisher, Platinum SuperFi PCR Master Mix from Thermo Fisher, and Titanium Taq DNA polymerase from Takara. MyTaq from Bioline is a low cost and high-performance polymerase and was used as a standard for comparison. We also tested and optimized the quantity of genomic DNA sample input. We found that Platinum SuperFi PCR mix produced higher PCR amplicon yield and superior accuracy for this application (data not shown). SuperFi was used for subsequent tests and studies. As reported previously, we found that the quantity of input template DNA has a substantial effect on the outcome of PCR reactions (Lorenz, 2012). The optimal DNA amount in our assays is between 70-140 ng per reaction, roughly equivalent to 10-20,000 cells (Figure S1A and B). Low amplification was observed with input cell numbers lower than this range (data not shown). PCR supplements have been shown to be effective in improving yields of specific PCR products at difficult loci (Lorenz, 2012). Given the presence of homopolymer stretches at many STR loci (Gettings et al., 2015), poor PCR amplification can result in higher error rates in NGS. Several tetramethylammonium (TMA) derivatives have been shown to increase the specificity of PCR and improve the yield of amplification products (DiLella and Woo, 1987; Kovarova and Draber, 2000). We tested two TMA derivatives, TMA chloride (TMAC) and TMA oxalate (TMAO). TMAC addition in PCRs made a modest improvement on reducing background in most STR loci compared to PCR with no additive (data not shown), while addition of TMAO to PCR

reactions increased the yield of specific PCR products (Figure S1A and B). Therefore, SuperFi plus TMAO was used for PCR1 in subsequent studies.

We also optimized the indexing reaction (PCR2). Dual-indexing has proven to increase the accuracy of multiplex sequencing and throughput on the NGS platform (Kircher et al., 2012). High-fidelity polymerases, such as SuperFi, are effective at reducing errors introduced in the indexing reaction (Figure S1B). Five representative STR loci are highlighted in comparison between standard and improved PCR conditions (Figure 1B) as well as the overall background improvement in all loci (Figure 1C). Moreover, it should be noted that amplification efficiency varies at different loci, as measured by total NGS reads at each locus (Figure S1C).

STR-NGS vs. STR-CE

We next compared the results of STR-NGS using our optimized conditions to STR-CE conducted by Cell Line Genetics (Madison, WI) for two diploid iPSC lines. The two sets of data are highly in agreement (Figure 1D). In both cell lines, noise, stutter, and allele percentages are shown for all 18 STR loci (Figure 1D). Stutters are defined as sequence reads that contain a number of STR repeat lengths that are smaller or larger (predominantly 1 repeat shorter) than the actual allele and with a frequency of less than 10%. Noise is defined as reads that are not alleles or stutters and are likely the product of sequencing errors or off-target amplification. Among all STR loci, 91-99% of the reads from STR-NGS corresponded to parent alleles and noise comprised less than 1% of the total reads. The stutters ranged from 0 to 3.6%, except D18S51 and D2S1338

which had higher stutter percentages. Together, we demonstrate that STR-NGS identifies human cell lines as accurately as the standard STR-CE method.

Sensitivity and multiplexing capability of STR-NGS. Early detection of cross contamination in a given culture is an important application of STR-NGS. We evaluated STR-NGS sensitivity in detecting a contaminating cell line sample among varying mixed ratios of BJFF.6 and AN1.1 cells, where AN1.1 represented the minor component of the mixtures. These two lines share STRs of the same length at several loci, therefore we focused on unique repeats found in AN1.1. We found that the AN1.1 fractions by read count correlate well with the expected ratios (Figure 2A). Next, we examined the D81179 locus specifically for its distinct repeat lengths in two cell lines. Given the major stutter of D81179 is 4 bp shorter than the parent allele (Figure 1B), it provided a good opportunity to identify the minor alleles from AN1.1. We found STR-NGS could detect a minor contamination level as low as 0.5% (1:200) within a sample (Figure 2B). Note the AN1.1 contaminating mixture has STR repeat lengths that are 1-2 repeats longer than the BJFF.6 repeats which allow them to be differentiated from stutters.

To reduce the cost and labor of the STR-NGS assay, we explored the feasibility of multiplexing the PCR reactions. We grouped different numbers of STR loci for multiplexed PCRs. Up to six loci can be amplified in one PCR reaction without compromising stutter and noise levels (Figure 2C). Stutter levels for D18S51 and D2S1338 remained higher than other loci regardless of PCR conditions, and dropout of the PentaD STR was observed during multiplexing reactions. Overall, we were able to

obtain satisfactory percentages of clean reads for all loci except for PentaD and conclude that STR-NGS is reliable in multiplexed format.

STR-NGS in mouse cell line authentication. With great effort from the scientific community, STRs have been identified that allow authentication of mouse cell lines from different mouse strains (Almeida et al., 2019; Almeida et al., 2014). To determine whether STR-NGS could be used for mouse cell line authentication, we performed STR-NGS on two previously reported mouse cell lines, MC3T3-E1 and NIH3T3 derived from different mouse strains (Almeida et al., 2014). First, both cell lines were used to optimize PCR performance and optimized primer sets are listed in Table S2. For MC3T3-E1 all loci matched the published STR-CE results. Of note, the STRs at loci 6-7, and 15-3, had higher stutter or noise percentages than the other STRs assayed (Figure 3A). Nonetheless, the combined stutter and noise for all STRs for MC3T3-E1 was less than 5% of the total reads. However, for NIH3T3 cells, minor fractions of different repeat lengths were observed at loci 4-2 and 18-3 (Figure 3A and 3B).

To investigate the reason for the discrepancy in STR repeat length in the NIH3T3 cells, we compared the read percentages and sequencing content of different repeat lengths in STR loci, 4-2 and 18-3. Stutter products in most loci usually comprise less than 5% of total reads in both human and mouse samples (Figure 1C and 3A). The read percentages for additional lengths were 15.60% for 4-2 and 13.13% and 16.40% for 18-3, which is unlikely the result of suboptimal assay conditions. One of the main advantages of the NGS based method is that the raw data contains the actual sequence of a given locus. We examined the STR-containing reads with complete microsatellite

sequences. At STR 4-2, our finding is consistent with results from a recent study (Almeida et al., 2019) which show 19.3 and 20.3 STR repeats for NIH3T3 cells (Figure 3C). At STR 18-3, we observed reads corresponding to all three repeat lengths – 17, 18, and 19. We hypothesized that the heterogeneity of the immortalized NIH3T3 cells might have another population harboring 17 and 18 repeats. To verify the 18 repeats did not come from stutter products of 17 and 19, we performed single cell sorting and analyzed single cell clones from parental NIH3T3 cells. Three out of the five independent clones carried 17 and 18 repeats at the locus. (Table 1 and Figure S2). Another advantage of STR-NGS is that some information can be inferred based on the number of sequence reads for each repeat length. For example, all five clones show an allelic ratio of approximately 2:1 for either the 17:18 or 17:19 repeats (Figure S2), which suggests that there are three copies of this locus.

Next, we expanded the number of total mouse STR loci to 15 based on a recent report (Almeida et al., 2019) and compared the results between STR-CE and STR-NGS using three different mouse cell lines – NIH3T3, CT26, and 4T1 used in that study. In all three lines, STR-CE and STR-NGS results matched at 12 of 15 loci tested (Table 2). To assess the differences among the three cell lines, we performed sequencing alignment with the data from parental cells and single cell clones. In NIH3T3 cells, STR-NGS was able to detect the alleles from different subpopulations present in the pool (Figure S3). Three unique subpopulations were identified by the presence of three different repeat lengths in STR locus 11-2 (Figure S3A). To investigate further, we isolated, expanded, and analyzed 12 single cell clones. Most clones were represented by repeat lengths of 15 and 17. However, five of twelve clones presented with 15,16,17 or 14,15,17 repeat

lengths (Figure S3B). A unique clonal population was also observed in 4T1 cells (Figure S4) where the 17-2 locus differed between STR-NGS and STR-CE. STR-CE only shows a repeat length of 15 in 4T1 cells compared to repeats of 14 and 15 given by STR-NGS (Table 2). Given that the stutter for 17-2 locus was below 6% in NIH3T3 and CT26 cell lines (Table 2), we believe a repeat of length 14 exists and is from the minor allele contributor (Figure S4B). Together, these data demonstrate that STR-NGS offers great accuracy on allele calling compared to the traditional STR-CE method and is a valuable alternative for mouse cell line authentication.

Discussion

STR profiling is the gold standard for human identity testing (Butler, 2007) and cell line authentication (Barallon et al., 2010), given the uniqueness of each individual and relatively easy assay format. STR-CE is widely used for separating STR markers based on size distribution; however, it has limited sensitivity and does not provide the STR sequence identity or context. Although NGS has been implemented for human identification in forensic genetics (Børsting and Morling, 2015),, cell line authentication has been limited to STR-CE. Because of heterogeneity within many cell lines, this lack of sensitivity may lead to cell line misidentification or delayed identification of cross contamination. Here, we present a simple and robust NGS-based solution for human and mouse cell line authentication using optimized amplification conditions for each STR loci and a novel STR analysis program called STRight.

To develop a robust STR-NGS workflow, we first evaluated different DNA polymerases, cell number, indexing PCR conditions, and additives on STR loci

amplification. After optimization, we achieved over 96% allele specific signals at most loci. Next, we compared the results from STR-CE and STR-NGS workflows on two diploid human iPSC lines. Overall, the STR-NGS and STR-CE results matched. Moreover, except for D18S51 and D2S1338, the average range of stutter and noise percentage observed per STR locus is below 5% and 1%, respectively, for STR-NGS compared to the stutter ratio of 20% for STR-CE which is often used for initial filtering (Brookes et al., 2012; Mikkelsen et al., 2014; Novroski et al., 2016). While sequencing errors were very minimal in most STR loci, D18S51 and D2S1338 are obvious underperformers in the current protocol with a higher proportion of reads containing errors in the ROI (data not shown). It is unclear at which point the errors are introduced, but a longer ROI and sequence context could be potential factors contributing to the stutter percentage and sequencing errors.

A previous study showed that in mixed samples, sequences from the minor fraction (down to 1%) were detectable by NGS (Fordyce et al., 2015). In contrast, STR-CE has a general detection limit for the minor contributor in 1:10 mixtures (Westen et al., 2009). In this study, we show that STR-NGS is sensitive in detecting a minor contributor among a series of DNA mixtures down to 0.5% (Figure 2B).

Multiplexed PCRs for STR typing are both time and cost-effective. Indeed, STR alleles are routinely analyzed by multiplexed PCRs followed by CE-based analysis (Butler, 2006, 2007). However, large variations in target loci lengths result in a bias towards loci with shorter amplicons and lead to an imbalanced locus coverage (Van Neste et al., 2012). While up to four or ten multiplexed PCRs for STR sequencing have been demonstrated for 454 or Ion Torrent sequencing platforms, respectively, high

noise ratios and poor performance on allele calling have limited its utility (Butler, 2007; Fordyce et al., 2015). In this report, we demonstrate that our optimized workflow for STR-NGS with TMAO additive is capable of multiplexing 6 STR loci with less than 5% stutters of total reads in 15 of 17 total human STR loci. Moreover, we recommend that underperforming loci, such as PentaD, be avoided in a multiplexed PCR setting. A careful selection of loci and further PCR multiplex optimization could further improve the utility of the STR-NGS approach.

Unlike STR forensic analysis, STR analysis is more complicated in human cell lines due to the heterogeneity and aneuploidy present in many immortalized and transformed lines. Indeed, STR profiles for a given human cell line may even vary from different sources (<https://web.expasy.org/cellosaurus/>), and genetic variation within cell lines has been well documented (Pourquier and Azzi, 2019). The need for cell line authentication in mouse cell lines has also risen in recent years (Almeida et al., 2019; Almeida et al., 2014), so much so that the American Type Culture Collection is now offering mouse STR profiling as a service. In this study, we demonstrate for the first time to our knowledge that STR-NGS can be used in mouse cell line authentication. In addition to generating correct genotypes, STR-NGS data resulted in the identification of sub-repeat variants from sub-populations of NIH3T3 cells that were undetected by the STR-CE method (Almeida et al., 2019; Almeida et al., 2014). In contrast to fragment length analysis by STR-CE, sequencing compositions in the ROI by STR-NGS provides additional detail for better interpretation of STR variation.

Conclusions

We developed a simple workflow that offers a reliable solution for genotyping the most frequently used STR loci in both human and mouse cell lines (Table 3 and 4) without the need for a specialized instrument made for STR. Multiplexing reactions provide the flexibility to run many samples at once and reduce both time and cost for the operation. STR-NGS improves allele calling efficiency, detection sensitivity, and allows better differentiation of artifacts from minor contributors than the standard STR-CE method. STR-NGS can be easily used to authenticate cell lines as well as sensitively detect cross contamination.

Acknowledgements

We thank Jane Kouranova for the initial primer design for mouse STR loci and Jed Lin for assisting NGS library preparation. We thank the DNA Sequencing Innovation Lab, The Edison Family Center for Genome Sciences & Systems Biology which provided NGS services.

References

Almeida, J.L., Cole, K.D., and Plant, A.L. (2016). Standards for Cell Line Authentication and Beyond. *PLoS Biol* 14, e1002476.

Almeida, J.L., Dakic, A., Kindig, K., Kone, M., Letham, D.L.D., Langdon, S., Peat, R., Holding-Pillai, J., Hall, E.M., Ladd, M., *et al.* (2019). Interlaboratory study to validate a STR profiling method for intraspecies identification of mouse cell lines. *PLoS One* 14, e0218412.

Almeida, J.L., Hill, C.R., and Cole, K.D. (2014). Mouse cell line authentication. *Cytotechnology* 66, 133-147.

ASN-0002, A.A. (2011). Authentication Of Human Cell Lines: Standardization Of STR Profiling.

Barallon, R., Bauer, S.R., Butler, J., Capes-Davis, A., Dirks, W.G., Elmore, E., Furtado, M., Kline, M.C., Kohara, A., Los, G.V., *et al.* (2010). Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues. *In Vitro Cell Dev Biol Anim* 46, 727-732.

Bornman, D.M., Hester, M.E., Schuetter, J.M., Kasoji, M.D., Minard-Smith, A., Barden, C.A., Nelson, S.C., Godbold, G.D., Baker, C.H., Yang, B., *et al.* (2012). Short-read, high-throughput sequencing technology for STR genotyping. *Biotech Rapid Dispatches* 2012, 1-6.

Børsting, C., and Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Forensic Sci Int Genet* 18, 78-89.

Brookes, C., Bright, J.A., Harbison, S., and Buckleton, J. (2012). Characterising stutter in forensic STR multiplexes. *Forensic Sci Int Genet* 6, 58-63.

Bruce Budowle, T.R.M., Stephen J. Niegzoda, and Barry L. Brown (1997). CODIS and PCR-Based Short Tandem Repeat Loci: Law Enforcement Tools. In 2nd European Symposium on Human Interaction.

Butler, J.M. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci* 51, 253-265.

Butler, J.M. (2007). Short tandem repeat typing technologies used in human identity testing. *Biotechniques* 43, ii-v.

Connelly, J.P., and Pruett-Miller, S.M. (2019). CRIS.py: A Versatile and High-throughput Analysis Program for CRISPR-based Genome Editing. *Sci Rep* 9, 4194.

Deforce, D.L., Millecamps, R.E., Van Hoofstat, D., and Van den Eeckhout, E.G. (1998). Comparison of slab gel electrophoresis and capillary electrophoresis for the detection of the fluorescently labeled polymerase chain reaction products of short tandem repeat fragments. *J Chromatogr A* 806, 149-155.

DiLella, A.G., and Woo, S.L. (1987). Hybridization of genomic DNA to oligonucleotide probes in the presence of tetramethylammonium chloride. *Methods Enzymol* 152, 447-451.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 5, 435-445.

Fordyce, S.L., Mogensen, H.S., Borsting, C., Lagace, R.E., Chang, C.W., Rajagopalan, N., and Morling, N. (2015). Second-generation sequencing of forensic STRs using the Ion Torrent HID STR 10-plex and the Ion PGM. *Forensic Sci Int Genet* 14, 132-140.

Gettings, K.B., Aponte, R.A., Vallone, P.M., and Butler, J.M. (2015). STR allele sequence variation: Current knowledge and future issues. *Forensic Sci Int Genet* 18, 118-130.

Hares, D.R. (2015). Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci Int Genet* 17, 33-34.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40, e3.

Kovarova, M., and Draber, P. (2000). New specificity and yield enhancer of polymerase chain reactions. *Nucleic Acids Res* 28, E70.

Lorenz, T.C. (2012). Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *J Vis Exp*, e3998.

Markowitz, M. (1957). Notes - A Convenient Method for Preparation of Quaternary Ammonium Salts. *The Journal of Organic Chemistry* 22, 983-984.

Mikkelsen, M., Frank-Hansen, R., Hansen, A.J., and Morling, N. (2014). Massively parallel pyrosequencing of the mitochondrial genome with the 454 methodology in forensic genetics. *Forensic Sci Int Genet* 12, 30-37.

Nardone, R.M. (2007). An Open Letter Regarding the Misidentification and Cross-Contamination of Cell Lines: Significance and Recommendations for Correction.

Novroski, N.M.M., King, J.L., Churchill, J.D., Seah, L.H., and Budowle, B. (2016). Characterization of genetic sequence variation of 58 STR loci in four major population groups. *Forensic Sci Int Genet* 25, 214-226.

Pourquier, P., and Azzi, J. (2019). [Genetic and transcriptional evolution alters cancer cell line drug response]. *Bull Cancer* 106, 9-10.

Puers, C., Hammond, H.A., Jin, L., Caskey, C.T., and Schumm, J.W. (1993). Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01[AATG] n and reassignment of alleles in population analysis by using a locus-specific allelic ladder. *Am J Hum Genet* 53, 953-958.

Ruitberg, C.M., Reeder, D.J., and Butler, J.M. (2001). STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res* 29, 320-322.

Shin, G., Grimes, S.M., Lee, H., Lau, B.T., Xia, L.C., and Ji, H.P. (2017). CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nat Commun* 8, 14291.

Van Neste, C., Van Nieuwerburgh, F., Van Hoofstat, D., and Deforce, D. (2012). Forensic STR analysis using massive parallel sequencing. *Forensic Sci Int Genet* 6, 810-818.

Van Neste, C., Vandewoestyne, M., Van Criekinge, W., Deforce, D., and Van Nieuwerburgh, F. (2014). My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. *Forensic Sci Int Genet* 9, 1-8.

Westen, A.A., Nagel, J.H., Benschop, C.C., Weiler, N.E., de Jong, B.J., and Sijen, T. (2009). Higher capillary electrophoresis injection settings as an efficient approach to increase the sensitivity of STR typing. *J Forensic Sci* 54, 591-598.

Figure Legends

Figure 1. STR-NGS optimization and performance. (A) Schematic representation of a STR locus with STR repeats, flanking regions at the start and end of the repeat region, and a targeted deep sequencing (DS) primer pair. Sequence from “5’ Flank” to “3’ Flank” is used for repeat length analysis. (B) Allele and stutter frequencies for five representative STR loci calculated as the percentage of the parent allele reads. Verified STRs indicated by a # above bar plot. (C) Total background (stutter + noise) in all STR loci in the comparison between standard and optimized PCR conditions. (D) STR profile for each locus examined by CE and NGS based methods in two diploid iPSC lines.

Figure 2. STR-NGS sensitivity on a mixed sample and optimized multiplexed conditions. (A) Observed allele fractions of informative STRs repeats are plotted against the expected ratio for given mixtures of 2 genomic samples. AN1.1 gDNA was diluted into BJFF.6 gDNA in a ratio of 1:1 to 1:1000. Expected allele fractions for diluted AN1.1 STRs correlate well with observed. (B) A mixture of two diploid cell lines was analyzed for STR D8S1179. M and N alleles indicate genotypes from the major and

minor components, respectively. Bar graph shows percentage of read counts for each STR repeat of both cell lines in different mixture ratios. (C) STR profile for each locus examined using multiplexed STR-NGS in BJFF.6 iPSC cells. (*) dropout allele.

Figure 3. STR-NGS in mouse cell lines. (A) STR profile in two mouse cell lines (NIH3T3 and MC3T3-E1) shows allele, stutter and noise fractions of total parent allele reads. Asterisk indicates the difference in STR repeat length between reference and STR-NGS. (B) The frequencies of different stutters and repeat lengths, and (C) sequence context of different STR repeat lengths in two STR loci (4-2 and 18-3) different between reference and STR-NGS in NIH3T3 cells.

Table 1. STR profiles of five NIH3T3 clones using STR-NGS

(*) difference in repeat length between STR-NGS and STR-CE method

Table 2. STR profiles of 3 mouse cell lines

(*) difference in repeat length between STR-NGS and STR-CE method

Table 3. Summary of human STR loci using STR-NGS

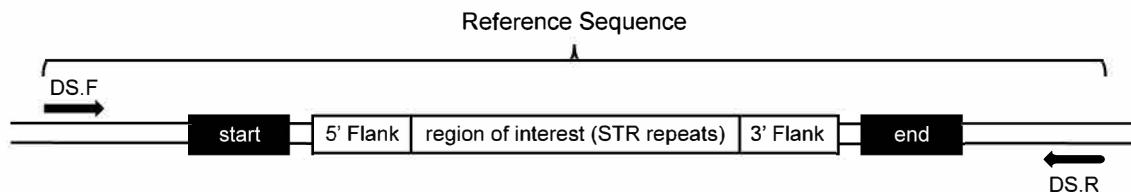
(*) Underperformer for human STR-NGS

Table 4. Summary of mouse STR loci using STR-NGS

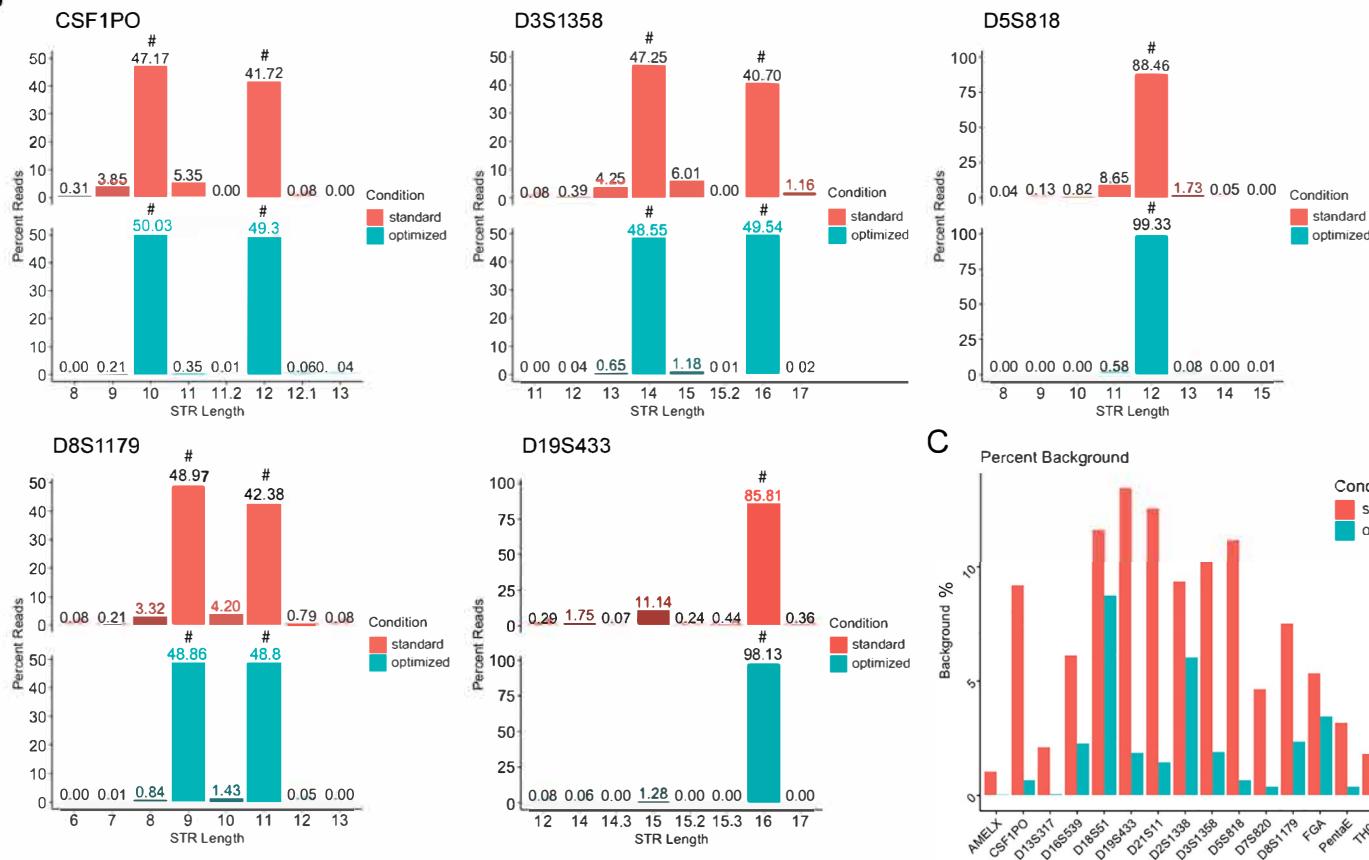
(*) Underperformer for mouse STR-NGS

Figure 1

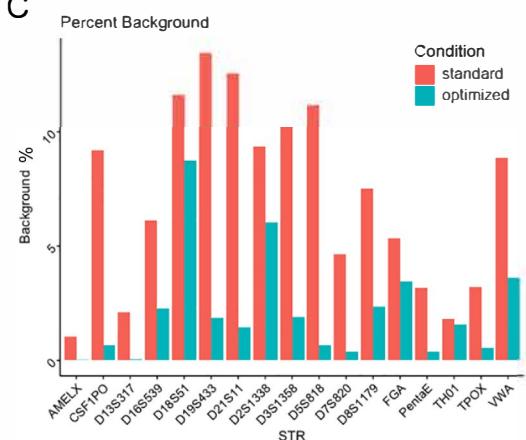
A



B



C

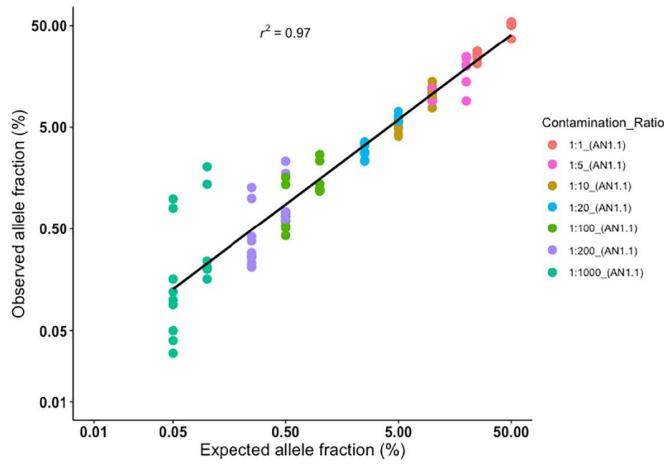


D

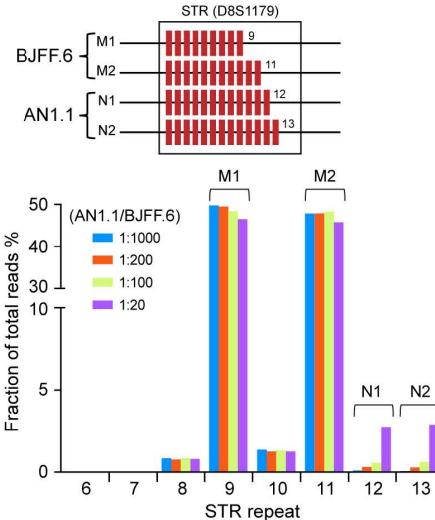
BJFF.6						AN1.1					
STR	STR_CE	STR_NGS	Allele (%)	Stutter (%)	Noise (%)	STR	STR_CE	STR_NGS	Allele (%)	Stutter (%)	Noise (%)
CSF1PO	10,12	10,12	99.33	0.6	0.07	CSF1PO	10,12	10,12	99.4	0.58	0.02
FGA	22,23	22,23	96.53	3.12	0.35	FGA	22,23	22,23	96.94	2.77	0.29
TH01	7,8	7,8	98.44	1.55	0.01	TH01	7,9,3	7,9,3	98.09	1.88	0.03
TPOX	10,11	10,11	99.44	0.56	0	TPOX	8,11	8,11	99.42	0.57	0.01
vWA	16,18	16,18	96.36	3.62	0.02	vWA	14,17	14,17	98	1.8	0.20
D3S1358	14,16	14,16	98.09	1.9	0.01	D3S1358	15,16	15,16	99.14	0.85	0.01
D5S818	12	12	99.33	0.67	0	D5S818	11,12	11,12	99.2	0.3	0.50
D7S820	11,12	11,12	99.6	0.4	0	D7S820	8,11	8,11	99.56	0.43	0.01
D8S1179	9,11	9,11	97.66	2.33	0.01	D8S1179	12,13	12,13	98.61	1.38	0.01
D13S317	8,9	8,9	99.94	0.05	0.01	D13S317	9,14	9,14	99.51	0.45	0.04
AMELX	X,Y	X,Y	99.98	0.01	0.01	AMELX	X	X	99.95	0	0.05
D18S51	17,19	17,19	91.24	8.55	0.21	D18S51	13	13	95.95	3.96	0.09
PentaE	7,17	7,17	99.63	0.31	0.06	PentaE	10,11	10,11	99.65	0.05	0.30
D16S539	9,13	9,13	97.73	2.27	0	D16S539	11	11	97.33	2.59	0.08
D19S433	N/A	16	98.13	1.38	0.49	D19S433	N/A	16,16,2	97.81	1.9	0.29
D2S1338	N/A	16,20	93.98	5.89	0.13	D2S1338	N/A	17,20	95.03	4.88	0.09
D21S11	29	29	98.55	1.4	0.05	D21S11	28,30	28,30	98.66	1.1	0.24
PentaD	12,13	12,13	99.19	0.58	0.23	PentaD	9,12	9,12	99.89	0.11	0.00

Figure 1. STR-NGS optimization and performance. (A) Schematic representation of a STR locus with STR repeats, flanking regions at the start and end of the repeat region, and a targeted deep sequencing (DS) primer pair. Sequence from “5’ Flank” to “3’ Flank” is used for repeat length analysis. (B) Allele and stutter frequencies for five representative STR loci calculated as the percentage of the parent allele reads. Verified STRs indicated by a # above bar plot. (C) Total background (stutter + noise) in all STR loci in the comparison between standard and optimized PCR conditions. (D) STR profile for each locus examined by CE and NGS based methods in two diploid iPSC lines.

Figure 2 A



B



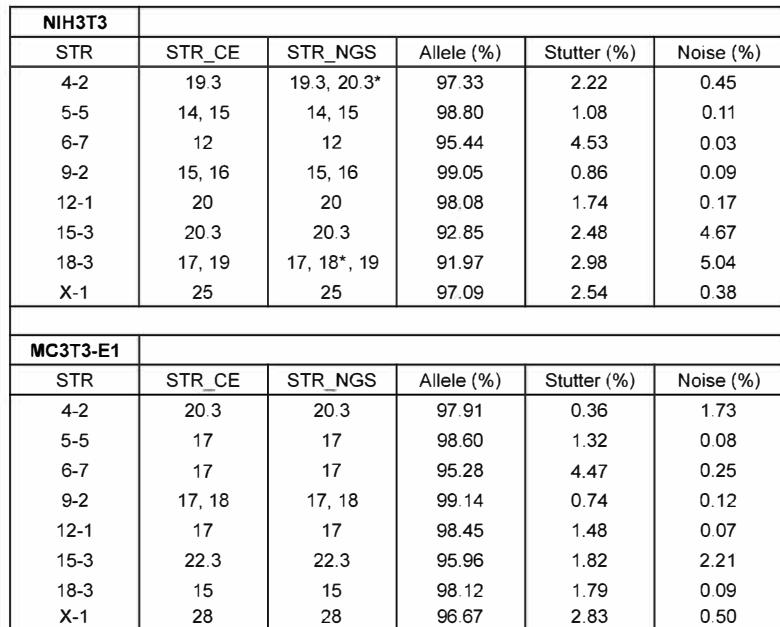
C

BJFF.6	STR	Allele (%)	Stutter (%)	Noise (%)
Group 1	VWA	96.19	3.68	0.13
	D19S433	98.39	0.11	1.50
	PentaD*	-	-	-
	FGA	96.26	3.06	0.68
	PentaE	96.67	0.76	2.57
	D2S1338	93.04	6.13	0.83
Group 2	CSF1PO	98.68	0.91	0.40
	TH01	98.46	1.52	0.02
	D18S51	90.29	8.01	1.69
	D3S1358	97.89	1.76	0.35
	D8S1179	97.74	2.19	0.07
	D13S317	99.88	0.11	0.01
Group 3	D7S820	99.38	0.54	0.08
	AMELX	99.92	0.07	0.00
	TPOX	99.44	0.51	0.05
	D5S818	99.00	0.13	0.87
	D16S539	96.70	3.02	0.28
	D21S11	98.58	0.24	1.18

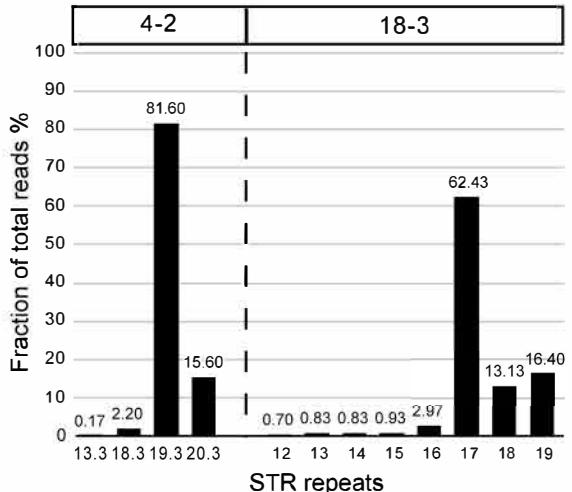
Figure 2. STR-NGS sensitivity on a mixed sample and optimized multiplexed conditions. (A) Observed allele fractions of informative STRs repeats are plotted against the expected ratio for given mixtures of 2 genomic samples. AN1.1 gDNA was diluted into BJFF.6 gDNA in a ratio of 1:1 to 1:1000. Expected allele fractions for diluted AN1.1 STRs correlate well with observed. (B) A mixture of two diploid cell lines was analyzed for STR D8S1179. M and N alleles indicate genotypes from the major and minor components, respectively. Bar graph shows percentage of read counts for each STR repeat of both cell lines in different mixture ratios. (C) STR profile for each locus examined using multiplexed STR-NGS in BJFF.6 iPSC cells. (*) dropout allele.

Figure 3

A



B



C



Figure 3. STR-NGS in mouse cell lines. (A) STR profile in two mouse cell lines (NIH3T3 and MC3T3-E1) shows allele, stutter and noise fractions of total parent allele reads. Asterisk indicates the difference in STR repeat length between reference and STR-NGS. (B) The frequencies of different stutters and repeat lengths, and (C) sequence context of different STR repeat lengths in two STR loci (4-2 and 18-3) different between reference and STR-NGS in NIH3T3 cells.

STR repeat	4-2	5-5	6-7	9-2	12-1	15-3	18-3	X-1
Clone #1	19.3, 20.3	14, 15	12	15, 16	20	20.3	17, 18*	25
Clone #2	19.3, 20.3	14, 15	12	15, 16	20	20.3	17, 19	25
Clone #3	19.3, 20.3	14, 15	12	15, 16	20	20.3	17, 19	25
Clone #4	19.3, 20.3	14, 15	12	15, 16	20	20.3	17, 18*	25
Clone #5	19.3, 20.3	14, 15	12	15, 16	20	20.3	17, 18*	25

Table 1. STR profiles of five NIH3T3 clones using STR-NGS
(*) difference in repeat length between STR-NGS and STR-CE method

NIH3T3					
STR	STR_CE	STR_NGS	Allele (%)	Stutter (%)	Noise (%)
4-2	19.3, 20.3	19.3, 20.3	98.47	0.94	0.58
5-5	14, 15	14, 15	98.09	0.98	0.93
6-7	12	12	94.65	5.26	0.09
9-2	15, 16	15, 16	99.07	0.82	0.11
12-1	20	20	98.08	1.74	0.17
15-3	20.3	20.3	96.17	0.73	3.10
18-3	17, 19	17, 18*, 19	98.12	1.28	0.60
X-1	25	25	97.97	1.58	0.46
1-1	10	10	99.37	0.57	0.06
2-1	9	9	99.67	0.29	0.04
3-2	14, 15	14, 15	98.22	1.42	0.36
8-1	15	15	97.86	1.97	0.17
11-2	15, 17	14*, 15, 16*, 17	99.56	0.00	0.44
17-2	13, 14	13, 14	94.23	5.46	0.31
19-2	11, 12	11, 12	99.26	0.70	0.05
CT26					
STR	STR_CE	STR_NGS	Allele (%)	Stutter (%)	Noise (%)
4-2	21.3	21.3	96.99	2.26	0.75
5-5	14	14	98.04	1.86	0.10
6-7	12	12	97.77	2.17	0.06
9-2	none	15	98.64	1.28	0.08
12-1	16, 17	16, 17	99.16	0.80	0.04
15-3	21.3, 22.3	21.3, 22.3	94.33	0.83	4.83
18-3	19, 20	19, 20	95.98	3.13	0.89
X-1	25, 26	25, 26	97.96	1.45	0.59
1-1	14, 15	14, 15	99.03	0.81	0.16
2-1	16, 17	16, 17	98.78	0.98	0.24
3-2	14	14	95.99	3.54	0.47
8-1	13	13	98.38	1.53	0.09
11-2	17	17	94.19	4.98	0.83
17-2	15, 16, 17	15, 16, 17	94.68	4.03	1.29
19-2	13	13	97.07	2.63	0.31
4T1					
STR	STR_CE	STR_NGS	Allele (%)	Stutter (%)	Noise (%)
4-2	21.3	21.3	98.31	1.31	0.38
5-5	14	14	98.89	1.04	0.07
6-7	12	12	98.27	1.68	0.06
9-2	none	15	98.96	0.94	0.09
12-1	16	16	98.90	1.03	0.07
15-3	22.3	22.3	94.94	1.97	3.08
18-3	18, 19	18, 19	97.35	2.14	0.52
X-1	25	25	97.82	1.78	0.40
1-1	15, 16	15, 16	98.97	0.86	0.17
2-1	16, 17	16, 17	99.02	0.85	0.13
3-2	14, 15	14, 15	98.39	1.27	0.35
8-1	13	13	98.53	1.41	0.06
11-2	18, 19, 20	18, 19, 20	95.49	3.26	1.25
17-2	15	14*, 15	97.86	1.65	0.49
19-2	13	13	97.48	2.40	0.12

Table 2. STR profiles of 3 mouse cell lines
(*) difference in repeat length between STR-NGS and STR-CE method

Human loci	STR-NGS (single rxn)	Multiplexing
Amelogenin	recommended	recommended
TPOX	recommended	recommended
D13S317	recommended	recommended
PentaE	recommended	recommended
D7S820	recommended	recommended
CSF1PO	recommended	recommended
D5S818	recommended	recommended
D21S11	recommended	recommended
TH01	recommended	recommended
D3S1358	recommended	recommended
D19S433	recommended	recommended
D8S1179	recommended	recommended
D16S539	recommended	recommended
PentaD	recommended	drop out
VWA*	stutter (slight above 3%)	stutter (slight above 3%)
FGA*	stutter (slight above 3%)	stutter (slight above 3%)
D2S1338*	higher stutter (>5%)	higher stutter (>5%)
D18S51*	higher stutter (>5%)	higher stutter (>5%)

Table 3. Summary of human STR loci using STR-NGS

(*) Underperformer for human STR-NGS

Mouse loci	STR-NGS (single rxn)
1-1	recommended
2-1	recommended
9-2	recommended
12-1	recommended
5-5	recommended
4-2	recommended
X-1	recommended
8-1	recommended
19-2	recommended
18-3	recommended
3-2	recommended
6-7	recommended
15-3*	higher noise (>3%)
11-2*	stutter (slight above 3%)
17-2*	stutter (slight above 3%)

Table 4. Summary of mouse STR loci using STR-NGS
(*) Underperformer for mouse STR-NGS

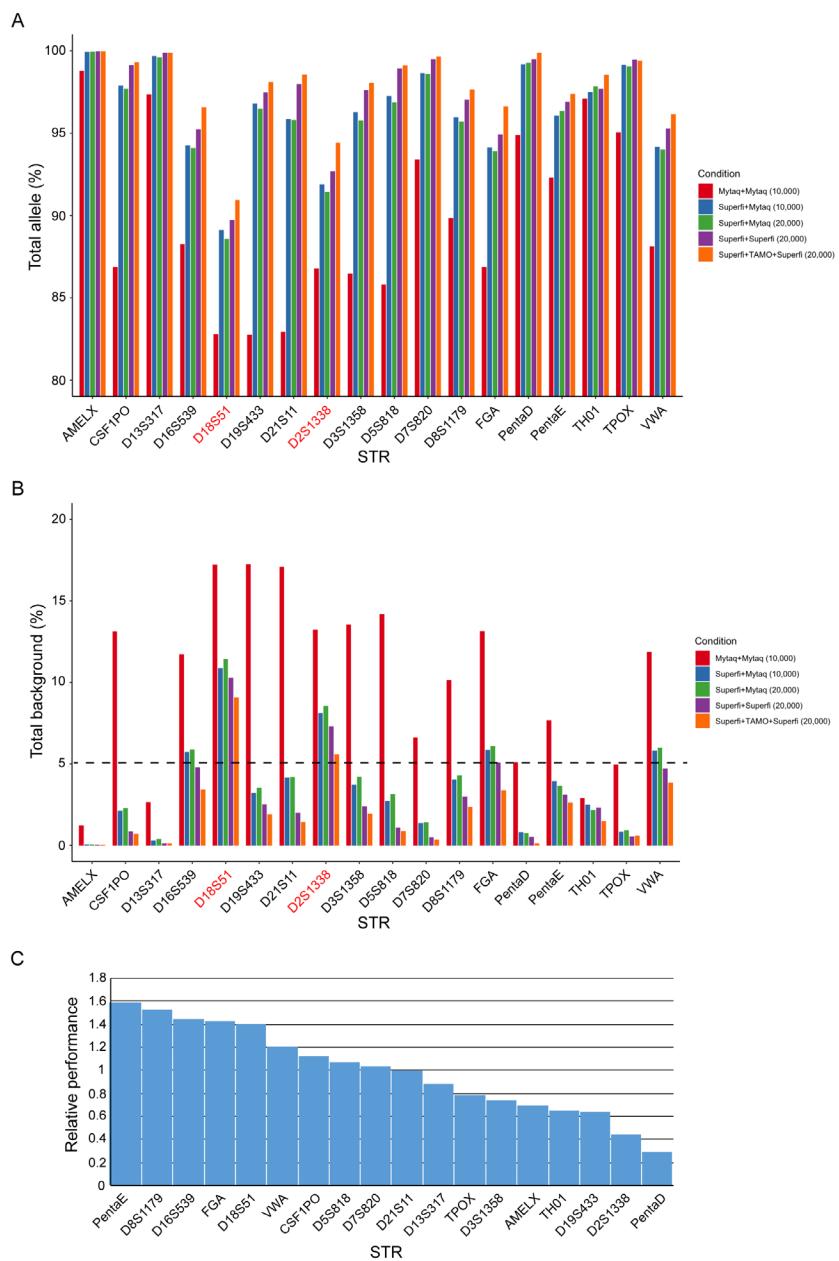


Figure S1. Human STR loci. (A) Percent of correct call on STR repeat length in each STR locus among different PCR conditions. Figure legend defines the polymerase used for each step of the NGS library construction (PCR1 and PCR2). The number of input cells is shown in parenthesis. (B) Percent of total background reads including stutter and noise in each STR locus among different PCR conditions. Underperforming STR loci with >5% background using the optimized PCR conditions are highlighted in red. (C) Relative STR locus performance (locus specific reads vs. mean of total reads).

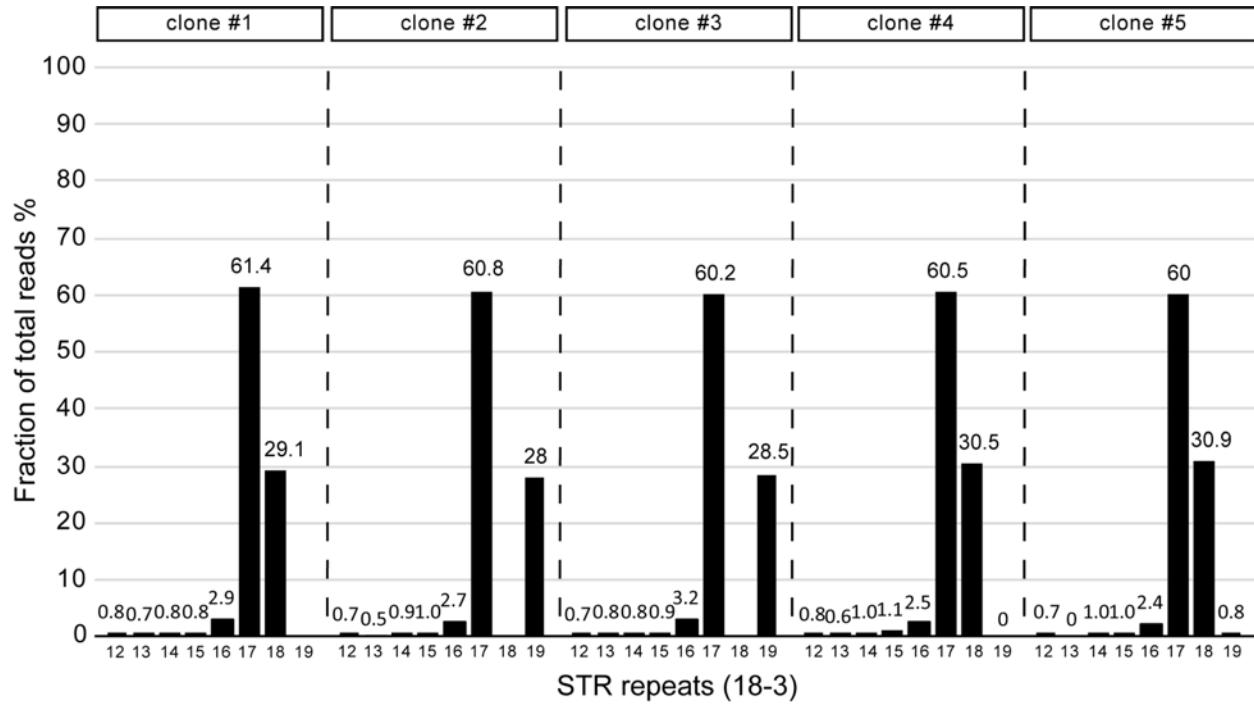
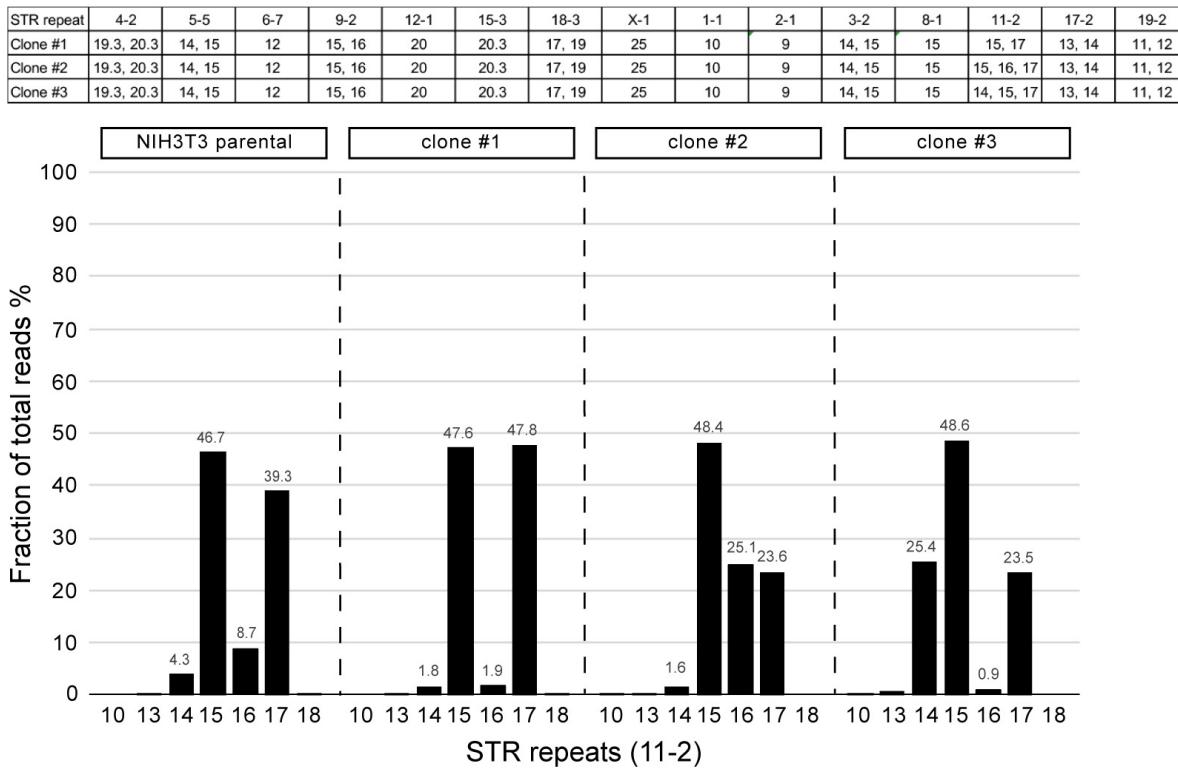


Figure S2. 18-3 STR locus in NIH3T3 cells. Bar graph displays the percent of each STR repeat length of 18-3 locus in single cell clones of NIH3T3 cell line.

A

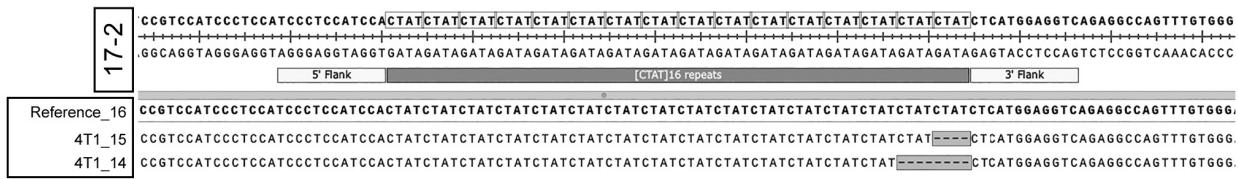


B

STR	STR_CE	STR_NGS_#1	STR_NGS_#2	STR_NGS_#3
11-2	15, 17	15, 17	15, 16, 17	14, 15, 17
Percentage	-	58% (7/12)	25% (3/12)	16.6% (2/12)

Figure S3. 11-2 STR loci in NIH3T3 cells. (A) Table reports STR repeat lengths of 15 mouse STR loci in single cell clones representing three different subpopulations from NIH3T3 cells. Bar graph demonstrates the percent of each STR repeat length of 11-2 locus in three single cell clones and parental cells. (B) Proportion of three different populations reported for STR repeat at 11-2 locus.

A



B

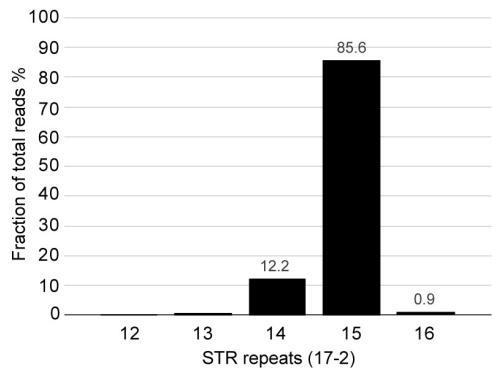


Figure S4. Sequencing alignment of 17-2 STR locus. (A) Sequence alignments of representative NGS reads among different STR repeat lengths for 17-2 STR locus of 4T1 cell line. (B) Percentage of each STR repeat length in 17-2 locus 4T1 cell line.

Table S1. Human STR primer sets.

STR locus	Primer Name	Primer Sequence	Amplicon size (bp)
AMELX/Y	AMELX-Y.DS.F	Adapter1 + ACACAGGCTTGAGGCCAAC	~ 217
	AMELX-Y.DS.R	Adapter2 + ATCCTGGGCACCCCTGGTTAT	
CSF1PO	CSF1PO.DS.F	Adapter1 + TGAGTCTGCCAAGGACTAGC	~304
	CSF1PO.DS.R	Adapter2 + CACCACTGCCATCTTC	
D3S1358	D3S1358.DS.F	Adapter1 + GCTTGAGCCCAGGAGTTGA	~282
	D3S1358.DS.R	Adapter2 + ATGGGCATGCTGCCATATT	
D5S818	D5S818.DS.F	Adapter1 + GCTTCTAATTAAAGTGGTGTCCC	~255
	D5S818.DS.R	Adapter2 + TCTCAGAGGAATGCTTAGTGCT	
D8S1179	D8S1179.DS.F	Adapter1 + GCCAGAAACCTCTGTAGCCA	~335
	D8S1179.DS.R	Adapter2 + ACCGTATGTATTCTGTTCCAGT	
D13S317	D13S317.DS.F	Adapter1 + GATGGGTTGCTGGACATGGTA	~265
	D13S317.DS.R	Adapter2 + AGAGAGCTTGAATTGTTGGTCAA	
D21S11	D21S11.DS.F	Adapter1 + ATTCTTCAGCTTAGATGGTCTG	~268
	D21S11.DS.R	Adapter2 + AGTCAATGTTCTCCAGAGACAGAC	
FGA	FGA.DS.F	Adapter1 + AAGGCTGCAGGGCATAACAT	~299
	FGA.DS.R	Adapter2 + CCTCTGACACTCGGTTGAGG	
TH01	TH01.DS.F	Adapter1 + TCTAGCAGCAGCTCATGGTG	~282
	TH01.DS.R	Adapter2 + AAGGTTCTGAGTGCCAAGG	
TPOX	TPOX.DS.F	Adapter1 + ACTGGCACAGAACAGGCACCTAGG	~244
	TPOX.DS.R	Adapter2 + GGAGGAACTGGGAACCACACAGGT	
VWA	VWA.DS.F	Adapter1 + GCTGAGATGTGAAAGCCCTA	~326
	VWA.DS.R	Adapter2 + ACTAGATACAATAGAGATAGATAGATAGAT	
D2S1338	D2S1338.DS.F	Adapter1 + GGCAATTCTACTGGCCCCT	~284
	D2S1338.DS.R	Adapter2 + CAGAATGCCAGTCCCAGAGG	
D7S820	D7S820.DS.F	Adapter1 + GAGACGGGTTTCACCATGT	~256
	D7S820.DS.R	Adapter2 + TCCACATTATCCTCATTGACAGA	
D16S539	D16S539.DS.F	Adapter1 + GGGGTCTAAGAGCTTGTAAAAAG	~340
	D16S539.DS.R	Adapter2 + AAAGTAGGTGGTAAAACAGCCT	
D18S51	D18S51.DS.F	Adapter1 + GGAGGAGTTCTTGAGCCCAG	~306
	D18S51.DS.R	Adapter2 + CGTCAGCCTAAGGTGGACAT	
D19S433	D19S433.DS.F	Adapter1 + GTTGAGGCTGCAAAAAGCTATAA	~326
	D19S433.DS.R	Adapter2 + GGGTTCTAGGAATCAATCTCCT	
PentaE*	PentaE.F	ACCAACATGAAAGGGTACCAATA	~474/342

	PentaE.DS.F	Adapter1 + ACATGGTGAAACCCCGTCTC	
	PentaE.DS.R	Adapter2 + TACTCATTACCTTGCATGCATG	
PentaD	PentaD.DS.F	Adapter1 + GAAGGTCGAAGCTGAAGTGA	~332
	PentaD.DS.R	Adapter2 + GGTATCTTTGAAGAGTAGAACT	

Table S2. Mouse STR primer set.

STR locus	Primer Name	Primer Sequence	Amplicon size (bp)
4-2	4-2.DS.F	Adapter1 + AAGCTTCTCTGGCCATTG	~ 237
	4-2.DS.R	Adapter2 + TTCATAAACTTCAAGCAATGACA	
5-5	5-5.DS.F	Adapter1 + CGTTTACCTGGCTGACACA	~341
	5-5.DS.R	Adapter2 + GGTTTAAACTCAATACCAAAACAA	
6-7	6-7.DS.F	Adapter1 + AGTCCACCCAGTGCATTCTC	~278
	6-7.DS.R	Adapter2 + GGCTTTAATTCCCCAGAATCTGT	
9-2	9-2.DS.F	Adapter1 + GGATTGCCAAGAATTGAGG	~228
	9-2.DS.R	Adapter2 + TCCTGAGTTGTGGACAGGGTTA	
12-1	12-1.DS.F	Adapter1 + AGCTACGTAATATGAGAGGGTTT	~325
	12-1.DS.R	Adapter2 + ATCAGAATGAACCAATGGTCAAGA	
15-3	15-3.DS.F	Adapter1 + TCTGGCGTGTCTGTCATAA	~196
	15-3.DS.R	Adapter2 + TTCTCAGGGAGGAGTGTGCT	
18-3	18-3.DS.F	Adapter1 + CAAGGGCCAGACTTGGACAATA	~285
	18-3.DS.R	Adapter2 + TAGCTCAGGGTAGAGAGTATGC	
X-1	X-1.DS.F	Adapter1 + GGGAGCATCACCTCTCAAAGATA	~363
	X-1.DS.R	Adapter2 + ATGGGATTGTGTGGTATCCTACT	
1-1	1-1.DS.F	Adapter1 + CCCCTTCACTCCTCATTCCA	~355
	1-1.DS.R	Adapter2 + CTCCCTTGCTGTACACTCCG	
2-1	2-1.DS.F	Adapter1 + AGTCTTTGTGGCTGGACCGAG	~259
	2-1.DS.R	Adapter2 + GAGCAATCACTGGCCTTGGTAG	
3-2	3-2.DS.F	Adapter1 + AAGCTGTAACATATTGTCTGTC	~301
	3-2.DS.R	Adapter2 + GACATGTGATAGAGGATGGATAG	
8-1	8-1.DS.F	Adapter1 + ATGAAGCCCCAAGCTCCATC	~291
	8-1.DS.R	Adapter2 + GCCTAGAACTGGGCCTGG	
11-2	11-2.DS.F	Adapter1 + GGTGGCTACAACCATCAGT	~255
	11-2.DS.R	Adapter2 + ATCCACGTCTTGAGGGCTG	
17-2	17-2.DS.F	Adapter1 + TGTGCTCCAGAACCATCCTG	~305
	17-2.DS.R	Adapter2 + AGCACAAACTGAGAGAGCCT	

19-2	19-2.DS.F	Adapter1 + TTGGTCTGGTTAGTCCCTGG	~296
	19-2.DS.R	Adapter2 + AATGGATTTAAAAAGTGAGGTTCCC	

Adapter1: 5' CACTCTTCCCTACACGACGCTCTCCGATCT 3'

Adapter2: 5' GTGACTGGAGTTCAGACGTGTGCTCTCCGATCT 3'

PCR index primer (forward):

5' CAAGCAGAAGACGGCATACGAGAT [i7 indexes] GTGACTGGAGTTCAGACGTGTGCTC

PCR index primer (reverse):

5' AATGATACGGCGACCACCGAGATCTACAC [i5 indexes]
ACACTCTTCCCTACACGACGCTCTTC 3'